mathematics

MDPI

Article

The Real-Time Depth Estimation for an Occluded Person Based on a Single Image and OpenPose Method

Yu-Shiuan Tsai [1], Li-Heng Hsu [1], Yi-Zeng Hsieh [2,3,4,*] and Shih-Syun Lin [1,*]

[1] Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung City 202, Taiwan; ystsai@email.ntou.edu.tw (Y.-S.T.); l1l12288tw@gmail.com (L.-H.H.)
[2] Department of Electrical Engineering, National Taiwan Ocean University, Keelung City 202, Taiwan
[3] Institute of Food Safety and Risk Management, National Taiwan Ocean University, Keelung City 202, Taiwan
[4] Center of Excellence for Ocean Engineering, National Taiwan Ocean University, Keelung City 202, Taiwan
* Correspondence: yzhsieh@mail.ntou.edu.tw (Y.-Z.H.); linss@mail.ntou.edu.tw (S.-S.L.)

Received: 7 July 2020; Accepted: 4 August 2020; Published: 10 August 2020

Abstract: In recent years, the breakthrough of neural networks and the rise of deep learning have led to the advancement of machine vision, which has been commonly used in the practical application of image recognition. Automobiles, drones, portable devices, behavior recognition, indoor positioning and many other industries also rely on the integrated application, and require the support of deep learning and machine vision. As for these technologies, there is a high demand for the accuracy related to the recognition of portraits or objects. The recognition of human figures is also a research goal that has drawn great attention in various fields. However, the portrait will be affected by various factors such as height, weight, posture, angle and whether it is covered or not, which affects the accuracy of recognition. This paper applies the application of deep learning to portraits with different poses and angles, especially the actual distance of a single lens for the shadowed portrait (depth estimation), so that it can be used for automatic control of drones in the future. Traditional methods for calculating depth using images are mainly divided into three types: one—single-lens estimation, two—lens estimation, and three—optical band estimation. In view of the fact that both the second and third categories require relatively large and expensive equipment to effectively perform distance calculations, numerous methods for calculating distance using a single lens have recently been produced. However, whether it is the use of traditional "units of distance measurement calibration", "defocus distance measurement", or the "three-dimensional grid space messages distance measurement method", all of these face corresponding difficulties and problems. Additionally, they have to deal with outside disturbances and process the shadowed image. Therefore, under the new research method, OpenPose, which is proposed by Carnegie Mellon University, this paper intends to propose a depth algorithm for a single-lens occluded portrait to estimate the actual portrait distance for different poses, angles of view and obscuration.

Keywords: depth estimation; openpose; occluded person

1. Introduction

Many applications (such as autonomous vehicles, drones) need to calculate the depth (i.e., the actual distance from the camera to the object) to the object in front. However, most objects tend to be difficult to estimate due to occlusion. The occlusion of objects may result in the disappearance of people or features that need to be identified due to various external factors such as weather, light, obstacles, and overlap.

In response to market-oriented technology, one of the issues is the positioning [1]. In addition, many prior art applications require depth measurement, the actual distance from the camera to the object.

For example, "automobiles" need to accurately measure the distance from the vehicle in front, but also to distinguish between different types of movement and depths of background, the human body, objects, etc., in order to determine traffic speed. The time required to stop or continue with the movement, therefore, needs to be limited in the automobiles to achieve real-time performance.

There are several ways to estimate the depth. For example, we can use the specific target of the human body and image size of the target [2] to roughly calculate the actual distance of the person. However, the part size of the object to be estimated must be known in advance, and the object to be estimated must not be excessively skewed. Using a defocusing approach to calculate the depth of the object [3] often takes more time to calculate the depth of multiple objects. The position of the line infers the image vanishing point to calculate the depth, which requires pre-establishing the image comparison database. Problems such as low recognition ability cause system failure. Hardware devices using methods such as time-of-flight are expensive, and the degree of freedom is poor due to device limitations [4–7].

The "drone" requires real-time distance and depth calculations to effectively maintain the flight distance or to meet other advanced requirements such as smart following and human body recognition. "Smart medical" also needs to analyze the subject's affected part or specimen image and then use the neural network model trained well in advance to determine whether there are any differences, so it is also necessary to obtain the depth and distance information of the patient's affected part to identify the injured part location and type. "Smart long-term medical care" also needs to determine whether the behavior of the patient or long-term subject is abnormal by obtaining the distance and depth of the image to prevent falls or other accidents.

However, the more complex calculations and forecasts have to pay the corresponding price. All of the operations that are mentioned above require higher hardware and software support to achieve real-time performance and recent market demand has shown that automobiles, drones and portable devices need to be delivered faster. The technical support of the disguised hardware directly limits the development and application related to the development of the machine's vision.

Because the neural network plays a big part, it requires a lot of hardware and software support but it is subjected to the physical limitations of the operating environment. Attention should be paid to how to simplify the calculation and operation mode without affecting the performance and reducing the hardware and software body burden, which has been an important goal in research. If there is a need to identify whether a combat situation has occurred, we can look in the image and then automatically return it to the police units to deal with the outcome at that particular time. Therefore, in order to identify in a single camera, it is a difficult task as the image plane of the crowd demonstrates the collective action of a show of hands, but there is lack of depth and distance information. Hence, we cannot determine whether the crowd has raised their hands or punched forward in the image, because in the single image when there is no depth distance information, there is no difference between the features of "hand dance" and "fist attack". Thus, if effective distance and depth information is added, it can effectively improve the precision of various neural networks, and smoothen the behavior identification process in quasi degree. Additionally, in order to reduce the burden on hardware and software and accelerate the operation speed, it is necessary to set up the environment in a different physical condition.

In this study, we developed a method for estimating the depth of a person using a single lens. In addition to being able to effectively estimate multiple people at the same time, depth can also be estimated for the person being occluded. Because drones are limited by the load, and most of the drone batteries can only fly for 30 min, photographic equipment that is too heavy may not be mounted on the drone. Besides, most of the deep learning applications require a lot of software and hardware support.

Therefore, how to effectively simplify the calculation and operation mode and reduce the burden of software and hardware without affecting the performance is important.

2. Related Works

2.1. Depth Detection

Distance (depth) detection occupies a very important position in the field of intelligent machines. It allows the machine to immediately understand its working environment to avoid collision accidents. For example, the patented drone safety warning mechanism in Google's patent Implementation model [8], the real-time multi-obstacle detection system with stereo vision fusion laser scanning equipment [9], Industry 4.0 Internet of Things combined with multiple perception analysis technologies to improve manufacturing efficiency and quality [10], simultaneous positioning and Simultaneous Localization and mapping (SLAM) [11,12], single-lens full-mirror depth prediction model technology [13,14], etc., were proposed. It can be seen that people are trying to integrate deep learning methods so that computers can learn the optical, acoustic, sensing components through machine learning. By the sampling and analysis of various data, the machines can do their best to simulate human judgement ability or stereoscopic perception.

2.2. OpenPose

OpenPose [15] is a humanoid recognition technology proposed by CMU in 2017, by combining convolutional pose machine (CPM) [16], real-time multi-person pose estimation [17] and hand key point detection using multi-view bootstrapping [18]. The core architecture of the three papers has been developed and the three main objectives mentioned in the text are related to pose estimation. Here are the challenges which we have focused to answer.

1. Multi-body detection for a single image;
2. A single image overlapped on a human body to detect the test;
3. Real-time operations.

It is beneficial for us to add depth and distance calculations, as this information will improve the limitations present in the existing environment based on image and behavior recognition, and the outcome can be applied to more existing platforms.

