International Journal of u- and e- Service, Science and Technology
Vol.8, No. 12 (2015), pp.171-182
http://dx.doi.org/10.14257/ijunesst.2015.8.12.17

Dynamic Gesture Recognition based on Image Sequence

Lihua Tian[1,2], Liguo Han[1] and Xiujuan Guo[3]

[1]*College of Geoexploration Science and Technology Jilin University, Changchun 130012, China*
[2]*College of Optical and Electronical Information, Changchun University of Science and Technology, Changchun 130012, China*
[3]*Jilin Jianzhu University, Changchun 130012, China*
[1,2]*lihua_tian18@sina.com* , [1]*68854058@qq.com and* [3]*779523836@qq.com*

Abstract

This paper proposed an algorithm for 3D hands tracking on the learned hierarchical latent variable space, which employs a Hierarchical Gaussian Process Latent Variable Model(HGPLVM) to learn the hierarchical latent space of hands motion and the nonlinear mapping from the hierarchical latent space to the pose space simultaneously. Nonlinear mappings from the hierarchical latent space to the space of hand images are constructed using radial basis function interpolation method. With these mappings, particles can be projected into hand images and measured in the image space directly. Particle filters with fewer particles are used to track the hand on the learned hierarchical low-dimensional space. Then the Hierarchical Conditional Random Field, which can capture extrinsic class dynamics and learn the relationship between motions of hand parts and different hand gestures simultaneously, is presented to model the continuous hand gestures. Experimental results show that our proposed method can track articulated hand robustly and approving recognition performance has also been achieved on the user-defined hand gesture dataset.

Keywords*: human action analysis, image sequence, gesture recognition*

1. Introduction

Human hand motions as well as its posture, behavior has abundant semantics. Human hand tracking and gesture recognition in image sequence is one of the most active research topics in the computer visual field [1-2]. Since human hand movements are characteristic of diversity and polysemy, and human hands are too complicated articulatory objects, hand tracking and gesture recognition are very challenging study concerns [3-4].

Gestures are various postures or actions generated by human hands. They have rich semantic information. Gesture includes static gesture, which refers to posture or single hand shapes, and dynamic gesture, which means hand movements consisted of a series of postures. Three-dimension hand tracking and consecutive dynamic gesture recognition promise wide application demands, such as in the advanced man-machine interaction or machine assembling, operant dynamic gesture is a refined human hand three-dimension behavior, without any mark nor any worn auxiliary device like data glove, indication light source. Through merely computer vision, theoretical computer model is established, transforming primitive images to semantic description, laying foundation for the building of man-machine interactive languages [5-6]. Here we introduce one three-dimension human hand tracking method based on hierarchical manifold learning, recognizing successive dynamic gestures in the learned hierarchical latent variable space. We use data glove and single common camera to constitute the experiment system. By data glove acquiring the hand movement data and creating low-dimension manifold space of three-

ISSN: 2005-4246 IJUNESST

dimension gesture motions, together with the utilization of Condensation technique to track those motions, we propose one dynamic gesture modeling technique which can be applied in human hand hierarchical latent variable space. (Hierarchical Conditional Random Field).

2 Human Hand Tracking Method

Hand tracking and dynamic gesture recognition were concerned and studied widely by researcher home and abroad [7-9]. Some typical literatures overviewed research work and progress in the field at different stages [10-14]. In the field of tracking human hand and similar articular objects like human body, there're two major research methods: one based on appearance and the other based on model.

2.1. Method based on Appearance

Representation-based processing technique is also named the one based on view, data driving or down-top method. Independent of Priori knowledge, it gets motion information directly from image sequence. Marr stated that the main task of visual process is to restore quantitatively from image's reflective scenes the profile and spatial position of 3D objects. Say specifically human hand tracking, it requires predefining a group of gesture collection, from which fetching the only gesture descriptor. Based on the mapping from image feature space to gesture space, we can estimate directly gesture. Generally, such image features include dot, line, corner or texture area *etc*. The application of this method needs establishing the mapping relationship between image characteristics and human hand posture. However, in hand moving process, hand's expressive sharp changes make the mapping certainly highly non-linear. This method doesn't needs calculating 3D gestures, instead, it requires learning and training in plentiful datasets which describe any possible gesture.

2.2. Model-based Method

This method is recalled model-driving or top-down approach. To a large degree, it depends on the constructed model or Priori knowledge. The method of this type builds the model of target tracking problem in Bayesian theoretical frame, that is, it is a process of seeking persistently the biggest posterior probability of target state based on the known priori probability of target status and acquisition of new measurement. It regards visual tracking as one best guessing or reasoning process. The key to this method is to solve posterior probability. When system noises are in Gaussian distribution, also coefficients of state transition probability function and observation function are linearly related, it's possible to employ Kalman filtering to obtain posterior probability. At this moment, both prior and posterior distribution must accord to Gaussian distribution. When coefficients of the two functions are not linearly associated, we can use extended Kalman filtering to get posterior probability.

