World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:1, 2008

On-line Recognition of Isolated Gestures of Flight Deck Officers (FDO)

Deniz T. Sodiri, and Venkat V S S Sastry

Abstract**—** The paper presents an on-line recognition machine (*RM*) for continuous/isolated, dynamic and static gestures that arise in Flight Deck Officer (FDO) training. *RM* is based on generic pattern recognition framework. Gestures are represented as templates using summary statistics. The proposed recognition algorithm exploits temporal and spatial characteristics of gestures via dynamic programming and Markovian process. The algorithm predicts corresponding index of incremental input data in the templates in an on-line mode. Accumulated consistency in the sequence of prediction provides a similarity measurement (Score) between input data and the templates. The algorithm provides an intuitive mechanism for automatic detection of start/end frames of continuous gestures. In the present paper, we consider isolated gestures. The performance of *RM* is evaluated using four datasets - artificial (W_TTest), hand motion (Yang) and FDO (tracker, vision-based). *RM* achieves comparable results which are in agreement with other on-line and off-line algorithms such as hidden Markov model (HMM) and dynamic time warping (DTW). The proposed algorithm has the additional advantage of providing timely feedback for training purposes.

Keywords**—** On-line Recognition Algorithm, Isolated Dynamic/Static Gesture Recognition, On-line Markovian/Dynamic Programming, Training in Virtual Environments.

I. Introduction

RECENT advances in technology have put computers at the centre of daily life. Yet, lack of naturalness in the interaction methods with computers still encumber users. From that perspective, gesture, one of most used means of the communication among humans, has been investigated for potential interaction scheme in some domains in the recent decades.

In the context of human computer interaction, a gesture is defined as "... expressive, meaningful, body motion -i.e., physical movement of the fingers, hands, arms, head, face or body with the intent to convey information or interact with the environment." [19].

The present study is motivated by a need to recognize automatically Flight Deck Officer (FDO) gestures for training purposes. FDOs are in charge of ensuring craft and maintaining operational status and readiness. For example, safe conduct of flight deck operations for helicopter such as launching and recovering on board are some of their responsibilities. This study aims to remove the role of the instructor, by automatically recognizing FDO's gestures to provide natural means to interact with the virtual environment during training sessions. In addition to that, a feedback has to be provided to the trainee about his/her performance for training purposes.

Gesture recognition problem is akin to temporal pattern recognition problem. It has common properties with other temporal pattern problems such as speech and hand writing recognition. For these problems and gesture recognition, a wide range of recognition techniques have been proposed with various success rate. Neural network [5], [20], [12], [21], dynamic time warping, [11], [4] , hidden markov model [10], [16], [13] and some other ad hoc methods [8] are among these techniques. But most of these efforts do not readily lend themselves to on-line recognition. During the last decade, Hidden Markov Models (HMM) and its variations, hybrid and extensions thereof, have attracted a huge attention. Subsequent to the development of HMM toolkits such as HTK [18], several applications have been developed in temporal pattern recognition domain.

In this paper, the authors propose an on-line recognition machine (*RM*) for dynamic and static gesture under a generic recognition framework. *RM* consists of classical pattern recognition components such as preprocessing, modelling/analysis, language and recognition algorithm. The characteristic features of the proposed *RM* are : its ability to address inter/intra personal spatial and temporal variance, ability to deal with both dynamic and static gestures in a continuous or segmented gesture streams, determine the start and the end of a gesture as part of recognition task and construct a base for additional feedbacks, assessments for training purposes. The recognition algorithm conceptually is an on-line template matching technique and it involves aspect of dynamic programming technique and Markovian process.

The remainder of the paper is organized as follows: A formal definition of problem and related issues are presented in Section 2. Then, an overview of the proposed recognition machine and its components are elaborated in Section 3. The components of *RM* are detailed under two subsections - gesture modelling/analysis and recognition algorithm. In Section 4, a comparison of the proposed algorithm with HMM and DTW and other possible techniques for the components are discussed. The performance of the proposed algorithm is evaluated using four data sets - artificial data set [9] , FDO data (tracker and visions-based) sets and hand motion data set ,Yang [6], in Section 5. In the last section, conclusion and future work are presented.

II. Definition of the Problem

The task at hand needs to address some of the following issues: Spatial and temporal variance; repeatability and con-

Deniz T Sodiri is a PhD student in Defence Academy of UK, Cranfield University, Swindon, UK (e-mail: d.turan@cranfield.ac.uk).

