Real-Time Gesture Recognition, Evaluation and Feed-Forward Correction of a Multimodal Tai-Chi Platform

Otniel Portillo-Rodriguez, Oscar O. Sandoval-Gonzalez, Emanuele Ruffaldi, Rosario Leonardi, Carlo Alberto Avizzano and Massimo Bergamasco.

PERCRO, Perpcetual Robotics Laboratory.
Scuola Superiore Sant'Anna. Pisa Italy

Abstract. This paper presents a multimodal system capable to understand and correct in real-time the movements of Tai-Chi students through the integration of audio-visual-tactile technologies. This platform acts like a virtual teacher that transfers the knowledge of five Tai-Chi movements using feed-back stimuli to compensate the errors committed by a user during the performance of the gesture. The fundamental components of this multimodal interface are the gesture recognition system (using k-means clustering, Probabilistic Neural Networks (PNN) and Finite State Machines (FSM)) and the real-time descriptor of motion which is used to compute and qualify the actual movements performed by the student respect to the movements performed by the master, obtaining several feedbacks and compensating this movement in real-time varying audio-visual-tactile parameters of different devices. The experiments of this multimodal platform have confirmed that the quality of the movements performed by the students is improved significantly.

Keywords: Multimodal Interfaces, real-time 3D time-independent gesture recognition, real-time descriptor, vibrotactile feedback, audio-position feedback, Virtual Realty and Skills transfer.

1 Introduction

The learning process is one of the most important qualities of the human being. This quality gives us the capacity to memorize different kind of information and behaviors that help us to analyze and survive in our environment. Approaches to model learning have interested researches since long time, resulting in such a way in a considerable number of underlying representative theories.

One possible classification of learning distinguishes two major areas: Non-associative learning like habituation and sensitization, and the associative learning like the operant conditioning (reinforcement, punish and extinction), classical conditioning (Pavlov Experiment), the observational learning or imitation (based on the repetition of a observed process) [1], play (the perfect way where a human being can practice and improve different situations and actions in a secure environment) [2], and the multimodal learning (dual coding theory) [3].

Undoubtedly, the imitation process has demonstrated a natural instinct action for the acquisition of knowledge that follows the learning process mentioned before. One example of multimodal interfaces using learning by imitation in Tai-chi has been

applied by the Carnegie Mellon University in a Tai-Chi trainer platform [4], demonstrating how through the use of technology and imitation the learning process is accelerated.

The human being has a natural parallel multimodal communication and interaction perceived by our senses like vision, hearing, touch, smell and taste. For this reason, the concept of Human-Machine Interaction HMI is important because the capabilities of the human users can be extended and the process of learning through the integration of different senses is accelerated [5] [6] [7]. Normally, any system that pretends to have a normal interaction must be as natural as possible [8] [9].However, one of the biggest problems in the HMI is to reach the transparency during the Human-Machine technology integration.

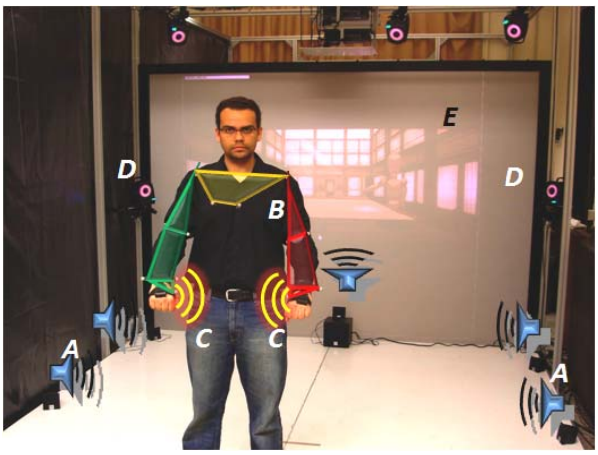

Fig. 1. Multimodal Platform set up, A) 3D sound, B) Kinematics Body C) Vibrotactile device (SHAKE) D) Vicon System E) Virtual Environment

In such a way, the multimodal interface should present information that answers to the "why, when and how" expectations of the user. For natural reasons exists a remarkable preference for the human to interact multimodally rather than unimodally. This preference is acquired depending of the degree of flexibility, expressiveness and control that the user feels when these multimodal platforms are performed [9]. Normally, like in real life, a user can obtain diverse information observing the environment. Therefore, the Virtual Reality environment (VR) concept should be applied in order to carry out a good Human-Machine Interaction. Moreover, the motor learning skills of a person is improved when diverse visual feedback information and correction is applied [10].

