Indonesian Journal of Electrical Engineering and Computer Science
Vol. x, No. x, June 2016, pp. 1 ~ 9
DOI:

VISUAL-BASED FINGERTIP DETECTION FOR HAND REHABILITATION

Dayang Qurratu'aini [1], Ali Sophian*[1], Wahju Sediono[1], Hazlina Md Yusof[1], Sud Sudirman[2]
[1]Department of Mechatronics Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia
[2]School of Computing and Mathematical Sciences, Liverpool John Moores University, Liverpool, UK

*Corresponding author, e-mail: ali_sophian@iium.edu.my

Abstract

This paper presents a visual detection of fingertips by using a classification technique based on the bag-of-words method. In this work, the fingertips are specifically of people who are are holding a therapy ball, as it is intended to be used in a hand rehabilitation project. Speeded Up Robust Features (SURF) descriptors are used to generate feature vectors and then the bag-of-feature model is constructed by K-mean clustering which reduces the number of features. Finally, a Support Vector Machine (SVM) is trained to produce a classifier that distinguishes whether the feature vector belongs to a fingertip or not. A total of 4200 images, 2100 fingertip images and 2100 non-fingertip images, were used in the experiment. Our results show that the success rates for the fingertip detection are higher than 94% which demonstrates that the proposed method produces a promising result for fingertip detection for therapy-ball-holding hands.

Keywords: *Fingertip Detection, SURF, K-mean clustering, Bag of Words*

1. Introduction

Medical applications, including hand rehabilitation for stroke survivors, have benefited from the advances in technology for many years. The exploitation of computer vision in this application field has not been spared and has been the subject of many research works. Although computer vision technology has been advancing rapidly throughout the years, there are still some difficult challenges that relate to vision-based approach for fingertip detection that need to be overcome. The challenges that need to deal with are (1) the non-rigid nature of hands possessing a high degree of freedom that makes it difficult to match various shapes of fingers with a set of images, (2) there is a variety of orientation and appearance of finger; thus it is difficult to detect the shape and posture of the fingers accurately and robustly, and (3) slight differences may lead to substantial error in the case of fingertips that belongs to the same person (Wu & Kang, 2016). These challenges get even more significant when commercial vision systems are used, instead of those of industrial grade.

In this paper, a potential solution using machine learning is to be used in hand rehabilitation. One of the widely practiced rehabilitation exercise is by asking the patient to squeeze a flexible exercise ball in his/her hands repetitively (Jaber; R.; Hewson; F., D.; J., 2012). The balls have various levels of resistance to accommodate the various levels of limitation of the patients' hands. However, one of the challenges is to measure objectively or quantitatively the progress that has been made if any. Machine-vision-based system may offer a non-intrusive way of measurement of fingertip position. Some present rehabilitation is assisted by machine vision based system involves the interaction between human and virtual world. Detection and tracking of fingertip are essential in a recognition of fingertip in a contactless position measurement.

There have been works on fingertip detection using machine vision by other researchers. However, most of the work done are having the fingers fully extended with no hand occlusion. An engine development for fingertip detection in real-time that is targeted at mobile devices for the Natural User Interfaces (NUIs) (Baldauf, Zambanini, Fröhlich, & Reichl, 2011); system development that is capable of detecting fingertip in a reliable manner in complex environment under different light conditions, different scenes without any markers (Yang, Jin, & Yin, 2005); Feng et al. (2012) used Kinect sensor for fingertip detection for writing-in-the-air character recognition system; an approach that allows the detection of hand and fingertip with or without illumination in cluttered background (Brancati, Caggianese, Frucci, Gallo, & Neroni, 2015).

Received -; Revised -; Accepted -

This paper is organized as follows. The related work on fingertip detections is reviewed in Section 2. In section 3, the proposed algorithm for fingertip detection the experimental results are presented and discussed. Finally, the summary of the work is presented in the concluding section.

