Computers, Materials & Continua
DOI:10.32604/cmc.2022.024026
Article

Tech Science Press

Enhance Egocentric Grasp Recognition Based Flex Sensor Under Low Illumination

Chana Chansri and Jakkree Srinonchat*

Faculty of Engineering, Rajamangala University of Technology Thanyaburi, Pathum Thani, 12110, Thailand
*Corresponding Author: Jakkree Srinonchat. Email: jakkree.s@en.rmutt.ac.th
Received: 30 September 2021; Accepted: 01 November 2021

Abstract: Egocentric recognition is exciting computer vision research by acquiring images and video from the first-person overview. However, an image becomes noisy and dark under low illumination conditions, making subsequent hand detection tasks difficult. Thus, image enhancement is necessary to make buried detail more visible. This article addresses the challenge of egocentric hand grasp recognition in low light conditions by utilizing the flex sensor and image enhancement algorithm based on adaptive gamma correction with weighting distribution. Initially, a flex sensor is installed to the thumb for object manipulation. The thumb placement that holds in a different position on the object of each grasp affects the voltage changing of the flex sensor circuit. The average voltages are used to configure the weighting parameter to improve images in the image enhancement stage. Moreover, the contrast and gamma function are used to adjust varies the low light condition. These grasp images are then separated to be training and testing with pre-trained deep neural networks as the feature extractor in YOLOv2 detection network for the grasp recognition system. The proposed of using a flex sensor significantly improves the grasp recognition rate in low light conditions.

Keywords: Egocentric vision; hand grasp; flex sensor; low light enhancement

1 Introduction

Hands are the priority for humans that allow us to collaborate with the matters and the surroundings, correspond with others and carry out daily activities like dining, cleaning and dressing. Focused on their significance, the computer vision researchers have attempted to analyze hands from various aspects: determine the position of the hand in the image [1], analyze the hands from multiple perspectives: localizing them in the images are investigated in any types of actions [2–4], as well as interact with the computer and the robot [5–7]. Wearable cameras allow hands to be examined from a first-person perspective, known as egocentric or First-Person Vision (FPV) in computer vision [8–12], to challenge object detection and identifying activities. The essential characteristic of egocentric vision is providing a first-person perspective of the scene by laying a forward-facing wearable camera on the chest or head. This wearable camera offers a person-centric view and is optimally set to capture

information arguably more relevant to the camera wearer [13]. Since then, the egocentric vision is now being applied to more applications, including video summarization [14,15] and it also extends to the realm of healthcare [16]. The egocentric vision has several advantages compared to third-person point of view, where the camera position is often fixed desultory by the user. The camera wearer affects the movements, attentions, and activities as the camera will record whatever is in front of the user. Hands and objects being manipulated tend to appear in the center of the image and reduce hand obscuring. These pros make it very interesting to develop new accesses for studying hands. Modeling the relationship between hand gestures and object characteristics can provide additional information with a model that perceived the liaison between hands and objects. There are also examining the interact with objects by hand for other proposed models of neural networks as a transmitter for 3D objects and acts from RGB images and recognize objects and actions of the user [17]. The temporal network incorporates bi-directional long short-term memory to model the long-range dependencies to predict the actions in object manipulation tasks [18]. By the way, regarding the egocentric vision, researchers still challenge a significant problem that the camera is not secure but moves along with the human body. This movement produces rapid movements and immediate diversities, which can significantly distort the quality of the recording. Also, the sudden illumination changes can significantly reduce the quality of the image. The vague images and increment of disturbance due to the camera sensor makes it troublesome to detect and recognize a hand grasp, which is also tough to distinguish the subject from the background. Hence, restoring the composition of the image in low light conditions is a difficult task. Formerly, there has not been any previous egocentric research that has developed low-light detection and recognition. This article investigated hand grasp recognition in first-person vision in the dimmed environments or nighttime environments conditions. In order to boost the effectiveness for the detection and recognition of hand gestures, the proposed system added the flex sensor, which has the advantage of being an easy to use and low-cost device, is used as an additional parameter to work with Adaptive Gamma Correction with Weighting Distribution (AGCWD) [19]. The flex sensors are arranged on the thump of the hand to track the finger movement and combine with the AGCWD fusion technique. Then, the proposed system recognize the grasp with Deep Convolutional Neural Networks (DCNN) as the feature extractor and detection that has emerged as a valuable tool for computer vision tasks.