OpenPose has four steps, and each step has two branches, one for calculating the heat-map and one for calculating vector maps. With the information of the heat-map and vector maps, we can know all the key points in the image, and then through the PAF's similarity calculation, these key points can be mapped to each person and the joint points of the person can be constructed. Through the number 0 to 17, a total of 18 nodes are used to mark human joints, and different body parts are colored to distinguish left and right and front and back, which can effectively distinguish most overlapping or obscured images.

2.3. Image Occlusion Problem Introduction

As early as 1990, Shimojo and Nakayama first proposed, in their paper, the unrecoverable nature of image masking and the difficulty of distinguishing it [19]. We referred to Faster's study in 2015, which proposed a rapid cycling neural network (Faster R-CNN) as a case study [20]. Using this type of neural network for human body identification experiments, the results can be classified into the following categories, based on the existing occlusion cases.

- Partial shielding

When mechanical vision recognizes a specific person's image, it is necessary to train a large number of images of the object in advance in order to capture, learn the features of the object and use it

to build a model and then classify the object. For example, in humanoid identification, we need to provide advanced input related to a large number of distinctive features (such as facial features) of humanoid images, like neural network architecture, which can be learned. However, the learning model based on complete humanoid features has a few limitations. As the features are unknown, masked image can occur, misjudgment will reduce the accuracy of identification, and there can be loss of identification features if we are unable to capture humanoid images properly.

- Partial features overlap and cover (not completely covered)

Due to various external factors such as weather, light, obstacles, overlap, etc., the features of the person or object that needs to be identified can partially disappear. The neural network can only identify the remaining features, although it can still identify the trained images; however, it will cause a decrease in recognition rate and an increase in error. The images located farther away will also suffer from identification failures and a large number of misjudgments will happen due to missing features. Due to the lack of the image portion features, the probability of identification was reduced. Additionally, due to partial feature masking, it can be seen that in the case of overlapping images, the blocked person has a poor recognition score. Moreover, the recognition rate of the person's image who is located farther away is significantly lower than that of the unclosed similar images.

- Completely covered

The condition of complete shielding is more severe than partial shielding due to the lack of some features caused by external conditions. When training an object model, machine vision will automatically acquire different image features for learning according to the different neural network architectures and the type used. Therefore, some feature values will not be selected as the main identification conditions of the object. Because of the complete lack of some features of the image within the estimated image, the case of full feature recognition will occur when the known feature value is completely occluded, resulting in problems such as image misjudgment and recognition disappearance.

- Features completely overlapping and missing (completely occluded)

Because the image model, prior to training, did not have a feature for capturing all the details of training, so we need to cover the face shield and overlap. If the distance is too far from the case, object features will be completely lost and the resulting image will not be justified directly. It will be unrecognizable and have other issues. Compared to partially occluded images, full-featured images are considered too difficult to solve this problem. This can be seen in the following example where there is mapping of complete feature obscuration (image misjudgment). As half-body is in front of the car, there is lack of training in detecting the image features and only the vehicle score can be identified. The upper half of the feature is completely occluded by the cow. The remaining feet features, classified as humans by neural networks, can also be seen from far and near distance maps. People who are far away lose their training recognition features because of the distance, so that they cannot be displayed at all.

3. The Proposed Methods

This section will test basic knowledge and process architecture. It consists of a detailed explanation of the hierarchical blocks consisting of blocks used for different experimental techniques in conjunction with the operation of each other. It includes the fine quasi-oriented basis to calculate the estimated amount of depth and distance figures and also present real-time effects.

The architecture is shown is Figure 1. First, the person is photographed, and then linear effect and skeleton feature extraction are performed. Then, the shoulder coordinates (#2, #5) required by the distance formula are extracted from the eighteen output coordinate points, as the input of the distance formula. After calculating the distance value of each frame of the image, it performs the return and

presentation action, and performs the stability and line smoothing processing on the output distance data to stabilize the output and improve its accuracy.

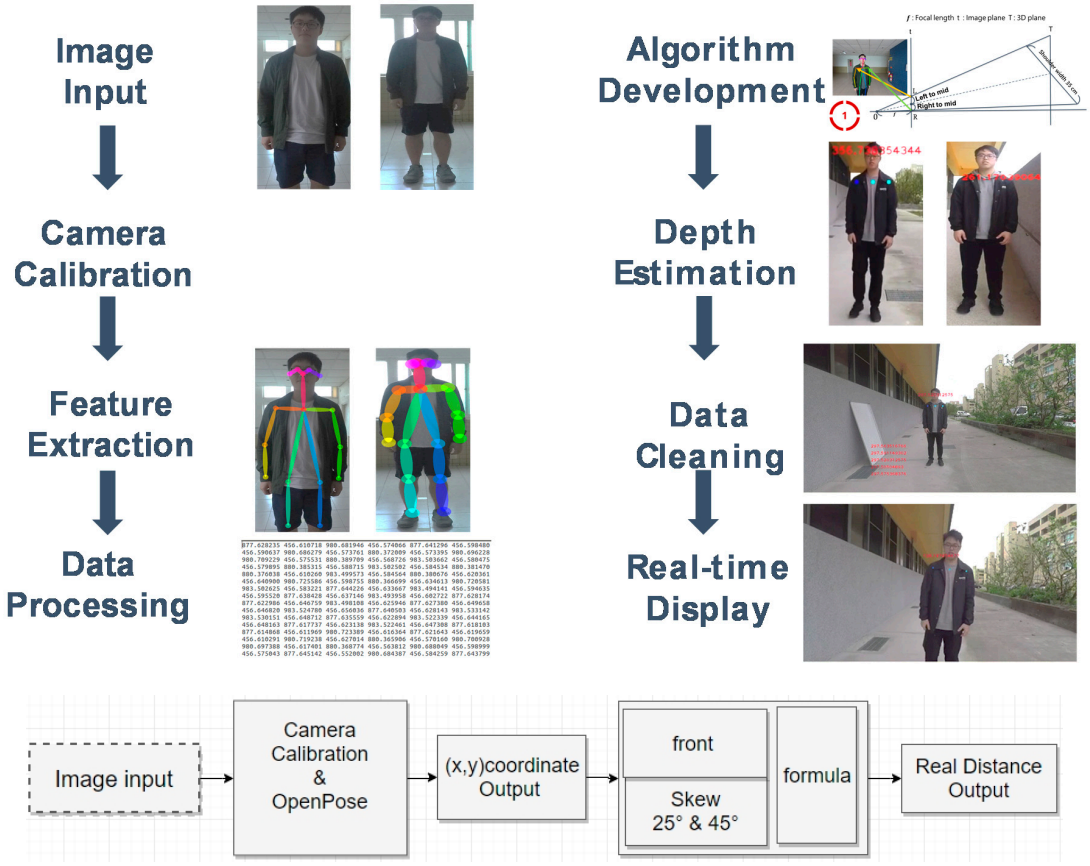

Figure 1. System architecture flow chart.

The first block—image input (image input): the photographic equipment "GoPro Hero5" is used to shoot the person to be estimated, and OpenPose image recognition material is used to facilitate the calculation of the subsequent distance formula.

The second block—camera calibration (camera calibration) and OpenPose feature extraction: the camera calibration part uses the built-in "linear" field of view (FOV) mode of "GoPro Hero5", which can directly perform real-time correction on the image. It works on pre-processing to obtain good undistorted images, and uses input from OpenPose for skeleton construction and joint feature point extraction.

In this experiment, "GoPro Hero5" was selected as the main camera for conducting the experiment. It was chosen because it has a high resolution of up to 4K and contains "HyperSmooth" and has three fields of view lens modes (FOV) to choose from, providing high stability and shooting accuracy to enhance the performance. The volume and weight of the shooting equipment is also one of the main areas of focus of this paper. The question of how to reduce the burden of hardware equipment and further improve the freedom of application to different vehicles is also an important part that we have considered. Therefore, "GoPro Hero5", which has the following features—single lens with light weight, long endurance, waterproof and shockproof, low price, easy to obtain, high image quality, and built-in camera correction and high stability performance—became the first choice to perform this experiment. High performance features with light weight and durability, as well as a camera with built-in linear (liner) shooting, saved the time required for the calibration process. Overall, it was able to simplify the experimental workflow, speed up the formula computing speed and also help to improve the degree of freedom in the future, which can be applied to other platforms for development.

