http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet

Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (9): 1809-1816, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.1809.1816

Visual Hand Pose Estimation Based on Hierarchical Temporal Memory in Virtual Reality Cockpit Simulator

Zhou Lai, Gu Hongbin and Niu Ben
College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract: Hand pose estimation is foundation of Human-Computer Interface (HCI) in virtual reality cockpit simulator but it is a challenging problem due to the variation of posture appearance, especially only from single camera. This study proposes a novel visual hand pose estimation method based on Hierarchical Temporal Memory (HTM) which is a biologically inspired model consisting of a hierarchically connected network of nodes. A database containing synthetic images generated by graphics software Pose8 and real images captured by camera is built to train the HTM network. The trained HTM network is used to classify the hand gestures and estimate the wrist parameters of input images. Subsequently, the classification result of HTM is utilized to identify hand motion sequence which is predefined and the finger parameters are acquired by searching the concrete position of input images in the sequence. Experimental results show that the proposed method possesses the characteristic of accurate rendering of the virtual hand applied in HCI and the ability to reconstruct hand postures in a virtual reality cockpit simulator.

Key words: Virtual reality, human-computer interface, poses estimation, hierarchical temporal memory

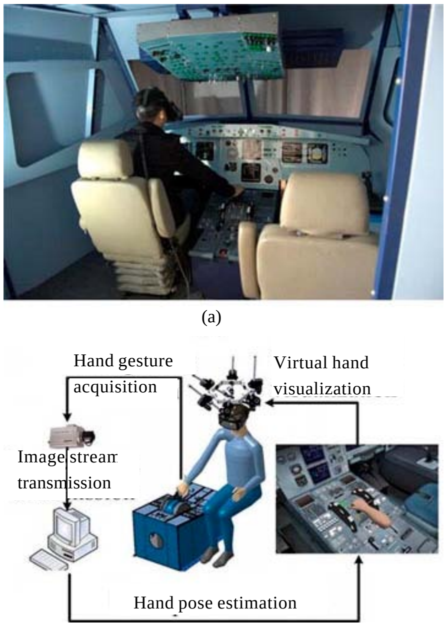
INTRODUCTION

Hand pose estimation has attracted much attention for developing nature and intuitive human-computer interface systems, especially in virtual reality cockpit simulator which trains airline pilots (Gu *et al*., 2009) and vehicle drivers (Salzmann and Froehlich, 2008). However, the hands are unable to be seen by trainees because of the simulation scene is provided by Head Mounted Display (HMD). Then the data glove is always used as interactive device in despite of it has a high purchase cost and affects immersion. As the development of computer vision, vision-based hand pose estimation enables to reflect human-centered characteristic which has been widely used in virtual reality interaction.

Recently, visual hand pose estimation can be divided into two main categories: model-based method (Stenger *et al*., 2001; Lu *et al*., 2003; Lin *et al*., 2004) and appearance-based method (Rosales *et al*., 2001; Wu *et al*., 2005). The former searches for the pose including wrist and finger parameters which map the 2D projection images to the 3D hand model. Although high estimate accuracy is achieved, it would take a long time to complete the parameters searching in high-dimensional pose space. The later estimate pose by matching an input image with a set of labeled hand pose images. In order to estimate precisely, it needs a large number of image exemplars and efficient searching algorithm. Guan *et al*. (2006) utilized Isometric Self-organizing Map (ISOSOM) to map high-dimensional image space to low-dimensional manifold space. Ge *et al*. (2008) used Distributed Locally Linear Embedding (DLLE) to reduce the image dimension and classify the image, then further searching was applied to estimating pose parameters. The global wrist parameters could not be estimated in above methods. Hawkings and George (2006) represented HTM based on memory-prediction theory of brain function which had been applied to image retrieval (Bobier and Wirth, 2008), object recognition (Bundzel and Hashimoto, 2010) and other computer vision fields. Kapuscinski (2010) realized hand gestures recognition with wrist rotation using HTM.

