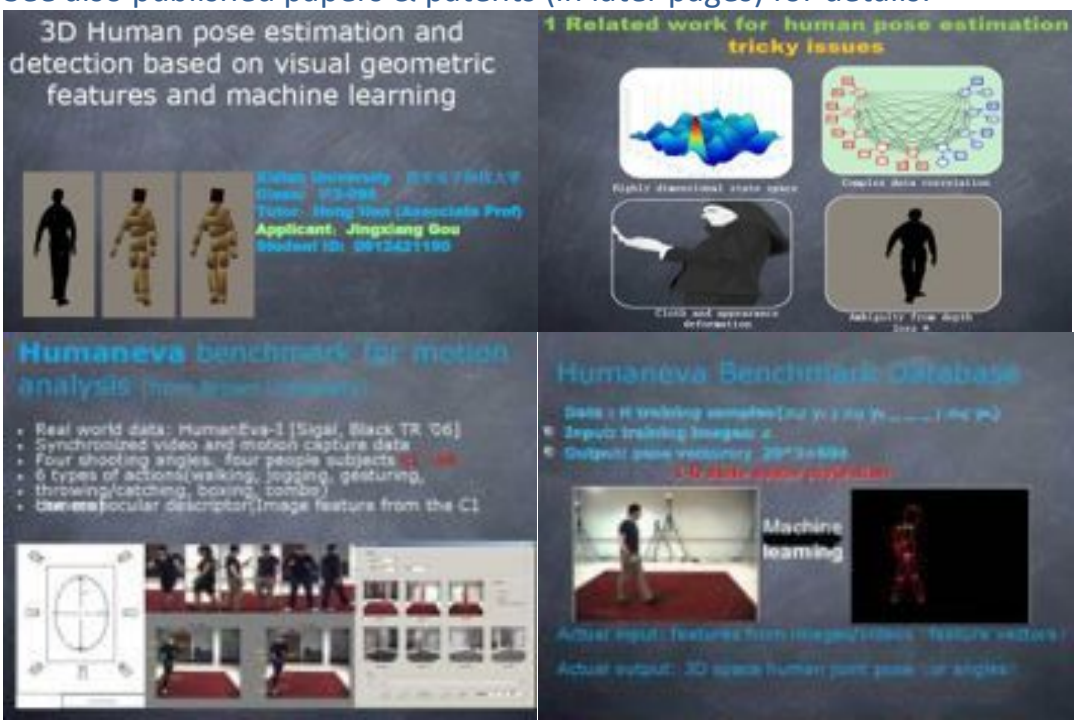
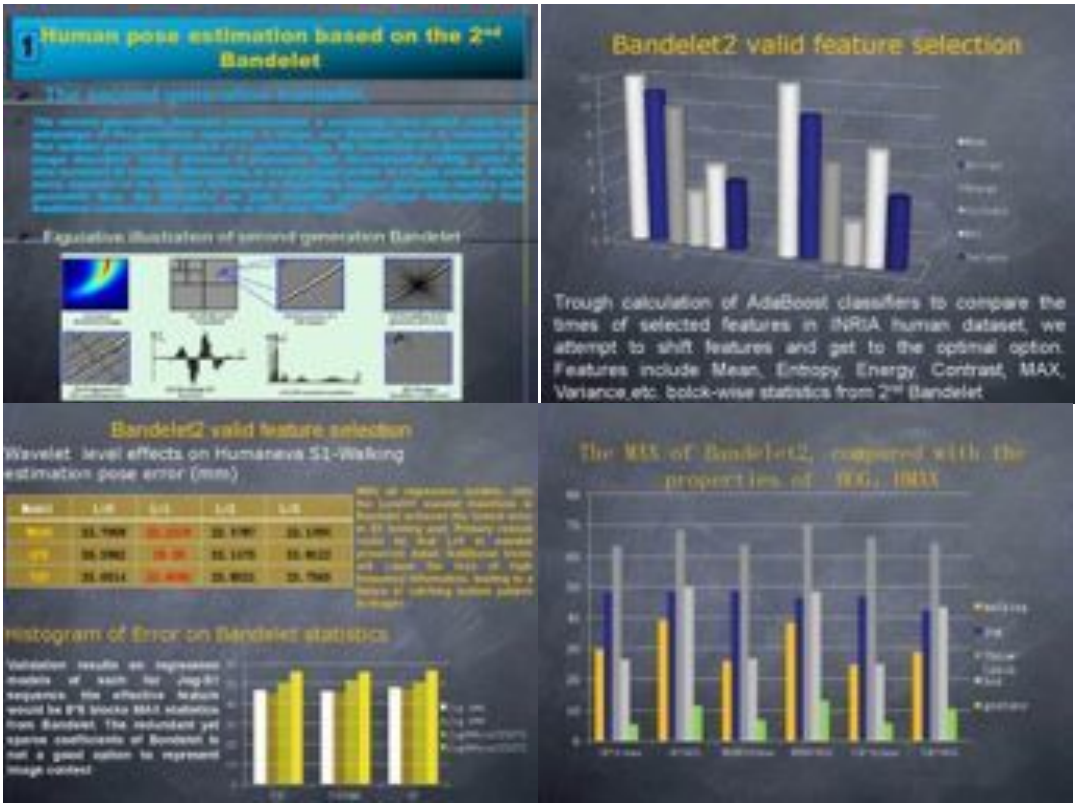
Academic work relates to HCI and human motion analysis:

Brief:

Following projects are done during my master study in Xidian university, where we collaborate in machine learning, computer vision and explore the technical basics for future (on-going) HCI manners. Hong Han is my tutor and Minglei Tong is another responsive person for the National Science foundation for these serials of works. So the substantial work is (mainly) done by our 3 students (me, guangjie feng, Rui Wang), where we all did coding and I did many ideation in research, presentation as well as debugging work. Also, I can verify my substantial writing work in the original working, where we have to be named after the tutor, so I know now it's hard to prove the validity of my words.(I have all the original copies before publishing, again.)

See also published papers & patents (in later pages) for details:





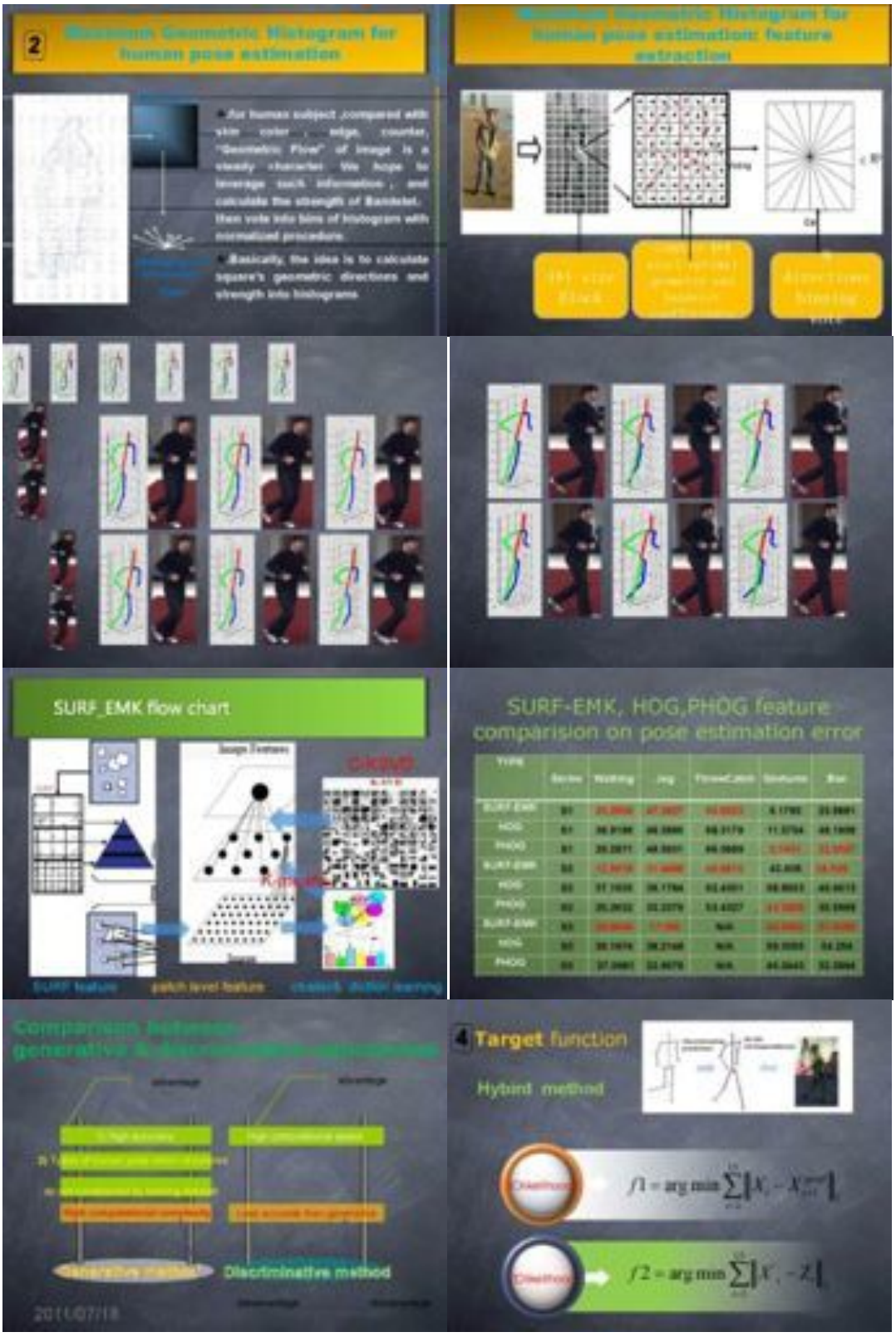
[1]. Hong Han, Minglei Tong*, Jingxiang Gou, Rui Wang, Guangjie Feng, “Discriminative Human Pose Estimation Based on the Bandelet 2 Image Descriptor”, Proceedings of the Sixth International Conference on Image and Graphics (ICIG 2011), p 679-84, 2011. Database: Compendex.（EI）