In recent years, many scholars based on the model of target tracking (especially human motion tracking) introduced the nonlinear dimensionality reduction techniques. Wang *et al.*, [15] used Isomap manifold space moving target, the linear model based on K-nearest neighbor approximation establish manifold space to high dimensional state space mapping, using the condensation algorithm in the manifold space sampling, the sampling particles are mapped to high dimensional state space medium test. Tian et al [16] using Gaussian process latent variable model (Gaussian process latent variable model, GPLVM) get human state space of low dimension manifold space, in the manifold space using particle filter tracking the movement of the human body. Hou *et al.*, [17] used to confine the GPLVM (back constrained GPLVM, BCGPLVM) get low dimensional manifold human motion, They use a Markov model with variable Length in the manifold space (Variable Length Markov Model, VLMM) to establish the motion of dynamic constraint,

and the constraint is used as a priori information to guide the particle filter sampling, better tracking results.

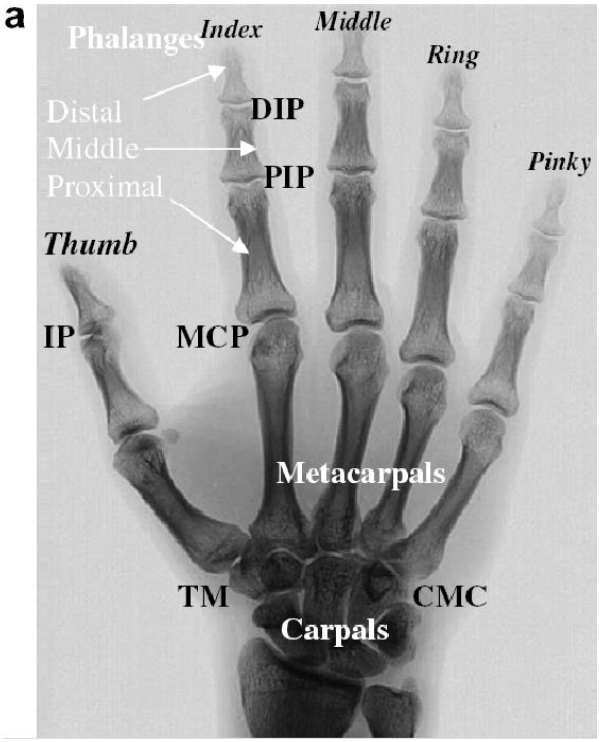
3. Hierarchical Latent Variable Space of Human Hand Movement

Human hands are composed of fingers and palms. Palm motions relate to the changes of hand global position and direction. Finger movements relate to the hand local movements and particular changes. They are our major concern in the paper. As shown in Figure 1(a), one hand has totally 27 pcs of bones, of which 8 locate in hand wrists and other 19 forming palm and fingers. Each bone can be seen as one rigid body. Bones are mutually connected by joints. Each joint has one or more degree of freedom (DoF). In hand anatomy, each joint is in the name of its position, such as Carpometacarpal (CMC), a joint linking metacarpal bones and wrist; Metacarpophalangeal (MCP), connecting each finger with the palm; Interphalangeal (IP), bridging each section of phalanxes. IP can be subdivided into proximal IP (PIP) and Distal Interphalangeal(DIP), according to the distance from the palm.

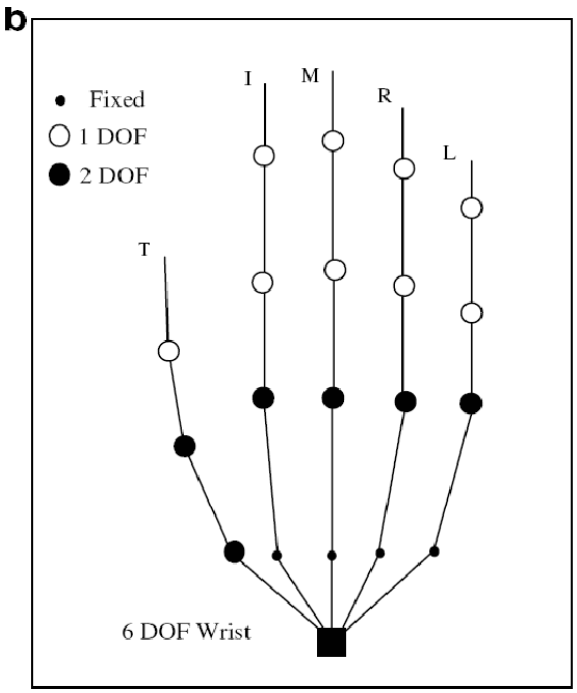
In human body kinematics, five MCPs each has two free angles; nine IPs each has 1 free angle; CMCs are too complicated especially thumb's CMCs, hardly creating model for them [18]. The human hand bone model with 27 free angles in figure1 (a) is the most representative, applied broadly in many human hand motion analyses.