2 Method and Propose Algorithm

The objective of this article is to regcognize the hand grasp from the routine hand works. Offers an egocentric vision system to detect different hand gesture and automatically learns the image capture structure from the big data captured via the wearable camera and the flex sensor attached to the thumb finger. The images are captured using a head-mounted camera. The hand grasp performs 18 different postures. In each action, the issue is handling some object. The posture of the grasp comprises a particular five objects. For the different postures, some objects may use the same, and the output voltage of the flex sensor uniquely describes an object's grasp. Dataset consists of 3600 images for the image of grasp training. It has 18 grasp postures that have actions overlapping in daily life. The proposed summarization process consists of four stages: image enhancement, grasp factor computing, YOLOv2 network, and grasp evaluation. The architecture of the proposed method is described in the following Fig. 1. This concept is implemented by using a combination of flex sensor information from Arduino and the image from first-person perspective of the scene, then enhance the input image with the AGCWD. Then, finding hands in egocentric frames is an instantiation of one particular object detection task. The real-time object detection, "You Only Look Once (YOLO)", the algorithm that is one of the most effective technique [20,21], which show the high speed with great accuracy among

the many deep-learning algorithms. The YOLOv2 [22,23] has been used to detect and recognize the frame's grasp posture. The experimental environment is in the Window 10 64bit operation system and the processor is Intel Core i7–8700, memory is 16GB, GPU NVIDIA GeForce GTX 1070. The YOLOv2 was trained with MATLAB R2021a platform.

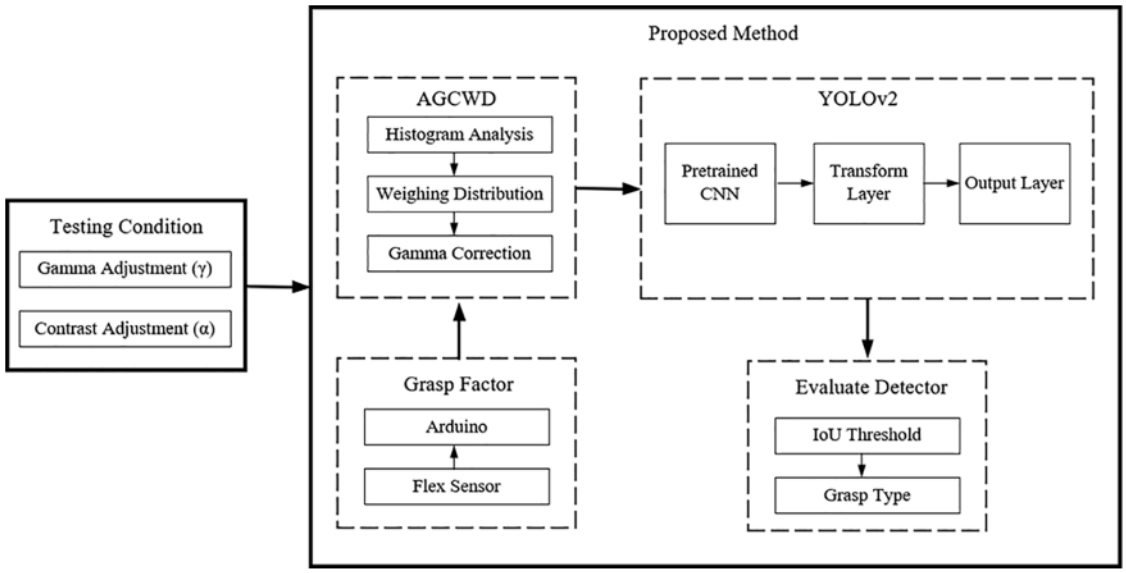

Figure 1: The architecture of the proposed summarization method

2.1 Grasp Type

The hand grasp type is vital for figuring the hand gesture due to the holding nature of the hand during control. Numerous studies have examined the classification of grips into discrete sets of types to help study manual grasping. We improve the classifiers to categorize the eighteen different comprehension types selected from the most widely used comprehension taxonomy [24]. The catch type has been chosen to cover other standard classification criteria based on the function, shape of the object, and the fingers' joints. According to work , a total of eighteen types have a high daily frequency of use [25,26]. Thus, the selected gestures can be used to analyze large amounts of manipulation tasks and possible for automatic recognition from image appearance. The grasp types provide information about how the hands are holding the objects during manipulation. Only the sole grasping cannot classify the delicate actions without details from the object being handled. In this Research, four volunteers were used to grasp objects in 18 postures, as Large Diameter (LD), Medium Wrap (MW), Small Diameter (SD), Fixed Hook (FH), Index Finger Extension (IFE), Thumb Index Finger (TIF), Thumb 2 Finger (T2F), Writing Tripod (WT), Tip Pinch (TPC), Power Sphere (PWS), Power Disk (PWD), Tripod (TPD), Lateral Tripod (LT), Parallel Extension (PE), Extension Type (ET), Literal Pinch (LP) and Ring (RN).

2.2 Flex Sensor

The flex sensors change the resistance depending upon the amount of bending on the sensors, mainly dealing with angle displacement measurement.The feature of flexible sensors produces resistance output related to the bending radius when the sensor is bent. The larger the radius, the higher the result, the more significant the change in deflection, the higher the resistance variation [27]. The flex sensor can be applied to the thumb as one of the features for grasp recognition because it is used in every grasp [28]. The thumb can either be removed or carried off. In the abducted part, the thumb can obstruct the fingertips. The abducted position allows to either apply forces on the fingers' side or move the finger out of the way. Such being the cased, the thumb has to be seized, as otherwise, the thumb cannot act against the fingertips, which will tilt the flex sensor according to that hand will manipulate an object cause a change in resistance. In this way, the voltage output will be send through the analog input ports on the Arduino MEGA 2560. by the internal digital convertor. Although the numerous hardware selections are available, the Arduino is the most popular due to the flexibility and user-friendly interface at a low cost. The thump grasp with a flexible sensor, each gestures on the object, and the flexible sensor's resistance must change according to the operation of the thumb.The signal of voltage output change on the flex sensor in each grasp posture is shown in Fig 2.