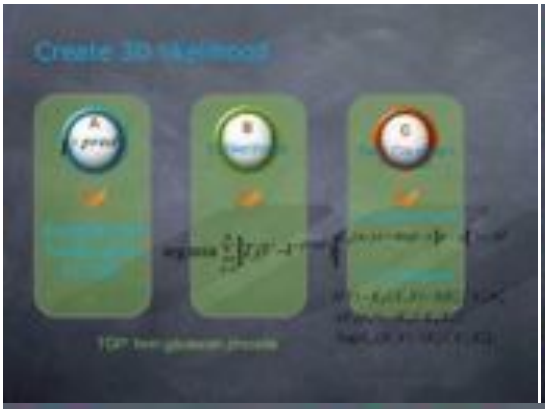
Link：http://www.computer.org/csdl/proceedings/icig/2011/4541/00/4541a679-abs.html

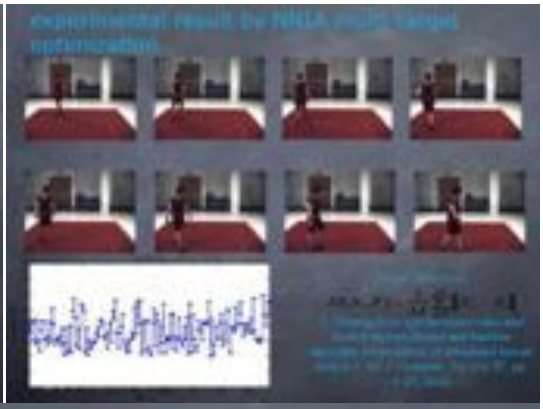
[2]. Hong Han, Jingxiang Gou, “A Sparse Geometric Feature for Visual Detection and Estimation”, Neurocomputing 2011 special issue on image representations, accepted in Aug, 2012. (SCI) Code: *NEUCOM-D-11-01209R2*

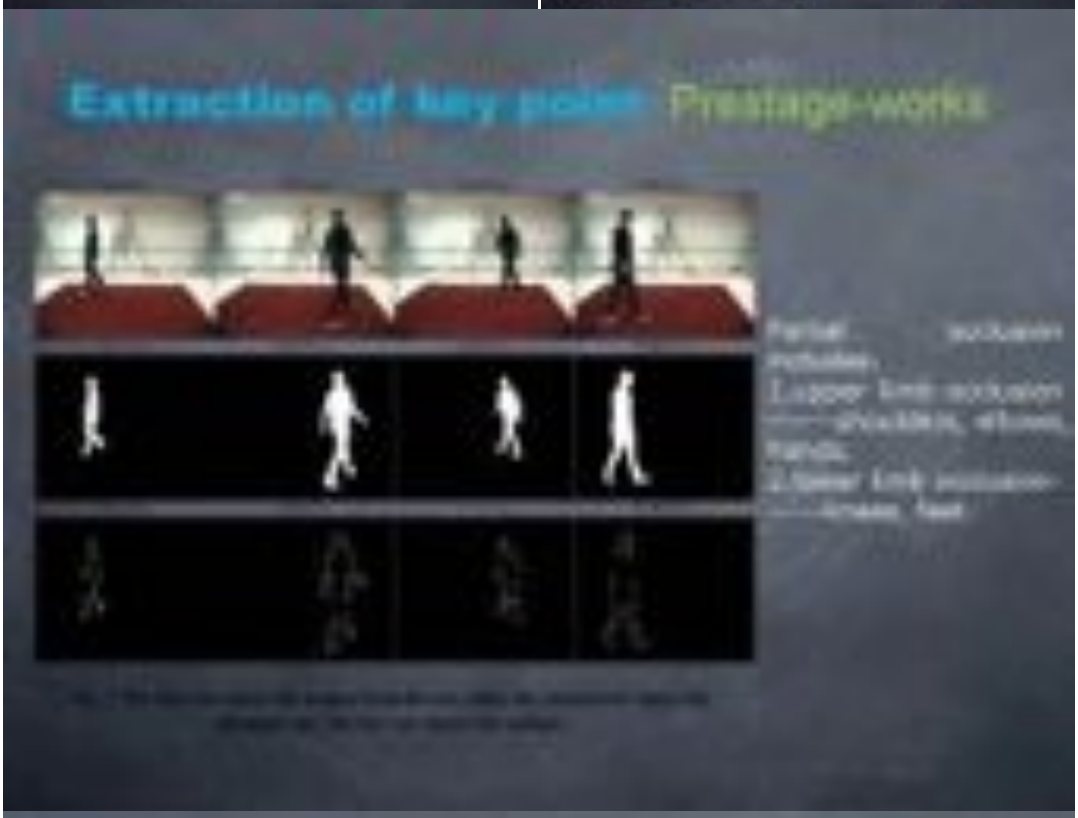
[P1].Hong Han, Jingxiang Gou, Guangjie Feng, Rui Wang, “Human Pose Estimation and Image Representation Based on the Bandelet2 Image Descriptor”, Chinese National Patent, Application No: 201110148496.3 . Accepted.

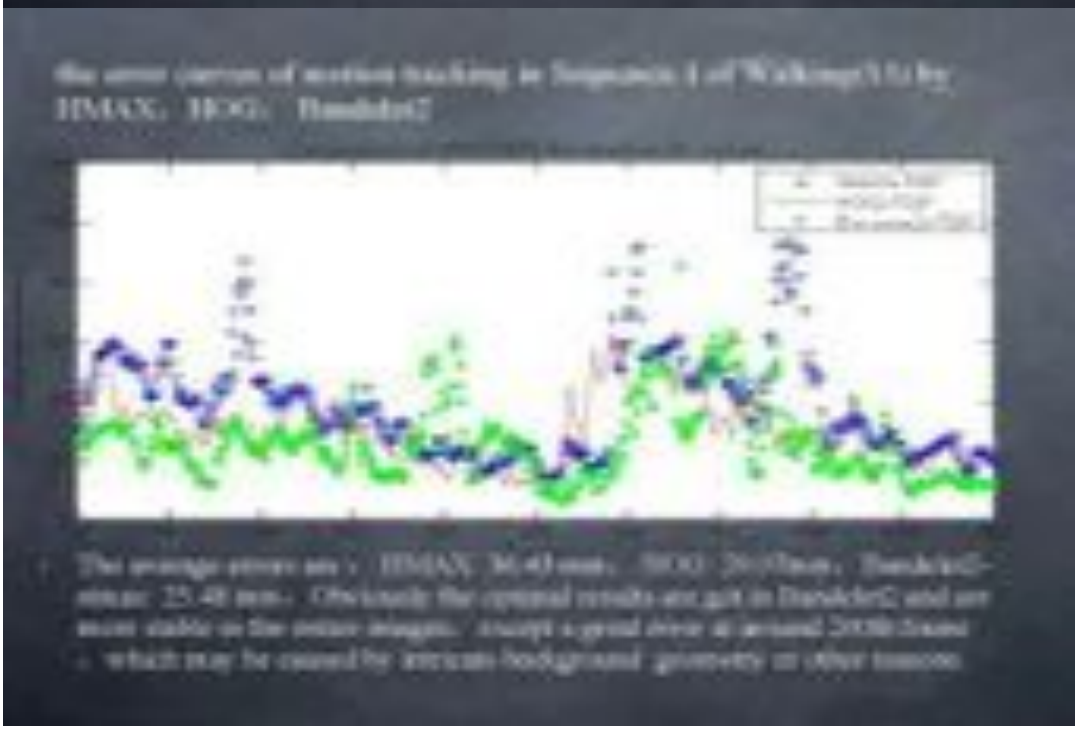
Above works: Human tracking system based on VSMM and Bandelet is proposed, which effectively performed video motion analysis and tracked human pose in sub-real time with generative approach with accuracy and visual results. The paper of 1st stage core idea was published in 2011 Science China (Non-credited author), see my editing for the paper in below portfolio. Responsibility: RA, for video material collection, filming, experimental implementation; Observation of the algorithm, data mining; paper revision, relative workloads and conclusion reflection. File in patents.


2
Basically, the idea is to calculate square's geometric directions and strength into histograms
SURF_EMK flow chart
SURF-EMK, HOG,PHOG feature comparision on pose estimation error
4 Target function
Hybird method
Discriminative method


