

# A NEW APPROACH DEDICATED TO REAL-TIME HAND GESTURE RECOGNITION

Nguyen Dang Binh, Enokida Shuichi, Toshiaki Ejima

Intelligence Media Laboratory, Kyushu Institute of Technology

680-4, Kawazu, Iizuka, Fukuoka 820, JAPAN

{ndbinh, shuichi, toshi}@mickey.ai.kyutech.ac.jp

## Abstract

We introduce a new Pseudo 2-D Hidden Markov Model (P2DHMM) structure dedicated to the time series recognition (T-CompP2DHMM). The T-P2DHMM allows it to do temporal analysis, and to be used in large set of hand gestures movement recognition systems in unconstrained environments. Additionally, robust and flexible hand gesture tracking using an algorithm that combines two powerful stochastic modeling techniques: the first one is pseudo two dimension hidden Markov model (P2DHMM) and the second technique is the well-known Kalman filter. Our work also present a feature extraction method based on the joint statistics of a subset of DCT coefficients and their position on the hand. Using feature extraction method along with the T-CompP2DHMM structure was used to develop a complete vocabulary of 36 gestures including the America Sign Language (ASL) letter spelling alphabet and digits. The results are effectiveness of the approach.

**Keywords:** gesture recognition, Pseudo 2-D HMM, time series recognition, Kalman filter.

## 1. INTRODUCTION

Computer recognition of hand gestures provides a more natural human-computer interface. The sign language is undoubtedly the most grammatically structured and complex set of human gestures. In American Sign Language (ASL), the use of hand postures [5] is very important to differentiate between many gestures. Thus, a fast and reliable method to extract the hand postures changes from the video sequence is very important in ASL recognition systems. The hand is complex object. The main challenge in hand gestures detection and recognition is the mount of variation in visual appearance. For example, hands vary in shape, size, coloring, and in small details such as the head lights, grill, and tires. In addition to its rigid transformation it has 14 joints which mean that number of possible configuration. Visual appearance also depends on the surrounding environment. Light sources will vary in their intensity,

color, and location with respect to the hand. The appearance of the hand also depends on its pose; that is, its position and orientation with respect to the camera. The basic idea lies in the real-time generation of gesture model for hand gesture recognition in the content analysis of video sequence from CCD camera. To cope with all this variation, we combines two powerful stochastic modeling techniques: the first one is pseudo two dimension hidden Markov model (P2DHMM) and the second technique is the well-known Kalman filter approach with hand detectors as described in Section 2. We use the combined use of time "spatialization" and P2DHMM in the proposed T-CompP2DHMM model for hand gestures recognition in Section 3. The next section presents the result of experiments. Finally, the summarize contribution of this work in the conclusion section.

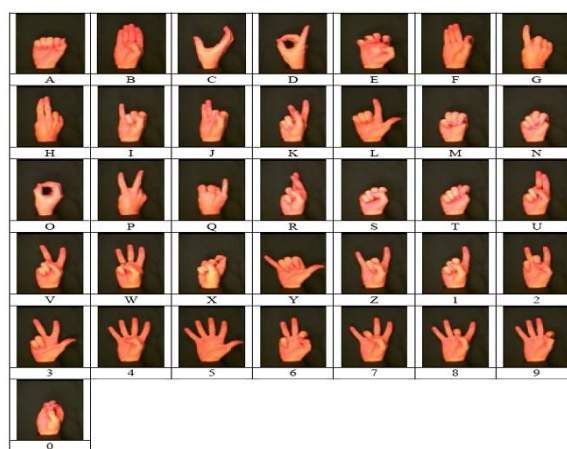


Fig. 1 The ASL gesture set to be recognized

## 2. HAND TRACKING

The approach suggested in this article for tracking of hand gesture is one of the first attempts to use a statistical shape mode for tracking. The statistical model is represented by a so-called Pseudo 2-D Hidden Markov Model (P2DHMM) [2]. Additionally, this P2-DHMM is combined with a Kalman filter for motion prediction. As

will be shown later in more detail, such an approach has the following advantages:

- The statistical shape model is able to exploit some a priori-knowledge about hand's shape. It retains some of advantages of this approach while being able to expand its flexibility and robustness.

- At the same time, the advantage of the model-free approach can be exploited to automatically learn the features, which are relevant for the problem. Thus, it combines the advantages of model-based and model-free approach.

- The system does not rely on any motion information has several important advantages. One of them is the capability of tracking hands people independent of the fact that they are moving or not.

- The advantage that the tracking is possible in presence of other moving objects in the background. There, it is also explained that another advantage of the P2DHMM approach is the exploitation of the automatic scaling capabilities of HMM in general, that become important in zooming operations due to the fact that the tracked object may change its size drastically.

- The HMM multi-stream technique, it is easily possible to combine various features, such as e. g skin color or hand shape and to give those features a different weighting.

- Using the previously mentioned capability, it is possible to either design the system for hands person-independent model or for hands-specific model.

We develop a real time hand tracking method based on the P2DHMM and Kalman filter, which is robust and reliable on hand tracking in unconstrained environments and then the hand region extraction fast and accurately. We need to consider the trade-off between the computation complexity and robustness.