Dr. Venkat V S S Sastry is a senior lecturer and director of scientific computing in Defence Academy of UK, Cranfield University, Swindon, UK (e-mail:v.v.s.s.sastry@cranfield.ac.uk).

nectivity; start/end frame detection [13]. While spatial variance accommodates shape, rotational and translational variations in space, temporal variance accounts for velocity changes. In addition to these variations, in a continuous gesture stream, like in a sign language, consequently multiple repetition of gestures or transition from one gesture to another gesture, makes the recognition task non-trivial as it involves detection of completion of a gesture (segmentation). Specifying the start and end frame in advance or during performance is also a burden. It interferes with the naturalness of the interaction. On the other hand, an automatic prediction of start/end frames of gestures makes problem more challenging . Note that, for example, in speech recognition, silence is used as a delimator for start and end of a word. In gesture recognition, spatial and temporal properties of unintentional or undefined movement and genuine gestures are potentially similar. In addition to these, the problem is sensitive to environmental noise as well.

Taking into account these issues, the problem can be stated as a five-tuple *(C, L, H, F, B)*. C accounts for gesture models with cardinality of ϖ. Thus $C = (C_1, C_2, C_3, \cdots C_{\varpi},)$. Length or period of gesture models are represented with the set L ($L = (l_1, l_2, l_3 \cdots l_{\varpi})$). Each gesture may have different period. A class C_i, consists of η number of channels ($H_{i,j}$) each of which constructed with a sequence of a feature f_j from the feature or alphabet set ($F = \{f_1, f_2, f_3 \cdots f_\eta\}$). For convenience, features at a certain time is referred to as a frame in this article and sequence of frames determine a gesture. B is an η dimensional input bands or channels which consist of the historical set of incremental frames (b_t). Each cell or unit in the band, contains one frame. Since, the frames are obtained incrementally, at a time *t*, only the present and previous data on the band are accessible. Thus, $B = \{b_1, b_2, b_3, \cdots, b_t\}$.

Similar to Pavlovic's [14], a temporal class or gesture can be defined as follows :

A temporal class C_i, is a trajectory of frames in the form of channel templates ($H_{i,1\cdots\eta}$) in a η dimensional feature space F , over a defined time interval l_i.

In the present study, the gestures of interest are either static or dynamic. Static gestures are those that have certain poses or configuration where trajectories remain approximately same for the period (l_i). On the other hand, dynamic gestures are the motion whose trajectories vary spatially with time. Using the above notation, the problem can be stated as:

Given a sequence of input frames (and hence B) incrementally, develop an algorithm or a recognition machine (RM) to recognize the gestures to which it belongs.

III. Recognition Machine

Recognition machine is implemented according to the classical pattern recognition framework [14]. Figure 1 illustrates the components of RM. A brief outline of the recognition machine is as follows: The recognition machine (RM) has nine interacting components. RM is fed by a sequence of input frames or input band B, of which properties are defined in the problem statement. Subsequent to acquiring data incrementally from the band ($b(t)$) at each discrete time t, data is pre-processed. Then, pre-processed data (x), is matched with all

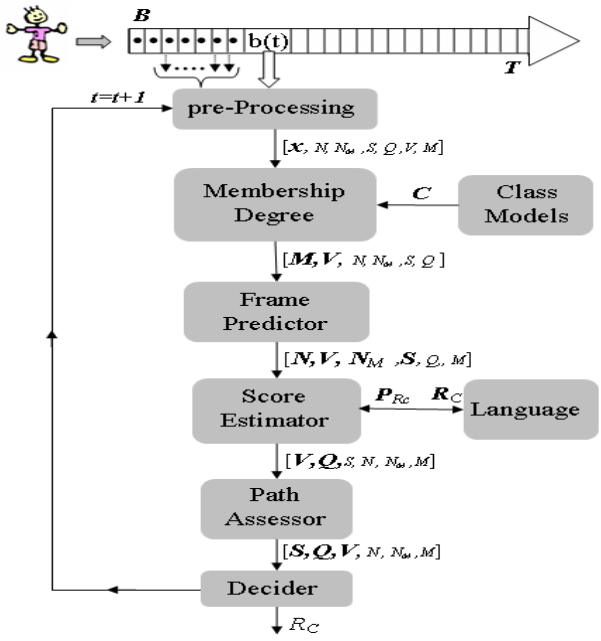

Fig. 1. Components and flow diagram of a recognition machine (C=Class Models; b, X=raw and processed current input frames; M=Membership degrees; V=Current Predicted Induces of Frames; N=Next Predicted Indexes; N_M=Membership Degree of Next Indexes; S=Scores; Q=Path Assessors; R_C=Recognized Class)