For instance the tactile sensation, produced on the skin, is sensitive to many qualities of touch. Lieberman and Breazeal [11] carried out, for first time, an experiment in real time with a vibrotactile feedback to compensate the movements and accelerate the human motion learning. The results demonstrate how the tactile feedback induces a very significant change in the performance of the user. In the same line of research Boolmfield performed a Virtual Training via Vibrotactile Arrays[12].

Another important perception variable is the sound because this variable can extend the human perception in Virtual Environments. The modification of parameters like shape, tone and volume in the sound perceived by the human ear [13], is a good approach in the generation of the description and feedback information in the human motion.

Although a great grade of transparency and perception capabilities are transmitted in a multimodal platform, the intelligence of the system is, unquestionably, one of the key parts in the Human-Machine interaction and the transfer of a skill. Because of the integration, recognition and classification in real-time of diverse technologies are not easy tasks, a robust gesture recognition system is necessary in order to obtain a system capable to understand and classify what a user is doing and pretending to do.

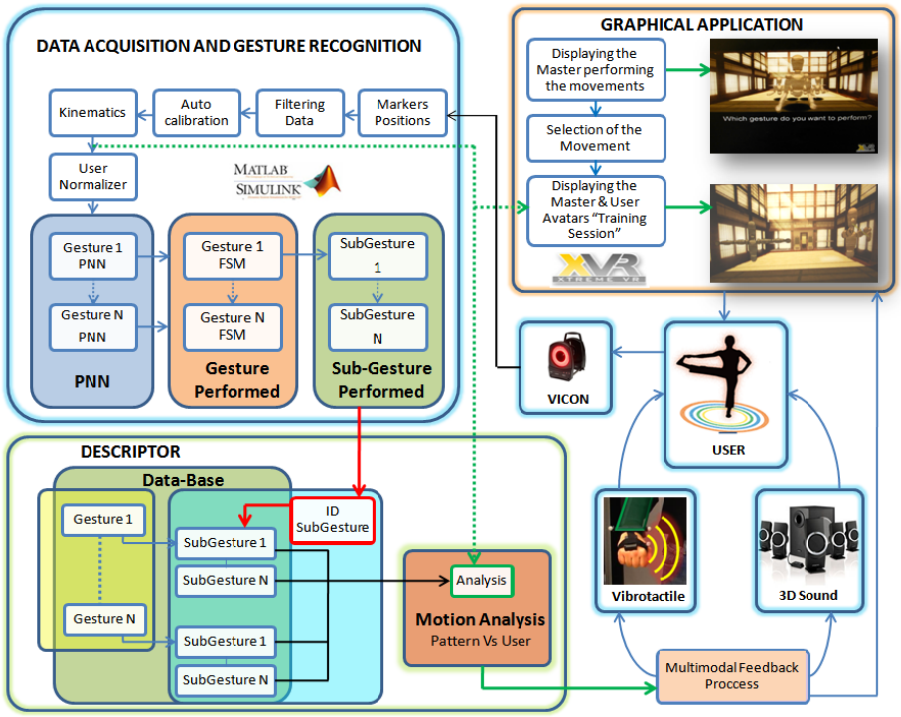

Fig. 2. Architecture of the Multimodal Tai Chi Platform System

2 System Implementation

This paper presents a multimodal interface that teaches to novel students, five basic tai chi movements. Each movement is indentified and analyzed in real time by the gesture recognition system. The gestures performed by the users are subdivided in n-states (time-independent) and evaluated step-by-step in real time by the descriptor system. Finally, the descriptor executes audio-visual-tactile feed-back stimuli in

order to correct the user's movements. Fig. 1 presents the interface that is composed by: The hardware and software of the 3D tracking optical system (VICON), the gesture recognition system and the description of motion (both running in Matlab Simulink), a graphical scenario developed in XVR, a 3D sound system and the wireless vibrotactile devices (SHAKE). The general architecture of the multimodal platform is shown in Fig. 2.