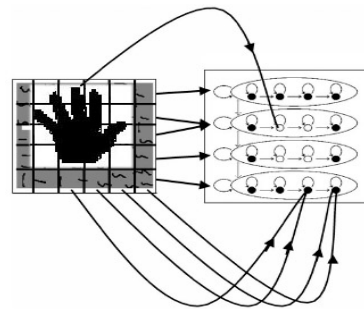
### 2.1. Basic Hand Tracking Algorithm

We propose a novel tracking method for problem using two powerful stochastic modeling techniques, namely P2DHMM and Kalman filter. The key feature of our algorithm is the fact that it makes use of two powerful stochastic modeling techniques, namely P2DHMM and Kalman filter. The input of the Kalman filter relies on the information provided by a complex shape model of the hand's person of which the structure has been automatically learned and acquired by the P2DHMM. The dynamic information need for tracking is solely generated by the Kalman filter. While the Kalman filter obtains its input information from the P2DHMM, the Kalman filter itself feeds its output information back to the P2DHMM and improves in this way the shape detection procedure of the P2DHMM. This optimal feedback between these two modules is another reason for the powerful performance of the approach. By letting only Kalman filter be responsible for the dynamic information of the tracking process, and relying in the measurement process

completely on shape and color information, the tracking procedure becomes entirely independent of other disturbing motions in the background. We must first extract hand region in each input image frame in real time.

### 2.2. Measurement Vector Generation with P2DHMM

P2DHMM generates a measurement vector that is uses as input to the Kalman filter. The components of this vector are the center of gravity of the hand person detected in the image and the width and height of the bounding box. The following steps are carried out for that purpose: firstly, the image is processed with a DCT-based feature extraction method that is adopted from [12]. An overlap between adjacent sampling windows improves the ability of the HMM to model the neighborhood relations between the windows. The result of the feature extraction is two-dimensional array of vectors. This array is presented to P2DHMM as shown in Fig 2.



**Fig. 2** Stochastic model of a two-dimensional object use a P2DHMM

Such a P2DHMM can be considered as a 2-D stochastic model of an object in an image. It models the occurrence of a feature vector sequence which can be derived from that object of the object is pre-processed in the same manner as described above [3]. The parameters of the P2DHMM consists of the transition and output probabilities of the various HMM states and can be learned in order to model different objects. The learning of hand shape persons can be accomplished in the following way: Several hundred image of hands with appropriate pre-processing are presented to the P2DHMM for learning the structure of the hand by applying parameter estimation methods. Since the P2DHMM can be considered as an elastic model, it is capable of modeling the hand in various positions. In order to be capable of locating a hand with a flexible environment with complex background, the following important step is carried out: the P2DHMM is trained with static image that show a hand within a complex environment, and not isolated or in front of a uniform background. While the HMM parameters of the system are learned successfully, several of the P2DHMM states will be assigned to the background, and other states will be assigned to the hand regions. The actual tracking procedure starts with the

presentation of the first frame of the tracking video sequence to the trained P2DHMM. If an image containing a hand is presented to the above displayed especially trained P2DHMM, the Viterbi algorithm can be used again in order to compute the dark background states and blocks assigned to the white hands states, thus obtaining the hand's shape.

The center of gravity of the hands are computed from the segmentation result obtained from Viterbi algorithm by simply calculating the appropriate moment from the blocks inside the black marked area indicating the hand (as show in Fig. 2). The coordinates of this center of gravity, denoted as x and y, and the size of the bounding box of the segmentation, denoted as w (width) and h (height) serve as the measurement input the Kalman filter.

### 2.3. Combination of P2DHMM Output with Kalman Filter and Measuring Trajectories

In order to describe the moving hands and to represent the result of the tracking procedure, we use Kalman filter to predict hand location in one image frame based on its location detected in the previous frame. First, we measure hand location and velocity in each image frame. Hence, we define the state vector as  $x_t$ :

$$x_t = (x(t), y(t), v_x(t), v_y(t), w(t), h(t))^T \quad (1)$$

Where  $x(t)$ ,  $y(t)$ ,  $v_x(t)$ ,  $v_y(t)$  shows the location of hand ( $x(t)$ ,  $y(t)$ ), the velocity of hand ( $v_x(t)$ ,  $v_y(t)$ ) and the width and height of hand  $w(t)$ ,  $h(t)$  in the  $t^{\text{th}}$  image frame. We define the observation vector  $y_t$  to present the location of the center of the hand detected in the  $t^{\text{th}}$  frame. The measurement vector  $y_t$  consists of the location of the center of the hand region. The state vector  $x_t$  and observation vector  $y_t$  are related as the following basic system equation:

$$x_t = \Phi x_{t-1} + G w_{t-1} \quad (2)$$

$$y_t = H x_t + v_t \quad (3)$$

Where  $\Phi$  is the state transition matrix,  $G$  is the driving matrix,  $\Phi$  is the observation matrix,  $w_t$  is system noise added to the velocity of the state vector  $x_t$  and  $v_t$  is the observation noise that is error between real and detected location. Here we assume approximately uniform straight motion for hand between two successive image frames because the frame interval  $\Delta T$  is short. Then  $\Phi$ ,  $G$ , and  $H$  are given as follows:

$$\Phi = \begin{bmatrix} 1 & 0 & \Delta T & 0 & 0 & 0 \\ 0 & 1 & 0 & \Delta T & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (5)$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

The  $(x, y)$  coordinates of the state vector  $x_t$  coincide with those of the observation vector  $y_t$  defined with respect to the image coordinate system. Also, we assume that both the system noise  $w_t$  and the observation noise  $v_t$  are constant Gaussian noise with zero mean. Thus the covariance matrix for  $w_t$  and  $v_t$  become