the channels of classes to obtain channel membership degree curves. In each class, channel membership degree curves are aggregated to obtain a final membership degree curve (M), which represents the membership degree of x to the class. In the frame predictor component, given the most recently predicted frame (V) and M , next frame (N) is predicted. Then, in the following component *(score estimator)*, scores (S) are estimated based on cumulative product of similarity factors (Θ), which consists of distance function (Ψ), and membership degree of predicted frames (M_N). In the final two components, some conditions are checked whether a recognition has emerged. The components of RM are further elaborated in the following two subsections: Analysis & Modelling and Recognition Algorithm. Note that, capital letters , such as V, M, N, S, M_N are used to denote a component related to all gestures. A subscript on these letters, V_i, M_i, N_i, M_{N_i} indicates the data related to the i^{th} class.

A. Analysis & Modelling

Analysing & Modelling part of RM is responsible for acquiring raw data from source, storing raw data on the input band (B), preprocessing, extracting feature and modelling classes. The main purpose of modelling stage is to build the templates for the gestures under consideration.

Recognition machine is fed by a η dimensional band (B). Typically, contents of B are obtained from source devices via input devices. Preprocessing component carries out smoothing, transformation and feature extraction tasks in that order. After smoothing the raw frames $b(t)$, necessary transformations are performed.

A template accumulates the trajectory of a class channel with two statistical parameters , mean (μ) and standard deviation (σ) at each discrete time step. It is assumed that at any discrete time, the underlying distribution is gaussian. The steps of constructing a template are described as follows: First step is to decide comprehensive and distinctive spatial and temporal feature vector(F). Due to temporal variance, training cycles have various lengths. Average length of all the training cycles of a class is used as the period of the class (l_i). Having estimated the period for the classes, then, all the training cycles are either stretched or compressed to the length of the period (L). In addition to that, sub events in the training cycles are aligned to occur at the same indexes during compression and stretching. Note that, these operations are performed only while constructing the templates. Finally, the aligned , stretched and compressed training cycles are used to construct the templates by using summary statistics (mean and standard deviation).

B. Recognition Algorithm

The recognition algorithm conceptually is an on-line template matching technique. The main idea behind recognition algorithm is to exploit sequential consistency of the input frames according to class models by using dynamic programming paradigm and Markovian process. Sequential consistency or so-called $Score$ (S) addresses similarity between the incremental input data and the class models. *Scores* employ similarity factors (Θ) for each class with an on-line sequential decision process which involves some predictions. The prediction process is a probabilistic estimation of the index of frames (N) in each class (C) which are spatially closest to the input frame (X) , given the most recently predicted frame index (V).

The following two metrics can be considered as similarity factors: A function of the distance ($\psi(.)$) between consecutive predicted frame index (N), and a membership degree of input frame to the predicted frames (M_N). The distance function ($\psi(.)$) utilizes the consistency along the sequence of predicted input frames index (N). A monotonic , steady incremental behaviour in the sequence of the predicted frame indexes points out consistency or similarity between the input frames and the class model of interest. In other words, small positive distances (Δ) between the consequent predicted frame indexes shows a possible recognition. A detailed and exemplified discussion of the distance function motivation can be found in [17]. In fact, the distance function is a type of radial basis function. Therefore, gaussian basis function ($e^{-\frac{\Delta^2}{2}}$) is used in this paper [1]. The similarity factors (distance function $\Psi(\Delta)$ and membership degree M_N) *score* of class C_i (S_i) are estimated as follows:

$$\begin{aligned} \Delta_{i,t} &= N_{i,t} - V_{i,t} \;\; ; \;\; \Psi(\Delta_{i,t}) = e^{-\frac{\Delta_{i,t}^2}{2}} \\ \Theta_{i,t} &= M_{N_{i,t}} \Psi(\Delta_{i,t}) = M_{N_{i,t}} e^{-\frac{\Delta_{i,t}^2}{2}} \\ S_i &= \prod_{t=1}^{T} \Theta_{i,t} \end{aligned} \quad (1)$$

Membership degree curves (M_i) estimation involves a partial on-line template matching operation ($M_i = P(X|C_i)$). It estimates the probabilities (M_i) of the input frame (X) belongs to the frames of each class model (C_i) in two stages. The first stage is a low level channel membership degree ($M_{i,j} = P(X_j|H_{i,j})$) estimation. The second phase is aggregation of channel membership degrees ($M_{i,1\cdots\eta}$) in order to obtain ultimate class membership degree (M_i). Membership degrees ($M_{i,j}$, M_i) are computed as follows:

$$M_i = \sqrt[\eta]{\prod_{j=1}^{\eta} M_{i,j}} \;\; ; \;\; M_{i,j} = e^{-\frac{(X - H_{\mu_{i,j}})^2}{2H_{\sigma_{i,j}}^2}} \quad (2)$$

where $H_{\mu_{i,j}}$ and $H_{\sigma_{i,j}}$ correspond to statistical mean and standard deviation parameters of the channels respectively. The parameter $M_{i,j}$ accommodates intra-membership degree redistribution. Intra redistribution regulates membership degrees among the indexes which are aligned during training phases because of temporal variances of sub events.

