# DYNAMIC HAND GESTURE RECOGNITION USING SPATIAL-REMPORAL FETURES

**Doan Huong Giang[1,2], Duy Anh Vu[1], Hai Vu[1], Thanh Hai Tran[1]**

[1] International Research Institute MICA HUST - CNRS/UMI - 2954 - INP Grenoble
[2] Industrial Vocational College Hanoi

*{huong-giang.doan, hai.vu, thanh-hai.tran}@mica.edu.vn, duyanhbkhn@gmail.com*

***ABSTRACT** - Hand gesture recognition has been studied for a long time. However it still is a challenge field. Furthermore, hand gesture recognition as a natural way to control devices in smart-house such as television, light, fan, camera, door often require high accuracy. This paper proposes a simple and effective technique to recognize a pre-defined hand gesture set. The defined dynamic gestures are represented in both hand shape changing during temporal dimension and direction of hand movements. In the proposed technique, we analyze Spatial - Temporal features that includes characteristics of a hand gesture such as cycle pattern, different in length, non-synchronization phase, After that we evaluate the proposed technique in term of recognition rate and computational time.*

***Keywords** - Human Computer Interaction, Hand gesture recognition, Spatial-Temporal hand gesture recognition, Principal Componel Analysis.*

## I. INTRODUCTION

Hand gesture recognition has been a very active research topic in the area of computer vision. It has been widely applied in a large variety of practical applications in real world which includes security and surveillance, content-based video analysis, Human-Computer Interaction (HCI) and animation. The mainly HCI, e.g recognizing hand/body gestures to control game consoles [1]. Samsung smart-TV can manipulate TV-functions using dynamic hand gestures. Omron introduces the smart-TV integrated facial and hand recognition. PointGrab [2] proposed an unique solution based on shape and motion recognition including gesture, face and user behavior analytic. Consumer electronics and mobile devices like WiSee system [3]. However, hand gesture recognition is still a challenging problem due to the complexity of hand shapes in gestures, multi-trajectory gestures, background condition and motion blurring and changing of light conditions. Recent research has been motivated to explore more efficient multi-modal gesture recognition methods [4][5][6][7]. Furthermore, how to recognize human gestures using multi-modal information in an efficient way is still an active topic. In this research, we try to solve some following problems:

- The first is proposing a hand gesture database. These gestures are to control equipments. They include on/off, go left and go right, increase, decrease dynamic hand gestures. A gesture represents a cyclic pattern of hand shape during hand movements. Each gesture can be implemented by different people therefore changing hand poses and its velocity is a critical issue.
- We propose to use multi-modal data for hand gestures recognition. The feature extraction consists of both spatial and temporal features. Firstly, we analyze spatial features of hand shape through a PCA model. The hand-path is extracted based on good-features points which are detected and tracked between frame-by-frame. The hand-path presents direction of hand. Our classification scheme evaluates similarity in term of both hand-shape and hand-path.

The general framework of the proposed approach is illustrated in Fig. 1. The rest of paper is organized as follows: Sec. II. briefly survey related works. Sec. III. describes the proposed framework. Sec. IV. describes the experimental results and finally Sec. V. concludes and suggests further research directions.

## II. RELATED WORKS

In term of the deploying applications using hand gestures, [8] proposes a static hand language recognition system to support the hearing impaired people; [9] uses hand postures to control a remote robot in mechanical systems; Similar systems have been deployed for game simulations such as [5][10]. The fact, there are uncountable solutions for a vision-based hand posture recognition system. Readers can refer good surveys such as [12][13] for technique details. Roughly speaking, according to the hand gesture recognition technique, some methods has been implemented that are Neural Network [14], Hidden Markov Models (HMMs) [15][16][17][18], Dynamic Time Warping (DTW) [19][20] and Conditional Random Fields (CRFs) [21][22][23], Or according to features combining unitizing that are none, late fusion [24] or early fusion. Another hand, according to the data inputs that the existing methods of gesture recognition or action recognition can be roughly divided into four categories: RGB video based, depth video based, skeleton data based and multi-modal data based.

In this study, we pursue a hand posture recognition system for controlling devices (e.g., televisions, lighting systems) in a smart-room. Therefore, we briefly survey recent trends that feasibly deploy to home appliances. Microsoft Xbox-Kinect is a success commercial product recognizing hand/body gestures to control game consoles [1]. Many technology companies launch smart-devices using like the Kinect sensors (e.g., Asus Kinect, softKinect). For instance, Samsung smart-TV can manipulate TV-functions using dynamic hand gestures. Omron introduces the smart-TV integrated facial and hand recognition. PointGrab [2] proposed an unique solution based on shape and motion recognition including gesture, face, and user behavior analytic. Increasingly, in-air gesture recognition is being incorporated into consumer electronics and mobile devices like WiSee system [3]. WiSee focuses on gesture recognition and shows how to extract gesture information from wireless transmissions. It promises new trend for home appliances because this technique can operate in non-line-of-sight scenario, that is a limitation of the vision-based system.

Controlling equipment always requires high accuracy. Normally, setting the commands to control that is cycle pattern and repetitive. In the fact, the user does not implement the standard dynamic hand gestures. There someone does not implement the true time (faster and slower) or another one does not implement standard phases (start, implement, stop). Different from above works, in this study we focus on resolve problems with our database. The proposed system effective combine the spatial and temporal features to recognize the dynamic hand gestures that changes both hand shape and temporal.