$\sigma_w^2 I_{4 \times 4}$  and  $\sigma_v^2 I_{4 \times 4}$  respect, where  $I_{4 \times 4}$  represent a  $4 \times 4$  identity matrix. Finally, we formulate a Kalman filter as

$$K_t = \bar{P}_t H^T (H \bar{P}_t H^T + I_{2 \times 2})^{-1} \quad (7)$$

$$\bar{x}_t = \Phi \{\bar{x}_{t-1} + K_{t-1} (y_{t-1} - H \bar{x}_{t-1})\} \quad (8)$$

$$\bar{P}_t = \Phi (\bar{P}_{t-1} - K_{t-1} H \bar{P}_{t-1}) \Phi^T + \frac{\sigma_w^2}{\sigma_v^2} Q_{t-1} \quad (9)$$

Where  $\bar{x}_t$  equal  $\bar{x}_{t|t-1}$ , the estimated value of  $x_t$  from  $y_0, \dots,$

$y_{t-1}$ ,  $\bar{P}_t$  equals  $\bar{\Sigma}_{t|t-1} / \sigma_v^2$ ,  $\bar{\Sigma}_{t|t-1}$  represents the

covariance matrix of estimate error of  $\bar{x}_{t|t-1}$ ,  $K_t$  is Kalman gain, and  $Q$  equals  $GG^T$ . Then the predicted location of the hand in the  $t+1$ th image frame is given as  $(x(t+1), y(t+1))$  of  $\bar{x}_{t+1}$ . If we need a predicted location after more than one image frame, we can calculate the predicted location as follows:

$$\bar{x}_{t+m|t} = \Phi^m \{\bar{x}_t + K_t (y_t - H \bar{x}_{t-1})\} \quad (10)$$

$$\bar{P}_{t+m|t} = \Phi^m (\bar{P}_t - K_t H \bar{P}_t) (\Phi^T)^m + \frac{\sigma_w^2}{\sigma_v^2} \sum_{k=0}^{m-1} \Phi^k Q (\Phi^T)^k \quad (11)$$

Where  $\bar{x}_{t+m|t}$  is the estimated value of  $\bar{x}_{t+m}$  from

$y_0, \dots, y_t$ ,  $\bar{P}_{t+m|t}$  equals  $\bar{\Sigma}_{t+m|t} / \sigma_v^2$ ,  $\bar{\Sigma}_{t+m|t}$  represents the covariance matrix of estimate error  $\bar{x}_{t+m|t}$ .

**Determining Trajectories.** We obtain hand trajectory by taking correspondences of detected hand between successive image frames. Suppose that we detect hand centroid in the  $t^{\text{th}}$  image frame  $t$ . We refer to hand's location as  $P_{t+1}'$ . First, we predict the location  $P_{t+1}$  of hand in the next frame  $(t+1)^{\text{th}}$  image frame  $t+1$  with the predicted location  $P_{t+1}'$ . Finding the best combination.

From the input information of the P2DHMM, contained in the vector  $y$ , the system estimates the state vector  $x$  and predicts in that way the information about bounding box, contained in the last two dimension of  $x$ . The third and fourth dimension of  $x$  deliver the velocity of the hand.

### 2.4. Interaction Between Kalman Filter and P2DHMM

The P2DHMM approach allows the elegant incorporation of additional feature, such as e.g. colors or textures in the hand segmentation procedure, by using multi-stream techniques. In this case, different features are derived from each frame of the image sequence (for instance DCT-based features, color features and texture features).

Each different feature type leads to a different feature stream, if all frames of the sequence are processed. The states of the P2DHMM model the occurrence of each feature with a different probability of the combined features in a certain state are computed as the product of the probabilities generated by each feature's density function. Weighting factors can be introduced in order to adjust the influence of the various feature streams. Consequently, the system can even be used to track hand in presence of other moving hands, if the hand to be tracked has been acquired previously by the P2DHMM parameters, which will automatically, learn the shape cues from the hand shape and color. An important points the fact that - while vector  $x$  is constructed from the vector  $y$  in the Kalman equations. The update of the vector  $x$  is used in return as input to the P2DHMM in order to improve estimation of the vector  $y$ , thus resulting into a cooperative feedback between the Kalman filter and the P2DHMM. The complete interaction procedure between P2DHMM and Kalman filter is illustrated in Fig.3: on left site up site, a moving hand has been segmented, and the coordinates of the center of gravity serve as measurement signal for the Kalman filter which predicts a new state vector from this measurement input and the motion equation. On the right upper side, this leads to a new bounding box, which can be derived from the updated state vector (inner black rectangle). This area is enlarged and thus yields an image fraction shown on the right lower side (black-white bold rectangle), which serves as search area for the P2DHMM. From there, the loop is closed by yielding a new segmentation which generates the new measurement signal in the upper left part of Fig.3.