Frame predictor component predicts possible position of the input frame in the class templates given the membership degree curve (M_i) and most recently predicted frame index. Index of the local maxima (N_i) travels within the membership degree curve from beginning to end with a monotonic and increasing order, if the input data belongs to the classes. The input frame creates a local maxima in the membership degree curves wherever the frame is closer to the template frames. This characteristic of membership degree curve, namely position of the local maxima, serves to predict possible frame index. In the cases of multiple local maxima in the membership degree curves , nearest local maxima in the neighbourhood of the most recently predicted frame index is considered.

On-line prediction and piecewise matching operation paves way to resolve issues of temporal variance and identify *start/end* of a gesture. For each input frame, corresponding frames in the class templates are predicted. Therefore, these operations enable to detect start and end frame of gesture and adapt to temporal variances.

Even though , *score* (S) is one of the major measurements indicating similarities, it does not accommodate any information in itself what time or in what condition it is appropriate to declare a recognition. Order of predicted indexes or the path of observed indexes would help more accurate declaration. These operations are employed in the *Path Assessor* component. It prevents premature or wrong recognition and provides auxiliary information to the *decider* component, in order to evaluate all status and declare a recognition if one has emerged.

It is stated that in a consistent recognition, the predicted frame index N_i must be in an order, namely follow a monotonic increasing path from beginning to end within the membership degree curve (M_i). In this study, it is assumed that, M_i is consolidated by four consecutive part or milestones, Q_i=$\{q_{i,1}, q_{i,2}, q_{i,3}, q_{i,4}\}$ which are referred to as *path* in the rest of the paper. Each part occupies a quarter of class period ($0 < q_{i,1} < 0.25{*}l_i < q_{i,2} < 0.5{*}l_i < q_{i,3} < 0.75{*}l_i < q_{i,4} \leq q_{i,4}$). This component ensures that, all the parts are observed with a monotonic increasing order from $q_{i,1}$ to $q_{i,4}$. Note that,

$q_{i,4}$ is followed by q_1 for continuous recognition. If any jump occurs in the path , for example from $q_{i,1}$ to $q_{i,3}$ or $q_{i,4}$ rather than $q_{i,2}$, score and path will be reset ($S_i = 0, Q_i = q_{i,1}$).

In addition to these, it is also expected that, a sufficient N_i (a threshold, at least, 10 % of class period, $l_i/10$) has to be observed in each part to build a confidence for the observed path. Therefore, this component also holds the number of $N_i \neq V_i$ observations (path age QA) for each path part ($qa_{i,1}, qa_{i,2}, qa_{i,3}, qa_{i,4}$).

Having accumulated current status (path assessor, scores), now, it can be decided whether or not a recognition has emerged. Following conditions have to be met for an on-line recognition (R_C): 1- The path (Q_i) has to be in a sequential order in terms of the predicted frame indexes and N_i must be in the final part ($Q_i = q_{i,4}$). 2 - The duration in each part QA_i must be greater than a threshold eg. 10 % of class period. 3- S_i has to be maximized among the classes of which the paths include the final part ($Q_j = q_{i,4}$)

IV. Discussion

A class C_i can be thought of as a chain of L_i states (s_j), each of which consists of η channels. Approaching the template as a chain of states enable us to make the analogy between the proposed recognition algorithm and widely used algorithms such as Hidden Markov Model (HMM) and Dynamic Time Warping (DTW). HMM is a stochastic finite state automate , in which emission of observations and transitions between states are expressed in a probabilistic manner [2], [15]. DTW is an off-line template matching algorithm, in which time dimension is warped monotonically and increasingly in a window bandwidth , in order to minimize the distance between input and reference template. The proposed algorithm can be reduced to Hidden Markov Models as a special case. For example, the distance function (Ψ) and the membership degrees curves (M_N) approximately correspond the transition and the emission probabilities in HMM , respectively.