## III. PROPOSED APROACH

### *A. Proposed framework*

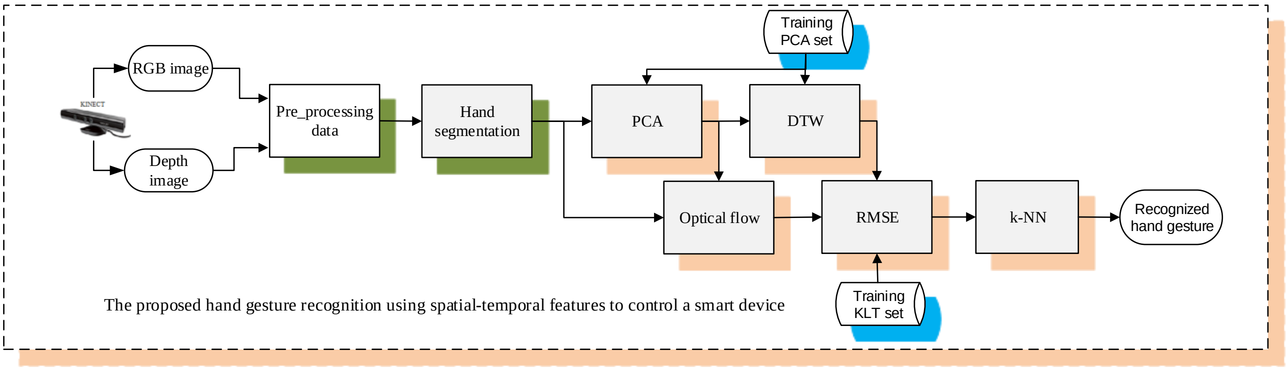


**Figure 1.** Proposed framework for hand gesture recognition

The smart room in MICA-HUST includes television, fan, camera and door. To control those equipments, we need mapping some common commands to hand gestures. Such common commands are Turn on/off, increase/decrease (volume), up/down (channel). We then design a new dataset of five dynamic hand gestures corresponding to these commands. Those dynamic hand gestures are cyclical and repeating pattern. Given a sequence of consecutive frames, we have to determine the label of gesture. We propose a framework that composes of the following main components is shown in Fig. 1. Pre-processing step does depth and RGB calibration because these informations are un-calibrated from the Kinect sensor. Hand segmentation step detects and segments hand region for every frame. Dynamic hand gestures representation by spatial-temporal features that extracts spatial and temporal features for representing dynamical hand gestures. For classification, a Knn technique is applied to predict the label of gesture. Details are presented in the following sections.

### *B. Pre-processing data*

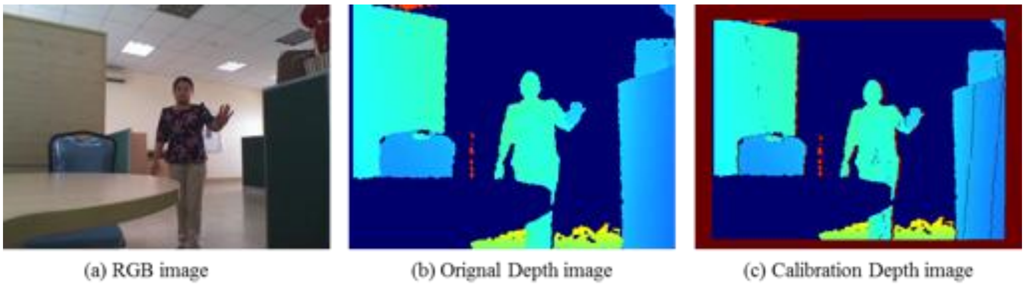
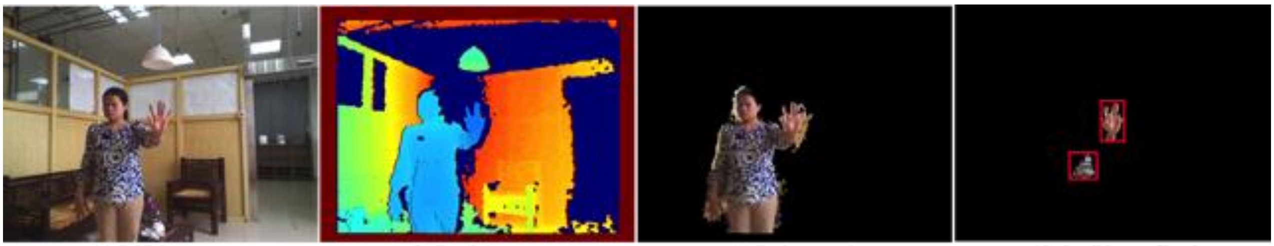

(a) RGB image (b) Orignal Depth image (c) Calibration Depth image

**Figure 2.** Calibration RGB-D images from the Kinect sensor

Depth and RGB images from the Kinect sensor are not measured from the same coordinates. In the past, there were many researches have mentioned this problems as [13]. In our work, we utilized calibration method of Microsoft to repair the depth images. The result showed in Fig. 2a is RGB image$(I)$, Fig.2b is original depth image and Fig. 2c show result of calibration depth image $(D)$.

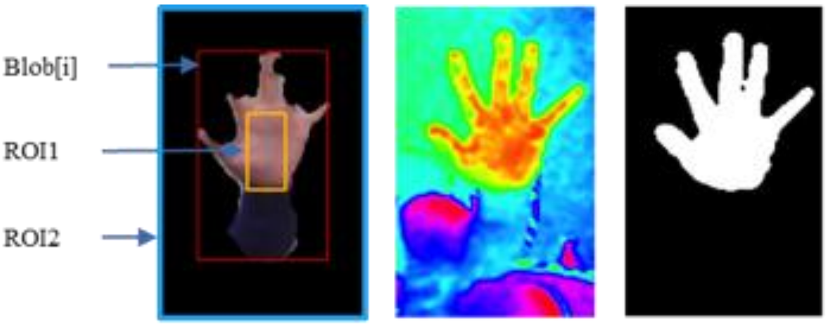
### C. *Hand segmentation*



(a) RGB image (b) Depth image (c) Body extraction (d) Candidates of hand

**Figure 3.** Hand region detection

Because the sensor and the environment are fixed, we firstly detect human regions using background subtraction techniques. Both depth and RGB images can be used for the background subtraction. The depth image is captured from the Kinect sensor that is less sensitive with illumination. Therefore, we use depth images for background subtraction. Among numerous techniques of the background subtractions, we adopt Gaussian Mixture Model (GMM) [36] because this technique have been shown to be the best suitable for our system. Figure 3(a-c) shows results of the background subtraction. Given a  region of human body (as shown in Fig. 3c), we continuously extract candidates of the hand (as shown in Fig. 3d) and a hand segmentation result $X$ (as shown in Fig. 4) is taken out after pruning hand region that is detail presented in [24] by us.