Since self-covering is often seen in human hand movements and every finger is not clearly distinctive, so it's pretty difficult to track finger motions. Finger movements can lead to changed angles among all fingers, between fingers and the palm, between knuckles of the same finger. By referring to the location of 5DT data glove sensor, we can create 3D human hand model. It is shown in Figure 2.

The model includes 14D free angles, which includes three kinds of intersection angles like angle between neighboring fingers (4D), angle among fingers and the palm (5D) and angle between the first and second knuckle of each finger (5D).



(a) Human Skeleton Anatomy



(b) the Kinematic Skeleton Model of Hand

Figure 1. Human Skeleton Model

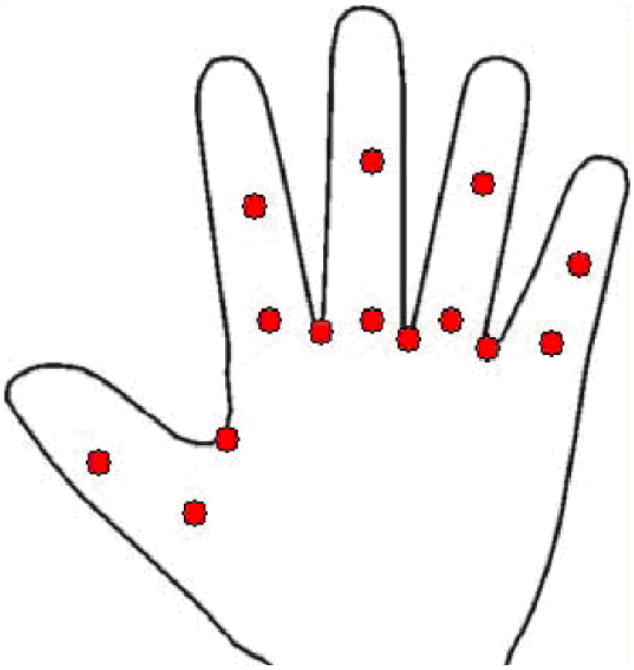


Figure 2. 14 Degree of Freedom Hand Model

4. 3D Human Hand Tracking

In human hand hierarchical latent variable space, we can use Condensation algorithm to track movements of hand and each part like fingers. One important point is how to transform particles onto image or image feature space when sampling particles are being observed. We use radius base function interpolation method to build the non-linear mapping from hierarchical latent variable space to image feature space. Besides, we need

track hand movements in each manifold space and learn the weight of each manifold space in hand tracking process, fusing multi-layer tracking results to get eventually 3D tracking results.

4.1. Non-linear Mapping from Hierarchical Latent Variable Space to Image Space

The model-based 3D human hand tracking method projects 3D model to high-dimension state or feature space for observation. If it's possible to establish smooth mapping from manifold space to hand image space, it's likely to project directly the predicted low-dimension samples to image space, avoiding errors and mistakes occurring in the course of feature extraction and matching. Radius base function interpolation method is effective for establishing smooth mapping from low-dimension space to high-dimension one [19]. Here we adopt it to build the mapping from each manifold space of hand motion hierarchical manifold space to hand image space.

Set input training set image $Z = \{ z_i \in R^D, i = 1, \ldots, N \}$, the latent variable of which in one manifold space is $X = \{ x_i \in R^d, i = 1, \ldots, N \}$; where D and d refers to respectively dimension of image space and manifold space; interpolation function is $f^k : R^d \to R$; k is the k dimension in input image space. Then we have:

$$f^k(x) = p^k(x) + \sum_{i=1}^{N} \omega_i^k \varphi(| x - x_i |) \quad (1)$$

4.2 Tracking Algorithm

By using the non-linear mapping from 3D hand motion hierarchical manifold space to its corresponding hand state space and image space, we can create 3D hand tracking method based on hierarchical manifold learning. The approach includes learning and tracking phase.