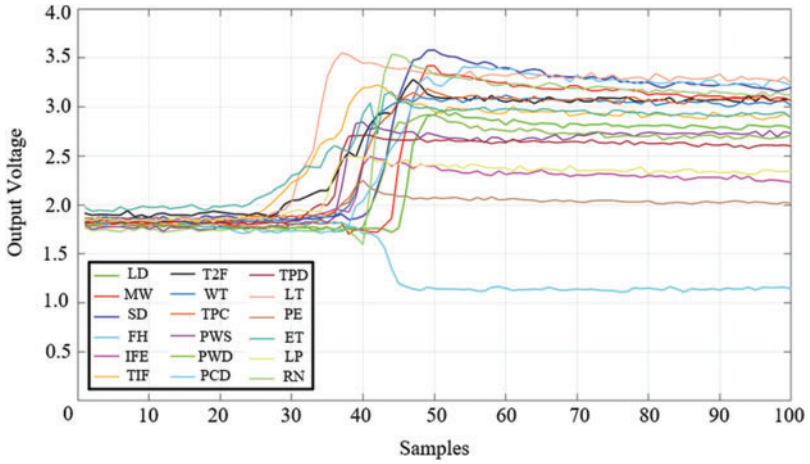

Figure 2: The flex sensor response in each grasp posture

2.3 The Combination Image Enhancement

The flex sensor is attached to the thumb to provide information on the movement and position of the finger, then give a data acquisition which convert into digital data with the Arduino. This information was used to configure the weighting parameter in the image enhancement section to enhance an image input. The flexible sensor signal was performed to show the response of the sensor. The Fig. 3, the start of the rest hand posture, then fingers are in the relaxed and will grasp the object when finger seat on the object and the voltage output is stable, the average voltage must be calculated, and the voltage output of the grasp could be varied depending on the griping posture and the shape of the object.

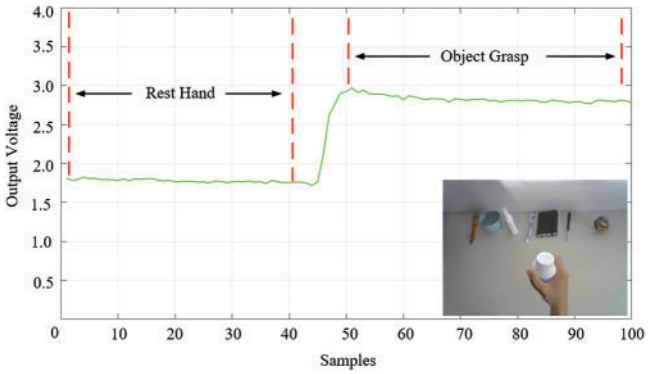

Figure 3: The signal of flex sensor circuit while bent with the object grasping

The flex sensor voltage output from the serial monitor is returned in bits from 0 to 1023 due to the built-in 10-bit ADC on the Arduino MEGA 2560 based on the voltage received from the circuit. The signal is collected by using the average voltage to calculate the grasp factor (gf) is obtained using the following Eq. (1) for each grasp posture as

$$gf = \frac{V_{avg}}{V_{fist}} \tag{1}$$

where V_{avg} is the average voltage at the object grasping, V_{fist} is average voltage at the clenched fist posture that is allowed the most bent of the finger due to no object supporting the fingers' grasp. The adaptive gamma correction method is gradual increases low intensities and avoids significant reductions in high intensity. The Weight Distribution (WD) function is also used to modify the statistical histogram and minimize its impact. The WD function can be calculated as the formula below

$$pdf_w(l) = pdf_{max}\left(\frac{pdf(l) - pdf_{min}}{pdf_{max} - pdf_{min}}\right)^{gf} \tag{2}$$

where gf is the adjusted parameter from grasp factor, pdf_{max} is the maximum probability distributions function of the statistical histogram and pdf_{min} is the minimum probability distributions function. Based on Eq. (2), the modified cdf is approximated by

$$cdf_w(l) = \sum_{l=0}^{l_{max}} pdf_w(l) / \Sigma pdf_w \tag{3}$$

where the sum of pdf_w is calculated as follows

$$\Sigma pdf_w = \sum_{l=0}^{l_{max}} \Sigma pdf_w(l) \tag{4}$$

Finally, the gamma parameter is modified as follows

$$\gamma = 1 - cdf_w(l) \tag{5}$$

Then, the image will be executed with AGCWD. The weight's function will depend on the average of the voltage output of the flex sensor circuit for each posture and doing normalization

with the voltage output of the clenched fist posture that to be between 0 to 1. The AGCWD offers an automatic image conversion technique that enhances the brightness of darkened images through gamma correction and probabilistic distribution of luminance pixels. This technique uses temporary data from the differences between images to simplify calculations to improve the image input. The flowchart of this procedure method, as shown in Fig. 4.