2. Fingertip Detection Algorithm

2.1 Bag of Words

Bag of words (BoW) model has been used in machine vision for around a decade. The model was originally applied in natural language analysis where a text document is represented in a histogram of words without considering the grammar and the order or the location of the words in the text. The model would build a dictionary consisting the vocabulary of words it has found in the texts that are fed into the model as the input. When it comes to the application in machine vision, the model has been popular due to its simplicity and effectiveness (Li, Dong, Xiao, & Zhou, 2016) and it is also widely known as bag of visual words and bag of features. The same researchers stated that traditionally BoW employs scale-invariant feature transform (SIFT) descriptors that reduces the dimensionality of the feature space.

To build the dictionary, also known as codebook, that consists of the visual words, the technique extracts these visual words from the training images – as illustrated by the flowchart in Figure 1. During the learning stage, a large set of images of different classes are used. From each image, extraction of keypoints is initially carried out. Subsequently, for each keypoint, feature descriptors are established which represent the features of the neighborhood of the keypoint. In the next step, for dimension reduction purposes, these descriptors are clustered into groups, which are called visual words. All the generated visual words from the training images are collected as the codebook, which is equivalent to a dictionary containing the vocabulary of words.

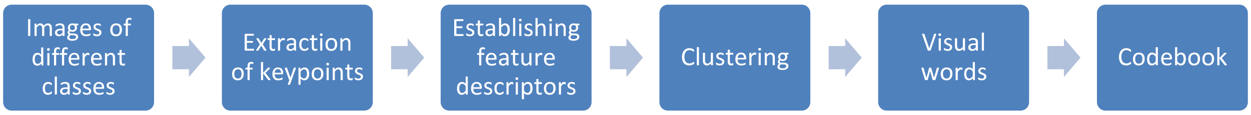


Figure 1 Extraction of features and generation of the codebook

During an image recognition stage, extraction of keypoints, defining feature descriptors and the clustering of the descriptors are carried out in generating the bag of words for the image, which is basically a histogram of the visual words that are present in the image, such as shown in Figure 2.

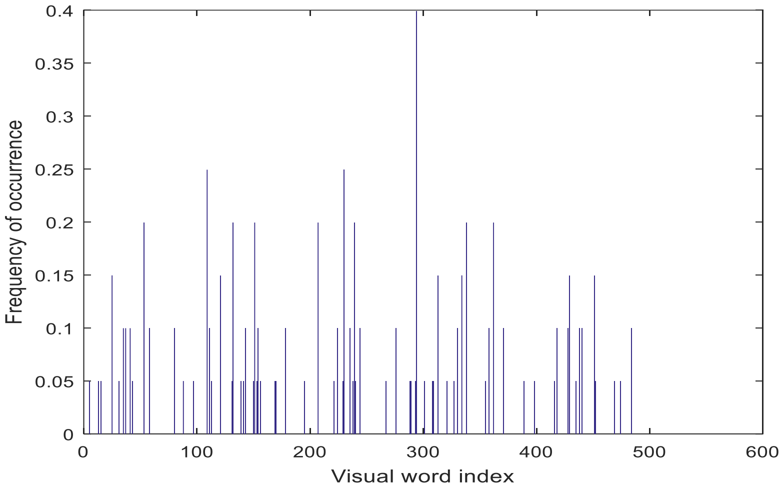


Figure 2 Histogram of Visual Word Occurrences

2.2 Speeded Up Robust Features (SURF)

SURF was introduced by Bay et al (Bay, Ess, Tuytelaars, & Van Gool, 2008), which has proven to be effective and popular, thanks to its repeatability, distinctiveness and relatively fast speed. In various comparative works, such as (Bauer, Sünderhauf, & Protzel, 2007), although SURF has lower number of identified features and slightly lower number of correct matches compared to its predecessor, Scale-Invariant Feature Transform (SIFT), however it manages to perform higher number of correct matches per given time, i.e. it's a faster method. SURF consists basically of four stages, which are integral image generation, approximated Hessian detector, descriptor orientation assignment and descriptor generation (A, Hebbar, Shekhar, Murthy, & Natarajan, 2015) . For achieving high speed, following its popularization by Viola and Jones (Viola & Jones, 2001), this detection uses integral images that reduce the number of mathematical operations. The integral image I_Σ is defined mathematically as the following

$$I_\Sigma(x,y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I(i,j) \tag{1}$$

