Proposing Gesture Recognition Algorithm Using HOG and SVM for Smart-Home Applications

Phat Nguyen Huu(✉), Tan Phung Ngoc, and Hoang Tran Manh

School of Electronics and Telecommunications, Hanoi University of Science and Technology (HUST), Hanoi, Vietnam
{phat.nguyenhuu,hoang.tranmanh}@hust.edu.vn,
Tan.PNCA190157@sis.hust.edu.vn

Abstract. Gesture recognition is one of the key aspects of robot communication systems. There are many image recognition techniques that are being developed to use in many different intelligent systems. In the paper, we perform the image processing techniques that include artificial intelligence technologies and deep learning in gesture recognition to apply for smart-home systems. We propose the gesture recognition model including the histogram of oriented gradient (HOG) and support vector machine (SVM) detection algorithms combining the kernel correlation filter (KCF) algorithm for tracking objects and a multi-layer convolution neural network (CNN) for classifications. The results show that the proposal algorithm is applicable for real environments with accuracy up to 99% per 6 seconds.

Keywords: Gesture recognition · Histogram of oriented gradient · Support vector machine · Kernel correlation filter · Convolution neural network

1 Introduction

With the development of automation applications, human-computer interaction (HMI) systems gradually become important and significant. HMI with smart-home control system is one of the core issues to create accuracy that is convenience and friendliness with civilian applications in the direction of getting closer to communication nature. Instead of living control, using voice controls or via gestures is developing.

Using gestures in HCI systems is an effective idea that helps people communicate in the real world. Gesture is the act of one or more combinations of different body parts that imply information. Hand gestures are commonly used body language comparing to other parts because of their flexibility. The various shapes and postures of hand bring a large amount of information to communicate in real environments.

N.-S. Vo et al. (Eds.): INISCOM 2021, LNICST 379, pp. 315–323, 2021.
https://doi.org/10.1007/978-3-030-77424-0_26

Therefore, we focus on new technologies in smart-home control using gesture language with limited hardware cost and quality in the paper. The goal of paper is to propose a gesture recognition algorithm using histogram of oriented gradient (HOG) and support vector machine (SVM) that is able to detect while minimizing noise and processing speed and reducing errors. Our static gestures are selected as on, off, up, and down. The dynamic gestures are as follows:

- Toggle state switching is hand from spread state upwards to re-grip state,
- Up order is hand from outstretched state up to left,
- Down order is hand from outstretched state up to right.

The rest of the paper includes five parts and is organized as follows. Section 2 presents several related works. Section 3 presents the proposal algorithm. Section 4 will evaluate the proposal model and analyze the results. In final section, we give conclusions and future research directions.

2 Related Work

The motion recognition problem can be solved by combining basic image processing, namely object detection, recognition, and tracking. There are many image processing algorithms that have been developed for detection and recognition. We divide into main groups, namely advanced machine learning and deep learning techniques.

Machine learning (ML) techniques are general terms commonly using with basic feature extraction methods from original data and then combining such as SVM, decision tree, nearest-neighbor to train identity models. There are several extraction techniques as follows

1. Viola – Jones's target detection technique [13] is the technique in real-time target detection based on Haar feature extraction. The technique is used in face detection.
2. Scale-invariant feature transform (SIFT) [10]: The special feature of SIFT is scale-invariant that will give stable results with different aspect ratios of image. Besides, it can be said that this algorithm is rotation-invariant to ensure the result with different rotation of object.
3. HOG [9] is calculated on a dense grid of cells and normalized the contrast among blocks to improve accuracy. It is used to describe the shape and appearance of objects.

Advanced deep learning techniques use multi-layered convolutional neural networks (CNNs) for training on labeled datasets. Several deep learning techniques are applied for object detection and recognition as follows:

1. Region proposals (R-CNN, Fast R-CNN, Faster R-CNN, cascade R-CNN) [8]: The method proposes areas capable of containing the object and performs identification to save computational capacity.