In present study, a two-step visual hand pose estimation method is proposed. Firstly, a database is built to train HTM network which is used to classify the hand gestures and obtain the wrist parameters with nearest neighbor searching. Secondly, predefined hand motion sequence is identified by classification result and the finger parameters are estimated according to location of sequence. The results including wrist and finger parameters could be sent to the computer of virtual reality cockpit simulator is shown in Fig. 1a and the interaction process is shown in Fig. 1b.

Corresponding Author: Gu Hongbin, College of Civil Aviation, Nanjing University of Aeronautics and Astronautics,
Nanjing 210016, China



(a)

(b)

Fig. 1(a-b): Virtual reality cockpit simulator and the interaction process

HIERARCHICAL TEMPORAL MEMORY

HTM is a machine learning model consisting of hierarchically connected network inspired by the mammalian neocortex. A HTM network is composed of several nodes in multi-level hierarchy, where the outputs of lower level become the inputs to higher level. Figure 2 shows the structure of the HTM network built in this study. The size of each input image is 64×64 pixels. The first-layer network contains 16×16 nodes, each of which receives a 4×4 pixel patch from input image. At the level 2, the nodes are arranged in an 8×8 grid where each node receives its input from 2×2 patch of child nodes. Similarly, the third-level hierarchy has 8×8 nodes and also receives four nodes. Finally, the single node at level 4 connects all nodes at level 3.

A node works in two modes: Training mode and inference mode. During the training mode, the node memorizes and groups the spatial patterns coming from children nodes which is called spatial pooling in HTM terminology. A row m and column n node at level q (q = 1, 2, 3, 4) is denoted as N^{q}_{mn}. The sequence used in training has F frame images and the steps of spatial pooling are as follows:

Step 1: The level 1 node N^{1}_{mn} (Fig. 2) at time t receives 4×4 pixel patch which is expanded to a 16-dimensional input vector $Si^{1}_{mn}(t)$

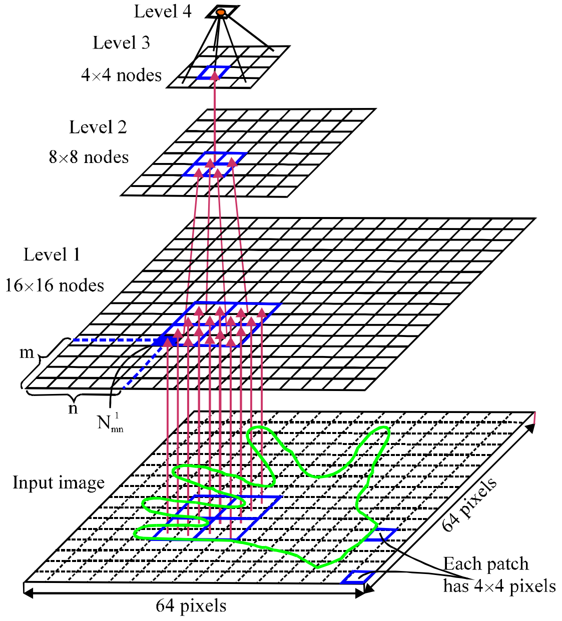


Fig. 2: HTM network with four levels

Step 2: The maximum number of spatial pooling centers is specified as Th^{1}_{num} and the distance which an input pattern differs with a stored pattern is specified as Th^{1}_{dist}. The spatial pooling centers $S^{1}_{mn} = \{S^{1}_{mn}(1), S^{1}_{mn}(2), \cdots, S^{1}_{mn}(Th^{1}_{num})\}$ could be obtained with Alg. 1. ($S^{1}_{mn}[t-1]$ represents the patterns which have been stored till t-1)

Alg. 1: Acquisition of spatial pooling centers

```
for m←1, 16 do
    for n←1, 16 do
        for t←1, F do
            if distance between $Si^1_{mn}(t)$ and element
    of $S^1_{mn}[t-1] > Th^1_{dist}$
            and the number of elements of
$S^1_{mn}[t-1] < Th^1_{num}$
            Add $Si^1_{mn}(t)$ to $S^1_{mn}$ ;
        else
            Do not add $Si^1_{mn}(t)$ to $S^1_{mn}$ ;
        end if
    end for
```