3.1. Image Correction

With this photographing apparatus, GoPro Hero5, different resolutions can be utilized under conditions of the same particle mirror as it has three different photographic field modes of selection. These are linear, wide-field and the superview modes (from left to right). The first mode is linear, where the field-of-view function performs algorithm calculations through the on-camera chip (PG1), actively corrects the camera's internal parameters and external field of view and presents an undistorted and distorted image effect. Through this built-in function, we can save the complicated process of pre-correction of photographic equipment and it also improves the accuracy and speed of the distance formula.

The third block represents the feature point coordinates (Coordinate) integrated capture: through OpenPose build feature nodes and with our application, the output from the eighteenth can capture the desired distance formula shoulder coordinates (#2, #5) as the distance formula input.

Through the use of "OpenPose", we can construct the humanoid backbone. We can find the available human body features and select the calibration. In order to calculate the depth and distance of the human body, we assess the calibration point, which has four candidate conditions that are as follows: 1. Invariability. 2. The difference should be small. 3. The features should be obvious and easy to distinguish. 4. It needs to be conducive to the calculation of the distance formula.

Based on the requirements of the calibration point, the selected features cannot be easily distorted by factors such as distortion or deformation due to human movements and cannot be so different among different races and categories that they are generated as a totally different human bodies. Hence, distance deviation should be taken into account. Therefore, according to the physical properties of the human body structure, the "shoulder width" will not easily change too much regardless of age, gender, height, weight and the shoulders will not be too excessive when someone is walking, running, etc. The physical features of distortion are considered and thereby, the width of the shoulder is kept horizontal to a certain extent with the camera, which is very conducive for depth calculation and distance estimation. Therefore, through the "OpenPose" humanoid feature recognition technology, after capturing the shoulder coordinates (#2, #5) as seen in Figure 2, the midpoint coordinate (#1) can be calculated through the formula and, finally, through the triangular similarity [1] and image imaging principles, we can calculate the depth and distance of the human body.

(**a**) (**b**) (**c**)

Figure 2. (**a**) Skeleton feature point; (**b**) human recognition; (**c**) human shoulder feature. Through OpenPose's humanoid feature recognition technology, the person's shoulder coordinates are captured and the person's depth and distance are calculated.

The fourth block represents the distance judgment (formula): through the shoulder coordinates (#2, #5) obtained from OpenPose as the input of the distance formula for calculation, the formula will use the shoulder width as the feature calibration point to judge the reality of the person and the camera distance. Different people's skew, distortion, etc. will have different physical meanings, and, based on that, four different situations will be produced (Case 1–4).

3.2. Distance Formula Calculation

Shooting images through a single lens will produce four imaging possibilities. According to the principle of image imaging, the image plane (t) formed by the origin (O) and the focal length (f), the angular distance between the shooting object and the real object plane (T), will be different due to the estimation of the shoulder offset and position of the person, which refers to the distance information. Assuming the object is facing the middle of the camera (dashed line of the image), the skewness of the different image planes on the left and right of the center line will cause different distance information, as shown in the following pictures (Figure 3). Case 1 and 2 illustrate that the left side of the image center line and the shoulder spacing is different. Case 3 and 4 illustrate another possibility on the right side of the center line of the image, with different shoulder spacing. The distance information will have a difference in sign due to the offset, but as long as the absolute value is taken into account, the four issues can get the same answer through the formula that we have used.

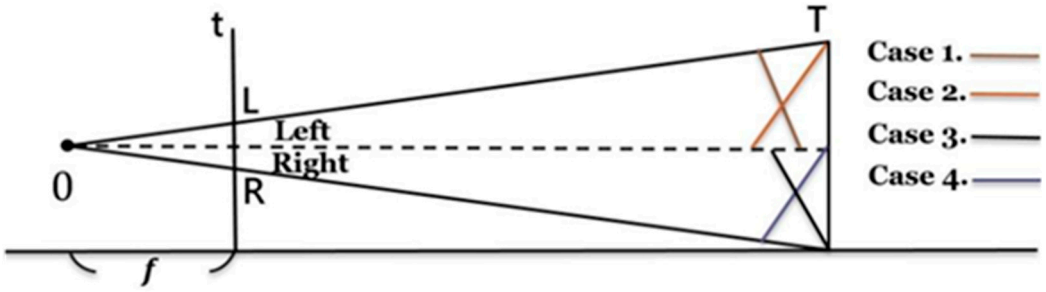

Figure 3. Schematic diagram of distance estimation. Taking pictures through a single lens will produce four imaging possibilities.

Before calculating the formula, we must obtain the focal length well in advance and the focal length can be seen through the following figure (Figure 4a). Based on the principle of pinhole imaging, we know that the left side of the pinhole is the internal plane of the image. The other side is reality, so as long as the shoulder coordinates (#2, #5) are obtained first by OpenPose, the image distance (d) can be effectively achieved and finally the person arrives at the figure according to the known shoulder actual width of 35 cm and the default camera setting. The straight line distance can be used to find the internal value of the focal length smoothly. Figs. 4b to 4e show four situations according to the tilt direction of the person's centerline of the lens and the person's shoulder facing the lens.