Hessian matrix is used on the integral image for the localization and scaling of interest points, which particularly looks for blob-like structures where the high determinants of the matrix are present. The Hessian matrix *H(X, σ)* in an image's point X at scale σ is defined as follows:

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{yx}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \tag{2}$$

where $Lxx(X,\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\delta^2}{\delta x^2}\, g(\sigma)$ with the image I at point X, and similarly for $Lxy(X,\sigma)$, $Lyx(X,\sigma)$ and $Lyy(X,\sigma)$.

Following interest point detection, SURF identifies an interest point descriptor around each interest point, which includes the dominant orientation. Each region around the interest point is split into subregions. For each sub-region, a vector is defined by using Haar wavelet responses. These vectors form the descriptor.

2.3 K-Means Clustering

K-mean clustering is one of the methods for image segmentation, which is the classification of an image into distinct groups. Before applying this unsupervised learning technique, an initial enhancement is applied to the image for image improvement. A subtractive clustering method generates centroids which is based on the potential value of data points. In other words, subtractive cluster is used to generate the initial centers which is used in K-mean algorithm for the data points (Dhanachandra, Manglem, & Chanu, 2015)

In this work, it is used to associate the generated descriptor to the right cluster, which is also known as visual world in the bag-of-words technique. By using this clustering, the classification stage, which is the next step, will deal with lower data dimension that, in turn, helps in gaining a higher processing speed.

2.4 Support Vector Machine (SVM)

SVM is a supervised learning method that is used for regression and classification (Dardas & Georganas, 2011). It carries out classification by creating a multi-dimensional hyperplane which divides the data into two groups optimally. This makes SVM classifier model closely associated with neural networks. The SVM classifier model uses a sigmoid kernel function, which is similar to the two-layer perceptron of neural network.

3. Experiment and Data Gathering

3.1 Experimental Setup

In this work, a commercial high-density (HD) Logitech C615 webcam with a resolution of 1920 x 1080 pixels has been used. An example of image captured by the webcam is as shown in Figure 3. Figure 3 is an example of an image of a hand holding a therapy ball. The images are captured while the webcam facing upwards which is facing a light-emitting source in the ceiling. The glare from the light source contributes to the variation of intensity in each captured image.

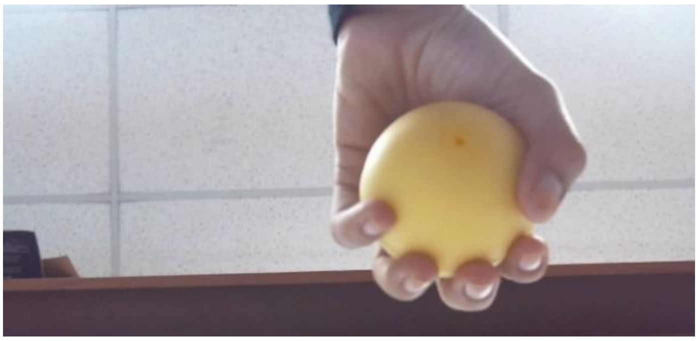


Figure 3 Example of A Captured Image of a Therapy-ball-holding Hand

The setup for the data image gathering is illustrated in Figure 4**Error! Reference source not found.**. The blue circles denote the position of the hands where the distance between adjacent blue circles is approximately 10 cm. The distance Y_{hand} denotes the perpendicular distance of the position of hands to the webcam. The webcam captured two images of the hand at each position.