Step 3: In order to save storage space, sparsification of the sorted patterns which preserves only maximum belief component from child node is considered. The spatial pooling centers of level q is obtained, $S^q_{mn} = \{S^q_{mn}(1), S^q_{mn}(2), L, S^q_{mn}(Th^q_{num})\}$, where the Th^q_{num} is maximum spatial pooling centers number of level Th^q_{num}

When spatial pooling is finished, the consecutive membership of the spatial pooling centers has also been recorded in a time sequence. HTM convert to temporal pooling which is divided into three steps:

Step 1: The frequency of each activated spatial centers is described with probability and considered as the input to the temporal pooling. The probability sequence is labeled as $Ti^q_{mn} = \{Ti^q_{mn}(1), Ti^q_{mn}(2), \cdots, Ti^q_{mn}(Th^q_{num})\}$, in which $Ti^q_{mn}(Th^q_{num})$ represents the activation probability of the Th^q_{num} th spatial centers

Step 2: Let the maximum element in Ti^q_{mn} as $Ti^q_{mnmax} = \mathrm{argmax} Ti^q_{mn}(k),\ k \in [1, Th^q_{num}]$. A temporal adjacency matrix Tb^q_{mn} of dimensions $Th^q_{num} \times Th^q_{num}$ has been formed under the assumption that the temporal transitions obey a first-order Markov process. The rows and columns of Tb^q_{mn} correspond with the spatial pooling centers activating at time t-1 and t, respectively and the elements of Tb^q_{mn} represent the probability of $Ti^q_{mnmax}(t)$ observed at time t with the condition that $Ti^q_{mnmax}(t-1)$ was observed at time t-1

Step 3: When the elements of Tb^q_{mn} are tending towards stability, Agglomerative Hierarchical Clustering (AHC) method (Gil-Garcia *et al*., 2006) is used to partition the matrix into a series of temporal clusters $Tc^q_{mn} = \{Tc^q_{mn}(1), Tc^q_{mn}(2), \cdots Tc^q_{mn}(Th_{tem})\}$, in which Th_{tem} is the maximum number of temporal clusters specified previously

After temporal clusters have been generated, the node switches its state from training mode to inference mode. In inference mode, considering the node N^q_{mn} receives input vector Ii^q_{mn} which is compare with each element of Tc^q_{mn}. The matching degree is served as inference output and is labeled as: $Io^q_{mn} = (Io^q_{mn}(1), Io^q_{mn}(2), \cdots, Io^q_{mn}(Th_{tem}))$, where, $Io^q_{mn}(j) = \exp(-D_d(Ii^q_{mn}, Tc^q_{mn}(j)) / \sigma^2),\ j \in [1, Th_{tem}],\ D_d(\ ,\ )$ is Euclidean distance between vectors and σ^2 is variance of Gaussian distribution. For the sake of simplifying, only the maximum of Io^q_{mn} is set to 1 and the others are set to 0. Figure 3 shows the simplified node inference process.