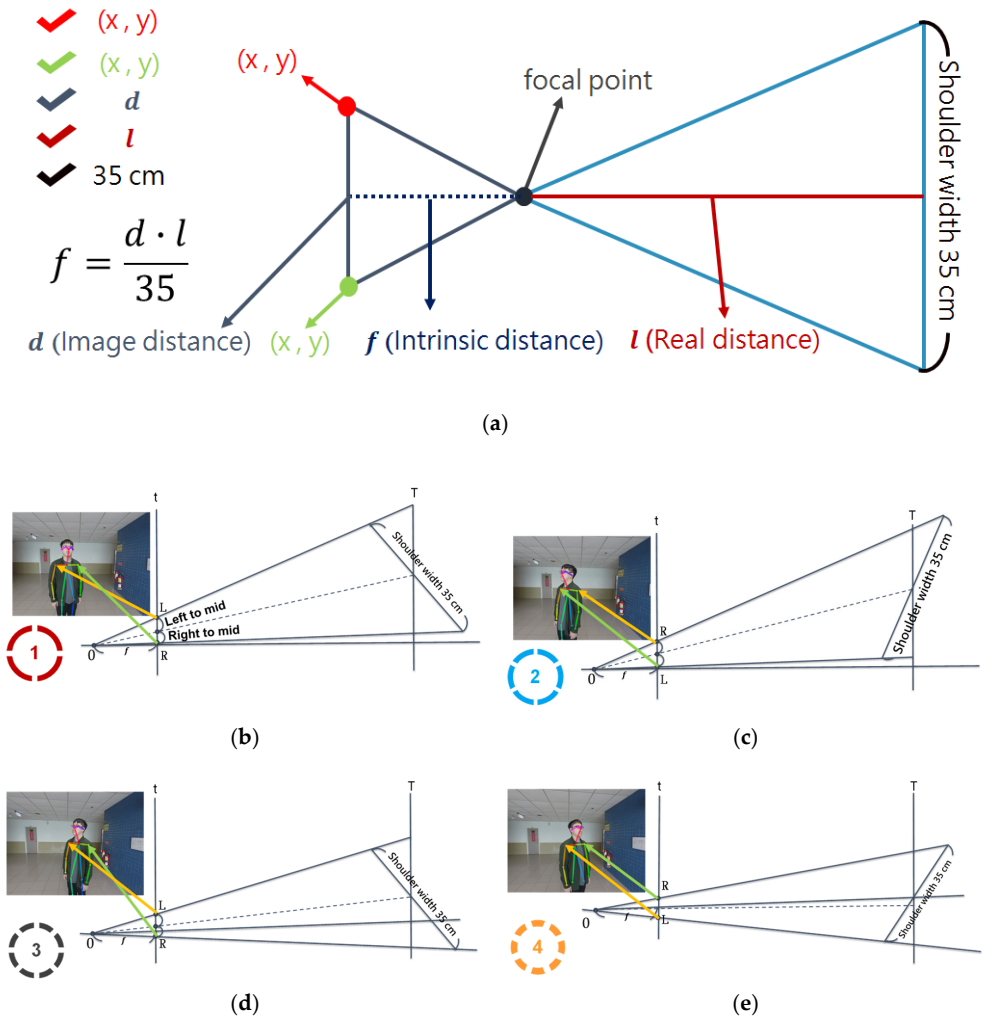

Figure 4. (**a**) Focal length calculation. (**b**–**e**) Four situations according to the tilt direction of the person's centerline of the lens and the person's shoulder facing the lens.

- **Case 1. The person is on one side of the axis of the photography, but the right shoulder is slightly forward.**

Facing the midline of the image plane (dashed line), the figure is shifted to the left and skewed, and the shoulder is slanted (a > b) (Figure 5).

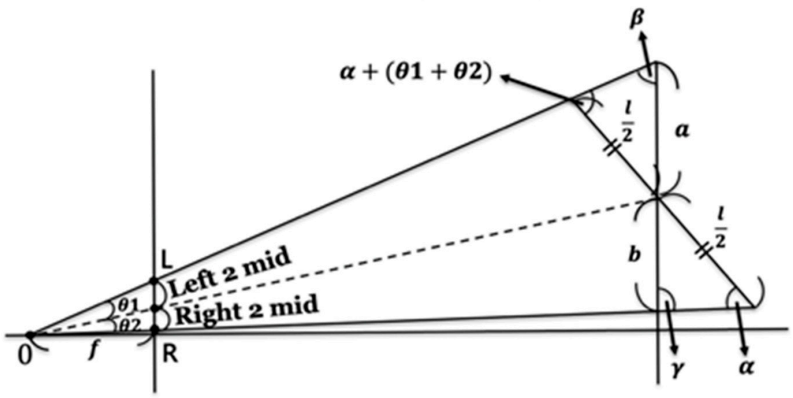

Figure 5. Schematic illustration of Case 1. The person is on one side of the axis of the photography, but the right shoulder is slightly forward.

It is known that l is the length of the shoulder, and, since the two angles α and β can be obtained according to the external angle formula, they are also known. Let $O = (0, 0)$ be the origin, $L = \left(L_x, L_y\right)$ and $R = \left(R_x, R_y\right)$ be the image positions of the two shoulders.

Suppose that $\gamma = 180^{\circ} - \angle ORL$, $\beta = \angle OLR$, a/b are known, where According to the inner product formula

$$\vec{L0} \cdot \vec{LR} = |\vec{L0}||\vec{LR}| cos\beta \tag{1}$$

In addition,

$$\vec{R0} \cdot \vec{RL} = |\vec{R0}||\vec{RL}| \cos(\pi - r) \tag{2}$$

Then, according to the sine theorem,

$$\frac{l/2}{sin\beta} = \frac{a}{sin(\alpha + (\theta 1 + \theta 2))} \tag{3}$$

And

$$\frac{l/2}{sin\gamma} = \frac{b}{sin\ \alpha} \tag{4}$$

Dividing Equation (3) by Equation (4), we can get

$$\frac{sin\gamma}{sin\beta} = \frac{a\ sin\ \alpha}{b\ sin(\alpha + (\theta 1 + \theta 2))} \tag{5}$$

or

$$cot\alpha = \frac{1}{sin(\theta 1 + \theta 2)}\left[\frac{a\ sin\ \beta}{b\ sin\ \gamma} - cos(\theta 1 + \theta 2)\right] \equiv \mathrm{A} \tag{6}$$

After solving α, we can substitute Equations (1) and (2) to find a and b.

$$\alpha = \cot^{-1} A \tag{7}$$

- **Case 2. The person is on one side of the axis of the photography, but the left shoulder is slightly forward.**

Facing the midline of the image plane (dashed line), the figure is shifted to the left and skewed, and the shoulder is slanted (b > a) (Figure 6).

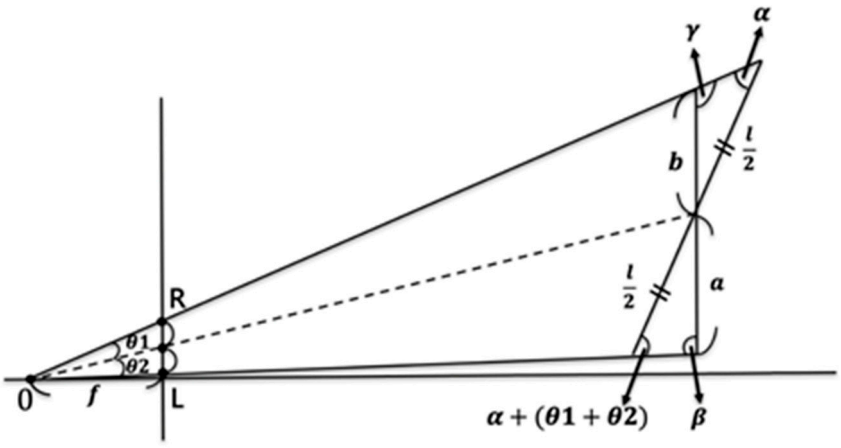

Figure 6. Schematic illustration of Case 2. The person is on one side of the axis of the photography, but the left shoulder is slightly forward.

According to the sine theorem,

$$\frac{l/2}{sin\gamma} = \frac{b}{sin\alpha} \tag{8}$$

And

$$\frac{l/2}{sin\beta} = \frac{a}{sin\ (\alpha + (\theta 1 + \theta 2))} \tag{9}$$

Dividing Equation (3) by Equation (4), we can get

$$\frac{sin\beta}{sin\gamma} = \frac{b\ sin(\alpha + (\theta 1 + \theta 2))}{a\ sin\ \alpha} \tag{10}$$

or

$$cot\alpha = \frac{1}{sin(\theta 1 + \theta 2)}\left[\frac{a\ sin\ \beta}{b\ sin\ \gamma} - cos(\theta 1 + \theta 2)\right] \equiv \mathrm{A} \tag{11}$$

After solving α, we can substitute Equations (1) and (2) to find a and b.

$$\alpha = \cot^{-1} A \tag{12}$$

Taking the longer side of the triangle as a and the shorter side as b, the diagonal angles of a and b are $\alpha + (\theta 1 + \theta 2)$ and α, respectively. We can conclude that Case 1 and Case 2 are consistent.

- **Case 3. The axis of the photograph passes through the person, but the right shoulder is slightly forward.**

Facing the midline of the image plane (dashed line), the figure is shifted to the right and skewed, and the shoulder is slanted (a > b) (Figure 7).