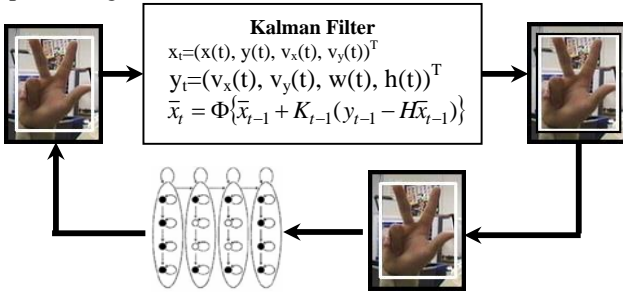


Fig. 3 Scheme of interaction between P2DHMM and Kalman filter

### 3. THE T-COMP2DHMM MODEL

The pseudo 2-D Hidden Markov Model (P2DHMM) model [2], we have developed to deal with real-time hand gesture recognition system [3]. The advantages of the improved P2DHMM structure is simplification, and efficient 2-D model that retains all of the useful HMM features and intelligent selection of training images of training stage, reducing the number of local minimum in P2DHMM training. One that the input space is pre-classified; a big problem is divided in several small ones. This philosophy allows a problem with a large number of

classes to be solved easily, reducing the training time and/or permitting a very good solution to be found, increasing the recognition rate. Taking these advantages and transporting it to a time varying space, we propose the T-ComP2DHMM shown by the Fig. 4.

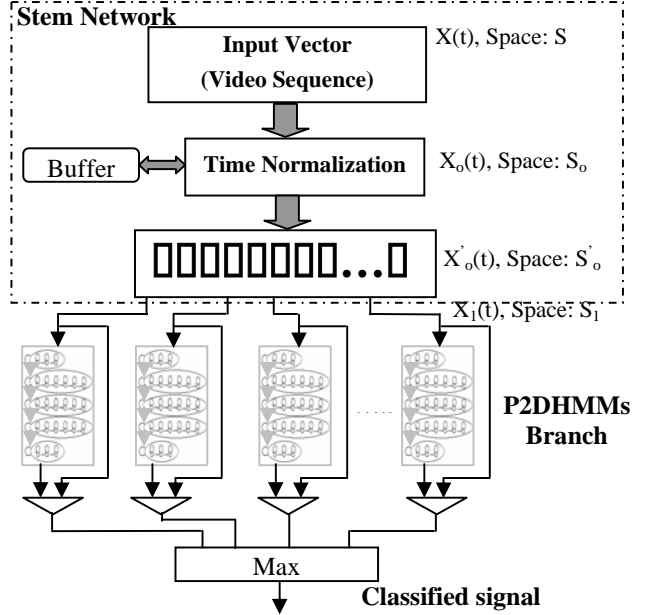


Fig. 4 The T-ComP2DHMM structure

#### 3.1. Stem Network

The Stem network receives an input vector and selects a subspace from the input space according to a similarity criterion. In a new T-ComP2DHMM structure the input space of the Stem network ( $S'_o$ ) is defined as a modified subspace of original input space  $S$ , and the P2DHMM input space  $S_1$  is the complementary subspace. This strategy permits a more efficient use of the spatial analysis capability of the layers. Given  $X$  a time sequence of  $N$  dimensional vectors  $x(t)$  belonging to the space  $S$ , with time length  $T(X)$  samples.  $x_o(t)$  is defined as sub-vector of  $x(t)$ ,  $x_o(t) \in S_o$ ;  $x_1(t)$  is a sub-vector of  $x$ ,

$$x_1(t) \in S_1 \text{ and } S_o \subset S \text{ and } S_1 \subset S \mid S = S_o \times S_1 \quad (12)$$

As we are dealing with temporal series, we need to create a space  $S'_o$  able to describe completely the time variations existing in the chosen subspace  $S_o$ . The vector  $\bar{x}(t)$  and the temporal length  $T(x)$  of the input signal can be random variable. So we propose the application of a Time Normalization procedure on the time independent space  $S'_o$ . This procedure is needed because a fixed dimension vector is due in the input of the LVQ that composes the Stem network. The dimension of the Stem input layer  $L$  is related to the original input time sequence  $X$  by

$$L = \text{Dim}(S_o) \times \bar{T} ; \bar{T} = N \mid N \in \mathcal{N} \text{ and } N \geq E\{T(X)\} \quad (13)$$

where  $\text{Dim}(S_o)$  is the dimension of the selected subspace  $S_o$  and  $\bar{T}$  is the first interger greater than the expected value of the time length  $T(X)$ . To obtain the vector  $X^1_o$

from the original time sequence  $X$ , we suggest the modeling of the time sampling series by  $\text{Dim}(S_o)$  continuous functions using cubic spline interpolation procedure. Given a  $N_o$  dimensional time series  $X_o^i = \{x_o^i(t_1), x_o^i(t_2), \dots, x_o^i(t_{T(x_o^i)})\}$  and defining a set of  $\text{Dim}(S_o)$  continuous associated function  $f_i(t) = \text{CubicSpline}(X_o^i)$  for  $i=1, 2, \dots, \text{Dim}(S_o)$ . Resampling each continuous function  $f_i(t)$  in  $T'$  sample points by  $\bar{x}_o^i(n) = f_i(n \times T_o)$ , where  $n=1, 2, \dots, \bar{T}$  and  $T_o$  is sample period in an appropriate time basis. The system input vector

$$X_o^i = [\bar{x}_o^i(1), \bar{x}_o^i(2), \dots, \bar{x}_o^i(\text{Dim}(S_o))], \bar{x}_o^i(1), \bar{x}_o^i(2), \dots, \bar{x}_o^i(\text{Dim}(S_o)) (2), \bar{x}_o^i(\bar{T}), \bar{x}_o^i(\bar{T}), \dots, \bar{x}_o^i(\text{Dim}(S_o))(\bar{T})] \quad (14)$$