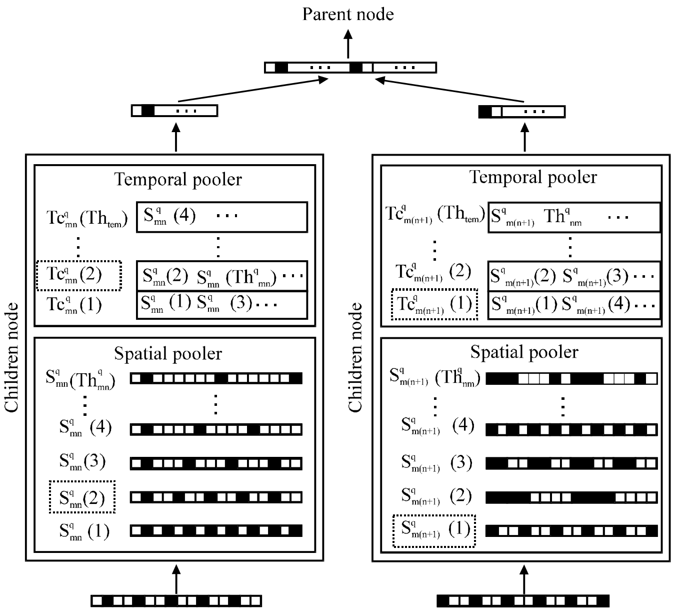


Fig. 3: Node inference process

CREATION OF DATABASE FOR HTM NETWORK TRAINING

Before HTM network training, a database containing synthetic images generated by graphics software Pose 8 and real images captured by camera is created. Figure 4a shows ten synthetic hand gestures under three varieties of illuminations. In Poser 8, large quantities of images with different wrist pose parameters are acquired by setting the angles of roll (ϕ), pitch (θ) and yaw (ψ) (Fig. 4b). Controllable angle ranges are limited above -30° and below 30° ($-30^\circ \leq \phi = \theta = \psi \leq 30^\circ$) and gesture images associated with wrist parameters are sampled every five degrees. Each gesture has 3×13×13×13 = 6591 synthetic images and Fig. 4c shows eight images at ultimate degrees.

To improve the classification accuracy, ten categories of gesture images captured by camera is also considered to HTM training. These images are taken under three different illuminations by ten participants and each gesture is sampled with ten varieties of wrist poses, thus the number of each gesture is 3×10×10 = 300. Figure 5 shows several images of one participant, rows a, b and c are images with different wrist poses but in the same illumination, rows c, d and e are captured in different illuminations. Finally, the total image number of created database is (6591+300)×10 = 68910.

HAND GESTURE CLASSIFICATION AND GLOBAL POSE ESTIMATION

Gesture images should be preprocessed before trained by HTM network. Preprocessing usually includes image size normalization and Gabor filter (Nikam *et al*., 2007). However, the real hand images also need to remove the background using skin color segmentation (Vezhnevets *et al*., 2003) before size normalization. The Gabor responses of normalized images and associated wrist pose parameters of synthetic images are imported together to train HTM network.