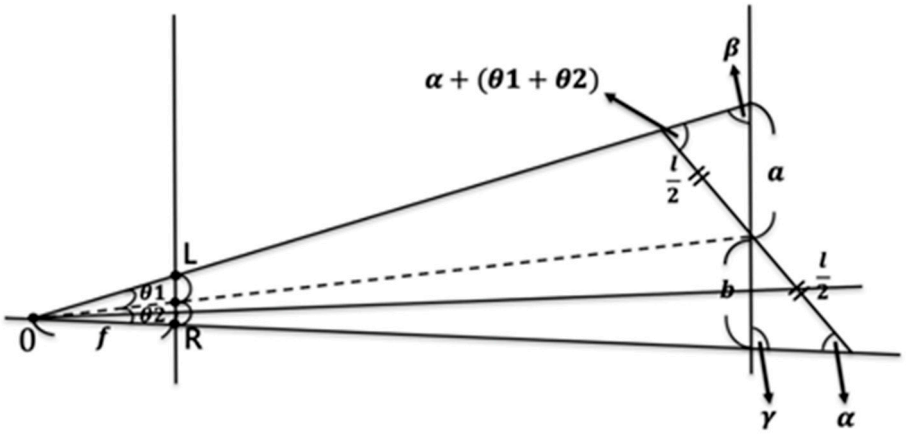

Figure 7. Schematic illustration of Case 3. The axis of the photograph passes through the person, but the right shoulder is slightly forward.

Notice that $\beta = \angle OLR$ and $\gamma = 180^{\circ} - \angle ORL$. According to the sine theorem,

$$\frac{l/2}{sin\beta} = \frac{a}{sin(\alpha + (\theta 1 + \theta 2))} \tag{13}$$

and

$$\frac{l/2}{sin\gamma} = \frac{b}{sin\ \alpha} \tag{14}$$

The following derivation is equivalent to Case 1.

- **Case 4. The axis of the photograph passes through the person, but the left shoulder is slightly forward.**

Facing the midline of the image plane (dashed line), the figure is shifted to the right and skewed, and the shoulder is slanted (b > a) (Figure 8).

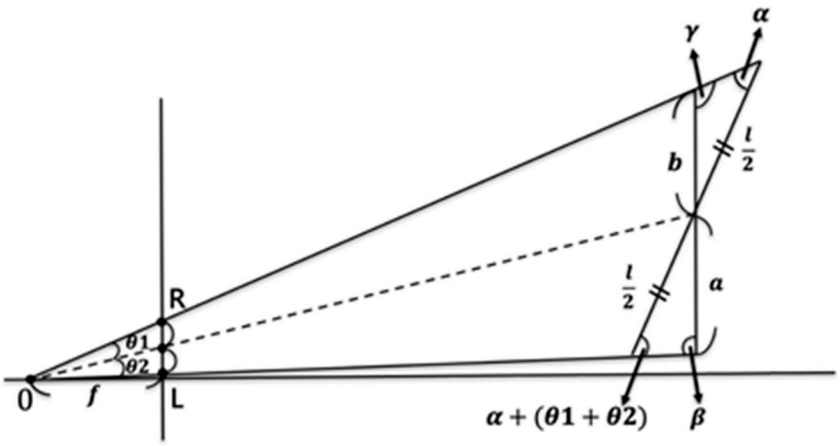

Figure 8. Schematic illustration of Case 4. The axis of the photograph passes through the person, but the left shoulder is slightly forward.

According to the sine theorem,

$$\frac{l/2}{sin\beta} = \frac{a}{sin(\alpha + (\theta 1 + \theta 2))} \tag{15}$$

and

$$\frac{l/2}{sin\gamma} = \frac{b}{sin\ \alpha} \tag{16}$$

The following derivation is equivalent to Case 2.

According to the above derivation, the above four situations can be calculated with the depth algorithm shown in Algorithm 1.

Algorithm 1. Depth Estimation.

Input. shoulder length l; Image position on both shoulders L; and R; focus length; Output: human depth d.
Step 1. Take longer edge as L, and another edge R
Step 2. $\beta = \angle OLR$, $\gamma = \pi - \angle ORL$, $\theta 1 + \theta 2 = \angle ROL$
Step 3. $A =$

$$\frac{1}{\sin(\theta 1 + \theta 2)}\left[\frac{a\ sin\ \beta}{b\ sin\ \gamma} - \cos(\theta 1 + \theta 2)\right]$$

$\alpha = cot^{-1}A$
Step 4. $a = \frac{l}{2}\frac{\sin(\alpha+\theta 1+\theta 2))}{sin\beta}$,

$$b = \frac{l}{2}\frac{sin\alpha}{sin\gamma}, d = \frac{a+b}{\overline{LR}}f$$

3.3. Real-Time Distance Display

The person's distance is calculated according to the formula. The formula measures the distance output in real time through the human body midpoint feature (#1) found in OpenPose. By using OPENCV, the processing of the image occurs after OpenPose captures the features and therefore, the distance of the person calculated by the formula can be displayed in real time. Using the middle point (#1) of the person as shown in the image (Figure 9) for distance calibration and tracking, the distance number can reach the real time by following the person's mobile presentation.

Figure 9. Human body real-time distance display. The figure shows the coordinates of the shoulder and midpoint of OpenPose, and displays the distance of the person in real time.

Through the output example as shown in the above figure, we can see the shoulder coordinates and the midpoint coordinate of the person grabbed by OpenPose. On the midpoint coordinate, there will be distance (cm) calculated by a set of formulas that can distinguish the distances of multiple people's images. We can also perform real-time human body location and tracking along with distance output presentation.

4. Experiment

This chapter will use the OpenPose humanoid architecture feature extraction capability, combined with the above four formulas, to use the shoulder spacing feature to determine the actual distance of the human body. It will include a quiz related to ten groups where the level is maintained by 3 m, 5 m, 7 m, 8 m, 9 m, 10 m, 12 m, 15 m, 17 m and 20 m. Each individual will be comprised of different shielding features (partially and completely masking) such that two will be the unmasked and there will be a display of front and back features of two human bodies. It will consist of information related to experiment's viewing angle, distance information of each human body including shaking, inclination and shoulder distortion data at each level. The entire experiment has six hundred distance (D) data.

4.1. Unmasked Image Test

The formula distance calculation is carried out for people with unoccluded features and experimental tests are conducted for multiple groups where the distance ranges from 3 m to 20 m, including multiple sets of data such as front, back and tilt angle. Each set of data has six hundred distance (D) information and also contains other information such as twisted shoulders, midpoint skew and movement of human bodies. In some special conducted figures, the image retains its full-featured human body as the image has not been occluded during the experiments, but some facial features will be missing when the person presents its back. We have used OpenPose to completely check the human body, its effective range limit and human body distance accuracy.

- **Feature part shadow image test**

In traditional image recognition, partial feature occlusion may cause misjudgment or complete failure of the pre-trained image model, which requires continuous retraining of the model. The shadow recognition of the image is also a major problem in computer vision, so the experiment in this section

focuses on the human body's half body and only retains the shoulder calibration feature, which is used to calculate the distance of the human body. The experiment is performed in the state of partial feature shadow to test the lack of half of the body in OpenPose, the effective range of the above mentioned human body features and accuracy of the human body's distance. The experimental data is tested among multiple groups from 3 m to 20 m. The data includes multiple groups' information such as the frontal view of the person and the angle of inclination. Each group of data has six hundred pieces of distance (D) information and contains the shoulder distortion of the human body. Information related to midpoint skew and human body movement is also included.

4.2. Partially Obscuring the Image of Flat Figures

The experiment contains images of the human body image portion including the experimental screening plane. OpenPose can be seen at a close distance of 3 m (Figure 10). It can be seen that due to the obstruction caused by the obstacle, the human body's foot is obviously not visible, but the upper part of the human body is not affected by the lack of the feature of the lower body. Overall, the distance is 3 to 7 m. We can see that, in the same front image plane, the figure has a high visibility distance and a fine distance quasi degree. However, the human body can be seen from the back surface in the second experiment by OpenPose. Due to the lack of human body features or when the human skeleton deviation is generated, the features are completely lost. It is also seen that the more the person's features are occluded, the larger the distance will produce a larger floating error value. Thus, OpenPose's effective scope only locates up to 3 to 10 m. The maximum distance is only 12 m, as long as the range does not exceed the scope of the marker. OpenPose finds the human body features rather than representing the unshielded flat front image of the human body only. The ruler recognizes the distance and its accuracy decreases significantly with the increase in the distance.