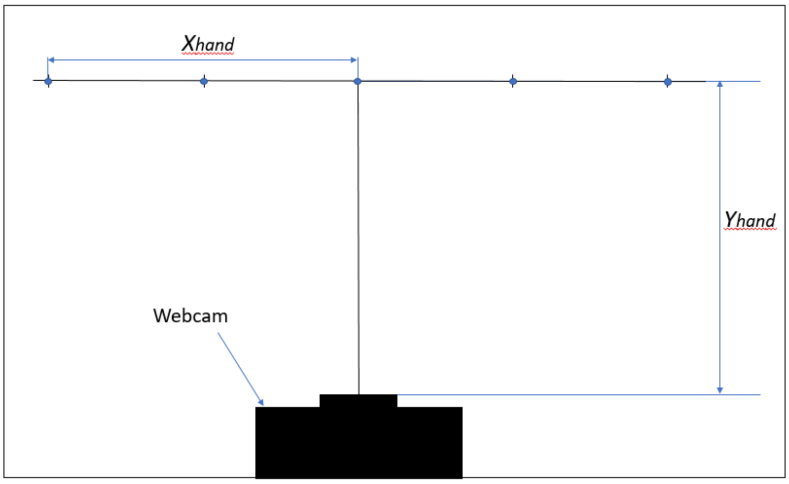


Figure 4 Experimental setup (Blue circles denotes the position of the hands in the experiment)

3.2 Image Data Gathering

For image data gathering, a few sets of images were captured. Different hand sizes, skin colors, and orientations from 10 different individuals (5 male and 5 female) were included in the captured image data. Examples of hands of different orientations are shown in Figure 5.

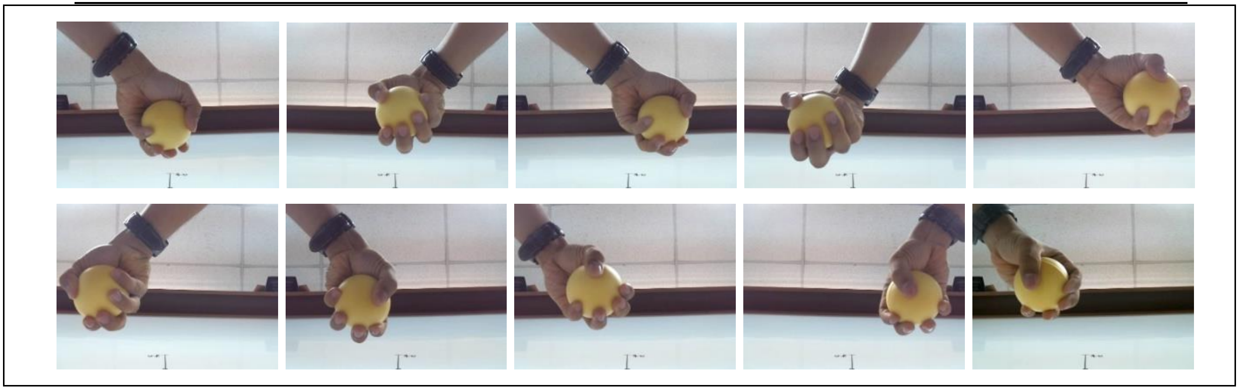
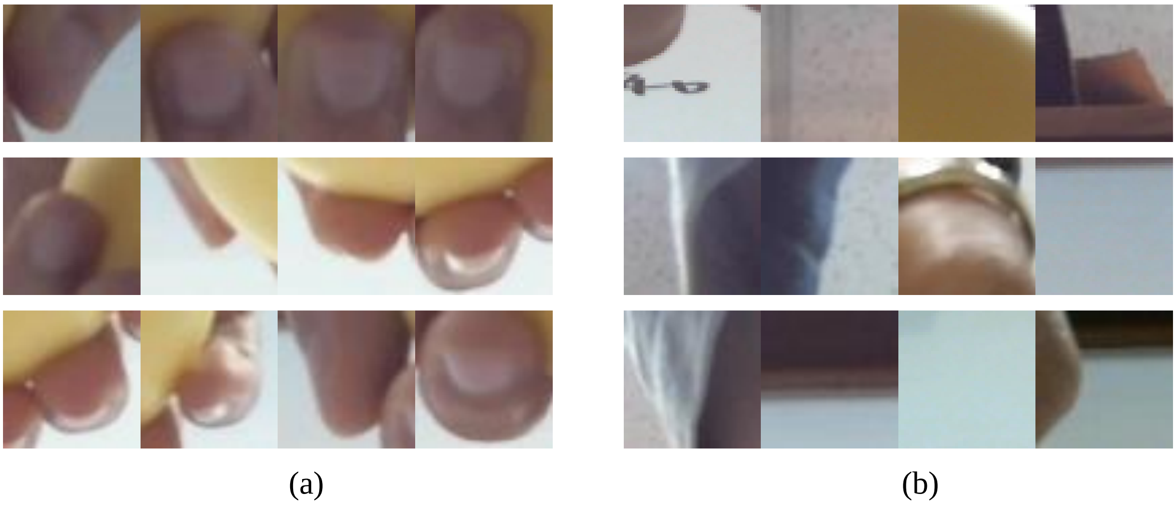