Create 3D skeleton


experimental result by MHLA multi-target optimization


Extraction of key poses Prestage-works

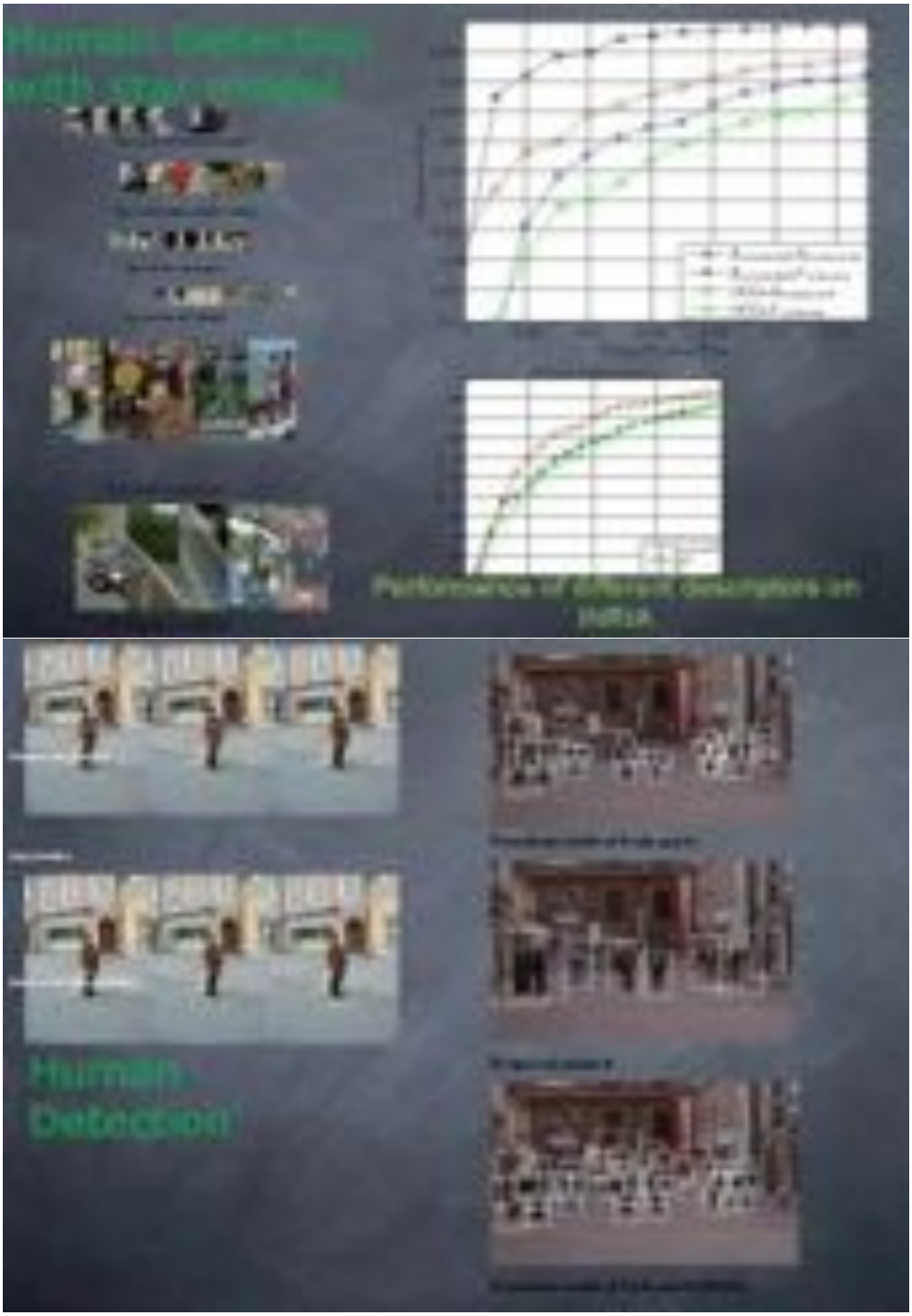

HMAX; HOG;

[3]. Hong Han, Guangjie Feng, Jingxiang Gou, Rui Wang, “A Multi-Objective Immune Algorithm for Human Motion Tracking”, 2011 International Conference on Multimedia Technology, ICMT 2011, p 59-62, 2011. (EI)

Link： http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6002110

[P2].Hong Han, Guangjie Feng, Jingxiang Gou, Rui Wang, “An Optimization Based Framework for Human Motion Tracking”, Chinese National Patent, Application No: 201110114626.1. Accepted.

Above work includes: Estimate and optimize human pose in 3D computer vision with an evolutionary algorithm. We constructed the likelihood between 3D skeleton and sketch of human skeleton, and reconstructed 3D motion. Several likelihoods such as contour, gray scale, edge and complexion are explored, which fully discovered the advantages of Quantum in multiple-object optimization. Responsibility: Assistant, adjustment and testing in algorithm robustness; observation and feedback on experiments, proposed opinions on possible accuracy. Indexing earlier stage materials, also collection and preparation for paper documentations.



[3]. Hong Han, Jingxiang Gou, “A Sparse Geometric Feature for Visual Detection and Estimation”, Neurocomputing 2011 special issue on image representations, accepted in Aug, 2012. (SCI)

[P3].Hong Han, Rui Wang, Jingxiang Gou, Guangjie Feng, “Pedestrian Detection Based on the SecondGeneration Bandelet Feature and the Star Model”, Chinese National Patent, Application No: 201110334700.0. Accepted.

Above works: Machine learning framework and optimization methods were both implemented in this HCI fundamental project. Multiple kinds of image representations for vision system were tested and improved, and advanced descriptors were initiated for both visual detection and tracking, we tried to push it further into future natural interface though novel theory. State-of-art result is achieved on international database HUMANEVA-I. Responsibility: core member for coding & algorithm design, data collection& processing, preprocessing video, image descriptions and database. Chief one in validation and algo-boosting, 2 projects as 2nd, and 2 as 1st student author (e.g. ICIG). Planning projects, calling routine meetings and leading presentations.

2011 Sixth International Conference on Image and Graphics

Discriminative Human Pose Estimation Based on the Bandelet2 Image Descriptor

Hong Han [*], Minglei Tong [†], Jingxiang Gou [*], Rui Wang [*], Guangjie Feng [*]

[*] Xidian University
Xi'an, 710071, China
Email: hanh@mail.xidian.edu.cn

[†] Shanghai University of Electric Power
Shanghai, 200090, China
Email: tongminglei@gmail.com

Abstract

In this paper, we address the recovering from monocular images focusing on designing a novel image descriptor derived from the second generation Bandelet transformation, noted as Bandelet2, to tackle with estimation accuracy combined with state-of-art prediction methods. The proposed Bandelet2 image representation could boost the accuracy for the final 3D pose prediction in monocular video images by information from geometric flow to characterize image context, especially for human body shapes and motions. We have compared our image descriptor with classic ones as HOG, HMAX, laterally tested among different regression methods on standard Humaneva-I motion capture dataset and showed 3D reconstruction results. Final statistics verifies competitive discriminatory effectiveness and precision of Bandelet2 descriptor in estimating 3D human poses from monocular images.

Keywords-*human pose estimation; Bandelet; discriminate approach;*

I. Introduction

Human pose estimation (or tracking) has been an intriguing topic in computer vision, with a wide span of ongoing and potential future applications such as video surveillance, HCI, personal attributes, action identification and retrieval, etc. However, due to intrinsic complexity of articulated objects as human body itself and recording quality of image sequences with complex video context variance, robust and satisfying solutions to real life challenges remains to be discovered.

Still, numbers of advances around this topic have been realized in recent years [1, 2, 5, 6, 7].Two main schools are well known in pose estimation problem as *generative approach* and *discriminative approach* .The former approach recovers human body pose by a synthesis-and-optimization step. Observation models are used to find correct perception in state space, searching and optimization algorithms within this framework are vital in determining a successful tracking [3,4]. Tough tracking may be precise under a delicate design, the whole process can be both arduous in optimizing the hidden states and time consuming in constructing an appropriate observation model. *Discriminative approach*, also known as *data-driven approach*, on the other hand, can be seen as using amount of data to learn a direct mapping correlation from the image space to the 3D pose state space. Prevalent ones are conducted in a supervised[1, 7, 8, 11, 14, 15] or semi-supervised[12] manner. Once trained on a dataset with sufficient actions and subjects with 3D motion capture data as labels[9], they are capable of predicting 3d poses for test image streams in a fast manner. Though might be less precise than conventional generative approaches, they are far more convenience since which do not need initialization, prior information, or camera parameter, also are not confined to any specific type of person in appearance.

To critically investigate all aspects behind this pose estimation frame work, every single piece cannot be trivialized. Such as original video recording quality, back ground clutter, image descriptor's robustness and predictor (also known as regression). In short, accurate designs of image descriptor and the predictor are of most significance in current concern of effective learning. Effective representative image features, whether global or local ones, have been proposed and applied to the pose estimation tasks [1, 7, 12]. However, most representations are mainly based on contours or edges of subjects, being unconvincing to descript image contents, which may still need lifting in order to better discriminate tinny subject differences.

The second generation Bandelet (Bandelet2) transformation is a promising image representation, which was originally devised for image/surface compression and denoising [17]. Bandelet2 image representation possesses high discriminative ability based on *geometric flow*, also being invariant to

978-0-7695-4541-7/11 $26.00 © 2011 IEEE
DOI 10.1109/ICIG.2011.124

rotation, illumination, or back ground clutter to a large extent. In this paper, we propose to follow the data-driven approach to solve human estimation problem with a novel descriptor specified on Bandelet2. Here we find a new tool of image representation to better depict image context. There is no need of background subtraction beyond regions of bounding box, which is computationally relaxed. Also, Representations with Bandelet2 are of low dimensions.

A. Related work

We conducted this research which mainly relates to the background of discriminative approaches, image descriptors, and a new trend in combining both discriminative and generative methods [10].

A successful image descriptor for human estimation tasks with discriminative approaches are generally categorized into local or global ones, former of which are mainly known for SIFT, Pyramid Match kernel (PMK) [7], also bag of words representations. Global features such as HOG [11, 14, 15], shape context [1], HMAX [12, 14], have shown great merit in measuring correspondence. All the above features are good, except most of them are based on silhouettes and which heavily relies on accurate background subtraction and/or silhouette extraction. Our Bandelet2 is also a dense global feature, but it circumvented above problems and made image presentation powerful.

The *discriminative approach* learns mapping functions varies from nearest-neighbors [15], sparse regression like RVM [1], conditional Bayesian mixture of experts [8, 11], Gaussian Process [7, 18], manifold embedding [12], structured prediction [14], etc. We used TGP and also basic Gaussian Process to show the effectiveness of Bandelet2 descriptor on the results.

As a feasible approach in pose estimation work, we believe that combination of discriminative and generative methods would further lead to a more precise estimation [16]. However, Due to our principle of simplicity and computationally relaxation, a discriminative approach is taken in this work.

The paper is organized as follows. In the next section we show detailed Bandelet2 theory. GP and TGP regression are reviewed in Sec.3. Sec.4 gives the experimental results.

II. Bandelet2 as Image Representation

A. Bandelet2 geometric image representation

The second generation Bandelet was proposed instead of describing the image geometry through edges, which are almost often ill defined, the image geometry is characterized by a *geometric flow* of vectors [17]. See Fig 1 for illustration. Motivated by this idiosyncrasy, we transform the Bandelet2 into image descriptor because its high discriminative ability, also invariant to rotation, illumination, or back ground clutter to a large extent. The main idea is to identify geometric features within a specific region to be *vector fields,* which indicate local orthonomal directions of gray scale change. In this sense, Bandelet2 is way different from non-adaptive methods of constructing bases. As we generally cannot predict in advance the geometric image regularity, Bandelet2 image representation is adaptively gained upon a specific image according to final applications. For this pose estimation frame work, we need to jointly learn parameters to the given task as well, see section 2.2.