(a) A candidate of hand (b) Malahanobis distance (c) Hand detection

**Figure 4.** Hand segmentation

### D. *Characteristics of hand gesture*

A gestural command is composed of three stages: preparation; performing; relaxing. At preparation phase, the user stays immobile. At performing phase, the user raises the right hand and moves according to a predefined trajectory while changing hand posture which divides into three states: start stage, implementing stage and stop stage. A dynamic hand gesture is a sequence of consecutive frames.  Let N is the number of classes of dynamic hand gestures to be considered.  Let denote a set $\mathrm{G} = \{\mathrm{G_i}|\mathrm{i} = 1, \dots, \mathrm{N}\}(N = 5)$ which $\mathrm{G_1}$ is turn on/off dynamic hand gesture class, $\mathrm{G_2}$ is increase dynamic hand gesture class, $\mathrm{G_3}$ is decrease dynamic hand gesture class, $\mathrm{G_4}$ is go_left dynamic hand gesture class and $\mathrm{G_5}$ is go_right dynamic hand gesture class. Each dynamic hand gesture class $\mathrm{G_i} = \{\mathrm{G_{ij}}|\mathrm{j} = 1,2, \dots, \mathrm{M}\}$ and each dynamic hand gesture $\mathrm{G_{ij}} = \{X_{ij}^{k}|\mathrm{k} = 1,2, \dots, \mathrm{L}\}$ is defined by postures, dynamic hand gestures changes in shape, direction, phase, and speed performing. However, they share following characteristics:

- Each gesture is performed in 3 stages: start, implementing and stop stage. Hand shapes at start and stop stages are the same as Fig. 5, Fig. 6. This characteristic reflects that the defined gestures are cyclical patterns.
- Gestures belonging a certain class could have different temporal length (e.g., as shown Fig. 5. Number of postures $X_k$ in gesture $G_4$ class is different in length event they are performed by one subject).
- The stages of gestures belong a certain class could be non-synchronized phase. A state of the first gesture is longer or shorter than second one. For example, initial stage of two gestures $G_{41}$; $G_{43}$ in Fig. 5 are non-synchronized. Gestures $G_{41}$ spends 5 frames to represent this stage, whereas $G_{43}$ spends only 3 frames to express.
- For each gesture class, the movement direction of hand is significantly distinctive from other ones. (Fig. 6).

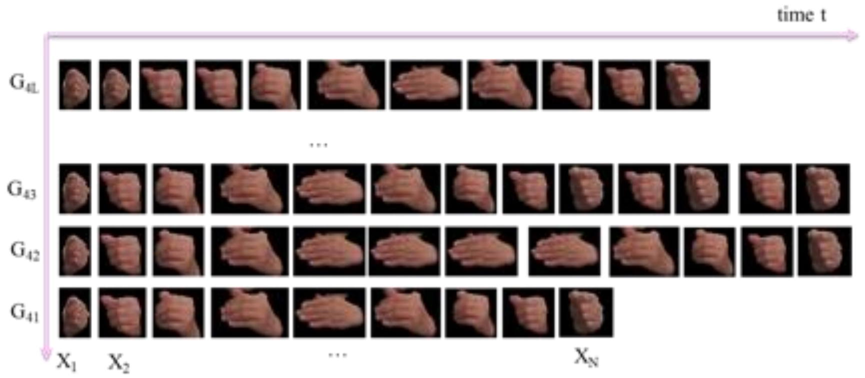
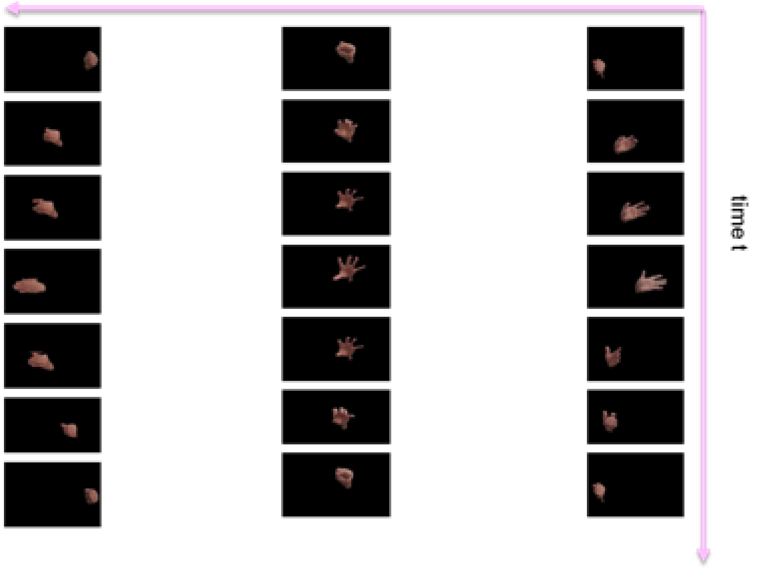


**Figure 5.** Dynamic hand gesture “Right to left” G4 class



(a) Go_left gesture (b) On_off gesture (c) Go_right gesture

**Figure 6.** Movement and changeable of three dynamic hand gesture in time

The defined gestures are discriminated in both characteristics: hand shape and movement of hand directions. Hand shapes represent a cyclical pattern of a gesture, whereas second one represents meaning of gestures. This type of gestures face to several issues: velocity of hand movement; non-synchronization of the stages, changes of hand shapes. To solve these issues, a hand gesture is characterized in term of both spatial and temporal features which will be presented in detail in the following section.

### *E. Spatial hand gesture*

Spatial features of a hand gesture which we focus on resolve some problems: changing of hand shapes, no synchronization and difference of length.