Figure 5 Images of Hand of Different Orientations

Then, the images of the fingertips and non-fingertips were cropped from hand holding ball images. The size of the cropped images for both fingertip and non-fingertip images is 50x50 pixels. Basically, non-fingertip images are images that do not contain any fingertip, instead they contain the background, the ball, the hand wrist, etc. All the cropped images are stored in two separate folders, one of which is for fingertip images and the other is for non-fingertip ones. Examples from both groups of images are shown in Figure 6. A total of 4200 images have been obtained that will be used for both classification training and validation.



(a) (b)

Figure 6 Examples of the images: (a) Fingertip and (b) Non-fingertip

3.3 Detection Validation Testing

By using the image data gathered, the classification machine was then trained and then the detection success rate was evaluated. Figure 7 captures how the experiment and evaluation were done step-by-step.

An array of image sets is constructed based on two main categories; fingertip and non-fingertip. The number of images per category as well as category labels was inspected. If the number of images are unequal per category, then it can be adjusted so that there will be equal number of images per category. The sets are then separated into training and validation sets. The splitting was randomized to prevent the results to be biased.

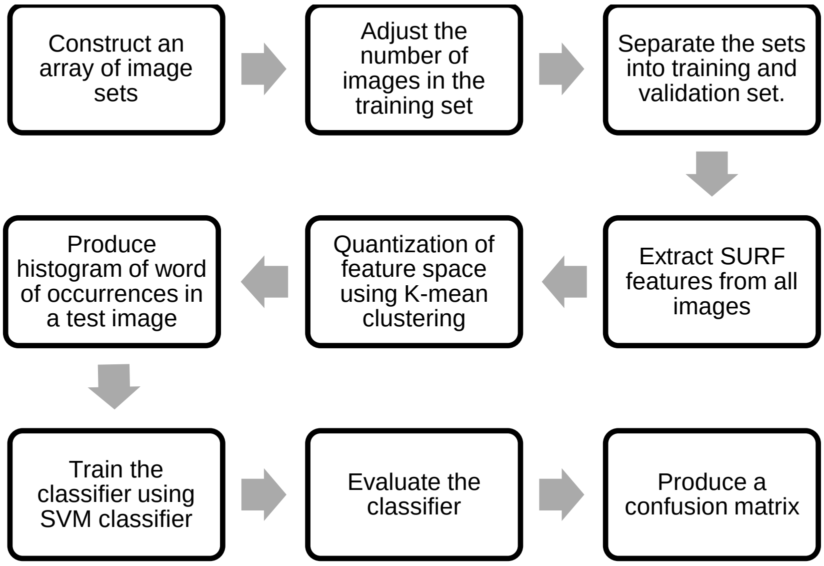

Construct an array of image sets
Adjust the number of images in the training set
Separate the sets into training and validation set.
Extract SURF features from all images
Quantization of feature space using K-mean clustering
Produce histogram of word of occurrences in a test image
Train the classifier using SVM classifier
Evaluate the classifier
Produce a confusion matrix

Figure 7 Category Classification Training

The bag of word technique is from the natural language processing adapted to computer vision. Images do not contain discrete words, therefore, SURF features from each image category must be collected into a visual 'vocabulary'. The visual vocabulary is constructed by reducing the number of features through quantization of feature space using K-mean clustering. Furthermore, the visual word occurrences in an image were counted by constructing a histogram to reduce the representation of an image as shown in **Error! Reference source not found.**. The encoded training images from both categories are fed into a classifier training process.