Figure 10. The 3-m flat figure half body occlusion (on flat ground). Under 3 m, it can be seen that the person's foot features are missing due to occlusion, but the upper part of the person is not affected by the lack of lower body features.

The experiment's result is shown in Figure 11. The performance at a close distance of 3 to 5 m is similar, which ranges from 7 m to 10 m. It was, however, seen that in the middle distance of the mentioned range, due to the lack of too many human body features, a larger error gradually occurs as the distance is lengthened, especially at 7 and 8 m because the OpenPose misjudgment makes the two sets of distance data very similar. While the long-distance data is distributed we saw that the recognition rate by OpenPose is too poor, so that it does not cover the effective functioning range

as only a portion of image is covered. However, its performance in human body occlusion is still much better than traditional image recognition. As for the recently popular neural network image training recognition method, it does not require repeated training of human bodies from different angles. The feature model greatly reduces the complexity of computing and equipment and has a better performance.

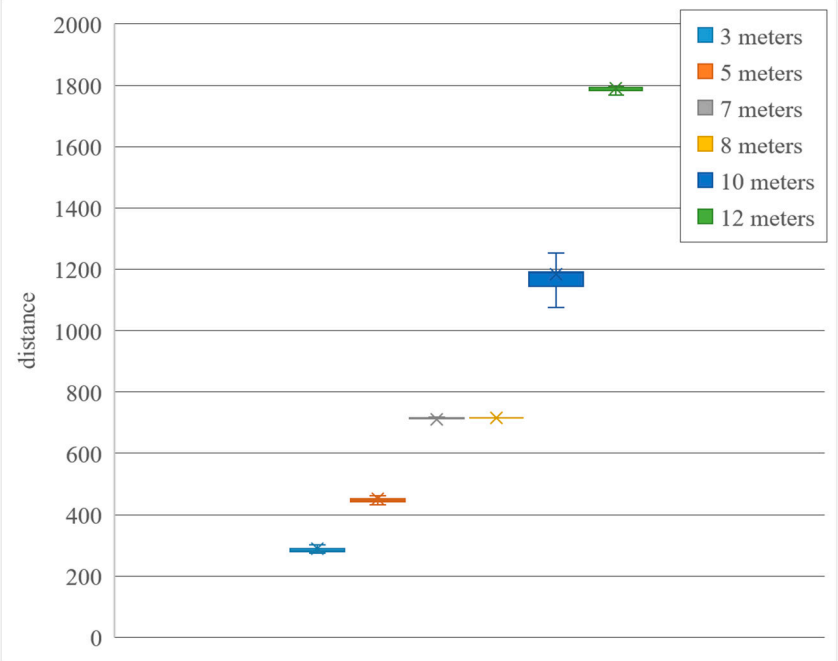

Figure 11. Flat figure half body occlusion statistics.

4.3. Partially Obscuring Oblique Figure Images

The experiments tested whether the distance information is calculated for the partially occluded person in the inclined plane (Figure 12). In addition to the calculation of the tilt angle of different image planes, some severe conditions of partial shadowing were also added to test if OpenPose's effective operating range and the accuracy of the distance of the human body formula are more accurate when it lacks more than half of the person's features and is located in the undulating image plane. This group of experiments contain data such as skewness and distortion of human bodies and are tested from 3 m to 15 m.

We can see from the unobstructed tilt experiment and the partial obscuration experiment that no matter under what conditions it is maintained, OpenPose can have good data performance as long as it is within a close range of 3 to 7 m. It can also be seen from the data that the errors are all less than 50 cm, however, compared with the general unmasked images, the experimentally masked experimental group's effective operating range is mostly between 3 and 12 m and the farthest recognition range is shorter. The main reason is that after adding multiple factors such as tilt, shadowing etc., OpenPose's human body feature technology is greatly disturbed, so that the distance performance under shadowing is poor.

The experiment's result is shown in Figure 13. The operation performance and effective range of OpenPose with many external disturbances have been greatly shortened and, due to the lack of human body features, OpenPose is used only for partial distances. The method of speculation and estimation establishes the human body's structure to form a small and uniform distance fluctuation within a uniform distance. At a long distance, due to the complete lack of human body features, the distribution of the distance data changes greatly. Therefore, under the multiple restrictions of tilt

and partial obscuration of human features, effective system operation and distance calculation can only be performed within a close range of 3 to 7 m.

Figure 12. The 3-m flat figure half body occlusion (on the slope).

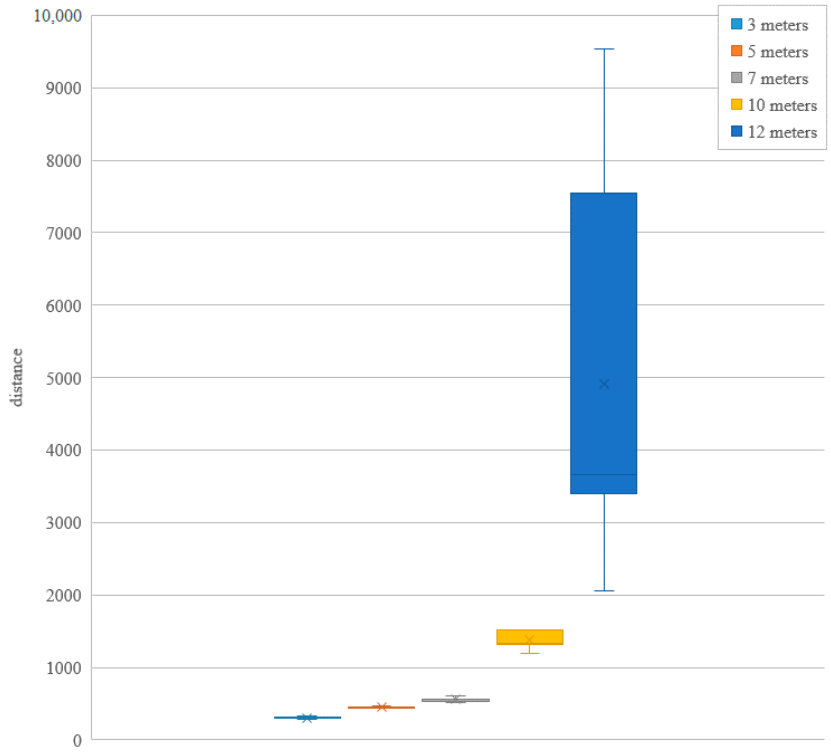

Figure 13. The slope half-length figure occluded chart. It is unstable at around 12 m.

4.4. Feature Completely Occluded Image Test

Compared with the partial feature masking of human bodies, full feature masking is challenging. Using traditional image recognition methods, when the features are completely occluded, it is a situation where spatial data and information cannot be recognized. In this section, we will use the human body feature to completely occlude the image in the experiment to examine how OpenPose uses the method of estimation and guessing to calculate the limb structure and how it coordinates

distribution of the shadowed human body under the completely missing human body features and calibration. Multiple sets of data related to high and low tilt angles were used and experimental tests are conducted with multiple sets of distances from 3 m to 20 m to test the effective range and limit of OpenPose.