#### *1. Hand shape in space features with PCA technique*

PCA (Principal Component Analysis) is popular technique for dimension reduction. After segmenting a hand image, hand region will be converted to a gray image. The segmented hand image is not same size, so it is resized into X $(p_{i,j}; i = 0 \div 64, j = 0 \div 64)$. To reduce image distortion, if size of image is larger than 64x64 pixels the image will be scaled to 64 pixels. If size of image is smaller than 64x64 pixels it will be scaled the same ratio and based on the center of the images. A row of X is template data, a column of X is feature. Data in rows of X are not similar in amplitude so normalization is implemented by the 1 standard deviation that presents as (1):

$$X^* = \{x_{ij}^*\};\ x_{ij}^* = \frac{x_{ij}-g_j}{\sqrt{n}\sigma_j};\ g_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \qquad (1)$$

$\sigma_j$ is standard deviation in column j of X matrix, n is rows number of X matrix, n is rows number of X matrix, $g_j$ is average value of column j of X matrix. Hand images $(X^*)$ is reshaped into matrix one row and 64x64cols that is Y $(q_i; i = 1 \div 4096)$ that data are still very big (4096 dimensions), as present as (2):

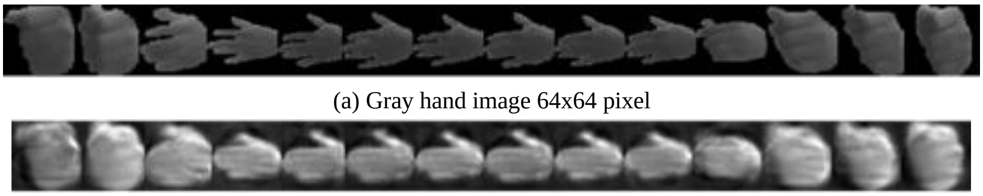
$$\mathrm{X} = \begin{bmatrix} \mathrm{x}_{1,1} & \mathrm{x}_{1,2} & \cdots & \mathrm{x}_{1,64} \\ \mathrm{x}_{2,1} & \mathrm{x}_{2,2} & \cdots & \mathrm{x}_{2,64} \\ \vdots & \vdots & \vdots & \vdots \\ x_{64,1} & x_{64,2} & \cdots & \mathrm{x}_{64,64} \end{bmatrix} => X^{*} = \begin{bmatrix} \mathrm{x}^{*}{}_{1,1} & \mathrm{x}^{*}{}_{1,2} & \cdots & \mathrm{x}^{*}{}_{1,64} \\ \mathrm{x}^{*}{}_{2,1} & \mathrm{x}^{*}{}_{2,2} & \cdots & \mathrm{x}^{*}{}_{2,64} \\ \vdots & \vdots & \vdots & \vdots \\ x^{*}{}_{64,1} & x^{*}{}_{64,2} & \cdots & \mathrm{x}^{*}{}_{64,64} \end{bmatrix} \Rightarrow \mathrm{Y} = [q_1 \quad q_2 \quad \cdots \quad q_{4096}] \qquad (2)$$

Therefore, using PCA to reduce correlation data between component from Y helps to reduce computational workload and still enough information that will be implement in training phase and testing phase:

- The first is training phase: a training hand gesture includes M hand postures $G_i = [Y_0 \quad Y_1 \quad \cdots \quad Y_M]^T$, a training hand gesture set includes N hand gestures: $G = [G_0 \quad G_1 \quad \cdots \quad G_N]^T$ that is input of PCA step and takes out feature space. In this research, PCA space is setup by 20 dimensions and feature space information PCA include: μ covariance matrix of $Y_i$ vector in $G_i$, eigenvalues is λ, eigenvectors is e and $H^*$ is projects of H in PCA space that fators of training hand gesture set will be saved out file. This file will be read when system restart to have the feature space.

- The second is testing phase: a gesture G has n postures $Y_i\ (i = 1 \div n)$, in PCA space this gesture will be presented $Y_i^*(i = 1 \div n)$ as presents (3) and result of hand gesture in PCA space illustrates as Fig. 7:

$$G = \begin{bmatrix} Y^1 \\ Y^2 \\ \cdots\cdots \\ Y^n \end{bmatrix} = \begin{bmatrix} q_1^1 & q_2^1 & \cdots\cdots & q_{4096}^1 \\ q_1^2 & q_2^2 & \cdots\cdots & q_{4096}^2 \\ & & & \\ q_1^n & q_2^n & \cdots\cdots & q_{4096}^n \end{bmatrix} \Rightarrow (PCA) \Rightarrow G* = \begin{bmatrix} Y^{*1} \\ Y^{*2} \\ \cdots\cdots \\ Y^{*n} \end{bmatrix} = \begin{bmatrix} q_1^{*1} & q_2^{*1} & \cdots\cdots & q_{20}^{*1} \\ q_1^{*2} & q_2^{*2} & \cdots\cdots & q_{20}^{*2} \\ & & & \\ q_1^{*n} & q_2^{*n} & \cdots\cdots & q_{20}^{*n} \end{bmatrix} \qquad (3)$$



(a) Gray hand image 64x64 pixel

(b) Hand image projects on PCA space

**Figure 7.** Go left hand gesture before and after projects on PCA space

*2. Synchronizing hand gestures with DTW technique*

As illustrated in Sec. D., for each dynamic hand gesture class, each person implements gestures with difference lengths. That has a difference number of postures and some gestures are not the same phase with other gestures (hand closing and hand opening state). The phase synchronization between two dynamic hand gestures that is necessary. Many techniques have been devised to perform these tasks as using DTW (Dynamic Time Wrapping), HMM (Hidden Markov Model),... Our proposed that matching is implemented by DTW algorithm that is applied to solve the problem enable matched two samples signal having with different lengths for small errors and real time.