The selection of the subspaces  $S_o$  and  $S_1$  is very important in the T-Comp2DHMM structure. The feature selection can be based on a class separability criterion that is evaluated for all of possible combinations of the input features. An Interclass Distance Measure criterion based on the distance between DCT vectors as follows:

$$J_s = \frac{1}{2} \sum_{i=1}^m P(w_i) \sum_{j=1}^m P(w_j) \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \delta(\bar{\xi}_{ik}, \bar{\xi}_{jl}) \quad (15)$$

where  $m$  is the number classes,  $P(w_i)$  is the probability of the  $i$ th class,  $N_i$  is the number of pattern vectors belonging to the class  $w_i$  and  $\delta(\bar{\xi}_{ik}, \bar{\xi}_{jl})$  is the DCT distance based measure from the  $k^{\text{th}}$  candidate pattern of the class  $i$  to the  $l^{\text{th}}$  candidate pattern of the class  $j$  defined by

$$\delta(\bar{\xi}_k, \bar{\xi}_l) = \left| \bar{\xi}_k - \bar{\xi}_l \right| = \sqrt{\sum_{j=1}^d (\xi_k^j - \xi_l^j)^2} \quad (16)$$

where  $d$  is the dimension of candidate space. In the T-P2DHMM context, it is needed to optimize the joint class separability of the Stem and P2DHMM. From the equation (15), the values of the class probability  $P(w_i)$  are possible to estimate only for the P2DHMM output. The classification criterion is based on the probability of error obtained for the P2DHMM. It can be modeled according

$$J_T = J_{\delta S} + \sum_{p=1}^m J_{\delta B_p} \quad (17)$$

where  $J_{\delta S}$  is the interclass distance measure for the Stem network and  $J_{\delta B_p}$  for the  $p^{\text{th}}$  P2DHMM, defined as

$$J_{\delta S} = \frac{1}{2} \sum_{i=1}^m P(\psi_i) \sum_{j=1}^m P(\psi_j) \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \delta(\bar{x}_{0_{ik}}, \bar{x}_{0_{jl}}) \quad (18)$$

where  $P(\psi_i)$  is the priori probability obtained for the  $i^{\text{th}}$  pseudo class designed by the Stem network training algorithm,  $\bar{x}_{0_{ik}}$  are the input vectors in the  $S'_o$  space, and  $m$  is the total number of vectors.

$$J_{\delta B_p} = \frac{1}{2} \sum_{i=1}^{m_p} P(w_{ip} | \psi_p) \sum_{j=1}^{m_p} P(w_{jp} | \psi_p) \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \delta(\bar{x}_{1_{ik}}, \bar{x}_{1_{jl}}) \quad (19)$$

where  $P(w_{ip} | \psi_p)$  is the conditional probability obtained for the  $i^{\text{th}}$  class of the  $p^{\text{th}}$  P2DHMM given the Stem.  $\bar{x}_{1_{ik}}$  are the input vectors allocated for the class  $i$  in the  $S_1$  space by the Stem network and  $m_p$  is the number of classes allocated to the Stem network. The set of features of  $\bar{x}_0$  and  $\bar{x}_1$  that maximize the class separability criterion given by equation (17) has chose to define the spaces  $S_o$ ,  $S_1$ , and  $S'_o$ . Doing this can required a huge amount of computation time to obtain an optimum solution, since the equation (17) must be evaluated for every  $C_{N_o}^N$  combination of the input space dimensions  $N$ .

### 3.2. Pseudo P2-DHMM Construction

Since hand images are two-dimensional, it is natural to believe that the 2DHMM, an extension to the standard HMM, will be helpful and offer a great potential for analyzing and recognizing gesture patterns. However a fully connected 2DHMMs lead to an algorithm of exponential complexity (Levin and Pieraccini, 1992). To avoid the problem, the connectivity of the network has been reduced in several ways, two among which are Markov random field and its variants (Chellapa and Chatterjee, 1985) and pseudo 2DHMM (Agazzi and Kuo, 1993). The latter model, called P2DHMM, is a very simple and efficient 2-D model that retains all of the useful HMM features. This paper focuses on the real-time construction of hand gesture P2DHMM. Our P2DHMM use observation vectors that are composed of two-dimensional Discrete Cosine Transform (2D DCT) coefficients. In addition, our gesture recognition system uses both the temporal and characteristics of the gesture for recognition. Unlike most other schemes, our system is robust to background clutter, does not use special glove to be worn and yet runs in real time. Furthermore, the method to combine hand region and temporal characteristics in P2DHMM framework is new contribution of this work. Use of both hand regions, features of location, angle, and velocity and motion pattern are also novel feature in this work.

**3.2.1 Description.** Pseudo 2-DHMMs in this paper are realized as a vertical connection of horizontal HMMs ( $\lambda_k$ ). However it is not the only one. In order to implement a continuous forward search method and sequential composition of gesture models, the former type has been used in this research. There are three kinds of parameters in the P2DHMMs. However, since the hand image is two-dimensional, we further divided the Markov transition parameters into super-state transition and state transition probabilities; each is denoted as

$$\bar{a}_{kl} = P(r_{t+1} = l | r_t = k), \quad 1 \leq k, l \leq N \quad \text{and}$$

$$a_{ij} = P(q_{t+1} = j | q_t = i), \quad 1 \leq i, j \leq M \quad (20)$$

where  $r_t$  denotes a super-state which corresponds to a HMMs  $\lambda_k$ , and  $q_t$  denotes a state observing at time  $t$ . The mode has  $N$  super-states and the HMMs  $\lambda_k$ , is defined as standard HMM consisting of  $M$  states.