2.1 Data acquisition

The motion of the Tai-Chi student was tracked with the VICON system. This system is an optical device which provides millimeter accuracy in the 3D space through the use of passive reflective markers attached to the body at 300Hz of sampling frequency. Sometimes, due to the markers obstructions in the human motion, the data information is lost. For this reason, the "cleaning algorithm" described in [14], was implemented. An inverse kinematics of fourteenth DOFs represented by the upper part of the body is computed. A calibration process is completely required in order to identify the actual position of the markers and adjust the kinematics model to the new values. Therefore, a fast (1ms) autocalibration process was designed in order to obtain the initial position of the markers of a person placed in a military position called "stand at attention". The algorithm checks the dimension of his/her arms and the position of the markers. The angles are computed and finally this information is compared with to the ideal values in order to compensate and normalize the whole system.

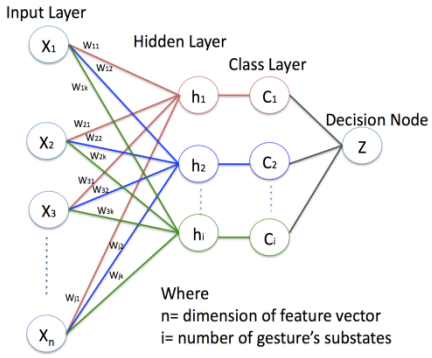

Fig. 3. PNN architecture used to estimate the most similar gesture's state from the current user's body position

2.2 Real-Time Gesture Recognition Process

In order to recognize the gesture performed by the user, a state space model approach was selected [15][16]. Normally, the principal problem to model a gesture in the state based approach, is the characterization of the optimal number of states and the establishment of their boundaries. For each gesture, the training data is obtained concatening the data of five demonstrations. A dynamic *k*-means clustering on the

training data defines the number of states and their spatial parameters of the gesture without temporal information [17]. This information from the segmented data is then added to the states and finally the spatial information is updated. This produces the state sequence that represents the gesture. The analysis and recognition of this sequence is performed using a simple Finite State Machine (FSM) [18], instead of use complex transitions conditions which depend only of the correct sequence of states for the gesture to be recognized and eventually of time restrictions i.e., minimum and maximum time permitted in a given state.

The novel idea is to use for each gesture a PNN to evaluate which is the nearest state (centroid in the configuration state) to the current input vector that represents the user's body position. The input layer has the same number of neurons as the input vector and the second layer has the same quantity of hidden neurons as states have the gesture. In our architecture (Fig. 3), each class node is connected just to one hidden neuron and the number of states (where the gesture is described) defines the quantity of class nodes. Finally, in the last layer, the class (state) with the highest summed activation is computed. A number of 12 variables were used in our configuration space: There are 2 distances between hands and 2 between elbows. 2 Vectors created from the XYZ position from the hands to the chest and 2 Vectors created from the XYZ positions from the elbows to the chest.

2.3 Real-Time Descriptor Process

The comparison and qualification in real-time of the movements performed by the user is computed by the descriptor system. In other words, the descriptor analyzes the differences between the movements executed by the expert and the movement executed by the student, obtaining the error values and generating the feedback stimuli to correct the movement of the user. Each pattern movement is characterized for a sequence of states which is formed by 18 variables and performs the comparison of the following information: *12 Angles:* Elbows(2), Wrists(4) and Shoulders(6), *2 Distances:* Distance between hands(1) and elbows(1) and *4 Positional vectors:* 2 vectors created from the XYZ position of the hands to the chest and 2 Vectors created from the XYZ positions of the elbows to the chest. Each state or subgesture is recognized in real time by the gesture recognition system during the performance of the movement. Using the classic feed-back control loop during the experiments was observed that the user feels a delay in the corrections. For that reason, a feed-forward strategy was selected to compensate this perception. In this methodology when a user arrives at one state of the gesture, the descriptor system creates n-substates and carries out an interpolation process to compare the actual values with respect to the values in the sub-state *(n+1)* of the pattern value, creating a feed-forward loop which estimates in advance the next correction values of the movement. The error is computed by:

$$\theta error = [P(n+1) - U(n)] * Fn \quad (1)$$

Where $\theta error$ is the difference between the pattern and the user, P is the pattern value, U is the user value, Fu is the normalize factor and n is the actual state.

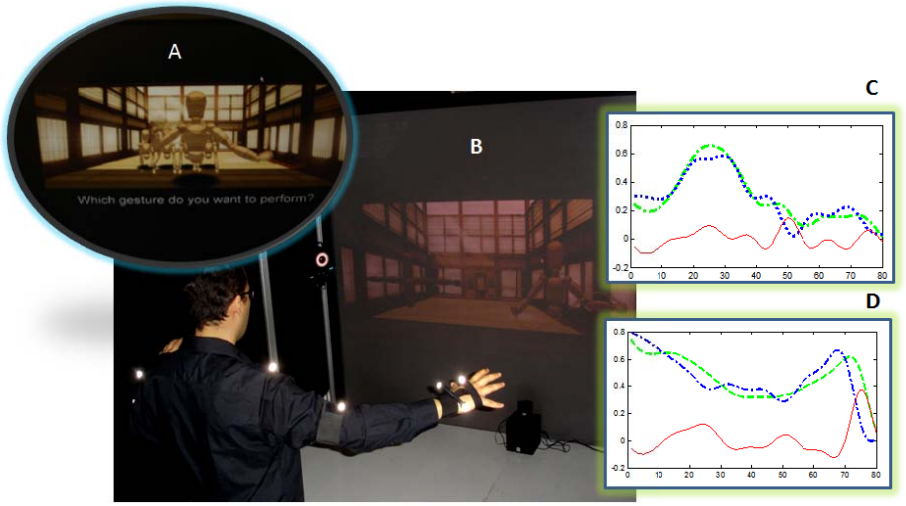

Fig. 4. VR environment , A) Initial Screen, 5 avatars performing Tai-Chi movements, B) Training session, two avatars, one is the master and second is the user. C) Distance of the Hands, D) Right Hand Position.

2.4 Virtual Reality Platform

The virtual environment platform which provides the visual information to the user was programmed in XVR. There are 3 different sequences involved in this scenery. The first one is the initial screen that shows 5 avatars executing different Tai Chi movements. When a user tries to imitate one movement, the system recognizes the movement through the gesture recognition algorithm and passes the control to the second stage called "training session". In this part, the system visualizes 2 avatars, one represents the master and the other one is the user. Because learning strategy is based on the imitation process, the master performs the movement one step forward to the user. The teacher avatar remains in the state*(n+1)* until the user has reached or performed the actual state*(n)*. With this strategy the master gives the future movement to the user and the user tries to reach him. Moreover, the graphics displays a virtual energy line between the hands of the user. The intensity of this line is changing proportionally depending on the error produced by the distance between the hands of the student. When a certain number of repetitions has been performed, the system finishes the training stage and displays a replay section which shows all the movements performed by the student and the statistical information of the movement's performance. Fig. 4 (A)(B) shows the virtual Tai-Chi environment.

2.5 Vibrotactile Feedback System

The SHAKE device was used to obtain wireless feedback vibrotactile stimulation. This device contains a small motor that produces vibrations at different frequencies. In this process, the descriptor obtains the information of the distance between the hands, after this, the data is compared with the pattern and finally sends a proportional

value of the error. The SHAKE varies proportionally the intensity of the vibration according to error value produced by the descriptor (1 Hz – 500 Hz). This constraint feedback is easy to understand for the users when the arms have reached a bad position and need to be corrected. Fig. 4 (C) shows the ideal distance between the hands (green), the distance between the hands performed by the user (blue) and the feed-back correction (red).