DTW method optimums match between the two time series with some constraints. Two sequence hand gesture are stretched non-linear along the time axis order to determine the similarity between them. The simplest version of DTW can be implemented by this presentation: two hand gestures $G_1\{X_{11}, X_{12}, \ldots X_{1n}\}$ and $G_2\{X_{21}, X_{22}, \ldots X_{2m}\}$ that length is n and m. DTW method will indicate that each $X_{1i}$ in the hand gesture G1 is matched positions in hand gesture G2. So, optimum matching pair is determined. Some results of synchronization as shown in Fig. 8. The PCA feature space has good performed information of the hand shape. Those results will be utilized that is an input for the temporal hand gesture step.

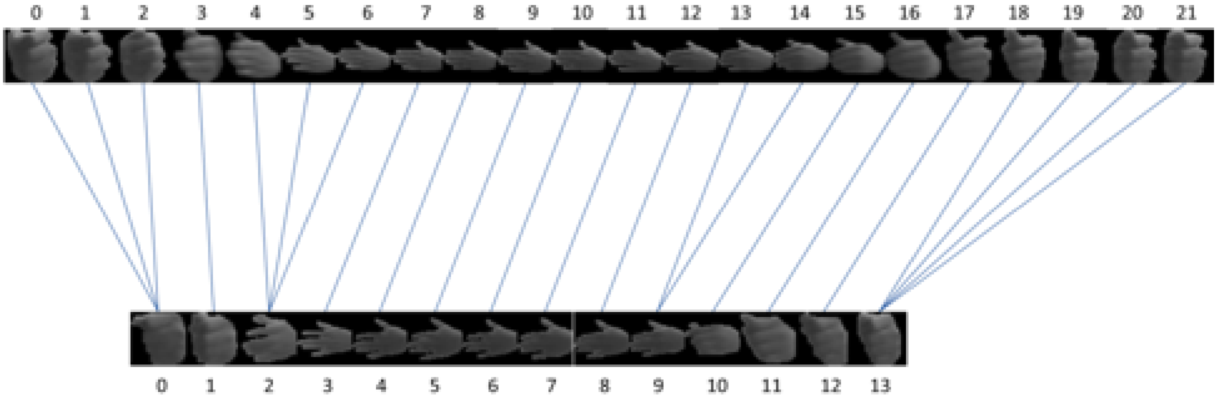
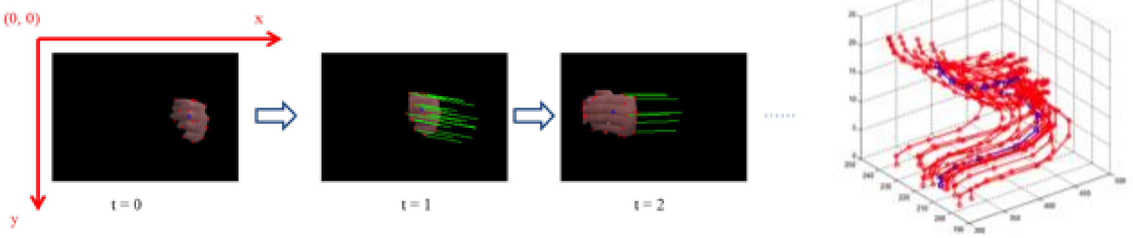


**Figure 8.** Result of DTW

### *F. Temporal features of hand gestures*

#### *1. Hand gesture trajectory with KLT*

Many proposed methods for action recognition in last years reply on temporal features. In our work, hand motion trajectory is implemented by KLT that combines the optical flow method of Lucas–Kanade [25] and the good feature points segmentation method of Shi–Tomasi [26]. The algorithm determines optical flow of Lucas-Kanade that based on three assumptions: the invariance of light intensity, the movement of hands in two consequence frames is small and Cohesion of space (the neighboring points on the same surface of the hand is the same motion). KLT help to trajectory of feature points of hand or calculates optical flow of hand between two sequence postures. At the first frame of hand gesture, feature points of hand posture will be segmented and this feature points will be trajectoried by the next posture to the end posture of gesture. So, each feature point creates a trajectory. If optical flow of two sequence postures less than 1, features is seem not movement and if optical flow of two sequence postures more than 50, features is seem not reliable. So this no movement points and no reliable points will be removed. If the feature points are less than threshold in a frame, that frame will segment some new feature points and this new feature points will trajectory in the next frame. Our research utilizes 20 feature points between two sequence posture hands and implements continuously from the start posture to the end posture of each gesture that is illustrated as Fig. 9(a).



(a) Optical flow (b) Trajectory

**Figure 9.** Optical flow and trajectory of the go-left hand gesture

Giving optical flows of postures in each gesture, trajectory of feature points is built. A gesture has n frame (n posture) $f_1, f_2,..., f_N$ , each $f_i$ has K feature points ($p_{i,j}$), average of K feature points on x and y is a point $\overline{p_{i,j}}$. Therefore, a dynamic hand gesture is presented by average of trajectory $T = [\overline{p^1_{i,j}} \quad ... \quad \overline{p^n_{i,j}}]$. Because the hand segmentation step has many noises on background, this proposed technique is more robust than trajectory that base on centroid point of hand postures. Fig.9(a) illustrates trajectories of 20 feature points and an average trajectory of the $G_5$ hand gesture in spatial-temporal coordinate. Red circles presents the feature point coordinates $p_{i,j}$ at frame t (t = 0 ÷ (n-1)). Blue square is presented by $\overline{p_{i,j}}$.