In the domain of gesture recognition for training purposes, the algorithm eliminates some issues of HMM such as training, decoding, evaluation [15]. Compared to speech, which is one of the main application area of HMM, a gesture trajectory is not as complex as speech. Therefore, unlike HMM, modelling of gesture data does not require *hidden* states which aims to represent unknown infrastructure. Gesture data or trajectories, roughly speaking, are well observable, unlike speech. Moreover, the proposed algorithm does not consist of training and modelling issues of HMM such as optimal number of states, topology, transition and emission probabilities . For example, in HMM, EM or Bauch-Welsher algorithm are used to estimate optimal transition probabilities, in a way to maximise the transition expectations. In this sense, the distance function directly employs the expectation which is that transition from a frame or state to the neighbourhood frames that are more probable than to the remote frames. Moreover, evaluation and decoding are run straight away in the proposed algorithm. Transparent decoding provides valuable feedback for training purposes and synthesis.

In addition to that, in on-line recognition, the proposed algorithm provides more control parameters (e.g. path assessors) to prevent premature or incorrect recognition, unlike HMM. Maximum likelihood criteria and some threshold mechanism are the only available methods when using HMM. It is worth noting that controlled recognition is critical for training and feedback. For example, in *Yang* and W_TTest2 experiments , it is observed that, while HMMs misrecognise some deformed and uncompleted gestures, the proposed algorithm rejects to any recognition , which is vital for a reliable training.

The proposed algorithm conceptually is a template matching technique in which time warping is employed in an on-line mode. In this sense, it is similar to dynamic time warping (DTW) apart from off-line mode. Recall that, DTWs make comparison between a reference and input template. But in the proposed algorithm, only an input frame X is compared with reference templates C_i. Moreover, in the proposed algorithm, since the distance operations are carried out over the membership degree curves (membership probabilities), the issue of common distance unit in DTW is eliminated.

The components of recognition algorithm have scope of further improvement. The task of some components can be carried out by other conventional algorithm. For example, the function of frame predictor component could be replaced by a function approximation algorithm such as RBN neural networks [1].

V. Experiments

In order to assess the performance of the proposed algorithm, four data sets are considered in this paper. The first data set is an artificial data set (W_TTest) which enables to perform parametric analysis. The remaining data sets come from real world applications involving user interactions in virtual environment (VE). The interaction gesture data set involves trajectories of hand motion while drawing shapes in a virtual environment. The final two data sets are related to FDO gestures which are gathered in two different ways , computer vision (FDO_CV) and tracker based (FDO_PT). Prior to conducting experiments, a PCA based similarity measurement (EROS) is applied to estimate intra disparity characterization of the data sets [22]. In this paper, we compare the performance of the proposed algorithm with HMM and DTW in an off-line fashion. But main emphasis is given to HMM in the experiments. Experiments are conducted in agreement with previously published studies [9], [6]. 10-fold cross validation scheme is used for training and testing.

DTW is implemented with 0.2 Sakoe-Chiba band windowing [3]. Class models of the recognition machine (C) are used as the reference templates in DTW and input templates are stretched or compressed to have identical length with the reference templates. HMM algorithm is applied using HTK toolkit [18]. Several configuration of states and topologies such as left to right (*lr*), left to right one skip(*lr1s*) and ergodic (*er*) are considered. EROS employs weighted Frobenious norms to the eigenvector and eigenvalues of principal components which are obtained from covariance matrices of temporal classes represented as matrices. In the evaluation part, EROS uses non

Fig. 2 Static and dynamic FDO gestures

parametric kNN neighbourhood scheme ($k = 1, 2, 3 \cdots 10$) to evaluate the disparity in data sets. Precision/Recall metrics in EROS accommodates proportion of k to the volume (recall) which consists of k number of samples of class of interest. High values of precision (% 100) indicates higher disparity in data set. Further information about EROS can be found in [22].

A. Synthetic data set_W TTest [9]

W_TTest is a parametic data set and consist of three classes A,B and C and each of which has three channels (α, β, γ). Period of classes are 100 time units. It must be noted that apart from a couple of frames, class A and B are identical to each other. In noisy circumstances, these distinctive frames could also disappear.