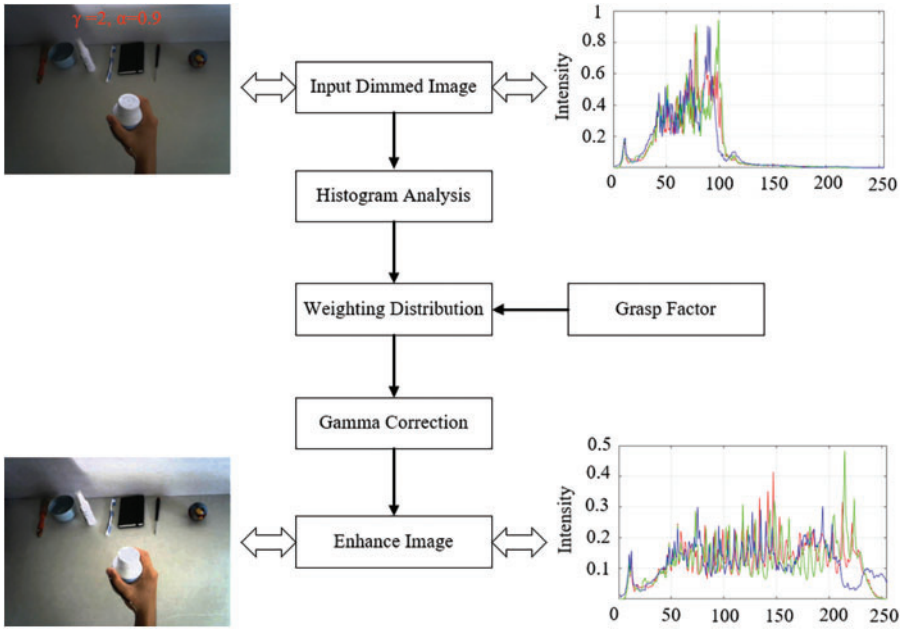

Figure 4: The flow chart of the AGCWD method

3 Experiments

This section first conducts experiments on synthetic data to demonstrate the advantages of the proposed flex sensor approach over traditional learning methods. Then, we apply our AGCWD algorithms to grasp dataset showing their effectiveness for recognizing hand activities. The data images are taken at the various low light condition needed for testing the proposed method, in working and testing the system, initializing the images importing used in the test into the system. Then, the image will be executed with AGCWD. The weighting parameter will depend on the grasp factor (*gf*). The experiment will determine the lighting conditions into three types, follow as 1) gamma adjustment γ as 1 to 6, 2) contrast adjustment α as 0.5 to 1, 3) mix contrast and gamma adjustment as shown in Fig. 5.

3.1 Image Testing Generation Method

The grasp dataset is collected the daytime images as ground truth, and the egocentric camera is recorded in the resolution of 640 × 480 pixels. The volunteer wears the camera with a strap mount belt on the heads. Then we generate a gesture to attain low-light images for each subject [29]. The details of both methods are described as follows.

Figure 5: The generated low light images for testing the proposed method

3.1.1 Gamma Transform

The gamma transform interprets the relationship between digital sensitivity and human eye sensitivity, providing many advantages on the one hand but adding complexity on the other hand. The output I_{out} is defined as

$$I_{out} = A.I_{in}^{\gamma} \tag{6}$$

where A is a constant determined by the maximum pixel intensity in the input image. Instinctively, when $\gamma > 1$, the mapping is weighted to the generation method's lower (darker) grayscale pixel intensity value. A uniform distribution of γ is used to darken the daytime image values from 1 to 6. In the experiment, the darkened image was caused by different values of γ. Daylight images will become darker after gamma conversion. However, gamma conversion with $\gamma > 1$ improves image contrast. However, it may not correspond to the effect of natural light falling.

3.1.2 Contrast Adjustment

The contrast is between an image's light and dark parts, making objects or details within an image more apparent. Reducing the contrast will keep the bright and dark areas close to the original. However, the overall image is getting better and starting to look washed out. The output I_{out} is defined as Eq (7), where α is a ratio between 0.5 and 1, which controls the contrast. The testing image has shown some examples of the resulting image, as Fig. 5.

$$I_{out} = \alpha.I_{in} \tag{7}$$

3.2 Image Enhancement Stage

The images, adjusted to various lighting conditions before YOLOv2, will be improved with AGCWD by using grasp factor to control the different weight distribution parameters for each hand posture. It depends on the voltage output of the flex sensor circuit, which generates the unique individual pattern. The pre-training CNN is used for the feature extractor to separate this uniqueness

for each grasp posture. The principles of calculation for the proposed AGCWD method is to calculate from gamma parameter via probability density to combine the simple forms of the transform-based gamma correction and the traditional histogram equalization. Although the original histogram is not directly used to create image conversion functions and improve image contrast. The AGCWD method can improve the brightness and produces acceptable clear images without restricted contrast, as shown in Fig. 6.