During the evaluation classifier's performance, the training set was tested and a near perfect confusion matrix was produced. The classifier evaluation step was also performed with validation set, which was not used during the training. The confusion matrix produced is a good indicator of how well the classifier is performing.

4. Experimental Results and Analysis

In this section, we assess the success rate of the detection algorithm. In the experiment, a total of 4200 images was used. The image data sets consist of 2 main subsets such as fingertip, and non-fingertip images, each with a resolution of 50 x 50 pixels. The set images are divided into three categories: training, validation, and unused sets. The splitting of the data sets was randomized to avoid biasing the results.

Table 1 shows the averaged success rate for the detection of fingertip and non-fingertip when the number of validation images varies from 100 to 2000 images. Based on Figure 8 that shows the graphical representation of the data in Table 1, we observed that the highest success rate for the fingertip is 95.6% and for non-fingertip is 92.4%, which is acceptably high. The trend also shows that if the number of training data is increased, a higher success rate can be obtained, especially for the non-fingertip detection.

Table 1 Averaged success rate from validation set

No. of training images	Averaged success rate (%)	
	Fingertip	Non-fingertip
100	93.8	83
500	94.2	87.4
1000	95.6	92
2000	94.4	92.4

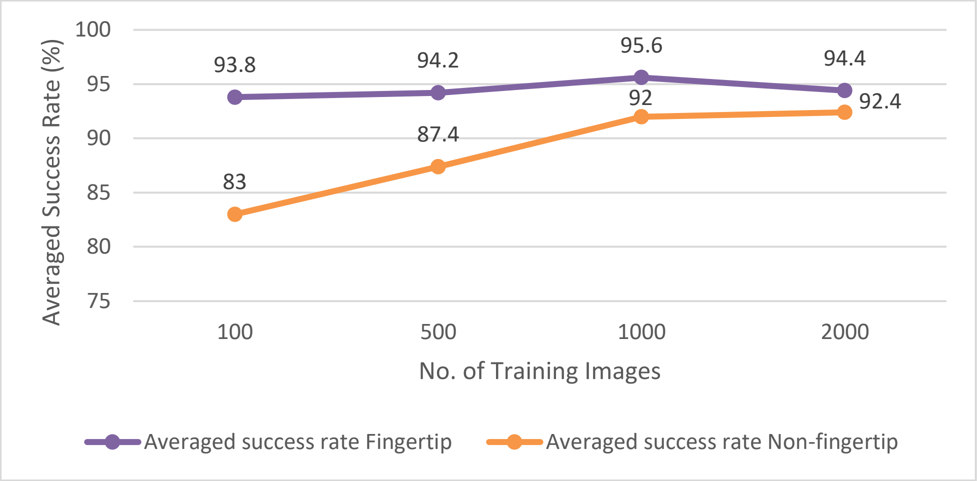
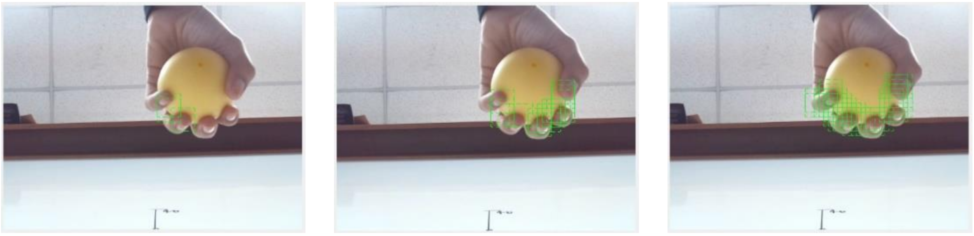


Figure 8 Graph of No. of Training Images vs Averaged Success Rate of the Detection of Fingertip and Non-Fingertip Using Validation Set

A histogram of visual word occurrences was generated during classification training as shown in Figure 2. The histogram forms a basis for training a classifier and for the actual image classification. In other words, it encodes an image into a feature vector. Each encoded training images in each category are fed into a classifier training. In the recognition stage, the image is represented by the visual words that will be distinguished by the classifier.