W_TTest addresses the following challenges: multiple channels, spatial variance, temporal variances in the form of periodic and sub event, gaussian noise and irrelevant channels. These challenges are controlled with following parameters : d, duration or periodic variance; c, variance in sub events' positions; h, variance in sub events' amplitude; g, noise level and *irrel*, irrelevant channels $A_\gamma, B_\gamma, C_\beta$ which seems to convey a message but in fact it is random and unrelated to the class. Apart from *irrel* , other parameters range in the interval of [0,1] where 0 indicates that the parameter of interest is *off*. *Irrel* parameter is either *on* or *off*. The noise is distributed uniformly and randomly in the data set. For a detailed definition of the dataset ,the interested reader is referred to [9]. Experiments are conducted over different values of noise levels $g = 0.1$ and $g = 0.2$ which are referred as W_TTest1 and W_TTest2 in the rest of paper. Other parameters are fixed as follows $h = d = 0.2, c = 0.1$, *irrel on*. Actually, W_TTest2 , due to high noise, accommodate unclassifiable samples (6-8 %) to check reliability of the algorithms. Both data set (W_TTest1, W_TTest2) consist of 1000 samples for each classes. Raw data is used as features in both experiments.

B. Gestures for Interaction in VE - Yang [6], [7]

Yang data set is a part of full body gesture data set compromising over 40 body motions [6], [7]. The gesture set consists of eight hand gestures of which is represented by three coordinates (x,y,z) at a given time. For each gesture, there are approximately 100 training samples. The quality of data set is very poor. Because of the shape of *circle*, *rectangle* and *triangle* gestures, there is a remarkable similarity in the data set. Each gesture is represented by following features: smoothed coordinate positions, their gradients, and angular velocity. Previous work on Yang data set achieves about 2.1 % recognition error [6].

C. Flight Deck Officer - Vision Based (FDO CV)

Computer Vision based gestures are collected via an average quality desktop web cam. Collected videos are pre-processed to extract the position of hands (x, y). Three different users performed the gestures. 18 out of over 40 FDO gestures are considered in the present paper (Figure 2, middle). These gestures accommodate all challenges which one would come cross during FDO's gesture recognition. Data set consist of four static gestures (Affirmative, Clean, Hold On, Negative), six dynamic gestures (Ahead, Back, Wave Off, Down ...) and eight hybrid gesture (Left, Right, Fire ...) , in which while one hand is static, the other hand is dynamic. The data set includes over 70 samples of each gesture. Each raw gesture is represented by stream of four coordinate data (x, y) for each hand. The coordinate data x, y and their gradients are used as feature vector.

D. Flight Deck Officer - Tracker Based (FDO PT)

Characteristics of this data set are similar to FDO_CV apart from a couple differences in the way data is collected, size of data set, and number of person performing the gestures. FDO_PT is collected via a tracker device (Polhemus FasTrak) of which two sensors acquire the position of hands in a three dimensional coordinate system (x, y, z). FDO_PT compromise about 150 samples for each class and these samples are collected only from a single person in different sessions. Similar to FDO_CV, each raw gesture is represented by stream of six coordinate data (x, y, z). But, for feature , coordinates of right and left hand (x, y, z) are transformed to angular features (α, β, γ) which correspond coordinate angles between axes and hand position in the local coordinate system. Hence, α, β, γ and their gradients are used as the feature vector.

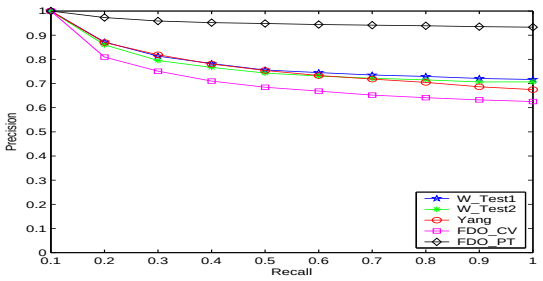

Fig. 3. Average Recall/Precision of data sets by using EROS ($k = 1, 2, 3, \cdots 10$)).

VI. Results

Prior to discussing recognition results, intra disparity of the datasets are analysed by EROS. For the sake of clarity and

TABLE I
HMM RECOGNITION ERROR RATES

	3lr	5lr	10lr	20lr	3lrs1	5lrs1	10lrs1	20lrs1	3er	5er	10er	20er
W_TTest1	0.6∓2.2	0.1∓0.5	0∓0	1.1∓3.6	1.4∓1.8	0.1∓0.5	2.3∓5.4	6.2∓11.9	18.2∓20.7	16.83∓25.7	17.1∓27	22.2∓38.5
W_TTest2	4.3∓3.7	3∓2.6	8.3∓7.6	14.1∓12.7	4.4∓3.7	2.9∓2.4	18.9∓25.2	19.8∓23.2	23∓23.9	24.8∓27	20.7∓19	16.3∓2.6
Yang	0∓0	0∓0	0∓0	0∓0	0∓0	0.3∓0.6	0.3∓0.6	0.2∓0.5	1.7∓1.4	1.3∓1.6	1.0∓1.6	1.6∓1.7
FDO_CV	1.2∓4.4	0.4∓2.0	0.4∓2.1	0.3∓1.8	1.2∓4.8	0.8∓3.8	0.4∓2.4	0.9∓4.1	1.6∓6.1	1.1∓3.7	1.9∓2.7	2.5∓4.2
FDO_PK	0.1∓0.3	0.1∓0.3	0∓0	0.2∓0.4	0∓0	0.1∓0.4	0.2∓0.5	0.2∓0.4	0∓0	0∓0	0.1∓0.3	0.2∓0.5