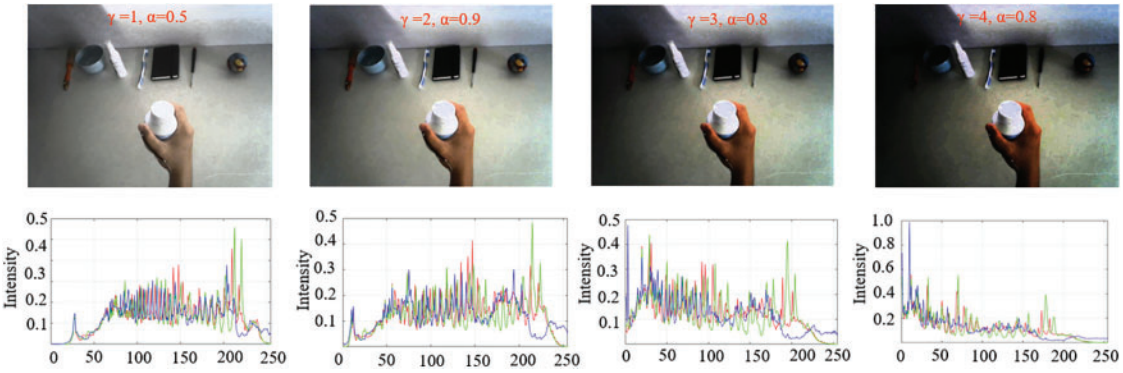

Figure 6: The AGCWD image enhancement output with histogram distributions

3.3 Testing Stage

The testing procedures of the proposed method in various lighting conditions will be divided into two part: The first is the normal light condition or $\gamma=1$, $\alpha=1$, and the second is the adjusted lighting condition (according to the gamma and contrast parameters) to test the robustness of the proposed technique. There will be a simulation of the light in the experiment into 3 cases: First, the contrast is fixed, $\alpha=1$ and then adjust gamma 1 step increments until $\gamma=6$. Second, the gamma is fixed, $\gamma=1$, and adjust the contrast 0.1 step increments from 0.5 until 1. Third, adjust both parameters simultaneously.

4 Results and Discussion

In this section, the proposed algorithm is implemented using AGCWD and flex sensor information have training and testing with various environment models. These were then tested for detection and recognition with deep learning. In the experiment, YOLOv2 is used with three pre-trained CNN for the feature extraction. These are VGG16 [30], ResNet 50 [31] and GooleNet [32]. The network learnable parameters using the stochastic gradient descent with momentum is 0.9, the initial learn rate is 0.001, use a mini-batch size with 16, the learn rate drop factor is 0.1, the learn rate drop period is 10 and the maximum number of epochs for training to 20. The input image is fed into the network processed by attribute extraction to separate the grasps' attributes. The ground-truth labels and drawn hand attributes are used as inputs to supervised learning to train comprehension classifiers for different comprehension classifications. The result of proposed method has a performance evaluation and then compare with the traditional learning methods, which training only the grasp image. The Intersection over Union (IoU) is used to measure the overlap of a predicted *vs.* actual bounding box for the hand grasp [33]. The confidence is obtained by multiplying two items. The first is an object in the pane, and the second is the intersection ratio of box and ground truth. If there is an object in the grid, the first item is 1. Otherwise, it is 0, and the latter is a general intersection ratio. Category conditional probability is

$P_r(Class_i|Object)$, there is an object in the pane, it is the probability of a particular category.

$$Probability = P_r(Class_i|Object)xP_r(object)xIoU_{pred}^{truth} \tag{8}$$

$$Probability = P_r(Class_i)xIoU_{pred}^{truth} \tag{9}$$

where $P_r(Object)$ represents the probability of the object existing in the current grid and IoU_{pred}^{truth} represents the IoU between the predicted box and the actual box. Most bounding boxes below the threshold will be removed. After testing with various lighting conditions, the obtained result from IoU = 0.5 is used to judge the efficiency of the proposed method. The evaluation [34] found that the VGG 16 provides the best results, achieving mean Average Precision (mAP) at score 0.856, which compute the average precision at condition γ=1, α=1 for each grasp class. As for the conditions whose gamma and contrast parameters are adjusted, the efficiency decreases consecutively. However, the adjustment by increasing the gamma will have a more significant impact on performance than adjusting contrast which has only slightly effect compared with to the original image and results in less impact on recognition ability. As shown in Fig. 7.

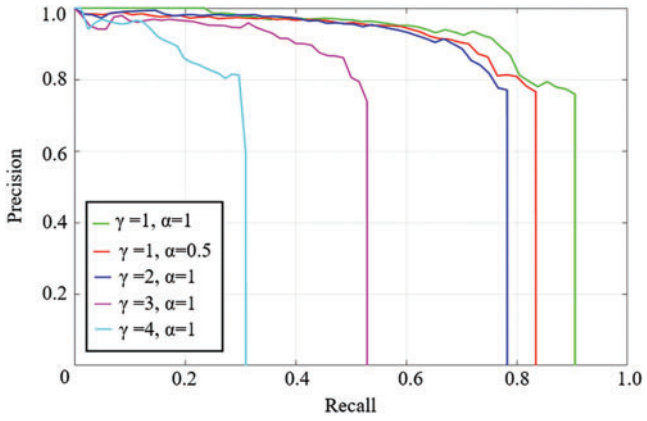

Figure 7: Precision-recall curve of VGG 16 at various ligth conditions

Considering to adjust the only single value at the time, either contrast or gamma, by maintaining the original value of another to observe the change in recognition efficiency in 3 methods: 1) Proposed Method 2**)** Traditional training 3) Fix all $gf = 1$ as the test results are shown below