4.2.1 Learning Phase

(1) Utilize HGPLVM to get the hierarchical low-dimension expression $[X_1, X_2, X_3, X_4, X_5]$ and X_6 of training set T_1; establish mapping $\mu_1, \mu_2, \mu_3, \mu_4$ and μ_5 from X_1, X_2, X_3, X_4, X_5 to its relative partial hand state space;

(2) Employ radius function interpolation method to construct coefficient matrix B_1 and B_6 of nonlinear mapping from $[X_1, X_2, X_3, X_4, X_5]$ and X_6 to image space;

(3) In each manifold space of hand motion hierarchical manifold space, utilize particle filter to track hand movements in training set T; particle filter's observation is realized by nonlinear mapping built by radius function interpolation method; tracking result $r_1 = [r_{11}, r_{12}, r_{13}, r_{14}, r_{15}]$ and r_2 is obtained;

(4) Use least square method to calculate confidence degree of r_1 and r_2, i.e. weight vector ω_1 and ω_2, in the formula as $f = \arg\min[(\omega_1 r_1 + \omega_2 r_2 - r_{truth})^2$.

4.2.2Tracking Phase

(1) In each sub-manifold space of hand motion hierarchical manifold space, use particle filter to trail hand movements in testing set T; particle filter's observation is realized by nonlinear mapping built by radius function interpolation method; tracking result $r_1^T = [r_{11}^T, r_{12}^T, r_{13}^T, r_{14}^T, r_{15}^T]$ and r_2^T is obtained;

(2) Calculation of the final tracking result $r^T = \omega_1 r_1^T + \omega_2 r_2^T$

The observed density function in the experiment is expressed as $p(z \mid y) \propto W(\Omega(\Omega(I\Theta(I, I_m))))$, where function $W(\ldots)$ calculates normalized weight of each particle; $\Omega(I)$ is dimension of input image space; I is input image at one time point. Function $\Theta(I, I_m)$ calculates the number of pixels with similar grey value of image I_m in relative image position which is acquired after input image I and particles at some time point are mapped to image space; I_m is hand image obtained after the current particle at one time point is mapped to image space.

5. Experiment Design and Discussion

5.1. Experimental Design

Particle filters are widely applied in target tracking field based on model, such as Condensation algorithm proposed by Isard et al [20], which is rather typical. Hou et al. used BCGPLVM model to create low-dimension expression of human body actions. With VLMM, they built dynamic model in manifold space to guide tracking. The method realized favorable tracking effects out of similar algorithms. Later we'll present comparisons of experimental results with the proposed new method here and the aforesaid two. We used 5DT CyberGlove to acquire 14-dimension data regarding human hands at different angles, together with related image sequence obtained with monocular camera. Human hand image resolution is 240x320. Through initial trimming, zoom and binarization, we normalized it to binary silhouette images of 60x70.

Figrue 3 gave 3x8 typical hand motion images, relative silhouette images and tracking images after Poser rendering. In the experiment, we chose 1632-frame data, of which the 3i+1 is used as training set T_1, the 3i+2 as training set T_2 and the 3i+3 as testing set T. Table 1 listed the design and time efficiency of three experiments.

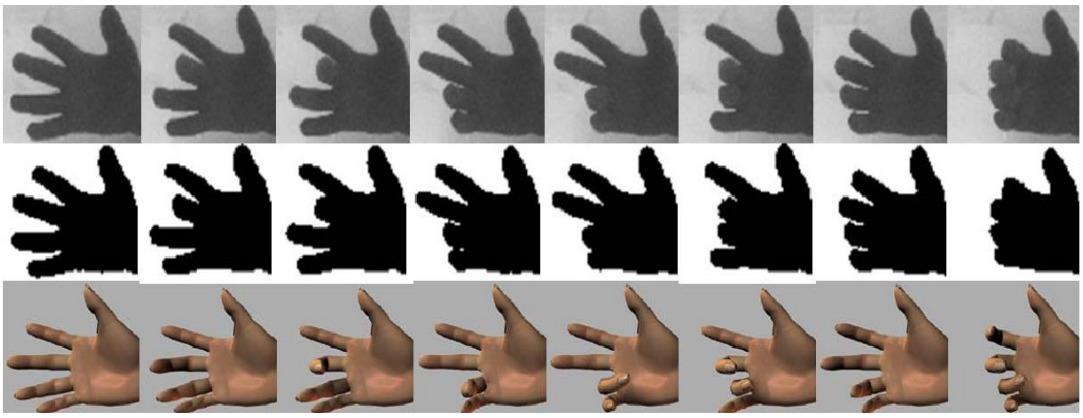


Figure 3. Comparative Image of Hand Image, Silhouette Image and Poser Image

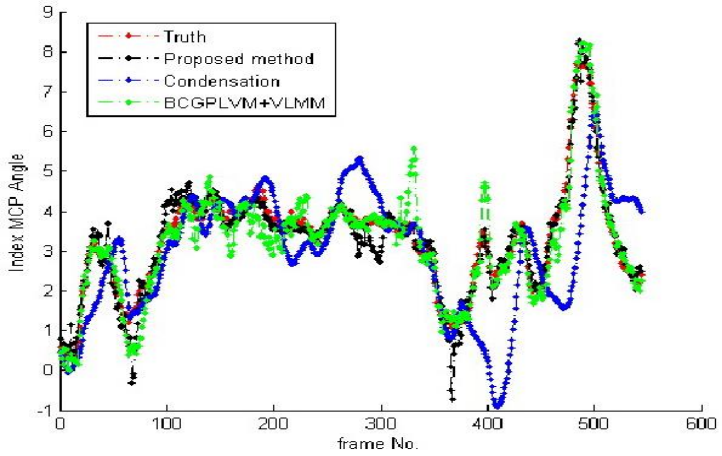
Table 1. The Experimental Design and Time Efficiency