#### *2. Similarity hand gestures*

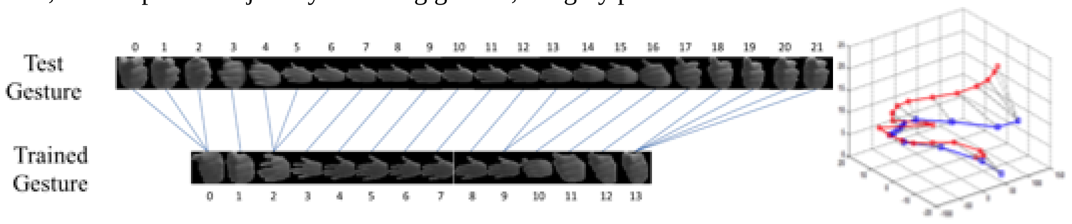
Giving average trajectory of a dynamic hand gestures, hand position in each gesture is not the same coordinate. Firstly, we have to normalize $T = [\overline{p^1_{i,j}} \quad ... \quad \overline{p^n_{i,j}}]$ on x and y dimension as (4):

$$T^* = [\overline{p^1_{i,j}} - (\bar{x},\bar{y}) \quad ... \quad \overline{p^n_{i,j}} - (\bar{x},\bar{y})] \qquad (4)$$

In hand gesture recognition, a training set is P and a test set is T that is similar estimated by RMSE distance which calculates an error at $p_{i,j}$ of P and T which is presented as (5):

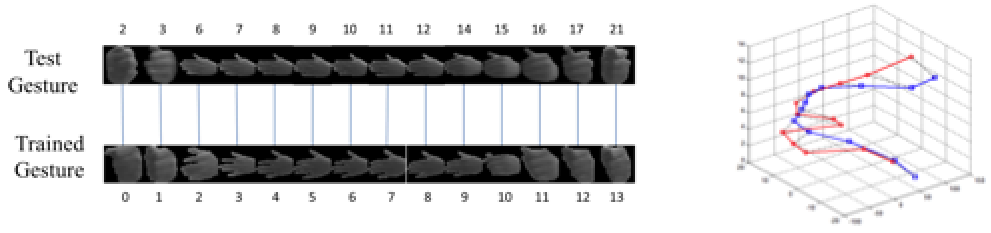
$$RMSE(T,P) = \sqrt{\frac{\sum_{i=1}^{n}(p_i(x,y)-q_i(x,y))^2}{n}} \qquad (5)$$

But length of T and P are not the same. Therefore a direction calculation does not implement (Fig.10a). Thanks to the link between postures (T, P) that are DTW results. Estimating of RMSE will become feasible (Fig.11a). The experimental results in Sec.IV will show that RMSE method is simple and clearly separation results. If RMSE value is smaller, T and P is more similar. Fig.10a. illustrates the removing link of two trajectories, the blue path is trajectory of training gesture, the red path is trajectory of testing gesture, the gray path are links DTW.



*(a)* *(b)*

**Figure 10.** All trajectory link of two hand gesture (T, P); (a) DTW results (b) RMSE results



(a) (b)

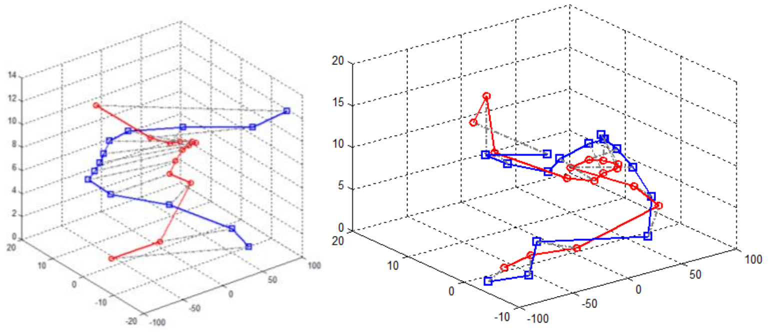
**Figure 11.** Trajectory link of two hand gesture (T, P) remove repetition link; (a) DTW results (b) RMSE results

## IV. EXPRIMENTAL RESULT

We evaluate the proposed method in the effectiveness utilizing of the spatial-temporal features that can be distinguished between two hand gestures based on three following assessments: (1) utilizing cross-correlation and the average squared error RMSE; (2) changes shape but not change trajectory; (3) change trajectory but not changes shape. Using five dataset to evaluate that includes on-off (G1), up (G2), down (G$_3$), go_left (G4) and go_right (G5). Each dataset includes 10 templates that are implemented by two people (P1, P2), each person implements one command in 5 times (L1, L2,…, L5). So dataset has 50 templates. Hand gestures are labeled by: [P+order number]_[L+times]_[class lable:ACx]. Ex: P1_L1_G4, P2_L3_G5,…

### *1. Utilizing cross-correlation and the average squared error RMSE*

The measurements are performed by the cross-correlation value and RMSE between hand gesture test T and hand gesture training P. If only using cross-correlation between T and P that will be very difficult to distinguish between the two hand gestures; while combined DTW allows standardized length T and P, the assessment showed that RMSE is easy to distinguish between two different hand gestures as Fig.12. Table 1, 2 illustrates results of cross-correlation and RMSE between some defined hand gestures.



(a) Difference hand gestures (b) Same hand gestures

**Figure 12.** DTW and RMSE results between two hand gestures

**Table 1.** P1_L3_AC5 with P1_L0_G4

| | | |
|---|---|---|
| Cross-correlation | 0.8522 | 0.6020 |
| RMSE | 83.1950 | |

**Table 2.** P1_L3_AC5 with P1_L0_G5

| | | |
|---|---|---|
| Cross-correlation | 0.8135 | 0.5921 |
| RMSE | 49.2284 | |

**Table 3.** Results between the testing (T) and training (P) using RMSE

| | P1_L1_G4 | P1_L1_G5 | P1_L3_G4 | P1_L3_G5 | P2_L2_G4 | P2_L2_G5 | P2_L3_G4 | P4_L3_G4 |
|---|---|---|---|---|---|---|---|---|
| P1_L0_G4 | **32.55** | 86.41 | **28.42** | 83.19 | **18.04** | 82.9 | **51.12** | **39.31** |
| P1_L0_G5 | 101 | **41.25** | 120.72 | **49.22** | 102.16 | **32.16** | 112.79 | 118 |

Table 1 illustrates that if only utilizing cross-correlation between T and P gesture that is difficult to distinguish them even the trajectory are identified; Because of the difference length of T and P trajectory. The cross-correlation calculating is not effectiveness. Our proposed method utilized DTW results to take out hand posture links and combine RMSE that clearly distinguish between two difference hand gestures. As shown in Table 3, based on the minimum values (average is 36.5) when comparing testing hand gesture with training hand gesture that has 100% accurate results.