Fig. 8. The results showed that γ adjustment had a more significant effect on recognizing efficiency than α, with a linear decrease when increased the γ because the adjustment affects the characteristics of the image, the brightness level of each RGB model changes more than the α adjustment. However, the grasp factor (gf) to help determine the weighting parameter of the AGCWD of each hand grasp creates especially image characteristics, allowing DCNN to improve recognition efficiency, which is noticed compared with the fixed grasp factor at 1 in all hand gestures. The proposed method had a higher recognition performance. Furthermore, the contrast and gamma adjustment retain the recognition performance almost similar to the usual light condition $\gamma = 1$, $\alpha = 1$ even testing with a reduction of $\alpha = 0.5$ as Fig. 9.

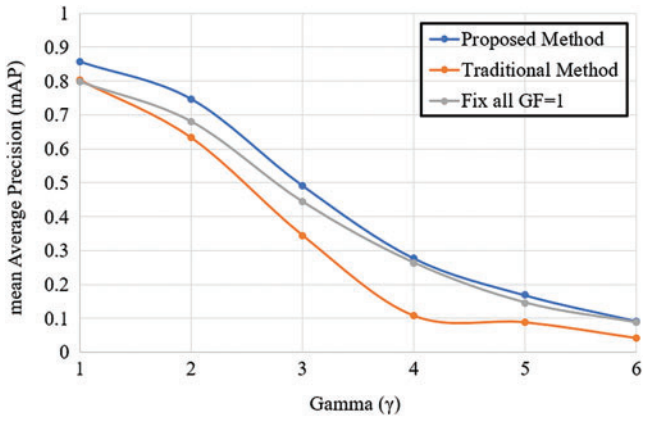

Figure 8: The grasp recognition results with gamma adjustment

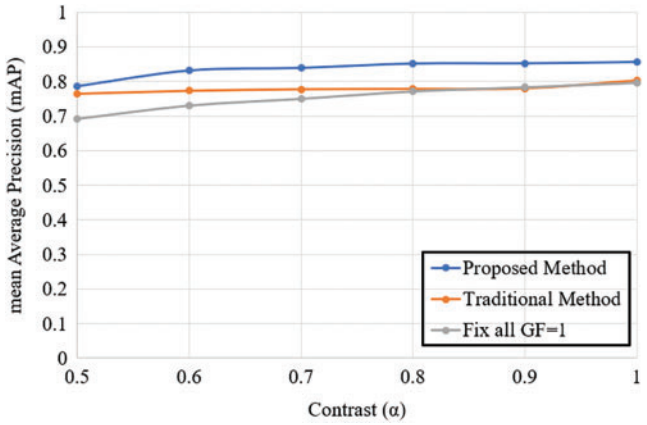

Figure 9: The grasp recognition results with contrast adjustment

The results of testing on the dataset are in Tab. 1. Starting with calculating the recognizing ability of each grasp in the dataset. Then, the overall grasp range score is calculated as the simple sum of the handle ability for each object. Therefore, the high accuracy rated such as Large Diameter, Power Sphere, and Tripod because the arrangement of fingers and the shape of the gripping objects are very different, which can recognize the hand gestures more than 0.9. Which excellently detected and recognized for all three pre-trained CNN. Also, the hand grasp is Thump2 Finger, Fix Hook, Tripod, and Lateral Tripod; these are difficult to recognize because the arrangement of fingers and the shape of the gripping objects are the same in some viewpoints. Thus, the flex sensor parameter can solve that problem. The flex sensor is attached to the thumb to provide information on the movement, and that is another parameter that enhances the ability to differentiate each hand grasp. The improved grasp recognition of flex sensor results is shown in Figs. 8 and 9. The results of normal conditions at $\gamma = 1$ and $\alpha = 1$ showed that recognition using YOLOv2 was the high performance in more than 0.8, there is an improvement of about 6% compared with traditional learning, especially VGG16 pre-trained CNN model that is highest recognition result.