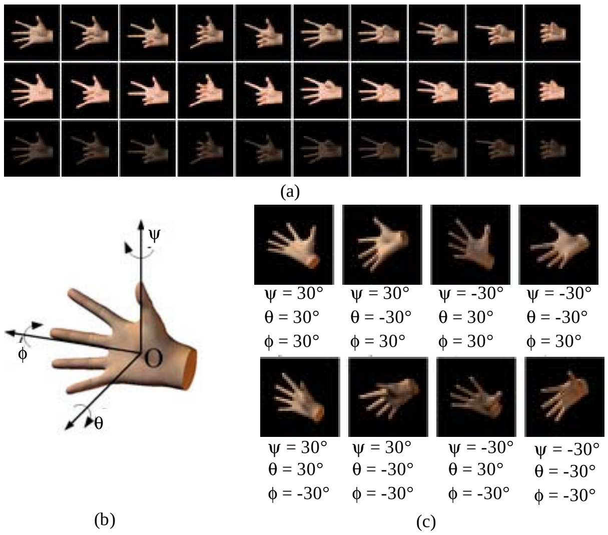


Fig. 4(a-c): Synthetic hand gestures

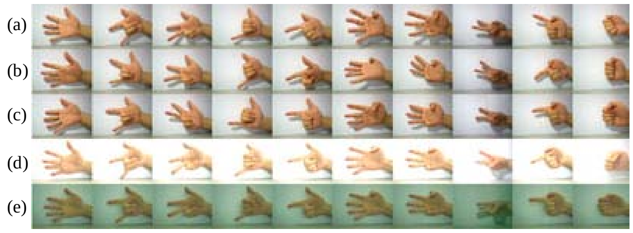


Fig. 5(a-e): Gesture images captured by camera

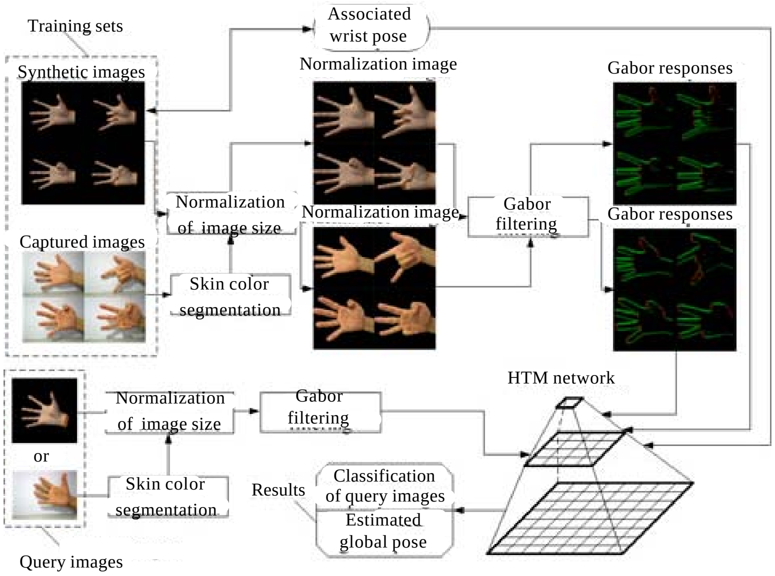


Fig. 6: Global pose estimation by HTM

According to the classification capability of trained HTM network, the category of query images which have been preprocessed could be obtained. While the estimated global poses are also achieved by searching the most similar images in synthetic database associated with wrist parameters. The detailed process is shown in Fig. 6.

LOCAL POSE ESTIMATION

The gesture categories of query images exported by HTM are utilized to identify hand motion sequence which is also generated by controlling bend of finger in Poser 8. Parts of one hand motion sequence can be seen in Fig. 7 and each hand motion sequence has 30 images, only 10 images are shown. In order to estimate the local pose of query images, corresponding positions in hand motion sequence have to been fixed by comparing HOG feature (Dalal and Triggs, 2005) distances among query image and sequence images. Suppose finger pose parameters of images in hand motion sequence is $\{\Phi_0, \Phi_1,..., \Phi_s\}$(in this study, s = 29). As the query image 1 in Fig. 7, its location is between a and b and corresponding finger pose Φ_x is also between Φ_a and Φ_b, $0 \le a < b \le s$. The equations below are subsequently used to estimate finger parameters more accurately.

$$\Phi_x = \Phi_0 + \frac{\Phi_s - \Phi_0}{s}\left(a + \frac{d_1}{d_1 + d_2}\right) \tag{1}$$

$$\Phi_x^{'} = \Phi_0 + \frac{\Phi_s - \Phi_0}{s}\left(b + \frac{d_1}{d_1 + d_2}\right) \tag{2}$$

where, d_1 is the nearest distance among query image and sequence images, d_2 is the second nearest distance. If the query image is closer to a than b, Eq. 1 is considered or else Eq. 2 is considered.

EXPERIMENTS

Results of experiments using the proposed pose estimation method are presented in this section. All experiments are executed on a PC with Pentium Dual-Core 2.8 Ghz CPU and 2 Ghz RAM. Numenta'NuPIC (George and Hawkins, 2009) package is used to construct and train the HTM network.

Experiment 1: Testing of HTM' classification performance. Parts of images are selected from synthesis database and real database to train the HTM network and the others are used for testing. In selection process, the number of each gesture should be the same. Table 1 lists classification accuracy of two selection method. For training purpose, method 1 selects 5000 synthesis images and 200 real images of each gesture (totally 10 gestures). In method 2, the number is 6000 and 250, respectively. From the results shown in Table 1, the amount of synthesis images classified accurately is 15321, the amount of real

Table 1: The classification accuracy of HTM

Selection methods	Training images	Testing images	Accurate classification	Classification rate (%)	Average classification rate (%)
Method 1					
Synthesis	5000×10	1591×10	15321	96.3	96.0
Real	200×10	100×10	917	91.7	
Method 2					
Synthesis	6000×10	591×10	5756	97.4	97.1
Real	250×10	50×10	466	93.2	