*2. Changes shape but not change trajectory*

This estimation is to check distinguishing of a dynamic hand gesture when the trajectory does not change but hand posture changes in time as Fig. 14. Clearly illustration is distinguishing of close-open-close hand gesture that is $G_1$ command. This dynamic hand gesture only changes hand shape but position of hand does not changing. RMSE value between G1 with G4 and G5 hand gesture is presented as Table 4. While RMSE value between $G_{11}$(on_off_1) and $G_{12}$(on_off_2) gesture is only 20.5 that helps to distinguish those gestures. Fig.13(a) illustrates hand shape of G1 gesture class and Fig.13(b) presents its trajectory.

**Table 4.** RMSE of the changing and no changing trajectory of hand gestures

| | $G_{11}$ | $G_{12}$ |
|---|---|---|
| G4 (P1_L0_G4) | 44.73 | 43 |
| G5 (P1_L0_G5) | 59.05 | 66.49 |

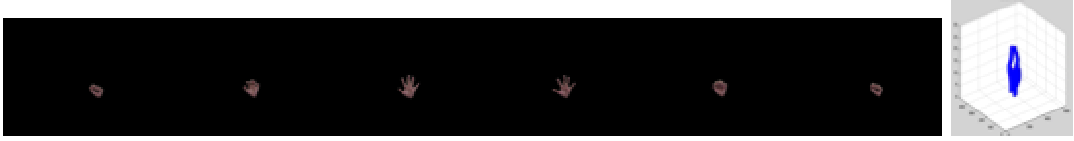

(a) (b)

**Figure 13.** "on_off" hand gesture (a) changing on hand postures; (b) no changing on trajectory

*3. Change trajectory but not changes shape*

This estimation is to check distinguishing of a hand gesture when the trajectory changes but hand shape does not change in time. Those dynamic hand gestures are go_left_1 and go_left_2. It changes only position and does not change hand shape that result is illustrated in Fig .14:

**Table 5.** RMSE of the changing and no changing hand shape of hand gestures

| | go_left_1 | go_left_2 |
|---|---|---|
| G4 (P1_L0_G4) | 63.06 | 58.59 |
| G5 (P1_L0_G5) | 125.66 | 117.84 |

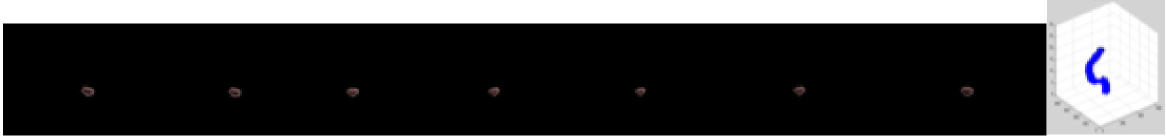

(a) (b)

**Figure 14.** "go_left_1" hand gesture (a) no changing on hand shape of go_left1; (b) Trajectory of go_left1

Trajectory of "go_left_1" and "go_left_2" are the same with G4 and not the same with G5. Therefore, RMSE values of those commands with G4 are smaller than with G5. Moreover, this values (63,06 and 58,59) are still more than average value of true hand gestures in table 3 (only 36.5). This shows that our system is still detecting and distinguishing if a person does not true implement a command.

Table 6 following presents confusion matrix of recognition results with fine dynamic hand gestures (left, right, up, down, on-off) that implements by 5 people and each dynamic hand gesture implements in 5 times. Each set has 25 dynamic hand gestures, evaluation method is 'Leave-p-out-cross-validation' method ($p$=5) to separate training and testing data. Utilizing Knn with RMSE distance to recognize dynamic hand gestures with mean time cost is 178 ± 12(ms) and the accuracy rate at 96% ± 2.5.

**Table 6.** Confusion matrix of five dynamic hand gesture recognitions

| Training<br>Testing | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| G1 | **25** | 0 | 0 | 0 | 0 |
| G2 | **1** | **24** | 0 | 0 | 0 |
| G3 | **2** | 0 | **23** | 0 | 0 |
| G4 | **1** | 0 | 0 | **24** | 0 |
| G5 | 0 | 0 | 0 | 0 | **25** |

## V. CONCLUSION

This report described a vision-based hand gesture recognition system. Our proposed method utilized basic techniques of vision-based to be applied to recognition of hand gestures with based techniques: PCA, DTW, optical flow and RMSE. Performa is relatively simple and effective. Initial results solve the requirement problems of our dynamic hand gesture recognition with accuracy approximates 96% and time rate approximates 178ms/frame. Thus, it is feasible to implement my recognition system to control the TV or indoor lighting system. In the future, we will continue improving, reducing time rate and perfecting the gesture recognition system to built a controlling system which using dynamic hand gesture to control equipments in a smart room.

## VI.ACKNOWLEDMENT

The research leading to this paper was supported by the National Project B2013.01.41 "*Study and develop an abnormal event recognition system based on computer vision techniques*”. We would like to thank the project and people involved in this project.

## VII. REFERENCES

[1] Microsoft Kinect for Windows, http://www.microsoft.com/enus/kinectforwindows., November 2013.

[2] PointGrab Company, PointGrab brings gesture control to home appliances, (2013).

[3] P.Qifan, S.Gupta, S.Gollakota, and S.Patel, “Whole-Home Gesture Recognition Using Wireless Signals”, In Proceedings of the 19th Annual International Conference on Mobile Computing and Networking, (2013).

[4] I.Bayer and T.Silbermann, “A multi modal approach to gesture recognition from audio and video data”, In Proceedings of the 15th ACM on International conference on multimodal interaction, 461-466, (2013).

[5] Q.Chen, C.Joslin, and N.D.Georganas, “A Dynamic Gesture Interface for Virtual Environments Based on Hidden Markov Models”, In Proceedings of IEEE International Workshop on Haptic Audio Visual Environments and their Applications, (2005).