Table 1: Accuracy rate comparison on grasp recognition at IoU = 0.5

Condition	Proposed method			Traditional learning		
	VGG16	ResNet 50	GoogleNet	VGG16	ResNet 50	GoogleNet
γ=1, α=1	86.28	80.84	76.51	81.21	77.99	70.95
γ=1, α=0.8	85.78	79.35	76.14	80.84	77.87	70.70
γ=1, α=0.6	83.44	76.76	74.04	79.72	76.63	70.45
γ=2, α=1	75.23	64.03	69.47	70.33	70.95	66.01
γ=2, α=0.8	69.84	60.69	67.49	69.34	69.96	65.76
γ=2, α=0.6	58.22	53.64	59.46	69.22	69.83	65.51
γ=3, α=1	50.43	40.42	55.37	34.73	45.61	49.20
γ=3, α=0.8	35.72	35.11	49.94	33.99	44.87	48.70
γ=3, α=0.6	27.69	28.43	43.26	33.49	44.01	48.45

When the image becomes darker by increasing γ, this will result in a faster reduction in recognition compared to adjust the contrast. For example, the testing condition $\gamma = 2$, $\alpha = 1$ decreases the recognition effect by about 10% in traditional learning, but the proposed method was able to keep the results as satisfactory as 75%. The recognizing reduction when compared between gamma and contrast adjustments, if we consider the image histogram of contrast adjustment was found to remain the similar, but it is shifted to the left more, making the image only darker. However, the characteristics are much the similar than adjusting the gamma, and an obvious example is condition $\gamma = 1$, $\alpha = 0.6$ the recognition result is still close to the condition $\gamma = 1$, $\alpha = 1$. Furthermore, when adjusting both, such as the condition $\gamma = 2$, $\alpha = 0.8$, the recognition result of VGG 16 was 70%, better than the traditional learning. However, when the γ was increased by 1 step to $\gamma = 3$, $\alpha = 0.8$, it was found that the recognition rate effect declined sharply. Evidence indicates that the proposed method has provided an increase in recognition efficiency of approximately 6%. It works great at $\gamma < 2$ and $\alpha > 0.5$. The VGG16 pre-trained CNN gives the best recognition results compared to GoogleNet and ResNet 50.

5 Conclusion

This paper showed how to detect and recognize hands grasp in egocentric vision by combining a flex sensor with image enhancement and the YOLOv2 architecture. This technique has combined the AGCWD and voltage output of the flex sensor circuit to address the low illumination condition, which makes it very difficult to perform detection and recognition for hand grasp, another of fingers arrangement and the shape of the gripping objects are the same in some viewpoint that causes difficulty to detect the hand in the scene. The flex sensor is attached to the thumb to provide information, which is a parameter to enhance the ability to differentiate each grasp posture. The proposed method can improve the grasp recognition rate from various condition models. The experimental results demonstrate that providing method can effectively grasp a wide range of different objects. The results also show that the technique can improve hand grasp recognition compared with traditional learning methods, increasing the recognition efficiency by approximately 6%, showing outstanding results at $\gamma < 2$ and $\alpha > 0.5$. This research is the first to utilize the flex sensor for hand grasp detection in egocentric systems, the proposed method has a not wide operating range. The future of work in egocentric system,

we are currently investigating a methodology to transform the flex sensor signal to image for multi-input CNN, which will improve the performance to expand the operating range at a wider luminance.

Acknowledgement: The authors would like to express sincere gratitude to the Signal Processing Research Laboratory, the Faculty of Engineering, Rajamangala University of Technology Thanyaburi, for insight and expertise.

Funding Statement: This research is supported by the National Research Council of Thailand (NRCT). NRISS No. 144276 and 2589488.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Xu, W. Cai, Y. Li, J. Zhou and L. Wei "Accurate hand detection from single-color images by reconstructing hand appearances," *Sensors*, vol. 20, no. 192, pp. 1–21, 2020.

[2] S. Huang, W. Wang, S. He and R. Lau, "Egocentric temporal action proposals," *IEEE Transaction on Image Processing*, vol. 27, no. 2, pp. 764–777, 2018.

[3] Y. Yan, E. Ricci, G. Liu and N. Sebe, "Egocentric daily activity recognition via multitask clustering," *IEEE Transaction on Image Processing*, vol. 24, no. 10, pp. 2984–2995, 2015.

[4] S. Noor and V. Uddin, "Using context from inside-out vision for improved activity recognition," *IET Computer Vision*, vol. 12, no. 3, pp. 276–287, 2018.

[5] Z. Wang, Z. Li, B. Wang and H. Liu, "Robot grasp detection using multimodal deep convolution neural networks," *Advances in Mechanical Engineering*, vol. 8, no. 9, pp. 1–12, 2016.

[6] Q. Gao, J. Liu and Z. Ju, "Robust real-time hand detection and localization for space human-robot interaction based on deep learning," *Neurocomputing*, vol. 390, pp. 198–206, 2020.

[7] O. Mazhar, B. Navarro, S. Ramdani, R. Passama and A. Cherubini, "A real-time human-robot interaction framework with robust background invariant hand gesture detection," *Robotics and Computer Integrated Manufacturing*, vol. 60, pp. 34–48, 2019.

[8] M. Cai, K. M. Kitani and Y. Sato, "An ego-vision system for hand grasp analysis," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 524–535, 2017.

[9] G. Kapidis, R. Poppe, E. V. Dam, L. P. J. J. Noldus and R. C. Veltkamp, "Egocentric hand track and object-based human action recognition," in *Proc. IEEE Smart World*, Leicester, UK, pp. 922–929, 2019.

[10] S. Singh, C. Arora and C. V. Jawahar, "First person action recognition using deep learned descriptors," in *Proc. CVPR*, Las Vegas, NV, USA, pp. 2620–2628, 2016.