Figure 9 shows typical results of the detection algorithm when the algorithm is applied scanning over a full image. The green detection box signifies part of the image where fingertips are detected. They show how the improvement has been achieved when a higher number of training images is used. The outputs shows a promising result in the detection.



(a) (b) (c)

Figure 9 Results of the detection algorithm;
(a)No. of training images=100, (b) No. of training images=1000, (c) No. of training images=2000

5. CONCLUSIONS

In this work, it has been shown that the method based on SURF and bag of words has been shown a good performance in detecting fingertips in images where a hand is holding a therapy ball that is normally used in a post-stroke hand therapy. The success rate was generally found to be increased when the number of training images were increased, especially in the correct identification of the non-fingertip, i.e. lower false positive detection rates. The success rate for the fingertip detection reached higher than 94% with the algorithm, which is reasonably high for the therapy applications, despite the use of commercial-grade cameras.

ACKNOWLEDGEMENT

This research has been supported by International Islamic University Malaysia through the research grant RIGS15-151-0151.

REFERENCES

A, V., Hebbar, D., Shekhar, V. S., Murthy, K. N. B., & Natarajan, S. (2015). Two Novel Detector-Descriptor Based Approaches for Face Recognition Using SIFT and SURF. *Procedia Computer Science*, *70*, 185–197. https://doi.org/http://dx.doi.org/10.1016/j.procs.2015.10.070

Baldauf, M., Zambanini, S., Fröhlich, P., & Reichl, P. (2011). Markerless visual fingertip detection for natural mobile device interaction. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 539–544. https://doi.org/10.1145/2037373.2037457

Bauer, J., Sünderhauf, N., & Protzel, P. (2007). COMPARING SEVERAL IMPLEMENTATIONS OF TWO RECENTLY PUBLISHED FEATURE DETECTORS. *IFAC Proceedings Volumes*, *40*(15), 143–148. https://doi.org/http://dx.doi.org/10.3182/20070903-3-FR-2921.00027

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359. https://doi.org/http://dx.doi.org/10.1016/j.cviu.2007.09.014

Brancati, N., Caggianese, G., Frucci, M., Gallo, L., & Neroni, P. (2015). Robust fingertip detection in egocentric vision under varying illumination conditions. *2015 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2015*. https://doi.org/10.1109/ICMEW.2015.7169798

Dardas, N. H., & Georganas, N. D. (2011). Real-time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques.pdf. *IEEE Transaction on Instrumentation and Measurement*, *60*(11), 3592–3607. https://doi.org/10.1109/TIM.2011.2161140

Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Computer Science*, *54*, 764–771. https://doi.org/10.1016/j.procs.2015.06.090

Jaber; R.; Hewson; F., D.; J., D. (2012). Design and validation of the Grip-ball for measurement of hand grip strength. *Medical Engineering & Physics*, *34*(9), 1356–1361.

Li, W., Dong, P., Xiao, B., & Zhou, L. (2016). Object recognition based on the Region of Interest and optimal Bag of Words model. *Neurocomputing*, *172*, 271–280. https://doi.org/http://dx.doi.org/10.1016/j.neucom.2015.01.083

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Wu, G., & Kang, W. (2016). Robust Fingertip Detection in a Complex Environment. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2016.2545401

Yang, D.-D. Y. D.-D., Jin, L.-W. J. L.-W., & Yin, J.-X. Y. J.-X. (2005). An effective robust fingertip detection method for finger writing character recognition system. *2005 International Conference on Machine Learning and Cybernetics*, *8*(August), 18–21. https://doi.org/10.1109/ICMLC.2005.1527822