W_TTest1 ($g = 0.1$) W_TTest2 ($g = 0.2$), Yang and FDO recognition results when using HMM with different states (3,5,10, 20) and topologies , left to right (lr), left to right skip 1 (lrs1) and ergodic (er). 10 fold cross validation scheme is applied to all data sets.

TABLE II
RECOGNITION ERROR RATES (%)

	RM	HMM	DTW	EROS
W_TTest1	0.93∓0.43	0∓0	4.73∓4.05	21.32∓8.90
W_TTest2	7.83∓2.32	2.9∓2.47	14.33∓7.46	22.54∓9.25
Yang	0.86∓1.00	0∓0	27.08∓27.07	22.60∓9.99
FDO_CV	1.91∓1.47	0.3∓1.8	5.63∓14.71	28.25∓11.50
FDO_PT	0.09∓0.14	0∓0	0.03∓0.01	4.76∓2.05

Recognition Error Results in Percentages (%) for online RM and HMM , DTW and EROS. For HMM, best results of the Table I is shown.

space, demonstration of intra class disparity of the datasets by EROS are omitted here. But the following observations are worth noting: There is remarkable similarity between class A and B in W_TTest1 and between *Rectangle* and *Triangle* gestures in Yang data set. Similar disparity results are also obtained for W_TTest2 data set. In FDOs, *Negative* and *Affirmative* gestures are similar due to common spatial and temporal properties. They are both static gestures and, except γ channel (z coordinate), other channels are same. Therefore, unique eigenvectors and eigenvalues are not formed for *Negative* and *Affirmative* gestures in EROS. Other gestures in FDOs are quite dissimilar. The figure 3 illustrates average cross disparity in all data sets in which the disparity in FDO_PT data set is higher compared to other data sets. It is worth noting that FDO_PT data set has higher disparity than FDO_CV. This difference is largely due to reduced number of channels in FDO_CV, combined with larger variation in number of users while collecting FDO_CV data set.

Table I shows HMM recognition error for all data sets. HMM experiments indicate that they perform better when state number is smaller (3,5, 10) and topology is left to right. It can be concluded that, left to right topology is more appropriate for gesture recognition task. Similarly, although RM is ergodic topology, its frame prediction component is biased to make prediction from left to right direction. Even though, HMM obtains comparable results as RM, during decoding of state sequence in HMM, small number of states, does not provide meaningful feedback which is critical for the training purposes.

Another important point of this study is that HMMs make strong assumptions during recognition decision. HMMs declare recognitions even in the cases where recognitions are impossible or unreliable due to high noise and missing data (table II). For example , in W_Test2 data set, because of high noise (g=0.2),in some cases (6-8 %), sub events emerge in the β channel of class B similar to class A, which make it impossible to segregate class A and B. Similarly, in Yang data set, becasue of high noise and missing data, some gestures barely can be classified by even a human. Even those , as the result tables indicate HMMs make over estimation and assign them to a class without considering the quality of the signal. In these circumstances, unlike HMMs, RM rejects to declare a defined class recognition and declares a non defined class recognition ($R_C = C_{NON}$). This advantage of RM is achieved by some heuristics along side with maximum likelihood criteria employed in the *path assessor* components.

Finally, the table II compares the recognition error (%) of the proposed algorithm (RM) with other algorithms. Best results of HMMs experiments from table I are shown in the table II. EROS column shows the average precision of cross disparity for all neighbourhoods ($k = 1, 2, 3 \cdots 10$) for the data set of interest.

The proposed algorithm (RM) achieves remarkable results compared to HMMs and other algorithm, considering that RM is an on-line algorithm and others are off-line, the performance are comparable. It is observed that in W_TTest2 data set, performance of RM is decreased because of high noise ($g = 0.2$), which deforms and diminishes the sub events of the classes.