4.5. Completely Obscuring the Image of Plane Human Bodies

The experimental image is an experiment associated with fully-featured occluded images of flat human bodies (Figure 14). Different from the partial feature masking as seen in the previous experiments, the full masking is conducted at a close distance of 3 m. As only the head features of the human body can be judged by OpenPose, so the human body's architecture is based entirely on the head. From the estimation of the swing and the angle and position, it can be seen that, after the masking, the human skeleton facing the camera lens is seriously distorted and shifted, which makes it difficult for humans to pose. However, with the close distance of 3 to 5 m, it can still perform better, but its error value is as high as about 20 cm, which is significantly ahead of previously conducted experiments. The farthest distance is also obtained due to the complete lack of features, which makes OpenPose almost impossible to identify the position of the person at 7 m and, therefore, a large amount of distance errors are generated. Therefore, the effective range for full-occlusion of a person's image is only 3 to 5 m. The farthest distance can only reach up to 7 m, however, which is partly the reason for the lack of features at the middle and long distances being related to image resolution and environmental factors. If the resolution and light source of the person in the image is improved, it will inevitably improve the complete shielding. This will result in an effective and accurate image.

Figure 14. OpenPose features of a 3-m flat figure with full body shadow. After being completely occluded, the skeleton of the person facing the camera lens is severely distorted and offset, forming a posture that is difficult for humans to pose.

The experiment's result is shown in Figure 15. Complete feature masking is undoubtedly a test to see the image outcome. With the loss of all the human bodies, there is also a serious external interference effect, which causes the error varied greatly in OpenPose. It can also be seen from the distribution chart that the distance between 3 and 5 m is accurate, but the error here is also greater. The distance of 7 m is obtained due to the misjudgment of OpenPose's human body estimation mechanism, resulting in 7 m. The above reported an error value, so its effective range is only distributed in the short range of 3 to 5 m and the farthest recognition distance is only 7 m. Although it has made a breakthrough performance presentation compared with traditional image recognition methods, image recognition with complete feature obscuration is still one of the most difficult goals that we need to overcome in this paper.

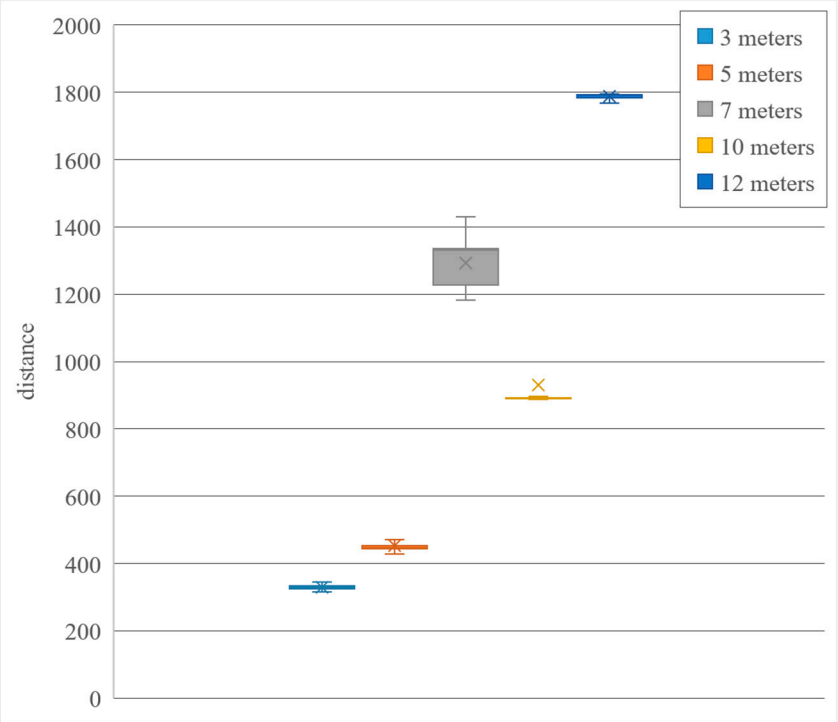

Figure 15. Statistics of the whole-body shadowing of plane human bodies.

4.6. Completely Obscuring the Image of Oblique Figures

The experimental image is a fully-featured oblique image experiment. This section is an advanced extension (Figure 16). If at the same time on the oblique image plane, the whole body of the person is occluded such that it is under the harsh environment of losing all features, high and low tilt errors are also added. We want to identify whether OpenPose can guess the human body's structure and effectively assess the human body's position and coordinates and then check its effective operating range and the accuracy of the human body distance formula to achieve the most difficult breakthrough in image recognition.

The seventh group consists of completely shielding experiments along with alignment features. A good close 3–5 m distance data performance of the case is taken into account. The inclination is from 3 m and systemic human body shielding by OpenPose is done wherein a FIG header will be apparent in addition to human body. In addition to the features of the department, plus the numerical error caused by the tilt, the features of the human bodies have shown below that the shoulders have completely disappeared. The longest estimation distance of OpenPose is 7 m, similar to previous experiments, but it can be unexpectedly seen from the picture that the person has a tilt at 7 m. The distance of the full-body masking is displayed as 644 cm, that is, the error is only about 56 cm. In some cases, OpenPose's estimation of the human body's features show unexpectedly excellent outcomes.

Under the interference of various factors such as tilt, full-featured shadowing and external interference (Figure 17), the stability of the human body's distance depends entirely on OpenPose's assumptions about the coordinates of the human bodies in the current environment. Therefore, in addition to the short range that is in between 3 and 5 m, the middle distance data also exhibits extreme floating estimation errors. In addition, its recognition limit is 7 m, which involves the general body cover features. Despite many serious interferences, OpenPose can still recognize the coordinates of people under certain conditions and then obtain spatial information through the distance formula. Compared with all the previous image recognition technologies, OpenPose has made a great amount of breakthroughs and improvements.

Figure 16. OpenPose features occluded by a 3-m tilted figure. In addition to the head, the features of the person below the shoulders have completely disappeared.

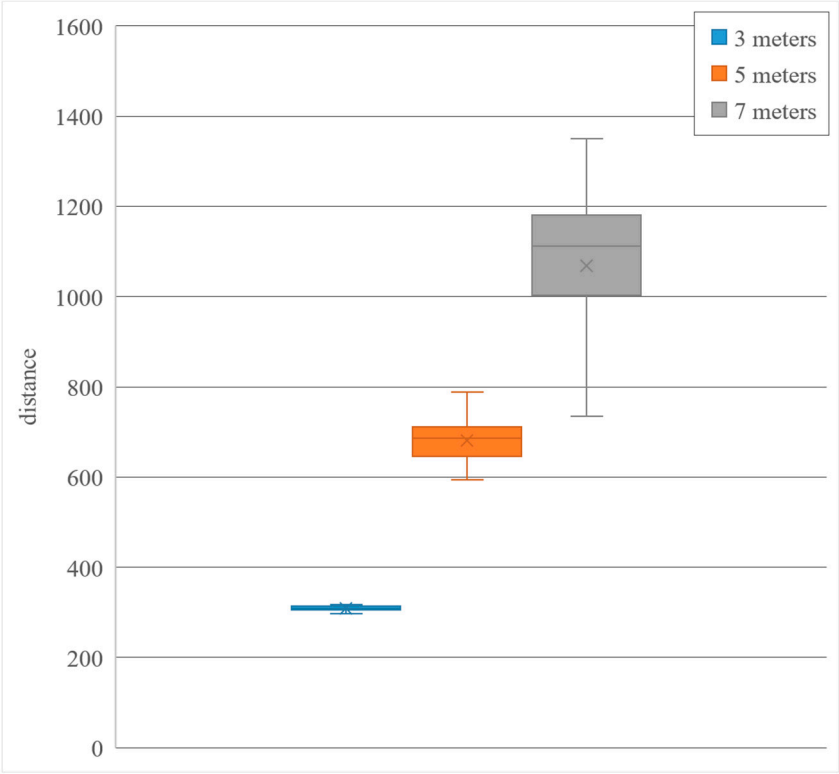

Figure 17. Statistical figure of full-body obscuration of slope figures.