2. Single shot multibox detector (SSD) [3] such as YOLO, Refinedet: The main idea of SSD comes from using bounding boxes by pre-initializing boxes at each location on image. The SSD will compute and evaluate information at each location to see if there are any objects. If there are any objects, it will determine which one it is. Based on the results, SSD will compute an amalgamation box covering the object.

Since the detection and recognition algorithms require a large amount of computation and the accuracy is not able to reach 100%, the object tracking techniques for gesture recognition are also widely applied to ensure the continuous real-time recording of location and to avoid interference in multi-subject environments. There are many tracking algorithms for image processing such as BOOSTING [2], MIL, KCF [5], TLD, MEDIANFLOW [6], GOTURN [14], MOSSE [1], CSRT [7,12]. Therefore, we are able to select the suitable algorithm.

3 Proposal Algorithm

The proposal solution for motion detection is developed and performed based on three main problems, namely hand detection, grip, and position. In the paper, we propose the gesture recognition algorithm using HOG and SVM based on our previous results [11] as shown in Fig. 1.

Details of the steps are as follows:

1. **Pre-processing** is an essential step in reducing noise and increasing reliability for computation. In the paper, processing steps include in resizing and synchronizing images, balancing histogram to reduce the light effect, eliminating noise by median filter.
2. **Hand region detection** is performed based on the HOG characteristic extraction [4] combining with SVM classifier. The HOG characteristics are the shape of object characterizing by the distribution of intensity and direction of pixel value (a gradient vector). The gradient vector represents the change of luminance pixel when it is in the corner and edge areas of object. Therefore, the HOG feature is an effective choice for hand posture.
   The HOG method is to use information about the distribution of intensity gradients or edge directions to describe local objects of image. Its operators are implemented by dividing an image into sub regions calling cells. We compute histograms for points within cell. Combining them together, we get a representation of original image. To enhance recognition performance, local histograms is able to be normalized by calculating an intensity threshold in area larger than the cell (blocks) and using them to normalize. The result is a feature vector that is more invariant to changes in lighting conditions. Details of the steps for extracting the HOG features are as shown in Fig. 2.
   To detect hands, the next step is to use SVM algorithm for classification or regression problems. We use the pyramid method and sliding window to address the areas of image obtaining at different scales. Finally, the results are processed by the non-maximum suppression (NMS) algorithm to eliminate unreliable or overlapping area.

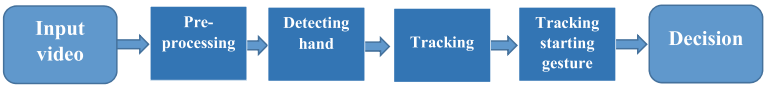

Fig. 1. Diagram of the proposal system.

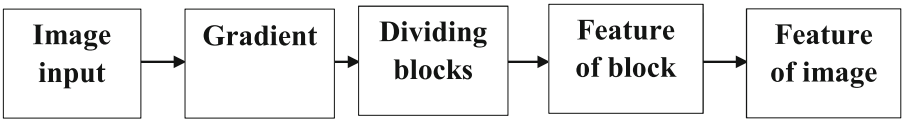

Fig. 2. Details of steps of HOG.

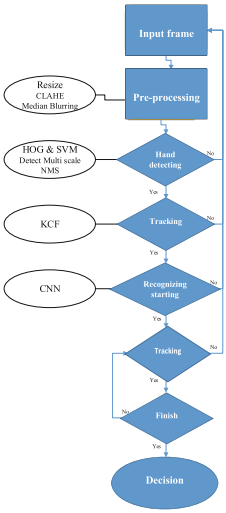

Fig. 3. Details of proposal algorithm.