VII. CONCLUSION

In this paper, we proposed an on-line recognition machine for gesture of FDO in the context of a training application. Recognition machine is based on the generic pattern recognition framework. Gestures are represented in a template form. Recognition algorithm is based on dynamic programming and markovian process and it conceptually implements an on-line template matching scheme. The algorithm predicts the index of an input frame in each class templates. Consistency in the sequence of prediction scores provides a merit for recognition. In addition, the prediction process paves way for automatic detection of start/end frames of gestures in a continuous stream by exploiting path heuristics.

The proposed algorithm (RM) is compared with HMM and DTW algorithm using a parametric artificial data set (W_Test) and three real word data sets (Yang, vision and tracker based FDO). Even though, RM is an on-line algorithm and uses limited historical data, it achieves comparable results on segmented data. Controlled declaration of recognition in the cases

of high noise and missing data provides an advantage over HMM. Moreover, RM provides more meaningful feedbacks (longer observed trajectory) by employing more observable states , unlike HMM which is generally successful on small configuration (3-5 states) and on hidden states. It is worth emphasizing that DTW are primarily designed for off-line recognition , while the RM algorithm has been designed to deal with continuous gestures. Preliminary results of continuous gesture recognition of RM are promising and RM achieves far better results compare to HMM, which will be presented in a future paper.

It is proposed to extend the present study for recognition of continuous gestures which forms part of gesture dialogues in the FDO training application.

REFERENCES

[1] Christopher M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1996.

[2] Herve Bourlardy and Samy Bengio. *The Handbook of Brain Theory and Neural Networks*, chapter Hidden Markov Models and other Finite State Automata for Sequence Processing. The MIT Press, second edition, 2002.

[3] E. Keogh C. A. Ratanamahatana. Everything you know about dynamic time warping is wrong. *Third Workshop on Mining Temporal and Sequential Data, in conjunction with the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[4] Andrea Corradini. Dynamic time warping for off-line recognition of a small gesture vocabulary. In *Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01)*, page 82. IEEE Computer Society, 2001.

[5] S. Sidney Fels and Geoffrey E. Hinton. Glove-Talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE Transactions on Neural Networks*, 4(1):2–8, January 1993.

[6] Yang-Hee Nam Jane Koh. Full-body motion recognition using principle component based target reduction. In *KIPS(Korean Information Processing Society) Proceedings*, volume Vol. 11 , no.1, pages 873–876, Korea, May 2004.

[7] Yang-Hee Nam Jane Koh, Eun-Woo Lee. Full-body motion recognition using multi-phase target reduction method. *HCI 2004(Korean)*, 2004.

[8] Yangsheng Xu Jie Yang. Hidden markov model for gesture recognition. Technical Report CMU-RI-TR-94-10, The Robotics Institute, Carneige Melon University, 1994.

[9] M. W. Kadous. *Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series.* PhD thesis, The University of New South Wales, School of Computer Science and Engineering, 2002.

[10] C. Lee and Y. Xu. Online, interactive learning of gestures for human/robot interfaces, 1996.

[11] H. Li and M. Greenspan. Continuous time-varying gesture segmentation by dynamic time warping of compound gesture models. 2005.

[12] Kouichi Murakami and Hitomi Taguchi. Gesture recognition using recurrent neural networks. In *CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 237–242, New York, NY, USA, 1991. ACM Press.

[13] Y. Nam and K. Wohn. Recognition of space-time handgestures using hidden markov model, 1996.

[14] Vladimir Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695, 1997.

[15] Rabiner L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77, (2):257–286, Feb 1989.

[16] Gerhard Rigoll, Andreas Kosmala, and Stefan Eickeler. High performance real-time gesture recognition using hidden markov models. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 69–80, London, UK, 1998. Springer-Verlag.

[17] D T Sodiri and V V S S Sastry. On the interpretation of gestures arising in flight deck officers training. In *Proceedings of the Thirteenth Conference on Behaviour Representation in Modelling and Simulation*, 2004.

[18] Thomas Hain-Phil Woodland Steve Young, Gunnar Evermann. *The HTK Book, 3.2.1*. Cambridge Research Laboratory Ltd, 2002.

[19] M Turk. *Handbook of virtual environments: Design, implementation, and applications*, chapter Gesture recognition, pages 223–238. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2002.

[20] P. Vamplew and A. Adams. Recognition and anticipation of hand motions using a recurrent neural network, 1995.

[21] Simei G. Wysoski, Marcus V. Lamar, Susumu Kuroyanagi, and Akira Iwata. A rotation invariant approach on static-gesture recognition using boundary histograms and neural networks.

[22] Kiyoung Yang and Cyrus Shahabi. A pca-based similarity measure for multivariate time series. In *Proceedings of the 2nd ACM international workshop on Multimedia databases*, pages 65–74. ACM Press, 2004.