	Training set	Test set	Space dimension	Number of particles	Tracking speed
Condensation		T	14	1000	9.8s
BCGPLVM+VLMM	T_1, T_2	T	3	200	1.8s
The paper method	T_1, T_2	T	1*5+2	30*5+50	2.1s

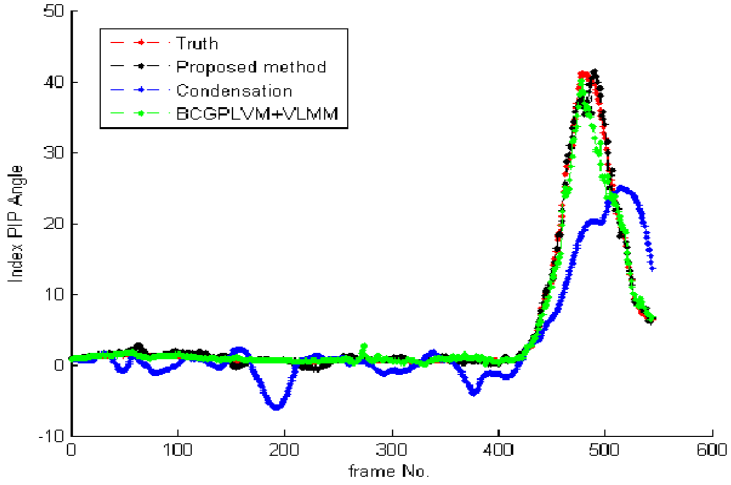
In the experiment, we set the dimension 1 of all low-dimension manifold space and tracking particle number 30 of each low-dimension manifold space, dimension 2 of high-dimension manifold space and tracking particle number 50 of high-dimension manifold space. BCGPLVM+VLMM method chose the summation of T_1 and T_2 as training set and T as testing set. With the method, the experiment finds when manifold space's dimension is 3 and particle number is 200, the tracking result is the best. Condensation needs to follow up human hand movements in 14-dimension status space. In the experiment, all three methods used the observation model here, which was initialized to the state of hand's full expansion. Particle filters' sampling noise variance σ^2 was estimated from training set. The PC CPU is dual core 3.4GHz, memory 1G. All algorithms were developed and ran in MatLab2010.

5.2. The Results of Experiment and Analysis

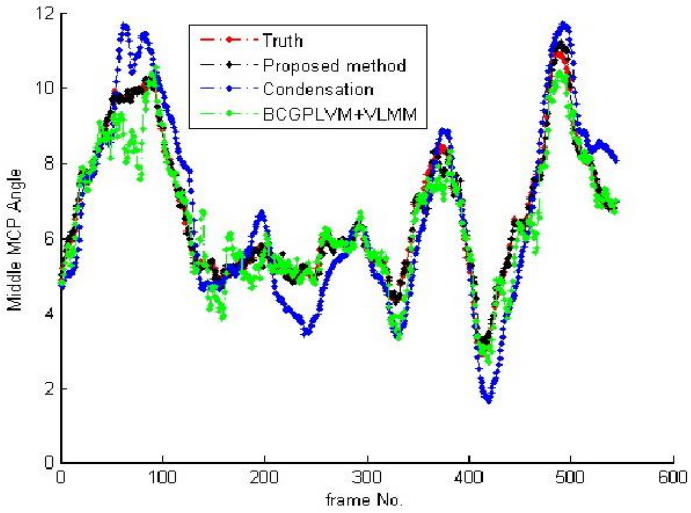
Eight pictures show tracking results of MCP and PIP's knuckle bending angles of four fingers in Figure 4. Apparently, tracking results of the proposed and BCGPLVM+VLMM methods were better than Condensation. At some time points where angles changed sharply, BCGPLVM+VLMM didn't get good tracking accuracy. But the proposed technique had better results. At extremely specific time point, the method here had bigger tracking errors, which however were soon modified, suggesting that our method has strong robustness.