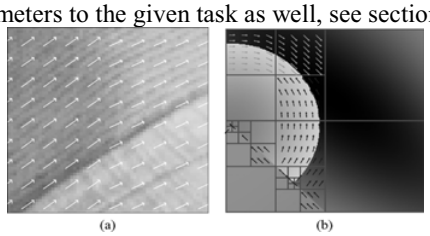


Figure 1 (a) geometric flow in a local region. (b) An adapted dyadic squares segmentation. Figures are from [17]

1). Constructional idea of Bandelet2 basis

1: the objective function is based on Lagrange method to optimize the best basis function. The objective function can be expressed as:

$$L(f_\theta, R) = \left\| f_\theta - \tilde{f}_\theta \right\|^2 + \lambda * T^2 (R_g + R_b) \qquad (1)$$

Where the first part is approximate error (squared error) and the second part is the computational complexity. See details in section 2.2.

2: adapt a quadtree division according to the objective function minimization principle, and CART law to emerge sub blocks bottom-up, the optimal quadtree division can thus be acquired, see also figure 3.1 for illustration. Details are discussed in [17].

3: determine the optimal geometric direction within each final sub-blocks of previous division in step2, perform orthogonal projection upon each optimal direction. Transform two dimensional wave functions into one dimension.

B. Bandelet2 image descriptor extraction

Detailed steps for extract image descriptors from video sequences with the Bandelet2 are as follows:

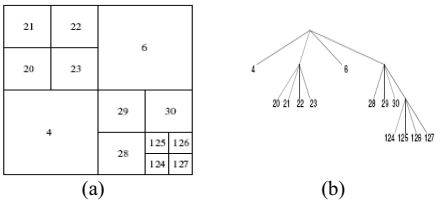
Step 1: For each image, the human body is detected within a bounding box and rescaled to a fixed size.

Step 2: Perform 2D wavelet transformation.

Step 3: According to the quadtree division and CART law, get the optimal devised sub-blocks (see part **2.1.1** and figure 2): For a square S of width L, compute the best geometry and Lagrange penalty value $L_0(s)$. And then let $L = 2L$, repeat the previous step. For each $L \times L$ square S, compute the optimal geometry and Lagrange penalty value $L(s)$. For a certain level square S of $L \times L$, its four children are (s_1 , s_2 , s_3 , s_4), and the combined Lagrange function is:

$$L(s) = L_0(s_1) + L_0(s_2) + L_0(s_3) + L_0(s_4) + L_0(s) \quad (2)$$

Let $L_0(s) = \min\{L(s), L^{'}(s)\}$.



(a) (b)

Figure 2 (a) A quadtree division. (b) Corresponding tree structure to leaf positions in (a). Both figures are from [17].

Step 4: Perform 1D discrete reordering of the sampling location referred in step 2. by projecting the sampling location along d (directions) and sorting the resulting 1D points in a line, which can defined as a 1D discrete signal f_d .Then perform a 1D discrete wavelet transformation of f_d , and get f_θ .

Step 5: Selection of the best geometry: For a given threshold T , the direction d which generates less projection error is to be chosen. R_b is the number of bits to encode the quantized coefficients $\{Q(t)\}$, $\tilde{f}_\theta$ is the quantized f_θ . We use a uniformed quantize term.

$$Q(t) = \begin{cases} 0 & |x| \le 0 \\ sign(x) * (q + 0.5) * T & qT \le |x| \le (q+1)T \end{cases} \quad (3)$$

To select the optimal geometry structure, we must choose the direction d that minimizes function (1), where R_g is the number of bits needed to code the geometric parameter d with an entropy coder, and λ is the empirical penalty factor 3/28.

Step 6: The resulting 1D wavelet coefficients corresponding to the best geometry d can be stored in a 2D matrix of the same size as S, which is called Bandelet2 coefficients.

Step 7：Calculate the max statistical geometric flow of Bandelet2 coefficients to form the final feature.

1) Parameter selection and valid sub-features. Parameters used specifically in our experiment are T , L , $j_{\min}$ and $j_{\max}$. T is the quantized threshold, L is the level of wavelet, $j_{\min}$ is the smallest scale division size, while $j_{\max}$ is the largest integration scale of bottom-up quadtree fusion. By experiments, we can determine T to be 15, L be 1, $j_{\min}$ be 2, and $j_{\max}$ be 2.

With a higher value of T, compressive rate rises while authenticity of image details declines, and in pose estimation work this largely affect the efficiency of Bandelet2 image descriptor in describing body contours, interior and exterior edges. As previous experiments and published works concerning Bandelet have suggested, neither a too high nor a too low value is preferable when attempting to fix the best direction of particular geometric flow. For the basic need to encode image descriptor, we then adopt the hard thresh (HT) to quantize the gate. T is set through experiment with training sets, with a uniform value of 15.

The 2D wavelet transformation was processed with different levels from L=0 to L=5 during previous stage of feature selection, and L=0 here means no 2D wavelet transformation. In accordance with our expectation of complexity does not necessarily ensure better performance, it turned out to be that only L=1 level wavelet is the most proper one. Explanation for the outcomes through further analysis would be: as the level of decomposition goes higher, the loss of feature representation in low frequency approximation outweighs far more than the gain of feature representation in high frequency detailed coefficients

The Bandelet2-coefficients are viable features. However, it's of expensive computational requirement due to its high dimensions. Besides Bandelet2 coefficients, due to the original robustness of geometric flow as image representation, it's entirely possible for statistical features such as energy sums (within each Bandelet2 block), coefficients sums, entropy and max geometric flow. Statistical features if managed well altogether would form a powerful image representation. However, correlation of weights and parameters for sub-features is difficult to learn, and some of these features alone are not intuitive enough to discriminate subtle change between adjacent images. Through experiments of feature extraction using each of the above mentioned features, we finally decide to take max geometric flow as the final image representation in Step7, noted as *stmax*. This *stmax* feature can well track the curve and shape change within a small block of image area, where the max geometric flow represents the most significant geometric structure, thus preserves both interior information and exterior information in image region that closely related to body shapes. Subtle motions will be discriminated through match of kernel dependency among image descriptors.

III. Prediction methods for pose regression

Due to the inherently ambiguity in estimation from a single view point, a superior predictor is urgently in need to solve issue between high dimensional image feature inputs and high dimensional structured outputs of human poses. In this section, we introduce the methods for prediction (or regression). We denote the image feature vector as $x \in R_k$, pose vector as $y \in R_d$.

A Gaussian Process Regression Review

Being a collection of random variables, any finite numbers of these have consistent joint Gaussian distributions, a Gaussian Process (GP) regression is a fully probabilistic method to model non-linear input-output dependencies [17]. It is a Bayesian approach which assumes a GP prior over functions:

$$p(f \mid X) = \mathrm{N}(0, K) \tag{4}$$

Where $f = [f_1, ..., f_n]^T$ is the vector with function values $f_i = f(x_i)$, $X = [x_1, ..., x_N]^T$, and K is a covariance matrix whose entries are given by a covariance function, $K_{i,j} = k(x_i, x_j)$. In practice, the function is commonly chosen as an RBF or Gaussian kernel. Because of the joint distribution on the prior, the posterior of prediction with a new test input is also a Gaussian conditioned on observed output, with mean and its covariance given by:

$$m(y^{(d)}) = Y^{(d)} K_X^{-1} K_X^x \tag{5}$$

$$\sigma^2(y^{(d)}) = K_X(x,x) - (K_X^x)^T K_X^{-1} K_X^x \tag{6}$$

Once we want to predict an output y, we can easily use (5) to directly infer from the distribution.

B Twin Gaussian Process Review

The GP regression is computationally intuitive yet unable to deal with multimodal mappings[6], which often arises in perceptual problems such as 3D from 2D inference. Twin Gaussian Process was originally introduced by Bo and C. Sminchisescu, to leverage information about correlations among output components in the predictor. The main idea of TGP is to extend existing GP with outputs of correlations, making the inputs and outputs relationship "structured". Given $Y^{(d)}$ and $y^{(d)}$ as training outputs and a testing prediction, combined together they form a Gaussian distribution $N_Y(0, K_{Y \cup y})$.

They measure the offset between the estimated Gaussian distribution of outputs with a new target and the corresponding input distribution, using Kullback-Leibler divergence (also called KL entropy):

$$D_{KL}(N_Y \,\|\, N_X) = -\frac{N}{2} - \log|K_{Y\cup y}| + \frac{1}{2} Tr\left\{ K_{Y\cup y} \begin{bmatrix} K_X & K_X^x \\ (K_X^x)^T & K_X(x,x) \end{bmatrix}^{-1} \right\} + \frac{1}{2}\log\left| \begin{bmatrix} K_X & K_X^x \\ (K_X^x)^T & K_X(x,x) \end{bmatrix} \right| \tag{7}$$

Because the output $y^{(d)}$ is unknown under this equation, in order to incorporate the observed input one into this offset, output targets (test data) are estimated by minimizing the KL divergence (9):

$$y^* = \arg\min_{y^{(d)}} [L(y^{(d)}) \equiv D_{KL}(N_Y \,\|\, N_X)] \tag{8}$$

Thus, prediction is finally made by minimizing the KL divergence w.r.t. outputs:

$$L(y) = K_Y(y,y) - 2(K_Y^y)^T K_X^{-1} K_X^x - \left[K_X(x,x) - (K_X^x)^T K_X^{-1} K_X^x \right] \times \log\left[K_Y(y,y) - (K_Y^y) K_Y^{-1} K_Y^y \right] \tag{9}$$

Final prediction on TGP adopts a non-linear optimization of the cost function (10). The terms (α and β) that do not depend on output are pre-computed in order to accelerate loss function and gradient evaluations. The function can then further be expressed as following:

$$L(y) = K_Y(y,y) - 2\alpha^T K_Y^y - \beta \log\left[K_Y(y,y) - (K_Y^y) K_Y^{-1} K_Y^y \right] \tag{10}$$

where $\alpha = K_X^{-1} K_X^x$, $\beta = K_X(x,x) - (K_X^x)^T K_X^{-1} K_X^x$.