**3.2.1 Evaluation algorithm.** Let us consider a  $t^{\text{th}}$  horizontal frame, observation future vector  $\vec{O}_t = \vec{o}_{1t}, \dots, \vec{o}_{st}$ ,  $1 \leq t \leq T$ . This is a one-dimension feature sequences like that of  $\vec{O}$  in

$$\Pr(\vec{O} | \lambda) = \sum_{\text{all } Q} \Pr(O, Q | \lambda) = \sum_{q_1, q_2, \dots, q_T} \pi_{q_1} b_{q_1}(\vec{o}_{1t}) \prod_{t=2}^T a_{q_{t-1}q_t} b_{q_t}(\vec{o}_{st}) \quad (21)$$

Here,  $q_t$  is one of the states from  $Q$ , the set of states, at time  $t$ . This is modeled by a HMM  $\lambda_k$ , with likelihood  $P(\vec{O}_t | \lambda_k)$ . Each HMM  $\lambda_k$  may be regarded as a super-state whose observation is a horizontal frame of states.

$$P_r(\vec{O}_t | \lambda_t) = \sum_{\text{all } Q} \Pr(\vec{O}_t, Q | \lambda_t) = \sum_{q_1, q_2, \dots, q_T} \pi_{q_1} b_{q_1}(\vec{o}_{1t}) \prod_{s=2}^S a_{q_{s-1}q_s} b_{q_s}(\vec{o}_{st}) \quad (22)$$

Now let us consider a hand region image, which we define as a sequence of such horizontal frames as  $\vec{O} = \vec{O}_1, \vec{O}_2, \dots, \vec{O}_T$ . Each frame will be modeled by a super-state or a HMM. Let  $\Lambda$  be a sequential concatenation of HMMs. Then the evaluation of  $\Lambda$  given feature sequence  $\vec{O}$  of the sample image  $X$  is

$$P(\vec{O} | \Lambda) = \sum_R P_1(\vec{O}_1) \prod_{t=2}^T \bar{a}_{r_{t-1}r_t} P_r(\vec{O}_t) \quad (23)$$

where it is assumed that super-state process starts only one from the first state. The  $P_r$  function is the super-state likelihood. Note that both of the Eqs. (22) and (23) can be effectively approximated by the Viterbi score. One immediate goal of the Viterbi search is the calculation of the matching likelihood score between  $\vec{O}$  and HMM. The objective function for an HMM is defined by the maximum likelihood as

$$\Delta(\vec{O}_t, \lambda_k) = \max_Q \prod_{s=1}^S a_{q_{s-1}q_s} b_{q_s}(\vec{o}_{st}) \quad (24)$$

where  $Q = q_1, q_2, \dots, q_s$  is a sequence of states of  $\lambda_k$ , and  $a_{q_0q_1} = \pi_{q_1}$ .  $\Delta(\vec{O}_t, \lambda_k)$  is the similarity score between two sequences of different length. The basic idea behind the efficiency of DP computation lies in formulating the expression into a recursive form

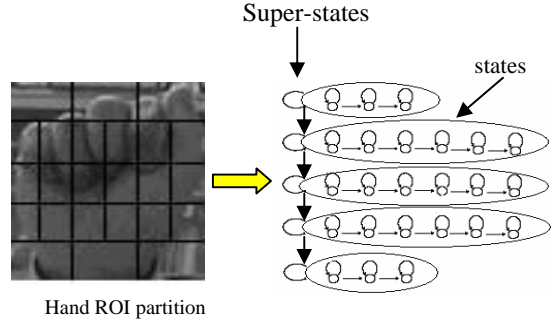
$$\delta_s^k(j) = \max_i \delta_{s-1}^k(i) a_{ij}^k b_j^k(\vec{o}_{st}), \quad j = 1, \dots, M_k, s = 1, \dots, S, k = 1, \dots, K$$

where  $\delta_s^k(j)$  denotes the probability of observing the partial sequence  $\vec{o}_{1t}, \dots, \vec{o}_{st}$  in model  $k$  along the best state sequence reaching the state  $j$  at time/step  $s$ . Note that  $\Delta(\vec{O}_t, \lambda_k) = \delta_S^k(N_k)$  where  $N_k$  is the final state of the

state sequence. The above recursion constitutes the DP in the lower level structure of the P2DHMM. The remaining DP in the upper level of the network is similarly defined by

$$D(\vec{O}, \Lambda) = \max_k \prod_{t=1}^T \bar{a}_{r_{t-1}r_t} \Delta(\vec{O}_t, \lambda_{r_t}) \quad (25)$$

that can similarly be reformulated into a recursive form. Here denotes the probability of transition from super-state  $r_1$  to  $r_2$ . According to the formulation described thus far, a P2DHMM add only one parameter set, i.e., the super-state transitions, to the conventional HMM parameter sets. Therefore it is simple extension to conventional HMM. Although simple in form, the time requirement is exponential. Thanks to the use of the DP technique, this can be computed in linear time in  $T$ . However when it comes to 2DHMM formulation, even the DP technique alone is not enough. One research direction is the structural simplification of the model, and the pseudo 2DHMM is one solution.