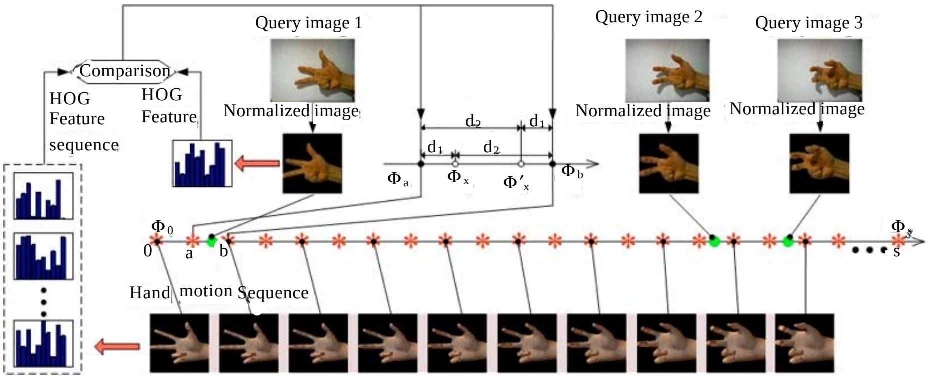


Fig. 7: Local pose estimation

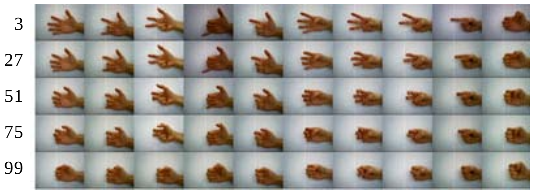


Fig. 8: Hand motion sequence used to test

images is 917 and the classification rate is respectively 96.3 and 91.7%, the average classification rate is 96.0% in method 1. By comparison, the amount of synthesis and real images classified accurately is respectively 5756 and 466, the classification rate is respectively 97.4 and 93.2% and the average classification rate is 97.1% in method 2. An important conclusion is obtained: much more training images improve the classification performance of HTM and method 2 is used to train HTM network in subsequent experiments.

Experiment 2: Testing of HTM' classification accuracy to hand motion sequence. Ten sequences were captured by camera and 100 frame images were randomly selected from each sequence. Figure 8 shows parts of test images and the lefts are frame numbers. Figure 9 shows the classification results, the horizontal axis represents the frame numbers and the vertical axis represents the number of accurate classification. If the vertical ordinate

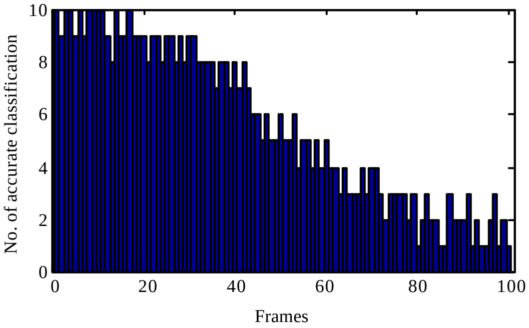


Fig. 9: Histogram of classification accuracy

is ten, it represents that the images in all ten sequences are completely classified at that frame time. On the contrary, if the ordinate is zero, none of the images could be accurately classified. From the figure, the conclusion can be drawn that HTM has less accurate accompanying with grasp of hand. Fortunately, it has no apparent influence to final estimation results, because the parameters of all ten gestures will be similar when hand grasping.

Experiment 3: Visualization of estimated results. Multigen-Paradigm' Creator which creates a hand model including 24 degrees of freedom (21 finger degrees of freedom and three wrist degrees of freedom) is adapted for purpose of visualization. As is shown in Fig. 10, the method in present study is also effective even rotation of wrist.

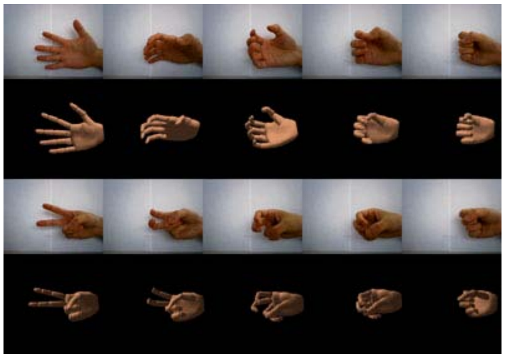


Fig. 10: Visualization of estimated results

DISCUSSION AND CONCLUSION

A visual hand pose estimation method has been proposed in this study, HTM network inspired biologically is trained by created database and used to obtain global wrist parameters and classify the hand gestures. Using the classification of hand gestures, the problem of local finger parameters tracking is converted into fixing the position of query image in hand motion sequence.