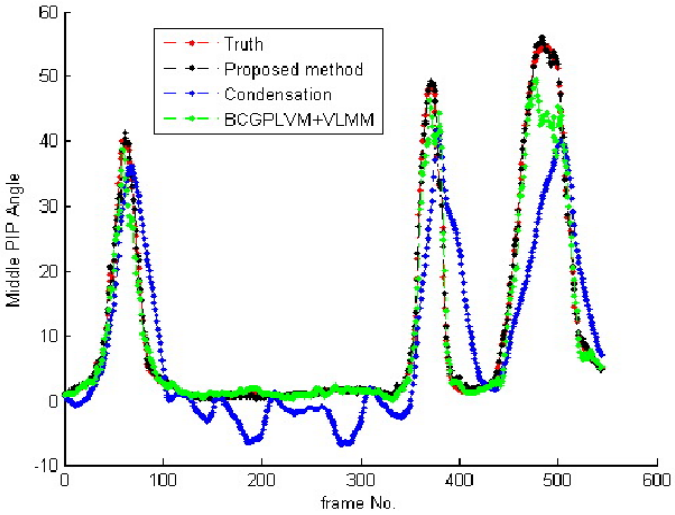
(a) Forefingers MCP



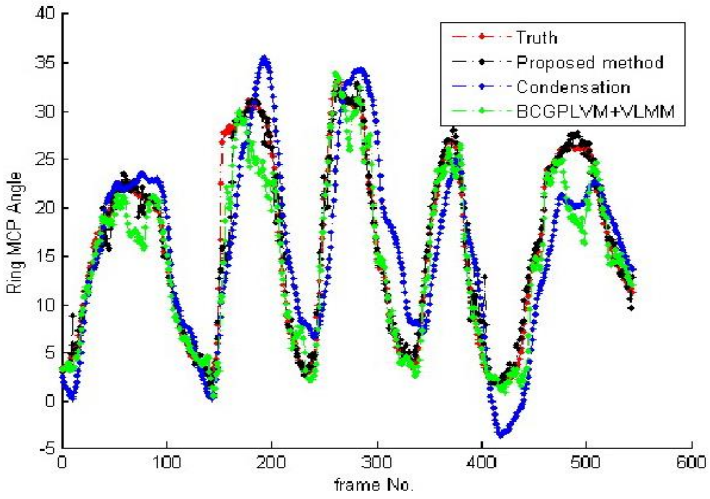
(b) Forefingers PIP



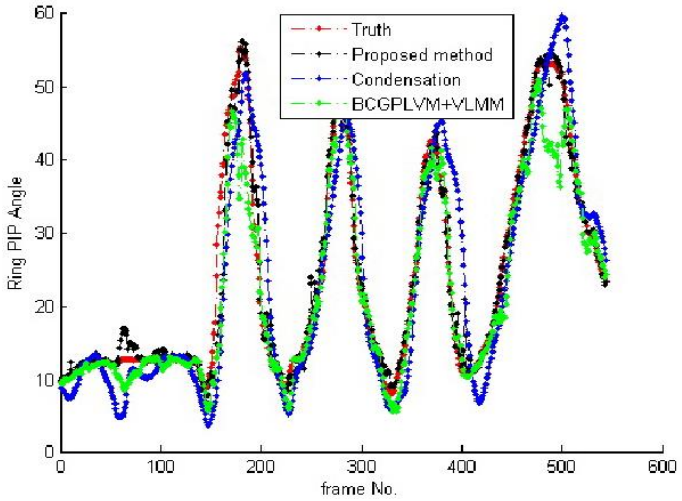
(c) Middle Finger MCP



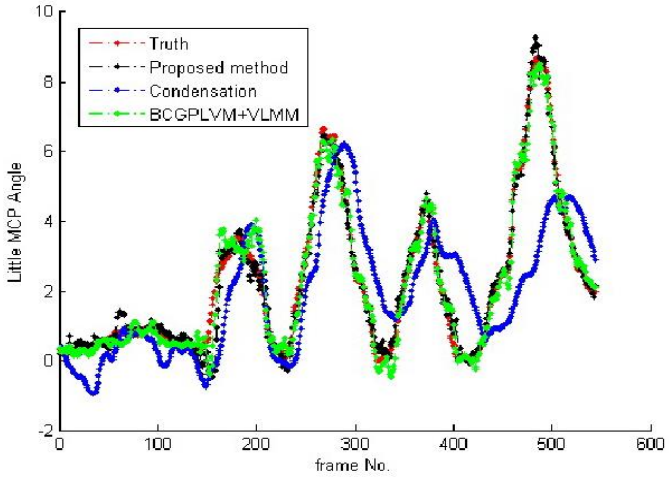
(d) Middle Finger PIP



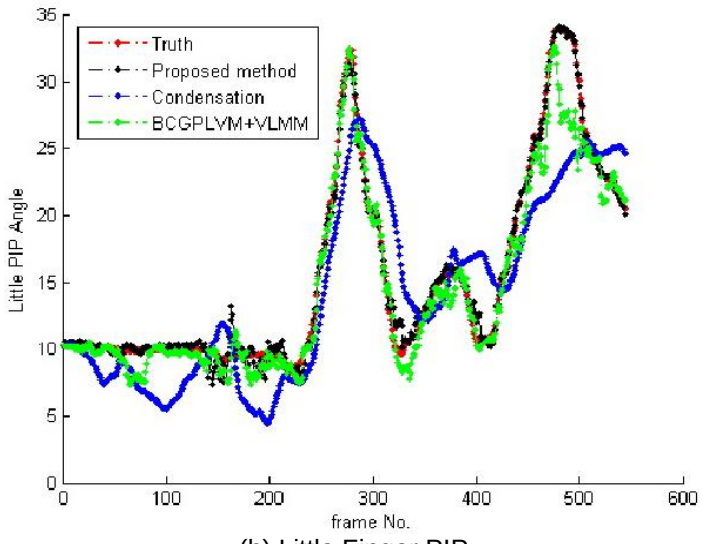
(e) Third Finger MCP



(f) Third Finger PIP



(g) Little Finger MCP

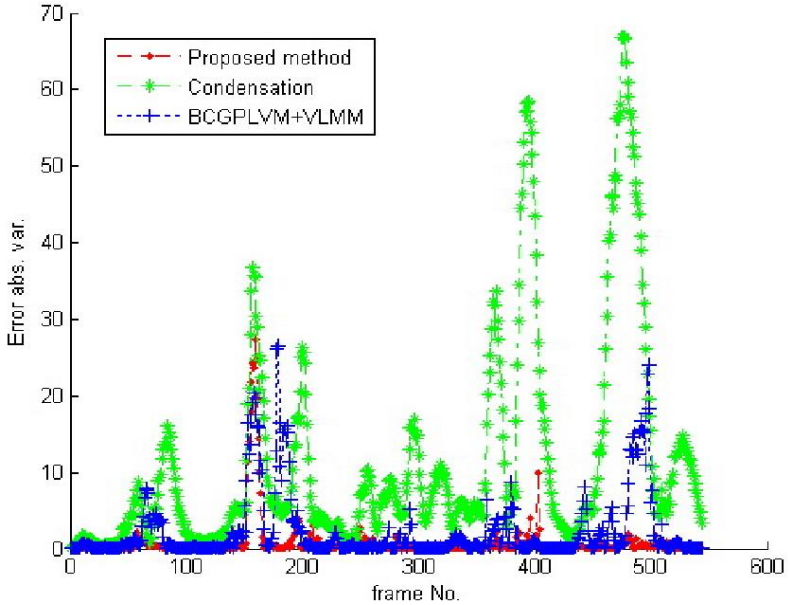


(h) Little Finger PIP

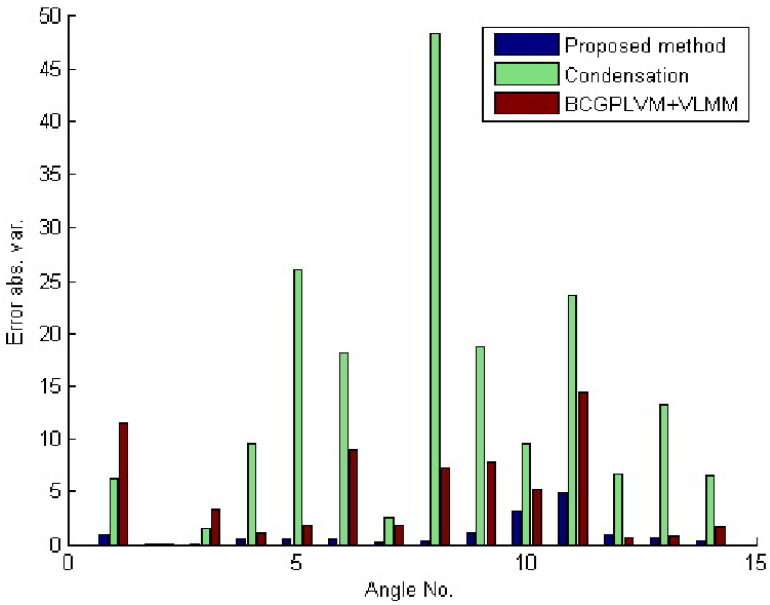
Figure 4. Comparison of the Tracking Results of the Three Algorithms

To compare tracking results of three methods, we analyzed tracking errors. Figure 5(a) showed the mean square errors of tracking errors of all fourteen joint angles at every moment. Condensation had the biggest mean square error and the proposed method had

the smallest. Fig. 5(b) presented in the whole tracking process, the mean square error of each joint. To display easily, we scaled up mean square error of the 1st, 6th and 9th joint by 5 times, and the 2nd, 3rd, 7th and 13th by 10 times. Noticeably, the proposed method had the minimal tracking errors.