**Fig. 5** P2DHMM

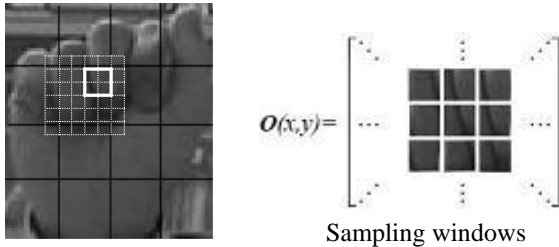
For each gesture there is a P2-DHMM, Fig. 5 shows a P2-DHMM model consists of 5 super-states and their states in each super-state that model the sequence of rows in the image. The topology of the super-state model is a linear model, where only self transitions and transitions to the following super-states are possible. Inside the super-states, these are linear one dimension hidden Markov model to model each row. The state sequence in the rows is independent of the state sequences of neighboring rows.

### 3.3. Representation of Visual Attributes by Subsets of DCT Coefficients

For hand gesture, our approach is to jointly model visual information that is localized in space, frequency, and orientation. To do so, we decompose visual appearance a long these dimensions. Below we explain this decomposition and in the next section we specify our visual attributes based on this decomposition. First, we decompose the appearance of the object into "parts" whereby each visual attribute describes a spatially localized region on the object. We would like these parts to be suited to the size of the features on each object. However, since important cues for hands at may size, we need multiple attributes over a range of scales. We will

define such attributes by making a joint decomposition in both space and frequency. We would like these parts to be suited to the size of the features on each object. Finally, by decomposing the object spatially, we do not want to discard all relationships between the various parts. We believe that the spatial relationship of the parts is an important cue for recognition. With this representation, each feature vectors now becomes a joint distribution of attribute and attribute position.

Using DCT coefficient as features instead of gray values of the pixels in the shift window where most of the image energy is found. They tend to be insensitive to image noise as well as image rotations or shifts, and changes in illumination. To create visual attributes that are localized in space, frequency, and orientation, we need to be able to easily select information that is localized along these dimensions. In particular, we would like to transform the image into a representation that is jointly localized in space, frequency, and orientation. To do so, we perform a DCT of image. DCT transform is not the only possible decomposition in space, frequency, and orientation. We use an overlap of 75% between adjacent sampling windows, we have also consider the neighboring sampling of a sampling window. Suppose we allow a deformation of up to  $\pm d$  ( $d$  is a positive integer) pixels in either X or Y directions. We have considered all the neighboring sampling within the distance  $d$  in order to detect a possible deformation. We use a shift window to improve the ability of the HMM to model the neighborhood relations between the sampling blocks.



**Fig. 6** The features extraction

To select the features which define the  $S_0$  subspace, the interclass separability measurement optimization procedure describe in the section 3.1 can be applied. In this work, we selected from the input vector , the two coordinates corresponding to the  $\vec{P}(t)$  hand position on the screen to compose the  $S_0$  subspace, generating two-space  $S_0$  and its complementary  $S_1$  space, to be used by T-CombP2DHMM model. It is efficient due to the natural uncorrelation existing in the hand posture, describable by  $S_1$ , and the hand trajectory, described by  $S_0$ . To obtain invariance of the motion to the camera relative position, we use the normalized velocity measurement  $\vec{P}(t)$  instead of absolute position, defined

$$\vec{P}(t_i) = \frac{\vec{P}(t_i) - \vec{P}(t_{i-1})}{|\vec{P}(t_i) - \vec{P}(t_{i-1})|} \quad \text{and} \quad \vec{P}(t_0) = 0 \quad (26)$$

Where  $\vec{P}(t_i)$  is the 2-D vector of the absolute position of the hand palm centroid in the screen at time  $t_i$ .

Using these definitions we are assigning the stem layer to analyze a normalized trajectory and the P2DHMMs to analyze fine hand postures variation for pre-selected trajectory.

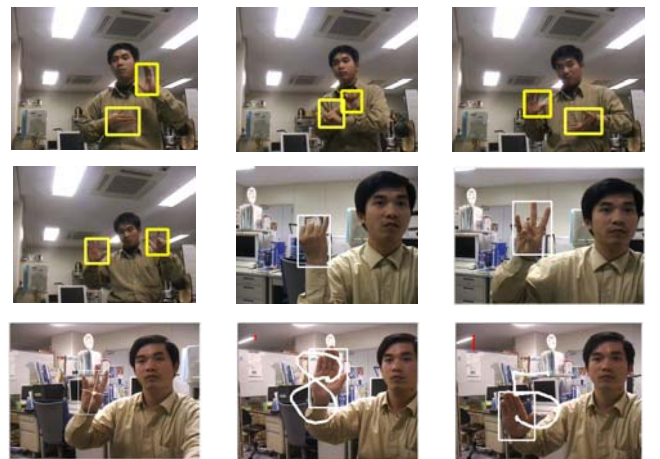
Each P2-DHMM is trained by hand gesture in the database obtained from the training set of each of the gesture using the Baum-Welch algorithm due to pre-analysis achieved by Stem network, which reduces the complexity of the problem. Then, the proposed T-CombP2DHMM structure has expected to be general and easily trainable.