The TGP is designed to ensure only training data that has both similar inputs and similar outputs with the query be of significance in the final estimate. In this sense, TGP with a structured framework differs much from classical regression methods such as ridge regression or GP regression, where the training data with similar inputs but different outputs will negatively affect estimation results in contradictory directions. Because of the highly computational requirement on the original TGP method, TGPKNN is proposed by Bo etc. in [14] with an intuitive notion that K nearest neighbors of a test input is found to estimate TGP.

IV. Experimental Evaluation

We validate the effectiveness of our approach mainly using real motion images from the Humaneva-I benchmark dataset from Brown University [9]. It contains four different human subjects (S1-S4) that mainly perform 6 actions (Walking, Jog, Box, Gesture, Throw/Catch, Combo) in a specific circumstance under calibrated camera sets. Video sequences are partitioned into training, validation, and test sets. Also, in test part of each subject, Combo sequences that contain walking, jogging and additional previously not appeared motions are included to evaluate estimation performance. Motion captured poses contain 10 parts:

head, torso, upper arms, lower arms, upper legs, and lower legs, 20 joints each represents a 3D camera free location (x, y, z), forming a 60d pose vector. During the experiments we use only one color camera information for monocular pose inference, be the case camera 'C1'. Pre-calculation for normalization in image serials is performed on each single human bounding box to overcome scale change in body size.

During experiments, we first use the training frames of each action to train parameters of corresponding subjects and tested on validation set. We compared with two types of state-of-the art image features: HMAX [12,14] (hierarchical descriptor on MAX operation) and HOG [14,15] (histogram of oriented gradients). Results are shown in S1 motions in table 1, also 3 different descriptors are validated on the same predictor of TGPKNN in Fig 3. Nearest neighbors are set to TGPKNN k=100， WKNN k=15 respectively for small training sets.

Table 1 Validation results on S1 between different models. HOG descriptors are extracted by R. Poppe's for comparison. Mean relative 3D error in *mm* per joint are given.

Model	***walking***	***jog***	***Throw-Catch***	***box***	***gesture***
GP+ *stmax*	29.55	48.15	**63.16**	26.44	**5.39**
GP+ HOG	38.92	48.37	68.32	49.79	11.38
WKNN+ stmax	25.86	48.48	63.83	26.72	6.69
WKNN+ HOG	38.36	46.38	69.79	48.31	13.17
TGP+ stmax	**24.66**	46.85	66.20	**24.92**	5.79
TGP+ HOG	28.71	**42.13**	64.19	43.25	10.38

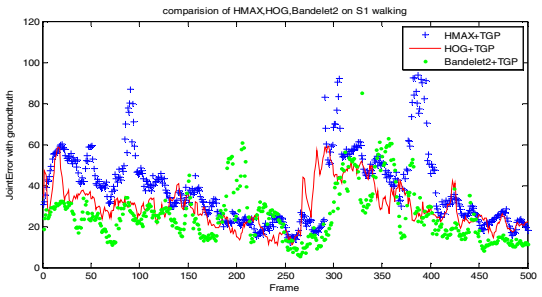


Figure 3 Results of mean joint error on validation part of continuous first 500 frames using different descriptors.

It can be seen from table 1 and Fig. 3, with the same predictor model, Bandelet2 descriptor outperforms HMAX descriptor and is more competitive, with most of the time better than HOG. The maxim error of pose estimation with Bandelet2 is much lower than that of HMAX. A few outliers in Bandelet2 may be due to invalid feature extraction. Box and Gestures fit well to the training set when on validation, but generalize normally worse on test sets.

On the test part, we first train separately on a small set e.g.: train on S1 training-validation part for walking and test on S1 test walking part. Noted that for the test sets, ground truth information is held back and results can only be gained through an online Humaneva evaluation system [9]. In our experiment, we set the nearest neighbors for k =100 for quantity evaluation. Evaluation results are shown in Table 1.

Table 2 Humaneva-I online evaluation results for test sequences, with our proposed method on GP, compared on HOG, HMAX models of Bo's work [14]. Both average error and SD in results are given. For space reason, only S1 is shown.

Model type	walking	jog	box	gesture	Ave-rage
Bandelet2	**59.05**	72.41	78.59	**24.93**	58.74
SD	***19.33***	*24.91*	*19.65*	***4.64***	\
HOG	62.1	**64.4**	**78.3**	26.8	57.9
SD	*25.9*	*21.1*	*21.7*	*11.2*	\
HMAX	65.3	69.6	90.8	30.8	64.13
SD	*22.3*	***19.8***	***15.6***	*6.7*	\

Though appearances also poses in the training and testing data have been relatively different, our method could still accurately infer the 3D poses (see table 2 and Fig. 4). Noted that results reported in table 2 of Bo's work are trained on all subjects and we have only trained on one single action series for each subject, still results are very competitive. It's clear that if training data is expanded better results would be achieved, we are going to explore more data during the next stage of our work. First row in figure 4 are original images from different subjects on test actions, second row are recovered 3D poses of the corresponding images.

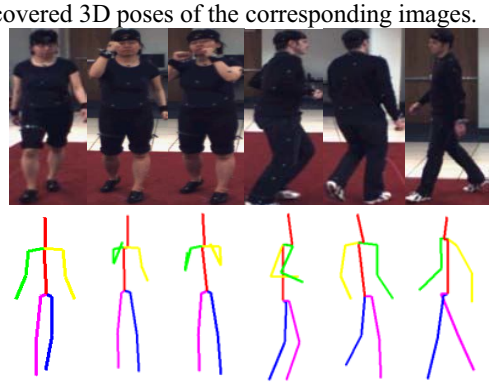


Figure 4 Results on test sets and recovered poses. Different colors represent different limb parts.

V. Conclusion

In this paper we presented an effective descriptor as image representation to inference high-dimensional, complex human motion poses. It's a novel of our attempt to convert the second generation Bandelet

transformation into an image feature with its statistical information besides Bandelet coefficients. Utilizing its representative character of geometric regularity other than traditional contours or edges, there is no need of background subtraction beyond regions of bounding box, which is computationally more relaxed.

We aim to tackle with the problem of 3D human pose estimation from monocular images by using a discriminative approach. Unlike classical image descriptors such as HOG, etc. that based heavily on contour or edge information of the human body, which more than often needs background subtraction and cannot strictly depict image context through only edge information, geometric flow enables Bandelet2 features to preserves both interior image information and exterior information that closely related to body shapes. Experiments reveal that Bandelet2 can preserve both invariance as well as selectivity in image. Results reported are only trained on each particular action, there is no doubt that by incorporating a large training set of variant actions and poses, the performance will be much better and robust [11,14]. In future work we aim to further exploration of Bandelet2 descriptors by fusion of other valid sub-features and effective exploration of training data to make better generalization on estimated pose.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61075041, 61001206 and 61001206), the Fundamental Research Funds for the Central Universities (No. K50510020009), the Key Project of Ministry of Education of China (No. 108115), and the National Research Foundation for the Doctoral Program of Higher Education of China (No. 2010020312005).

We thank Roland Poppe for giving great advice and generous support for making available materials from his work for testing and comparison.

References

[1] A. Agarwal, and B. Triggs,. "Recovering 3d human pose from monocular images." *IEEE transactions on PAMI, Number 1*, Vol.28, Jan 2006, pp. 44-58.

[2] L. Sigal and M. J. Black. "Guest Editorial: State of the Art in Image- and Video-Based Human Pose and Motion Estimation." *IJCV Vol87*, 2010, pp.1-3

[3] C. Sminchisescu and B. Triggs , "Hyperdynamics importance sampling." In *ECCV Vol. 1* , Copenhagen ,2002, pp.769–783.

[4] J. Deutscher , A. Blake , and I. Reid, "Articulated body motioncapture by annealed particle filtering." In *IEEE CVPR'00*, 2000, pp. 126-133.

[5] G. Shakhnarovich, P. Viola, and T. Darrell. "Fast pose estimation with parameter-sensitive hashing." In *ICCV 03 Vol 2*,Nice, France, October 2003, pp.750-759.

[6] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. "Discriminative Density Propagation for 3D Human Motion Estimation." In *CVPR 2005*,San Diego,CA, 2005.pp.390-397.

[7] R. Urtasun and T. Darrell. "Sparse probabilistic regression for activity independenthuman pose inference." In *CVPR'08*, Anchorage, AK, June 2008.pp.1-8.

[8] C. Sminchisescu, A. Kanaujia, and D. Metaxas. "BM3E: Discriminative density propagation for visual tracking." In *PAMI Vol 104*, 11,2007, pp.2030-2044.

[9] L. Sigal, A. Blan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion." In *IJCV Vol87*, 2010, pp.1-2

[10] L. Sigal, A. Blan, and M. J. Black, "Combined discriminative and generative articulated pose and non-rigid shape estimation." In *Advances in NIPS 20th*,Vancouver, Canada, December, 2008. pp. 1337–1344.

[11] L. Bo, C.Sminchisescu, A. Kanaujia, and D. Metaxas, "Fast algorithms for large scale conditional 3D prediction." In *IEEE conferenceon CVPR'08*, Anchorage, AK, June, 2008, pp.1-8.

[12] A. Kanaujia, C. Sminchisescu, and D. Metaxas. "Semisupervised hierarchical models for 3d human pose reconstruction." In *CVPR'07*, Minneapolis, MN, June, 2007. pp.1-8.

[13] Huazhong Ning, Wei Xu, Yihong Gong, and Thomas S. Huang, "Discriminative Learning of Visual Words for 3D Human Pose Estimation", In *IEEE conferenceon CVPR'08*, Anchorage, AK, June 2008, pp. 1-8.

[14] L. Bo, and C. Sminchisescu. "Twin Gaussian Processes for Structured Prediction." *In IJCV Vol. 87*, *no. 1-2*, 2010, pp. 28-52.

[15] R. Poppe. "Evaluating Example-based Pose Estimation: Experiments on the HumanEva Sets," In *CVPR 2nd Workshop on EHuM*, Minneapolis, MN, June 2007.