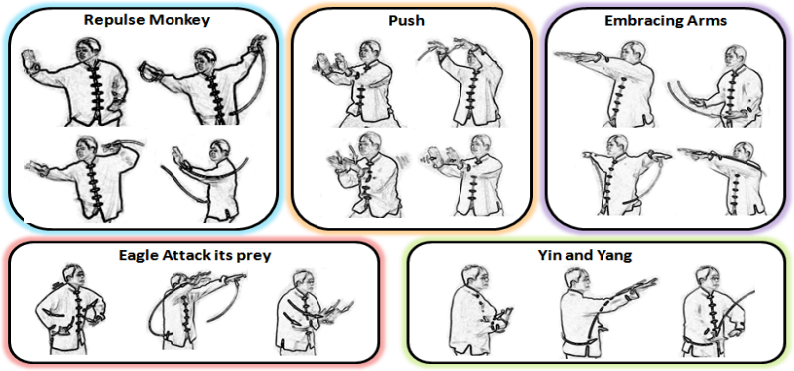

Fig. 5. The 5 Tai-Chi Movements

2.6 Audio Feedback System

The position of the arms in the X-Y plane is analyzed by the descriptor and the difference in position between the pattern and the actual movement in each state of the movement is computed. A commercial Creative SBS 5.1 audio system was used to render the sound through 5 speakers (2 Left, 2 Right, 1 Frontal) and 1 Subwoofer. In this platform was selected a background soft-repetitive sound with a certain level of volume. The sound strategy performs two major actions (volume and pitch) when the position of the hands exceeds the position of the pattern in one or both axes. The first one increases, proportionally to the error, the volume of the speakers in the corresponding axis-side (Left-Center-Right) where is found the deviation and decreases the volume proportionally in the rest of the speakers. The second strategy varies proportionally the pitch of the sound (100-10KHz) in the corresponding axis-side where was found the deviation. Finally, the user through the pitch and the volume can obtain information which indicates where is located the error and its intensity in the space.

3. Experimental Results

The experiments were performed capturing the movements of 5 Tai-Chi gestures (Fig. 5) from 5 different subjects. The tests were dived in 5 sections where the users performed 10 repetitions of the each one of the 5 movement performed. In the first section was avoid the use of technology and the users performs the movement in a traditional way, only observing a video of a professor performing one simple tai-chi movement. The total average error TAVG is calculated in the following way :

$$TAVG = \frac{1}{Ns}\sum_{s=0}^{Ns}\frac{1}{n}\sum_{i=0}^{n}(\theta_{Teacher} - \theta_{Student}) \quad (2)$$

Where *Ns* is the total number of subjects, *n* is the total number of states in the gesture and θ is the error between the teacher movement and the student.

Fig. 6 (A) shows the ideal movements (Master Movements) of the gesture number 1 and (B) represents the TAVG of the gesture 1 executed by the 5 subjects without feedback. The TAVG value the 5 subjects without feedback was around *34.79%* respect to the ideal movement.