For gesture recognition, The Viterbi algorithm is used to determine the probability of each hand model. The image is recognized as the hand gesture, whose model has the highest production probability. Due to the structure of the P2-DHMM, the most likely state sequence is calculated in two stages. The first stage is to calculate the probability that rows of the individual images have been generated by one-dimensional HMMs, that are assigned to the super-stages of the P2-DHMM. These probabilities are used as observation probabilities of the super-states of the P2-DHMM. Finally, on the level second Viterbi algorithm is executed.

## 4. EXPERIMENTAL RESULTS

### 4.1. Hand Detector

For our experiments, example is shown in the video clip that accompanies the paper and is available from the first author's website.<sup>1</sup> A few frames detector ran real-time from CCD camera are also shown in Figure 7.



**Fig. 7** The results of the hand detectors

<sup>1</sup> currently at

<http://www.fsai.kyutech.ac.jp/~ndbinh/Research/HandTracking.avi>

## 4.2. Results of T-CompP2DHMM Based Gesture Recognition

The training set consists of 36 hand gestures from vocabulary of 36 gestures including the ASL letter spelling alphabet and digits. Each one of the 36 gestures was performed 60 times by one person to create a database. The images of the same gesture were taken at different times. From the database, a set composed of 30 examples for each hand gesture is used to create the training set. The remaining 30 examples have reserved to the test set. Thus, each set has composed of 1080 images. The combined using of time “spatialization” and P2DHMM in the proposed T-CompP2DHMM model overcoming the classical approaches achieving a 98.5% of correct recognition rate. The example is shown in the video clip that accompanies the paper and is available from the first author’s website.<sup>2</sup>

**Table 1.** Recognition rates and complexities of HMMs for hand gestures recognition.

	Complexity	Recognition Rate
Classical 1D-HMM	$N_0 T_0^2$	85%
Optimized 1D-HMM	$N_0 T_0^2$	92%
Classical 2D-HMM	$(\sum_{k=1}^{N_0} N_1^{(k)})^2 T_0 T_1$	96%
P2D-HMM	$(\sum_{k=1}^{N_0} (N_1^{(k)})^2 T_1) T_0 + N_0^2 T_0$	98.5%
T-Com-P2HMM	$((\sum_{k=1}^{N_0} (N_1^{(k)})^2 T_1) T_0 + N_0^2 T_0) / M$	99.5%

$N_0$ = number of super states,  $N_1^{(k)}$  = number of states in the  $k$ 'th super state,  $T_0$  = number of vertical observations,  $T_1$  = number of horizontal observations,  $M$  = number of P2DHMMs branch )

## 5. CONCLUSIONS

This work presented a new feature extraction method using joint statistics of a subset of DCT coefficients to hand gestures, and introduced a new structure T-CompP2DHMM, dedicated to the time series recognition. The T-CompP2DHMM structure uses a Time Normalized Learning Vector Quantization in the Stem network and P2DHMM. We build the T-CompP2DHMM model based upon the P2DHMM, which allows it to do temporal analysis and to be used in large set of human movements’

<sup>2</sup> currently at

[http://www.fsai.kyutech.ac.jp/~ndbinh/Research/Real-timeDemo\\_NEW.wmv](http://www.fsai.kyutech.ac.jp/~ndbinh/Research/Real-timeDemo_NEW.wmv)

recognition system. The results obtained for a set of 36 different gestures of ASL show a 99.5 % of correct recognition rate. This results demonstrate that the joint use of time “spatialization” techniques, natural time processing techniques and P2DHMM given good results.

## 6. REFERENCES

- [1] T. Starner, and Pentland, “Real-Time American Sign Language Recognition from Video Using Hidden Markov Models”, TR-375, MIT Media Lab, 1995
- [2] O.E. Agazzi and S.S.Kuo, “Pseudo two-dimensional hidden markov model for document recognition”, AT&T Technical Journal, 72(5), pp. 60-72, Oct, 1993
- [3] N. D. Binh, E. Shuchi and T. Ejima, “Real-Time Hand Tracking and Gesture Recognition System”, Proceedings of International Conference on Graphics, Vision and Image Processing (GVIP-05), pp. 362 – 368, December, 2005
- [4] T. Kirishima, K. Sato, and K. Chihara, “Real-Time Gesture Recognition by Learning and Selective Control of Visual Interest Points”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 3, pp. 351 – 364, 2005
- [5] R. Lockton, A. W. Fitzgibbon, “Real-time gesture recognition using deterministic boosting”, Proceedings of British Machine Vision Conference, pp. 817 – 826, 2002
- [6] V.I. Pavlovic, R. Sharma, T.S. Huang, “Visual interpretation of hand gestures for human-computer interaction, A Review”, IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), pp. 677-695, 1997
- [7] J.Davis, M.Shah, “Recognizing hand gestures”. In Proceedings of European Conference on Computer Vision, ECCV, pp. 331-340, 1994
- [8] Hyeon-Kyu Lee, Jin H. Kim, “An HMM- based threshold model approach for gesture recognition”, IEEE Trans. Pattern Anal. Mach. Intell. 201(10), pp. 961-973, 1999
- [9] Ho-Sub Yoon, Jung Soh, Younglae J. Bae, Hyun Seung Yang. Hand gesture recognition using combined features of location, angle, velocity, Pattern Recognition 34, pp. 1491-1501, 2001
- [10] Pavlovic V.I., Sharma R., Huang T.S. Visual interpretation of hand gestures for human-computer interaction, A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7): pp. 677-695, 1997