[16] C. Sminchisescu, A. Kanaujia, and D.Metaxas. "Learning joint top-down and bottom-up processes for 3d visual inferenc." In *CVPR'06 Vol 2*, New York, June 2006, pp. 1743–1752.

[17] Erwan Le Pennec and Stéphane Mallat. "Sparse Geometric Image Representations With Bandelets". In *IEEE transactions on IP Vol. 14 no. 4*, April 2005.

[18] J. Qui.nonero-Candela and C. E. Rasmussen. "A unifying view of sparse approximate gaussian process regression." In *Journal of Machine Learning Research*, 2005.

A Multi-Objective Immune Algorithm for Human Motion Tracking

Hong Han, Guangjie Feng, Jingxiang Gou, Rui Wang
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China
Xidian University
Xi'an, China
hanh@mail.xidian.edu.cn

***Abstract*—A crucial challenge in human body tracking is the high degrees of freedom (up to around 40) to be recovered. A method based on multi-objective optimization algorithm is presented here to tackle this problem. In our multi-objective optimization based human body tracking framework, tracking is considered as two functions' co-optimization problem where the aim is to optimize the matching functions between the model projection and the image observation. Experiments on 3D human motion tracking demonstrate the effectiveness and computational efficiency of the proposed tracking method.**

Keywords-body tracking; multi-objective optimization algorithm; immune algorithm

I. INTRODUCTION

Due to a broad prospect of potential applications concerning human motion analysis, where many practical utilization areas including intelligent advanced human-computer interface, virtual reality etc., would be progressively boosted, the research zest on human motion tracking has augmented profoundly within the last decades. However, recovering 3D poses and motions is always considered as an ill-conditioned problem since the depth is missing in the original 2D image space. How to track people effectively as well as accurately will be an "everlasting" challenging topic in the research of computer vision

Two main streams of different strategies are most popular for human pose estimation, well known as generative and discriminative methods [1].While discriminative techniques are easier to implement and less time consuming, they need huge training data set and have lower accuracy compared with generative approaches. Human body tracking adopting a generative method is considered as a problem finding a best solution in high dimension pose space. For nowadays powerful off-the-shelf PC, real time marker-less people tracking approaches have already become prevalent. Especially, many researchers address the intelligent searching and optimizing strategies which proved to be effective and comparatively more accurate. Wachter et al. [2] estimated motion of articulated model utilizing extended Kalman filter (EKF). Deutscher et al. [3] proposed the annealed particle filter (APF) algorithm, searched within plausible space in a deterministic manner with constrained particles. Shen et al. [4] proposed an Evolutionary Algorithm with probability (PEA) to tracking human motion within highly dimensional space using Voxel data.

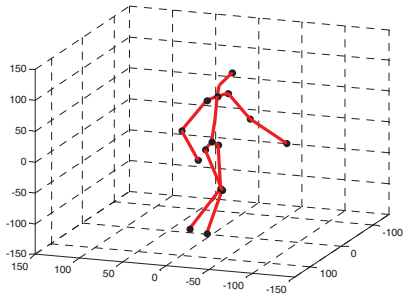


Figure 1. The human body model is represented as a 3-D kinematic tree

Because human body tracking can be considered as an optimization problem, in this paper we proposed a multi-objective optimization based approach for 3D human pose (as shown in Fig.1) estimation from monocular images. We proposed two likelihood functions to be optimized, which utilize more image information for tracking human body motion. We used Non-domination Neighbor Immune Algorithm (NNIA), proposed by M. Gong [5], as the multi-objective optimization algorithm. The method we employed here has the ability to generate a uniform distribution of representative Pareto optimal solutions on the Pareto Front (see [5] for details) and has lower computational complexity than other popular multi-objective optimization evolutionary algorithms such as NSGA-II [6]. Finally, because a set of Pareto-optimal solutions are found, a proper criterion is used to choose a final solution among the Pareto-optimal solutions.

The rest of the paper is organized as follows. The next section describes the related background for human body tracking. In section III, the framework of our algorithm is presented in detail. Experimental results are shown in Section IV, and Section V comes to a brief conclusion.

II. THE RELATED BACKGROUND

A. The human body model

The human body model used in this paper is defined as a set of n=15 joints in 3D (see Fig.1). The model has one root

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61075041, 61001202 and 61001206), the Fundamental Research Funds for the Central Universities (No. K50510020009), the Key Project of Ministry of Education of China (No. 108115), and the National Research Foundation for the Doctoral Program of Higher Education of China (No. 20100203120005).

978-1-61284-774-0/11/$26.00 ©2011 IEEE

joint and 14 parts: head, cervices, trunk, clavicle, humeruses, radiuses, femurs, and tibias, as shown in Fig.2.

We denote the coordinate of each individual joint by $v_i = [x_i, y_i, z_i]^T$, and the articulated pose configuration of human body can be represented by associating the three coordinates of 15 joints v_i together.

$$V = [v_1, v_2, \cdots, v_n]^T \tag{1}$$

Given the constraint of a bone length preservation are as follows:

$$\|v_p - v_q\| = l_{p,q} \tag{2}$$

Here the length of the bone between v_p and v_q is $l_{p,q}$. For $v_p - v_q$ is a linear transformation of V, So the constrains in (2) can be written to:

$$\|L_i V\| = l_i, \quad i = 1, 2, \cdots, m \tag{3}$$

Here L_i is a $3 \times 3n$ matrix. l_i is the length of bone i. m is the number of bones.

B. *Multi-objective optimization problems and NNIA*

Multi-objective optimization problems can be described as seeking to optimize a vector of functions such as:

$$\begin{cases} min \;\; y = F(x) = (f_1(x), f_2(x), \cdots, f_k(x),)^T \\ s.t. \;\; g_i(x) \le 0, \quad i = 1, 2, \cdots, q \\ \quad\;\; h_j(x) = 0, \quad j = 1, 2, \cdots, p \end{cases} \tag{4}$$

It is recognized that evolutionary algorithm is very fit for multi-objective optimization. There are variously famous algorithms which all plied in many real word problems. Such as Multi-objective Evolutionary Algorithm Based on Decomposition (MOEA/D)[7], Non-domination Sorting Genetic Algorithm II (NSGA-II). As one of the multi-objective optimization algorithm, Non-domination Neighbor Immune Algorithm (NNIA) can solution the multi-objective problems which NSGA-II, one of the most popular approaches, can tackle with.

Compared with NSGA-II, NNIA only sort in one front, then it can optimize faster. Considering the importance of real-time in computer vision community, we use NNIA in our tracking approach.

III. THE FRAMEWORK OF HUMAN MOTION TRACKING

In this paper, human motion tracking was considered as a multi-objective optimization problem, which was different

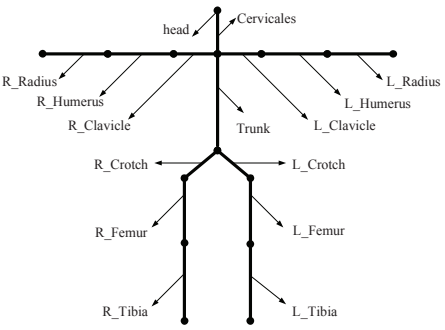


Figure 2. The 2D human body mode

from the other human tracking algorithms using the particle filter based on Bayesian framework. 2D image observation data is denoted as Z, and each individual joint can denote as $z_i = [x_i^{'}, y_i^{'}]^T$. In our work, the approach to find the observation values Z of the human body from the silhouette information is the same as [8]. Kalman filters are used to forecast the occlusion key points.

A. *The likelihood function of 3D human motion tracking*

For each candidate pose, corresponding image likelihood needs to be computed to evaluate how well the given body-pose matches the observed 2D images. The most commonly used features are edges and silhouettes, Grays, colors, and textures. Features above are utilized to generate a likelihood model solely as well as several combinations. In this paper, grays and key points are used to generate the following two likelihood functions.

1) *The likelihood function of distance*

When we aimed to estimate human motion with monocular image sequence, a projection model could be assumed. In this paper, a weak perspective projection model was used, and the default parameter of a root depth of field D was preset, therefore the 3D human body model could be projected onto 2D image plane, and then the projected points were matched with the corresponding detected points. Motion parameters which need to be resolved include the overall translation and rotation of the human body model. And human model is expressed as a matrix $V = [V_1, V_2, \cdots, V_n]^T$ which is described in more detail in section II. The projection coordinate of one joint point on the image figure is $v_i^{proj} = [u_i, w_i]^T$, the relation between 3D coordinates and the 2D coordinates can be formulated as:

$$[u_i, w_i, 1]^T = P_i V \quad i = 1, 2, \cdots, n \tag{5}$$

Here $P_i = T_i K$, and T_i is a transformation matrix dependent on the barycentric coordinate, K is internal camera parameter. Therefore distance function can be written as:

$$\arg\min_V f_1(V) = \sum_{i=1}^{n} \left\| v_i^{proj} - z_i \right\|_2 = \sum_{i=1}^{n} \left\| P_i V - z_i \right\|_2 \tag{6}$$

2) The likelihood function of Gray feature

Assuming that changes in image intensity are only due to translation of local image intensity, a parametric image motion between t and $t-1$ can be described by the following equation:

$$I_t(V_t^{proj}) = I_{t-1}(V_{t-1}^{proj}) \tag{7}$$

I_t is the image intensity at time t, which can be got by using sobel operator, and we denote $I_t(V_t^{proj}) = S_t V_t^{proj}$, then the gray likelihood function takes the form as:

$$\arg\min_V f_2(V) = \sum_{i=1}^{n} S_t V_t^{proj} - S_{t-1} V_{t-1}^{proj} = \sum_{i=1}^{n} S_t P_i V_t - S_{t-1} P_i V_{t-1} \tag{8}$$

B. tracking

Our optimization algorithm in this paper is Non-domination Neighbor Immune Algorithm (NNIA). The following of this section gives a description of our algorithm.

1) Immunic Representation

In Non-domination Neighbor Immune Algorithm an individual of the antibody population consists of Nbits. $G_1, G_2 \cdots G_N$ and each bit can assume as a joint v_i in the body model, while each individual can construct a human pose. N is the number of joints. Especially, in our model N=15 and the number of generations is 10 in our algorithm.