All hand pose parameters including wrist and finger parameters could be estimated and sent to the virtual reality cockpit simulator for HCI purpose. The time of each hand image processing is 0.122 sec and the computing speed basically meets the requirement for real-time computation. In addition, the estimation precision can still be improved by adding the constraints of fingers to reduce local degrees of freedom in the future work. Besides virtual reality cockpit simulator, the method could be widely used in other HCI systems, such as gesture recognition with wrist rotation and manipulator teleoperation control system.

REFERENCES

Bobier, B.A. and M. Wirth, 2008. Content-based image retrieval using hierarchical temporal memory. Proceedings of 16th International Conference on Multimedia, Oct. 26-31, ACM Press, pp: 925-928.

Bundzel, M. and S. Hashimoto, 2010. Object identification in dynamic images based on the memory-prediction theory of brain function. J. Intelli. Learn. Syst. Applic., 2: 212-220.

Dalal, N. and B. Triggs, 2005. Histograms of oriented gradients for human detection. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition, 1: 886-893.

Ge, S.S., Y. Yang and T.H. Lee, 2008. Hand gesture recognition and tracking based on distributed locally linear embedding. Image Vision Comput., 26: 1607-1620.

George, D. and J. Hawkins, 2009. Toward a mathematical theory of cortical micro-circuits. PLoS Comput. Biol., 5: e1000532-e1000532.

Gil-Garcia, R., J.M. Badia-Contelles and A. Pons-Porrata, 2006. A general framework for agglomerative hierarchical clustering algorithms. Proceedings of the 18th International Conference on Pattern Recognition, Aug. 20-24, IEEE., pp: 569-572.

Gu, H.B., D.S. Wu and H. Liu, 2009. Development of a novel low-cost flight simulator for pilot training. World Acad. Sci. Eng. Technol., 60: 685-689.

Guan, H., R.S. Feris and M. Turk, 2006. The isometric self-organizing map for 3D hand pose estimation. Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, April 2-6, IEEE., pp: 263-268.

Hawkings, J. and D. George, 2006. Hierarchical Temporal Memory: Concepts, Theory and Terminology. Whitepaper, Numenta Inc., Menlo Park, CA., USA.

Kapuscinski, T., 2010. Using hierarchical temporal memory for vision-based hand shape recognition under large variations in hand's rotation. Artificial Intelli. Soft Comput., 6114: 272-279.

Lin, J.Y., Y. Wu and T.S. Huang, 2004. 3D model-based hand tracking using stochastic direct search method. Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition, May 17-19, IEEE., pp: 693-698.

Lu, S., D. Metaxas, D. Samaras and J. Oliensis, 2003. Using multiple cues for hand tracking and model refinement. Proc. Comput. Soc. Conf. Comput. Vision Pattern Recognit., 2: 443-450.

Nikam, S.B., P. Goel, R. Tapadar and S. Agarwal, 2007. Combining gabor local texture pattern and wavelet global features for fingerprint matching. Proc. Int. Conf. Computat. Intelligence Multimedia Applic., 2: 409-416.

Rosales, R., V. Athitsos and S. Sclaroff, 2001. 3D hand pose reconstruction using specialized mappings. Proceedings of 8th International Conference on Computer Vision, July 7-14, IEEE., pp: 378-385.

Salzmann, H. and B. Froehlich, 2008. The two-user seating buck: Enabling face-to-face discussions of novel car interface concepts. Proceedings of the Virtual Reality Conference, Mar. 8-12, IEEE., pp: 75-82.

Stenger, B., P.R.S. Mendonca and R. Cipolla, 2001. Model-based 3D tracking of an articulated hand. Comput. Soc. Conf. Comput. Vision Pattern Recognit., 2: 310-315.

Vezhnevets, V., V. Sazonov and A. Andreeva, 2003. A survey on pixel-based skin color detection techniques. Proceedings of the Graphicon., Sept. 5-10, Russia, pp: 85-92.

Wu, Y., J. Lin and T.S. Huang, 2005. Analyzing and capturing articulated hand motion in image sequences. Pattern Anal. Mach. Intelligence, 27: 1910-1922.