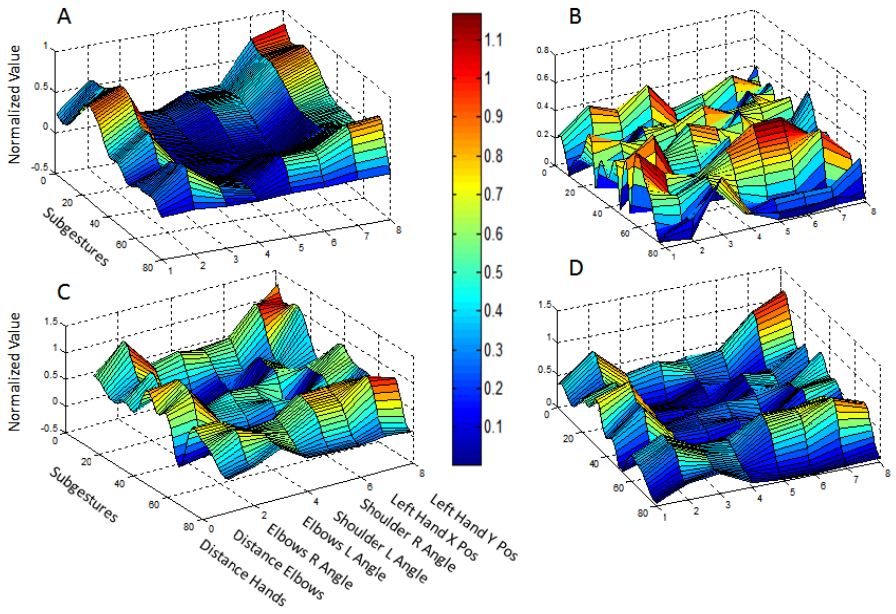

Fig. 6. Variables of Gesture 1, A) Pattern Movement, B) Movement without feedback, C) Movement with Visual feedback and D) Signals with Audio-Visual-Tactile feedback.

In the second stage of the experiments, the Virtual Reality Environment was activated. The TAVG value for the average of the 5 subjects in the visual feedback system presented in Fig. 6(C) was around 23.13%. In the third section the Visual-Tactile system was activated and the TAVG value was around *15.70%* respect to the ideal gesture. In the next stage of the experiments, the visual- 3D audio system was performed and the TAVG value for the 5 subjects in the audio-visual feedback system was around *16.36%* respect to the ideal gesture. The final stage consists in the integration of the audio, vibrotactile and visual systems. The total mean error value for the average of the 5 subjects in the audio-visual-tactile feedback system was around *13.38%* respect to the ideal gesture. Fig. 6 (D) shows the results using the

whole integration of the technologies. Finally, Fig. 7 presents an interesting graph where the results of the four experiments are indicated. In one hand, as it was expected, the visual feedback presented the major error. In the other hand the integration of audio-visual-vibrotactile feedback has produced a significant reduction of the error of the users.

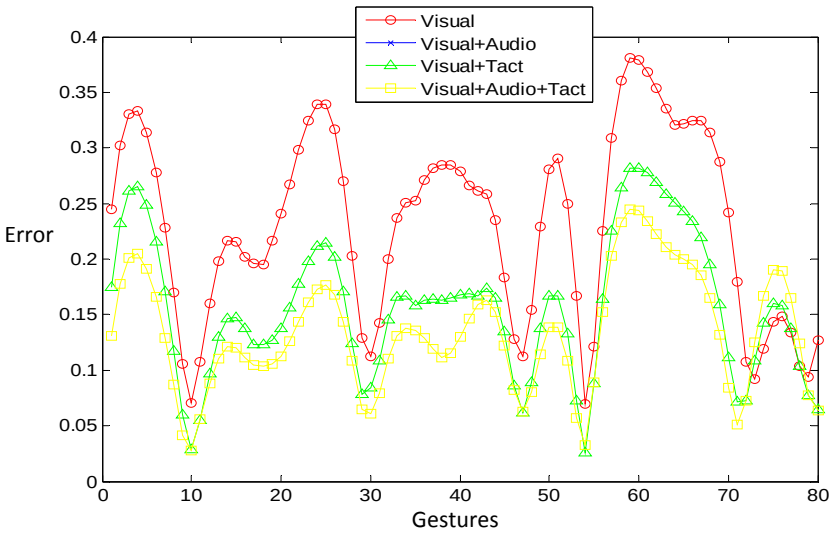

Fig. 7. Average Errors

4. Conclusions

A novel methodology of a real-time gesture recognition and descriptor used in a multimodal platform with audio-visual-tactile feedback system was presented in this paper. The aim to obtain a robust gesture recognition system capable to recognize 5 complex gestures and divide them in different subgestures was fulfilled. Moreover, the function of the real-time descriptor offers the possibility to analyze and evaluate, in a separate and integrate way, the behavior of movements from the different variables related to the feed-back system (audio, vision and tact). The results of the experiments have shown that although the process of learning by imitation is really important, there is a remarkable improvement when the users perform the movements using the combination of diverse multimodal feedbacks systems.