2) Objective Function and Optimization

As described above, considering the constancy constrain of bone length (3), the aim of optimizing the matching function (6)(8)can express as one multi-objective problems as follow:

$$\begin{cases} \arg\min_V F(V) = [f_1(V), f_2(V)]^T \\ f_1(V) = \sum_{i=1}^{n} \left\| P_i V - z_i \right\|_2 \\ f_2(V) = \sum_{i=1}^{n} S_t P_i V_t - S_{t-1} P_i V_{t-1} \\ s.t. \left\| L_i V \right\| = l_i, \quad i = 1, 2, \cdots, m \end{cases} \tag{9}$$

3) Select Criterion

In multi-objective optimization algorithm, the Pareto-optimal front is a set of solutions. In order to get one fitness pose we use a method to selection from two consecutive video frames. Since the human body poses in consecutive frames are similar and have an internal consistency, we adopt a subtraction scheme to discriminated ifferences between detection value at time and projection associate with pose at time.

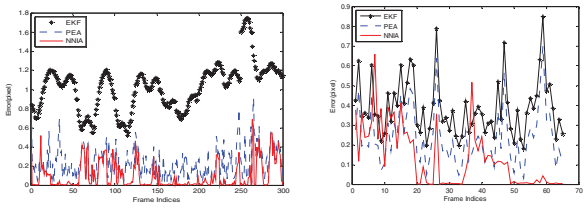


Figure 3. Comparasion betweem EKF ,PEA and our method using video1(left) and video2(right)

IV. EXPERIMENTAL RESULT

In order to evaluate the performance of the proposed algorithm, we have carried out a large number of experiments on video sequences. The first video, a rat of 30 frames per second, is a real person squares as seen in Fig.4. The image size is 320×240 and last 300 frames. The second video is created with commercial 3D rendering software Poser to render the real walk. The image size is 500×488 and have 65 frames seen again in Fig.5.

A. Error measure

The error measure we used in this paper is the method mentioned in [9] . The error between the 2 D projection associated with estimated pose V and detection key joint Z is expressed as average Euclidean distance (see Fig.3):

$$D(V, Z) = \frac{1}{M} \sum_{i=1}^{M} \left\| V_i^{proj} - Z_i \right\|_2 \tag{10}$$

B. Tracking result

We compare our approach with EKF (see [2] for detail)and PEA (see [4] for de tail). Because the criterion (see section III for detail) employed in our algorithm is based on the difference between concatenation frames, the accumulation error is the least and the performance is better than EKF. Due to combining the two objective functions (distance likelihood function and gray likelihood function) our method is more accurate. The recovered 3D human pose reflects the key points' consistency according to distance and gray, which certainly perform better than single objective optimization framework PEA used. EKF perform better in video 2, for the velocities of real person motion between upper limb and lower limb are different. EKF set the same velocity in the filter processing, showing better result in video 2 sequences.

PEA is also an optimization method for human body tracking, however it uses only one objective and shows better results than EKF. Compared with PEA, our method has more accurate results because the two objectives can reveal image information more exactly, sees Fig.4.

Fig.4 presents several sample images out of 300 frames, and the recovered 2D and 3D human pose of the first video.

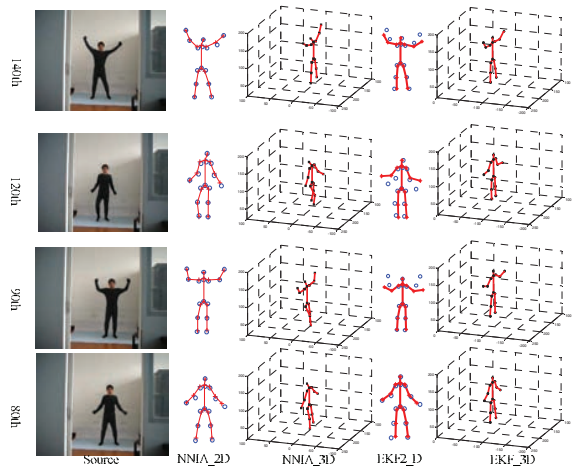


Figure 4. Comparasion our NNIA-based among tracking method and EKF. The first column is the frame in test video. The second and the fourth is 2D tracking result with our methed and EKF. The blue circle and red line outlet the detaction value and projection of 3D recovered pose, respectively . The third and the fifth is 3D tracking result using our method and EKF. The black dot and red line outlei teh key points and skeleton, respectively

Considering 2D results, our method can recover almost all of the human motion, while the EFK only tracks few frames. That is because our method utilized the information between two concatenate frames and therefore avoided error accumulation which seriously affected the tracking result of EKF. In term of 3D results, EKF's performance is worse than our approach.

Our method could recover the depth information and the vision effect is better. However the 3D results of EKF cannot recover most of the depth information. Compared with EKF, our method produces a better tracking result both in terms of error and robustness. Though better results have been recovered by our method, sometimes it is still unsatisfactory and cannot obtain all the best results in the experiment, such as in Fig.5, the left elbow cannot track with the body motion.

V. CONCLUSION

For the high dimensionality nature of human body itself, non-rigidness, close deformation and depth ambiguity, human motion tracking is a real challenge in computer vision field. In this paper, we considered tracking problem to be a multi-objective optimization problem and presented a novel tracking method, which introduces the multi-objective immune algorithm (NNIA) in to the searching strategies within high-dimensional space. The two likelihood functions' optimizing strategy rendered the tracking procedure more effective and robust. Experiments on people tracking using unclebrated video camera demonstrate the effectiveness and computation efficiency of our approach..

In our approach, the likelihood functions are based on key points and gray information. Actually there are many other features can be used to generate the likelihood function, such as edge, texture, color or more sophisticated features. We will investigate other possible likelihood functions and introduce

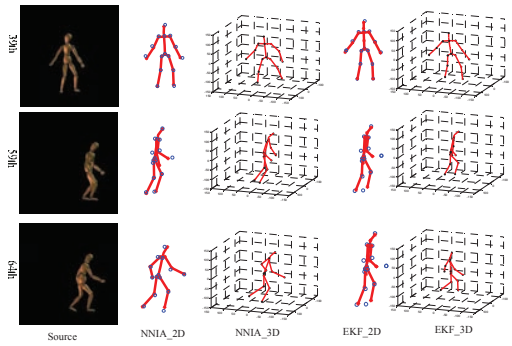


Figure 5. Comparasion our NNIA-based among tracking method and EKF. The first column is the frame in test video. The second and the fourth is 2D tracking result with our methed and EKF. The blue circle and red line outlet the detaction value and projection of 3D recovered pose, respectively . The third and the fifth is 3D tracking result using our method and EKF. The black dot and red line outlei teh key points and skeleton, respectively

abundant valid features of image into multi-objective optimization frame in the future.

ACKNOWLEDGMENT

We thank Zhichao Chen for discussions and generous support with key points extraction.

REFERENCES

[1] D. Kim, D. Kim, "A Fast ICP Algorithm for 3-D Human Body Motion Track, "Signal Processing Letters, vol. 17, pp. 402-405, April 2010.

[2] S. Wachter, H. Nagel, " Tracking Persons in Monocular Image Seuences," Computer Vision and Image Understanding, vol.74(3), pp. 174-192, 1999.

[3] J. Duetscher, A. Blake, I. Reid, "Articulated Body Motion Capture by Annealed Particle Filtering," cvpr, vol. 2, pp.2126, 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'00) - Volume 2, 2000

[4] S. Shen, H. Deng and Y. Liu, " Probability Evolutionary Algorithm based human motion tracking use voxel data," Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on. Pp. 44, June 2008.

[5] M. Gong, L. Jiao, H. Du, and L. Bo, "Multiobjective Immune Algrithm with Nondominated Neighbor-based Selection," Evoluationary Computation, vol. 16(2), pp. 225-255, 2008.

[6] K. Deb, A. Pratap,S. Agarwal, and T. Meyarivan, " A Fase and elitism multiobjective genetic algorithm: NSGA-II," Evolutionary Computaation, vol6(2), pp.182-297, 2002.

[7] Q. Zhang, H. Li, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition," Evolutionary Computaation, voll 1(6), pp.712-731, 2007

[8] M.Tong, " Three Dimensional Human Motion Analysis from Uncalibrated Monocular Image Sequences," PhD thesis (In Chinese). Shanghai Jiao Tong University.2007..

[9] L. Sigal, O. A. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," Evoluationary Computation, vol. 87, pp. 4–27,2010.

Sparse Geometric Features for Visual Detection and Estimation

Hong Han [*] , Jingxiang Gou[*] , Guangjie Feng[*], Rui Wang[*] ,

[*] Xidian University
Xi'an, 710071, China
Email: hanh@mail.xidian.edu.cn

Abstract

Image feature has been a crucial perquisite for computer vision researches such as pedestrian detection and human pose estimation. In this paper, we devised a novel image feature utilizing sparse geometric representation, to deal with image patterns that have to do closely with visual human detection and pose estimation. The speed up optimization here ensures a promising performance for effective feature extraction. Combined with state-of-art learning methods, this paper indicates how the proposed image features could boost the representativeness for specific vision tasks from monocular video images by leveraging geometric flow information to characterize image context, especially for human body shapes and motions. We have compared our image feature with classic global features such as HOG, HMAX, and conducted comprehensive experiments in detection and human pose estimation problems on benchmark datasets. Final evaluation verifies competitive discriminatory effectiveness and distinctiveness for our proposed feature in such visual tasks.