3. **Tracking object** uses KCF technique. KCF method includes the following steps:
   - Determination of grip area: can be the initial user-defined area or an area detecting by the previous frame,
   - Description of features: defines the characteristics of image area,
   - Regression training: The detecting ROI features will be added to form a dataset including past and present features for training model,
   - The results after regression training is a new model. The model is the basis for the next step.
4. **Object classification** is one of the typical problems of image processing that has achieved many achievements by applying deep learning techniques to multi-layered CNN. We select a CNN model including 2-dimensional convolution layers (Conv2D) with sizes 32, 64, 128 (pyramid structure), Relu activation function, and 3 layers of MaxPooling2D, respectively. With a flatten layer, the output is a dense layer of size 128 and 4. The model output is

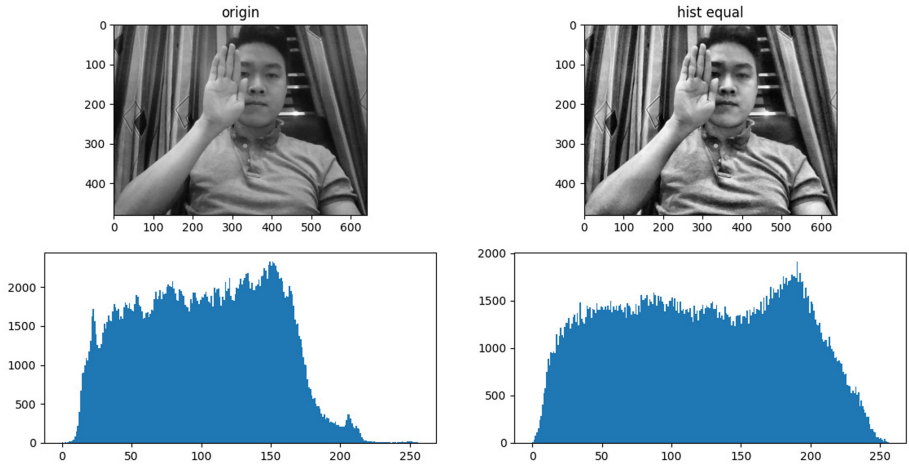

Fig. 4. Image results of adaptive histogram balancing process.

a list of expectations where each of them represents a classification. In the paper, we classify four postures for the four classifier outputs. Subject is classified based on ROI of image.

After detecting and recognizing the starting of gesture, subsequent frames will be continuously updated. The human hand will be categorized to find out what kind of ending or exiting gesture when the grip is lost.

As a result, we are able to implement each part of algorithm. The final step is to incorporate the algorithms into real-time frames from computer camera. Combining the algorithms increases the processing time since the program will not perform continuous object detection. We only perform at three frames to detect for one times. Details of the proposed algorithm are shown in Fig. 3.

4 Simulation and Results

4.1 Setup

Our program is built by Python on Jupyter Notebook platform. We use the OpenCV and Dlib libraries. The simulation is performed on a personal computer with a Core i5 4310 CPU configuration at 2 GHz without using GPU.

4.2 Training Result

Firstly, the images are pre-processed to increase reliability and accuracy as shown in Fig. 4.

The training samples will be performed on real time for smart indoor application. In the paper, the training sample is captured from the user of actual image.

Fig. 5. Results of samples using target detection.

The dataset with hand position is selected for training. Results are shown in Fig. 5.

For categorization of training dataset, we use 1000 images for one time. These images are only ROI regions containing the target without other details. Results are shown in Fig. 6.

Fig. 6. Results of images for the training dataset.

4.3 Detection Results

Besides, we perform to evaluate for detection results by HOG and SVM using images with many different backgrounds. The results are shown in Tables 1, 2,

Fig. 7. Result of detecting hand error.

Table 1. Result of training model.

Posture	Number of samples	Number of incorrect identification	Number of unrecognized samples	Identification time (Seconds/images)	Accuracy
Posture spread your arms up	1000	5	85	0.063471447	91.00%

Table 2. Result of detection model for **static gesture**.

Posture	Number of samples	Number of incorrect identification	Number of unrecognized samples	Identification time (Seconds/images)	Accuracy
Holding hands	1000	2	2	0.069411886	99.60%
Posture spread your arms up	1000	0	5	0.072152341	99.50%
Posture spread your arms left	1000	8	3	0.066520801	98.90%
Posture spread your arms right	1000	7	0	0.067326031	99.30%

and 3. When the background is constantly changing, the algorithm will fail. Results shown in Fig. 7.