(a) Each Time the Error Variance



(b) Each Joint Angle Error Variance

Figure 5. Three Methods of Tracking Error Results

6. Conclusion

In the paper we created with HGPLVM the low-dimension hierarchical latent variable space which reflected better the essence of human hand motions. In the space, we distinguished the robustness tracking of 3D movements and consecutive dynamic gestures. Different from other tracking methods in low-dimension manifold space, our method can track simultaneously motions of one whole hand and several hands. While tracking, we mapped directly low-dimension particles of particle filters to image space

and observed there. It avoided errors which maybe occur in feature fetching and matching periods. Experiments proved the method is simple and effective.

References

[1] L. Han, "Behavior analysis and recognition method in image sequences of people", Beijing Institute of Technology, **(2009)**.
[2] X. H. Song, "Research on three-dimension handtracking analysis method based on the relationships between the variables", University of Jinan, **(2012)**.
[3] D. D. Chen, "The research of gesture recognition in image sequences", Northeastern University, **(2012)**.
[4] T. F. Zhang, "Research on semantic gesture recognition algorithm and application based on a cognitive behavioral model", University of Jinan, **(2014)**.
[5] A. L. Shang, "The method of tracking 3D gesture based on the microstructure of state variable", University of Jinan, **(2013)**.
[6] H. L. Zhang, "Research and development of real time vision based gesture recognition system", Anhui University, **(2013)**.
[7] X. Y. Wang, X. W. Zhang and G. Z. Dai, "A for real-time interactive approach to tracking deformable hand gesture", Journal of software, vol. 18, no. 10, **(2007)**, pp. 2423-2433.
[8] L. Yang, "Research on gesture interaction method based on Wearable vision", Beijing: Beijing Institute of Technology, **(2005)**.
[9] T. L. Liu, "The study on tracking articulated hand based on graph model", Beijing: Beijing Institute of Technology, **(2005)**.
[10] Z. Q. Hou and C. Z. Han, "A survey of visual tracking technology", Journal of automation, vol. 32, no. 4, **(2006)**, pp. 603-617.
[11] Y. Wu and T. S. Huang, "Vision-based gesture recognition: a review", International Gesture Workshop, **(1999)**, pp. 103-115.
[12] Y. Wu and T. S. Huang, "Hand modeling, analysis, and recognition for vision-based human computer interaction", IEEE Signal Processing Magazine, **(2001)**, pp. 51-60.
[13] X. P. Zhang, "Renet, I wish with, Xu Guang you, forestry classes", Research on vision based gesture recognition, Electronic journal, vol. 28, no. 2, **(2002)**, pp. 118-121.
[14] A. Erol, G. Bebis, M. Nicolescu and R. D. Boyle, "Twombly X. Vision-based hand pose estimation: a review", Computer Vision and Image Understanding, vol. 108, **(2007)**, pp. 52-73.
[15] Q. Wang, G. Xu and H. Ai, "Learning object intrinsic structure for robust visual tracking", IEEE International Conference on Computer Vision and Pattern Recognition, **(2003)**.
[16] T. P. Tian, R. Li and S. Sclaroff, "Tracking human body pose on a learned smooth space", Boston: Boston University, **(2005)**.
[17] S. Hou, A. Galata and F. Caillette, "Real-time body tracking using a gaussian process latent variable model", IEEE International Conference on Computer Vision, **(2007)**.
[18] R. Urtasun and T. Darrell, "Local probabilistic regression for activity independent human pose inference', IEEE Conference on Computer Vision and Pattern Recognition, **(2008)**.
[19] T. Poggio and F. Girosi, "Network for approximation and learning", Proceedings of IEEE, **(1990)**.
[20] M. Isard and A. Blake, "Condensation-conditional density propagation of visual tracking", International Journal of Computer Vision, vol. 29, no. 1, **(1998)**, pp. 5-28.

Authors

Lihua Tian, She received her M.S degree from Changchun Institute of optics, Fine Mechanics and Physics, Chinese Academy of Sciences. She has been Ph.D student in College of Geoexploration Science and Technology Jilin University. She is an associate professor in college of optical and electronical information, Changchun University of science and technology. Her research interests include digital image processing.

Liguo Han, He received his ph.D degree from Jilin University. He is a professor in College of Geoexploration Science and Technology of Jilin University. His research interests include Inversion and imaging of the complex earthquake wave field digital image processing.

Xiujuan Guo, she received her Ph.D degree from Jilin University. She is a professor in Jilin Jianzhu University. Her research interests include Data processing and GIS.