4.7. Multi-Person Image Comprehensive Test

Finally, for a multi-person real-time image comprehensive experiment, estimation of the distance of people by shooting in the normal action environment in the general laboratory was performed. The environment may include an increased number of people or less than two people. There can be an excessive tilt in the person's position as seen from far to near and vice-versa. Various factors such as human body's movement, partial occlusion and complete occlusion are tested. OpenPose also detects how many people are there in the image when recognizing people. Through its numbering information, it can rewrite the frame to meet the real-time data, which can increase and decrease the

number of people who are recognizable and achieve the comprehensive experimental test results of one to eight projects.

After the experimental design, we can make some discussions and proceed to improve and optimize. The first step is to capture the human body's joint points. When the shoulder is skewed, there will be a slight deviation of the coordinate position, which will lead to some errors in the calculation of the shoulder midpoint coordinate (#1), resulting in a decrease in the accuracy of the distance calculation. Furthermore, the recognition distance output through the real-time image will cause the skew and error of the shoulder point due to the skew and distortion of the person, which will affect the continuous output distance value; it will appear to be too floating. Therefore, resolving the issue of how to effectively solve the problem of midpoint identification, smoothing of data, and deletion of misjudged data will be the goal of improvement and optimization. On the other hand, since this research discusses the distance calculation of a single image, continuous image output will be more conducive to stable image distance calculation. Therefore, statistics or filtering can also be performed on the output in the future. For example, deleting the outlier that is too deviated, or smoothing the continuous distance output, can effectively improve the stability and accuracy of the distance data output.

5. Conclusions

Different from the traditional image human body recognition technology, OpenPose has combined the simple hardware advantages of a single lens; it can solve a number of computer vision problems such as partial and complete feature masking, image plane tilt, conduct simultaneous recognition of multiple people and perform real-time calculation of distance images. Processing and solving procedures are used to calculate the distance, depth and spatial information of a single image. It used the recently popularized deep learning method to solve the situation that most traditional images cannot handle. It greatly simplifies the hardware equipment, effectively achieves low cost, results in high efficiency, stable and accurate performance. In the future, multiple technology transfers and applications can also be achieved by acquiring information about the space distance of the human bodies by the light weight single-lens photography equipment. For example, we can easily assess people's position at high-altitude by using drones, therefore reducing the operational burden of automobiles and mobile devices. By performing indoor positioning of images or using spatial depth information, we can increase the accuracy and selectivity of today's neural networks to obtain a better behavior recognition.

Author Contributions: Conceptualization, Y.-S.T., Y.-Z.H. and S.-S.L.; Methodology, Y.-S.T., Y.-Z.H. and S.-S.L.; Software, L.-H.H.; Validation, Y.-S.T. and L.-H.H.; Formal Analysis, L.-H.H.; Investigation, Y.-S.T.; Resources, L.-H.H.; Data Curation, L.-H.H.; Writing Original Draft Preparation, L.-H.H.; Writing Review & Editing, Y.-S.T., Y.-Z.H. and S.-S.L.; Visualization, L.-H.H.; Supervision, Y.-S.T.; Project Administration, Y.-S.T., Y.-Z.H. and S.-S.L.; Funding Acquisition, Y.-S.T., Y.-Z.H. and S.-S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Ministry of Science and Technology, R.O.C] grant number [MOST-109-2221-E-019-057-, MOST-108-2221-E-019-047-, MOST-108-2221-E-019-038-MY2, MOST-107-2221-E-019-039-MY2, MOST-108-2634F-019-001, MOST-108-2634-F-008-001, MOST-109-2634-F008-007, and MOST-109-2634-F-019-001, MOST 109-2634-F-009-015-]. This research was funded by [University System of Taipei Joint Research Program] grant number [USTP-NTUT-NTOU-109-01]. The research was also co-sponsored by Pervasive Artificial Intelligence Research (PAIR) Lab.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mateev, V.; Marinova, I.; Kartunov, Z. Gas Leakage Source Detection for Li-Ion Batteries by Distributed Sensor Array. *Sensors* **2019**, *19*, 2900. [CrossRef] [PubMed]
2. Nguyen, A.; Simard-Meilleur, A.; Berthiaume, C.; Godbout, R.; Mottron, L. Head circumference in Canadian male adults: Development of a normalized chart. *Int. J. Morphol.* **2012**, *30*, 1474–1480. [CrossRef]

3. Pentland, A.P. A new sense for depth of field. *IEEE Trans. Pattern Anal. Mach. Intell.* **1987**, *4*, 523–531. [CrossRef] [PubMed]
4. Deems, J.S.; Painter, T.H.; Finnegan, D.C. Lidar measurement of snow depth: A review. *J. Glaciol.* **2013**, *59*, 467–479. [CrossRef]
5. Massa, J.S.; Wallace, A.M.; Buller, G.S.; Fancey, S.J.; Walker, A.C. Laser depth measurement based on time-correlated single-photon counting. *Opt. Lett.* **1997**, *22*, 543–545. [CrossRef] [PubMed]
6. Smisek, J.; Jancosek, M.; Pajdla, T. 3D with Kinect. In *Consumer Depth Cameras for Computer Vision*; Springer: New York, NY, USA, 2013; pp. 3–25.
7. Tan, W.L.; Vohra, M.S.; Yeo, S.H. Depth and Horizontal Distance of Surface Roughness Improvement on Vertical Surface of 3D-Printed Material Using Ultrasonic Cavitation Machining Process with Abrasive Particles. In *Key Engineering Materials*; Trans Tech Publ: StafaZurich, Switzerland, 2017.
8. Cahill, P. Drone Safety Alert Monitoring System and Method. U.S. Patent Application 14/824,011, 11 February 2016.
9. Labayrade, R.; Royere, C.; Gruyer, D.; Aubert, D. Cooperative fusion for multi-obstacles detection with use of stereovision and laser scanner. *Auton. Robot.* **2005**, *19*, 117–140. [CrossRef]
10. Zhong, R.Y.; Xu, X.; Klotz, E.; Newman, S.T. Intelligent manufacturing in the context of industry 4.0: A review. *Engineering* **2017**, *3*, 616–630. [CrossRef]
11. Dissanayake, M.G.; Newman, P.; Clark, S.; Durrant-Whyte, H.F.; Csorba, M. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Trans. Robot. Autom.* **2001**, *17*, 229–241. [CrossRef]
12. Bailey, T.; Durrant-Whyte, H. Simultaneous localization and mapping (SLAM): Part II. *IEEE Robot. Autom. Mag.* **2006**, *13*, 108–117. [CrossRef]
13. Li, Z.; Dekel, T.; Cole, F.; Tucker, R.; Snavely, N.; Liu, C.; Freeman, W.T. Learning the depths of moving people by watching frozen people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4521–4530.
14. Kuznietsov, Y.; Stuckler, J.; Leibe, B. Semi-supervised deep learning for monocular depth map prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 6647–6655.
15. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 7291–7299.
16. Wei, S.-E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4724–4732.
17. Huang, K.S.; Trivedi, M.M. Robust real-time detection, tracking, and pose estimation of faces in video streams. In Proceedings of the 17th International Conference on Pattern Recognition(ICPR 2004), Cambridge, UK, 23–26 August 2004; pp. 965–968.
18. Simon, T.; Joo, H.; Matthews, I.A.; Sheikh, Y. Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. In Proceedings of the CVPR, Honolulu, HI, USA, 22–25 July 2017; p. 2.
19. Nakayama, K.; Shimojo, S.; Ramachandran, V.S. Transparency: Relation to depth, subjective contours, luminance, and neon color spreading. *Perception* **1990**, *19*, 497–513. [CrossRef] [PubMed]
20. Faster, R. Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *39*, 91–99.