According to CNN model above, the classifier output will be a 4-element sequence where each element represents a label. The elements have a value from 0 to 1. When the representative value of label is close to 1, the result of classifier will be more likely to be that label. We choose a limit of 0.85 that means the

Table 3. Result of detection model for **dynamic gesture**.

Posture	Number of test	Number of incorrect identification	Number of unrecognized samples	Accuracy
Switch state (on/off)	30	0	2	93.33%
Increasing	30	1	3	86.67%
Reducing	30	2	1	90%

label will be selected when the corresponding value is the largest and greater than 0.85. If there is no label with a corresponding value greater than 0.85, the result is counted as unrecognizable. In case, the result is counted as false identification.

We found that due to the limitations of experimental gesture samples, the effective proposal has not been completely evaluated. However, it is possible to evaluate through the accuracy of hand position detection steps and recognition of their starting and ending posture. The problems will be solved in the future.

5 Conclusion

In the paper, we have explored image processing and machine learning techniques to evaluate applicability in gesture recognition systems. Based on the techniques, we have synthesized and performed the gesture recognition model including detection algorithm using HOG and SVM, clinging algorithm by correlation filter with KCF, and the algorithm to classify the hand using CNN. The algorithm has been trained on datasets including labels with 1000 images. As a result, the algorithm ensures real-time processing speed at an acceptable level and high accuracy. Based on the obtaining results, we will develop and apply the method for smart-home applications.

Acknowledgement. This research is carried out in the framework of the project funded by the Ministry of Education and Training (MOET), Vietnam under the grant B2020-BKA-06. The authors would like to thank the MOET for their financial support.

References

1. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2544–2550 (2010)
2. Breiman, L.: Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, November 2000
3. Ning, C., Zhou, H., Song, Y., Tang, J.: Inception single shot multibox detector for object detection. In: 2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW), pp. 549–554 (2017)

4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893 (2005)
5. Dalei, L., Ruitao, L., Xiaogang, Y.: Object tracking based on kernel correlation filter and multi-feature fusion. In: 2019 Chinese Automation Congress (CAC), pp. 4192–4196 (2019)
6. Dattathreya, Han, S., Kim, M., Maik, V., Paik, J.: Keypoint-based object tracking using modified median flow. In: 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), pp. 1–2 (2016)
7. Feng, F., Wu, X., Xu, T.: Object tracking with kernel correlation filters based on mean shift. In: 2017 International Smart Cities Conference (ISC2), pp. 1–7 (2017)
8. Girshick, R.B.: Fast R-CNN. CoRR abs/1504.08083 (2015)
9. Lee, H.-J., Chung, J.-H.: Hand gesture recognition using orientation histogram. In: Proceedings of IEEE. IEEE Region 10 Conference, 'Multimedia Technology for Asia-Pacific Information Infrastructure' (Cat. No.99CH37030), TENCON 1999, vol. 2, pp. 1355–1358 (1999)
10. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
11. Nguyen Huu, P., Quang, T.M., Hoang Lai, T.: An ANN-based gesture recognition algorithm for smart-home applications. KSII Trans. Internet Inf. Syst. **14**(5), 1967–1983 (2020)
12. Torregrosa Olivero, J.A., María Burgos Anillo, C., Guerrero Barrios, J.P., Montoya Morales, E., Gachancipá, E.J., Andrés Zamora de la Torre, C.: Comparing state-of-the-art methods of detection and tracking people on security cameras video. In: 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), pp. 1–5 (2019)
13. Viola, P., Jones, M.: Robust real-time object detection. Int. J. Comput. Vis. - IJCV **57**, 137–154 (2001)
14. Wang, C., Galoogahi, H.K., Lin, C., Lucey, S.: Deep-LK for efficient adaptive object tracking. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 627–634 (2018)