5. Future Work

Once the multimodal platform has demonstrated the feasibility to perform the experiments related to the transfer of a skill in real-time, the next step will be focused in the implementation of a skill methodology which consists, in a brief description, into acquire the data from different experts, analyze their styles and the descriptions of the most relevant data performed in the movement and, through this information,

select a certain lessons and exercises which can help the user to improve his/her movements. Finally it will be monitored these strategies in order to measure the progress of the user and evaluate the training. These information and strategies will help us to understand in detail the final effects and repercussions that produces each multimodal variable in the process of learning.

6. References

1. Byrne, Richard and Russon, Anne: Learning by imitation: a Hierarchical Approach, Behavioral and Brain Sciences. vol. 21, pp. 667-721, (1998).
2. Spitzer, Manfred. The mind within the net: models of learning, thinking and acting. The MIT press, (1998).
3. Viatt, S. and Kuhn, K.: Integration and synchronization of input modes during multimodal human-computer interaction. Proc. Conf. Human Factors in Computing Systems CHI, (1997).
4. Tan Chua, Philo: Training for physical Tasks in Virtual Environments: Tai Chi. Proceedings of the IEEE Virtual Reality, (2003).
5. Cole, R. and Mariani, J.: Multimodality. Survey of the State of the Art of Human Language Technolgy. Carnegie Mellon University, Pittsburgh,PA., (1995).
6. Sharma, Rajeev, Pavlovic, Vladimir and Huang, Thomas: Toward Multimodal Human-Computer Interface. Proceedings of the IEEE, vol. 86, no.5, pp.853-869, (1998).
7. Akay, Metin, I. Marsic and A. Medl: A System for Medical Consultation and Education Using Multimodal Human/Machine Communication. IEEE Transactions on information technology in Biomedicine, Vol. 2, (1998).
8. Hauptmann, A.G. and McAvinney, P.: Gesture with Speech for Graphics Manipulation. Man-Machines Studies, Vol. 38, (1993).
9. Oviatt, Sharon: User-Centered Modeling and Evaluation of Multimodal Interfaces. Proceedings of the IEEE, Vol. 91, (1993).
10. Bizzi, E, Mussa-Ivaldi, F.A. and Shadmehr, R. : System for human trajectory learning in virtual environmets. US Patent No. 5,554,033, (1996).
11. Lieberman, Jeff and Breazeal, Cynthia.: Development of a wearable Vibrotactile FeedBack Suit for Accelerated Human Motor Learning. IEEE International Conference on Robotics and Automation, (2007).
12. Bloomfield, Aaron and Badler, Norman: Virtual Training via vibrotactile arrays.
Teleoperator and Virtual Environments, Vol. 17, (2008)
13. Hollander, Ari J. and Furness III, Thomas A.: Perception of Virtual Auditory Shapes.
Proceedings of the International Conference on Auditory Displays, (1994).
14. Qian, Gang.: A gesture-Driven Multimodal Interactive Dance System. IEEE International Conference on Multimedia and Expo ICME, (2004).
15. Bobick, Wilson: State-Based Approach to the Representation and Recognition of Gesture. Pattern Analysis and Machine Intelligence, IEEE Transactions, Vol. 19, (1997).
16. Farmer, J.: State-space reconstruction in the presence of noise. Physics, (1991).
17. Jain, A.K., Murty, M.N. and Flynn, P.J.: Data Clustering: A review. ACM Computing Surveys, Vol. 31, (1999).
18. Hong, Pengyo, and Turk, Matthew: Gesture Modeling and Recognition Using Finite State Machines. Proceedings of the Fourth IEEE International Conference on Automatic Face and recognition, (2000).