[11] M. Ma, H. Fan and K. M. Kitani, "Going deeper into first-person activity recognition," in *Proc. CVPR*, Las Vegas, NV, USA, pp. 1895–1903, 2016.

[12] M. Wang, C. Luo, B. Ni, J. Yuan, J. Wang *et al.,* "First-person daily activity recognition with manipulated object proposals and non-linear feature fusion," *IEEE Trasactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2946–2955, 2018.

[13] M. Cai, F. Lu and Y. Gao, "Desktop action recognition from first-person point-of-view," *IEEE Transaction on Cybernetics*, vol. 49, no. 5, pp. 1616–1628, 2019.

[14] H. A. Ghafoor, A. Javed, A. Irtaza, H. Dawood, H. Dawood *et al.,* "Egocentric video summarization based on people interaction using deep learning," *Mathematical Problems in Engineering*, vol. 2018, pp. 1–12, 2018.

[15] S. Sultan, A. Javed, A. Irtaza, H. Dawood, H. Dawood *et al.,* "A hybrid egocentric video summarization method to improve the healthcare for Alzheimer patients," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, pp. 4197–4206, 2019.

[16] G. Wilson, D. Jones, P. Schofield and D. J. Martin, "Experiences of using a wearable camera to record activity, participation and heath-related behaviours: Qualitative reflections of using the Sensecam," *Digital Health*, vol. 2, pp. 1–11, 2016.
[17] B. Tekin, F. Bogo and M. Pollefeys, "H+O: Unified egocentric recognition of 3D hand-object poses and interactions," in *Proc. CVPR*, Long Beach, CA, USA, pp. 4506–4515, 2019.
[18] M. Lu, Z. Li, Y. Wang and G. Pan, "Deep attention network for egocentric action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3703–3713, 2019.
[19] S. C. Huang, F. C. Cheng and Y. S. Chiu, "Efficient contrast enhancement using adptive gamma correction with weighting distribution," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1032–1041, 2013.
[20] Q. Zhou, J. Qin, X. Xiang, Y. Tan and N. N. Xiong, "Algorithm of helmet wearing detection based on AT-YOLO deep mode," *Computers, Materials & Continua*, vol. 69, no. 1, pp. 159–174, 2021.
[21] J. Kim and J. Cho, "Exploring a multimodal mixture-of-YOLOs framework for advanced real-time object detection," *Applied Sciences*, vol. 10, no. 612, pp. 1–15, 2020.
[22] S. Seong, J. Song, D. Yoon, J. Kim and J. Choi, "Determination of vehicle trajectory through optimization of vichicle bounding boxes using a convolutional neural network," *Sensors*, vol. 19, no. 4263, pp. 1–18, 2019.
[23] C. B. Murthy, M. F. Hashmi, G. Muhammad and S. A. A. Qahtani, "YOLOv2PD: An efficient pedestrian detection algorithm using improved YOLOv2 model," *Computers, Materials & Continua*, vol. 69, no. 3, pp. 95–119, 2021.
[24] T. Feix, J. Romero, H. B. Schmiedmayer, A. M. Dollar and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transaction on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2016.
[25] J. Z. Zheng, S. D. L. Rosa and A. M. Dollar, "An investigation of grasp type and frequency in daily household and machinic shop tasks," in *Proc. Robotics and Automation*, Shanghai, China, pp. 4169–4175, 2011.
[26] I. M. Bullock, J. Z. Zheng, S. D. L. Rosa, C. Guertler and A. M. Dollar, "Grasp frequency and usage in daily household and machine shop tasks," *IEEE Transaction on Haptics*, vol. 6, no. 3, pp. 296–308, 2013.
[27] A. Sreejan and Y. S. Narayan, "A review on applications of flex sensors," *International Journal of Engineering Technology and Advanced Engineering*, vol. 7, no. 7, pp. 97–100, 2017.
[28] G. Cotugno, K. Althoefer and T. Nanayakkara, "The role of the thumb: Study of finger motion in grasping and reachability space in human and robotics hands," *IEEE Transaction on Systems, Man and Cybernetics: Systems*, vol. 47, no. 7, pp. 1061–1070, 2017.
[29] G. Li, Y. Yang, X. Qu, D. Cao and K. Li, "A deep learning based image enhancement approach for autonomous driving at night," *Knowledge-Based Systems*, vol. 213, pp. 1–14, 2021.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, pp. 1–14, 2015.
[31] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, pp. 770–778, 2016.
[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.,* "Going deeper with convolutions," in *Proc. CVPR*, Boston, MA, USA, pp. 1–12, 2015.
[33] R. Padilla, S. L. Netto and E. A. B. Silva, "A survey on performance metrics for object-detection algorithms," in *Proc. IWSSIP*, Rio de Janeiro, Brazil, pp. 237–242, 2020.
[34] A. Furnari, S. Battiato and G. M. Farinella, "How shall we evaluate egocentric action regcognition?," in *Proc. ICCVW*, Venice, Italy, pp. 2373–2382, 2017.