[6] X.Chen and M.Koskela, “Online rgb-d gesture recognition with extreme learning machines”, ACM on International conference on multimodal interaction, pp.467-474, (2013).

[7] S.Escalera, J.Gonz‘alez, X.Bar´o, M.Reyes, O.Lopes, I.Guyon, V.Athitsos, and H.Escalante, “Multi-modal gesture recognition challenge 2013: Dataset and results”, In Proceedings of the 15th ACM on International conference on multimodal interaction, pp.445–452, (2013).

[8] J.Choi and B.Seo, “Robust Hand Detection for Augmented Reality Interface”, In Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry, (2009).

[9] F.Picard and P.Estraillier, “Motion Capture System Contextualization Application to Game Development”, In Proceedings of Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational, Serious Games, (2009).

[10] MPD.Silva, V.Courboulay, and A.Prigent, “Gameplay experience based on a gaze tracking system”, EURASIP Journal on Applied Signal Processing, (2007).

[11] K.Nandakumar, K.W.Wan, S.M.A.Chan, W.Z.T.Ng, J.G.Wang, and W.Y.Yau, “A multi-modal gesture recognition system using audio, video, and skeletal joint data”, In Proceedings of the 15th ACM on International conference on multimodal interaction, pp.475-482, (2013).

[12] X.Zabulis, H.Baltzakis, and A.Argyros, “Vision-based Hand Gesture Recognition for Human Computer Interaction”, Lawrence Erlbaum Associates, (2009).

[13] S.Rautaray and A.Agrawal, “Vision based hand gesture recognition for human computer interaction”, a survey, Artificial Intelligence Review (2012).

[14] X.Deyou, “A Network Approach for Hand Gesture Recognition in Virtual Reality Driving Training System of SPG”, International Conference on Pattern Recognitioin, pp.519-522, (2006).

[15] M.Elmezain, A.Al-Hamadi, and C.Michaelis, “Real-Time Capable System for Hand Gesture Recognition Using Hidden Markov Models in Stereo Color Image Sequences”, Journal of WSCG 16, pp.65–72, (2008),.

[16] M.Elmezain, A.Al-Hamadi, and J.Appenrodt, “A Hidden Markov Model-based continuous gesture ecognition system for hand motion trajectory”, International Conference on Pattern Recognition, pp.1–4, (2008).

[17] D.Kim, J.Song, and D.Kim, “Simultaneous Gesture Segmentation and Recognition Based on Forward Spotting Accumlative HMMs”, Journal of Pattern Recognition Society 40, pp.1–4, (2007).

[18] A.Oikonomopoulos, I.Patras, and M.Pantic, “Spatiotemporal salient points for visual recognition of human actions”, IEEE Transactions, pp.710–719, (2005).

[19] K.Takahashi, S.Sexi, and R.Oka, “Spotting Recognition of Human Gestures From Motion Images”, In Technical Report IE92-134, pp.9–16, (1992).

[20] J.Alon, V.Athitsos, Y.Quan, and S.Sclaroff, “A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence 31, pp. 1685–1699, (2009).

[21] H.Yang, S.Scharoff, and S.Lee, “Sign Language Spotting with a Threshold Model Based on Conditional Random Fields”, IEEE Transaction on Pattern Analysis and Machine Intelligence 31, pp.1264–1277, (2009),.

[22] J.Lafferty, A.McCallum, and F.Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling sequence Data”, Internation Conference on Machine Learning, pp.282–289, (2001).

[23] M.Elmezain, A.Al-Hamadi and B.Michaelis, “Discriminative models-based hand gesture recognition”, International Conference on Machine Vision, pp.123–127, (2009).

[24] Huong-Giang Doan, Hai Vu, Thanh Hai Tran, Eric Castelli, “Improvements of RGB-D Hand Posture Recognition Using an User-Guide Scheme”, In Proceeding(s) of the 7th IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and the 7th IEEE International Conference on Robotics, Automation and Mechatronics (RAM), (2015).

[25] Bruce D. Lucas and Takeo Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision”, Proc. International Joint Conference on Artificial Intelligence, pages 674–679, 1981.

[26] J.Shi and C.Tomasi, “Good Features to Track”, IEEE Conference on Computer Vision and Pattern Recognition, pages 593–600, (1994).

[27] J.Listgarten, “Analysis of sibling time series data: alignment and difference detection”, Ph.D. thesis, University of Toronto, (2006).

[28] J.Rob Hyndman, B.Anne Koehler, "Another look at measures of forecast accuracy", International Journal of Forecasting 22(4), pp.679 – 688, (2006).

# NHẬN DẠNG CỬ CHỈ ĐỘNG CỦA BÀN TAY SỬ DỤNG CÁC ĐẶC TRƯNG KHÔNG GIAN VÀ THỜI GIAN

**Đoàn Hương Giang, Vũ Duy Anh, Vũ Hải, Trần Thị Thanh Hải**

***Tóm tắt*** - *Nhận dạng cử chỉ tay đã được nghiên cứu trong thời gian dài. Tuy nhiên nó vẫn còn là một lĩnh vực còn nhiều thách thức. Hơn nữa, nhận dạng cử chỉ tay để điều khiển các thiết bị phòng thông minh như tivi, quạt, camera, cửa,… đòi hỏi phải có độ chính xác nhận dạng cao. Bài báo này đề xuất một mô hình đơn giản và hiệu quả để nhận dạng bộ cơ sở dữ liệu đã được định nghĩa để điều khiển các thiết bị điện. Việc phân tích các đặc trưng không gian và thời gian bao gồm tính chu kỳ lặp lại của các cử chỉ tĩnh trong mỗi cử chỉ động, sự khác nhau về độ dài của các cử chỉ động, sự không đồng bộ về pha giữa các cử chỉ động, sự thay đổi về hình trạng tay và các đặc trưng về hướng cũng như sự di chuyển của tay của bộ cơ sở dữ liệu đã định nghĩa. Sau đó nhận dạng các cử chỉ dựa trên các đặc trưng không gian và thời gian.*