Keywords-*geometric flow; sparse representation; visual detection; pose estimation; multiscale geometric analysis*

I. Introduction

Image features or representation has been a hot topic in computer vision, with a wide scope of direct impacting topics in detection, recognition, image retrieval and pose estimation, etc. Further real world applications can be expended into HCI, personal identification and intelligent surveillance, etc. The notion of image feature primarily concerns with the information that could provide structural context, from point, pixels to more complex ones such as curve, edge or high level semantics. There are numerous types of image features that could reveal the resemblance of patterns as well as distinguish the difference among large amount of picture sets. In general, too schools are bringing about highlights, local features and global features. Local features detect interest points with certain rules of invariance within near neighboring, or collect dense local points within a small region. The earliest research came from a particular pursuit for object tracking, where Harris and Stephens further proposed famous Harris detector to locate corners in moving objects [33].Famous ones of local features are mainly known for SIFT, Pyramid Match Kernel (PMK) [7], shape context [1], also other bag of words representations. Local features like SIFT [18] are rather appealing for visual analysis and recognition tasks, problems however lies in the time consuming process to learn a code book for matching correspondence, as well as the final image presentation itself that would form a large dimensional data for each frame. It's computational prohibitive for searching and comparison among large datasets. Another generally category of global features such as HOG [11, 13, 14], HMAX [12, 13], have shown great potential in measuring correspondence based on entire image frame and make available for faster batch calculation procedures. All the above features are good, except most of them are based on edge or silhouette information and thus heavily rely on accurate background subtraction and/or silhouette extraction. Our speed up feature with sparse geometry from Bandelet is yet a sparse global feature, and it circumvented above problems and made image presentation powerful with much lower feature dimension, both for detection tasks and human pose estimation tasks.

Human detection and human pose estimation However, due to intrinsic complexity of articulated objects as human body itself and recording quality of image sequences with complex video context variance, robust and satisfying solutions to real life challenges remains to be discovered. We can normally use representations with a certainty measure to ensure the

representation contains specific information in terms of edge orientation, color, pixel or sub-pixel neighborhood correlations, such as HOG and PCA-SIFT methods have brought about inspiring achievements. From another prospective, human body is unlike other kinds of objects that need to be matched, the former largely shows specific visual pattern like straight up standing and a star like physical structure. Though motion types are variant, but the most difficulty subjects to scale change and inner information within in image, other than rotation of direction, (a human subject is very unlikely to lean up sight-down like the situation with most work in object categorization). So in a computer vision work, when designing an algorithm, the choice of a specific feature representation towards a particular purpose can be very crucial. We need to find a balance between representativeness and computational demand when deciding whether to incorporate high-level of details or not, in other words, what kind of information is necessary and helpful to solve the work at hand and what are not. In this case, we are aiming to discover a proper feature from the second generation Bandelet transformation that is particular useful for identifying human body shapes and motions.

For visual human analysis tasks a rising school of approach is the *discriminative approach,* which is also known as *data-driven approach*. It can be interpreted as using amount of data and machine learning methodology to learn a general correlation from the training images to the labeled status of human objects, to determine whether or not the image contains a specific human object or the motion types/parameters of a human. For pedestrian detection works, we use a feature in conjunction with classification model[20], the learned rules can make confidence whether a image region contains human objects. As in human pose estimation work, prevalent ones are conducted in a supervised [1, 7, 8, 11, 13, 14] or semi-supervised [12] manner. Once trained on a dataset with sufficient actions and subjects with 3D motion capture data as labels[9], they are capable of predicting 3d poses for test image streams in a fast manner. Though learning approaches are generally less precise than conventional generative approaches, they are far more convenience since which do not need initialization, prior model, or camera parameter, also are not confined to any specific type of person in appearance.

To carefully investigate all aspects behind this human motion analysis frame work, every single piece cannot be trivialized. Such as original video recording quality, back ground clutter, image descriptor /detector's robustness and learning models (also known as classification/regression models). In general, there has been rising trend in designing effective image representation and the features are of equal significance compared with learning methods in current concern of effective learning. Effective image features, whether detectors or descriptors, global or local ones, have been proposed and applied to the human detection [20,21] and pose estimation tasks [1, 7, 12]. However, most representations are mainly based on contours or edges of subjects, being unconvincing to descript image contents when edges lose track within image region or glitches appear, which may still need lifting in order to better discriminate tinny subject differences.

A. Related work

The research mainly relates to the background of image descriptors, sparse geometric representation, multiscale analysis, visual pedestrian detection and human pose estimation, as well as a new trend in devising novel models for object recognition [24].

A critical problem that we encounter in image processing and presentation is that description for images heavily rely on the scale of the given image pattern. A multitude of image structures only exist within their limited regions of scales. Therefore, a variety of multi-scale image representations are developed, such as the image pyramid, quad-tree method, etc. Sparse representation with geometric regularity from the second generation Bandelet transformation is a promising way to perform multiscale analysis and image signal processing [2]. Original work with Bandelet relates to devising for image/surface/mesh compression [23] and denoising task [16] through the very step of bandelization to track directional geometric pattern upon each targeted scale. In this paper, our proposed speed up feature with the idea of Bandelet to form sparse image representation with high selectivity and distinctiveness. Our sparse geometric feature is based on geometric flow, also being invariant to scale change, illumination, or back ground clutter to a large extent. Here, we follow the discriminative approach to solve visual human detection and estimation problem with a novel descriptor specified on the second generation Bandelet. We have found a new tool of image representation to better depict image context for visual tasks. For detection work, the resulted feature for detection is low-dimensional, and scanning speed for localizing a human subject is impressively quick, while results are more accurate. For human pose estimation works, there is no need of background subtraction beyond regions of bounding box, which is computationally relaxed.

Also, Representations with Bandelet2 are of low dimensions.

Effective image representation whether be it detector or descriptor, will finally affect the results to a large extent. Most of these works are testifying their effectiveness on Caltech-101 or Pascal challenges, where objects in such database vary a lot in pose, illumination, rotation and scale. It's a bit different from our object to devise a feature that fits particularly for the purpose of identifying human and their motions.

As a feasible approach in human motion analysis work, we believe that the combination of discriminative and generative methods would further lead to a more precise estimation [15], that is to use inherent human body structural information as a prior condition for likelihood calculation of given resemblance scoring while a discriminative approach serve the function for finding significant parameters for data correlation. However, Due to our principle of simplicity and computationally relaxation to solely observe the effect of image feature, a discriminative approach is taken in this work. Training and testing covers the primary space of such works [26,27].

The paper is organized as follows. In the next section we review Bandelet theory and give detailed designing process of our proposed feature. Experimental selection of feature parameters is given in Sec.3. Sec.4 gives the experimental results on human detection and human pose estimation of benchmark datasets. In Sec.5 we came to the conclusion.

II. Geometric Bandelet as Image Representation

A. The second generation Bandelet for geometric image representation

The second generation Bandelet (shorted for Bandelet2), was proposed on the idea of multiscale analysis instead of describing the image geometry through edges, where the latter are often ill defined. The image geometry change in orientation is characterized by a *geometric flow* of vectors [23]. See Fig 1 for illustration of geometric flow. The very step of bandelization preserves both patterns inner and exterior along the geometry boundaries. Motivated by this idiosyncrasy of Bandelet, we transform the Bandelet into distinctive image feature because its highly discriminative ability, which is also invariant to scale change, illumination, or back ground clutter to a large extent. The main idea is to characterize geometric features within a specific region to be *vector fields,* which indicate local orthonormal directions of gray scale change. In this sense, Bandelet2 is way different from non-adaptive methods of constructing bases. As we generally cannot predict in advance the geometric image regularity, image representation from sparse geometric flows must be adaptively gained upon a specific image according to final applications. Which is time consuming and computational prohibitive for large-scale datasets feature extraction and correspondence matching. We have pre-indexed the bandelet coefficients after the wavelet transformation step to achieve much faster coding speed than the original Bandelization. For this pose estimation frame work, we need to jointly learn parameters to the given task as well, see section 2.2.

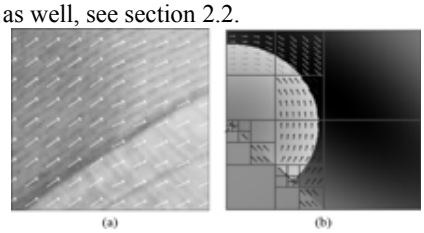


Figure 1 (a) geometric flow in a local region. (b) An adapted dyadic squares segmentation. Figures are from [16]

1). Constructing sparse Bandelet basis

1: the objective function is based on Lagrange method to optimize the best basis function, where the objective function can be expressed as:

$$L(f_\theta, R) = \left\| f_\theta - \widetilde{f}_\theta \right\|^2 + \lambda * T^2 (R_g + R_b) \qquad (1)$$

Where the first part is approximate error (squared error) and the second part is the computational complexity. Altogether they form a penalty evaluation for locating best local geometry. See details in section 2.2.

2: adapt a quadtree division according to the objective function minimization principle, and CART law to emerge sub blocks bottom-up, the optimal quadtree division can thus be acquired, see also figure 3.1 for illustration. More details are given in [35].

3: determine the optimal geometric direction within each final sub-block of previous division in step2, perform orthogonal projection upon each optimal direction. Thus two dimensional wave functions are transformed into one dimension.

B. Sparse geometric image feature extraction

Detailed steps for extracting image feature from image sequences with the Sparse Geometric transform are as follows:

Step 1: For each image, the human body is detected within a bounding box and rescaled to a fixed size.

(For images/datasets of the same size, just skip step 1).

Step 2: Perform 2D wavelet transformation.

Step 3: According to the quadtree division and CART law, and get the optimal devised sub-blocks (see part **2.1.1**): For a square S of width L, compute the best geometry and Lagrange penalty value $L_0(s)$. And then let $L = 2L$, repeat the previous step. For each $L \times L$ sized square S, compute the optimal geometry and Lagrange penalty value $L(s)$. For a certain level square S of $L \times L$, its four children are (s_1, s_2, s_3, s_4), and the combined Lagrange function is:

$$L(s) = L_0(s_1) + L_0(s_2) + L_0(s_3) + L_0(s_4) + L_0(s) \quad (2)$$

Let $L_0(s) = \min\{L(s), L'(s)\}$.

Here we use a speed up indexing method to rearrange the resulted 2D wavelet signal into 1D signal. The division of sub-blocks can be viewed as continuous partition in a line.

Step 4: Perform 1D discrete reordering of the sampling location referred in step 3. by projecting the sampling location along d (directions) and sorting the resulting 1D points in a line, which can defined as a 1D discrete signal f_d .Then perform a 1D discrete wavelet transformation of f_d, and get f_θ.

Step 5: Selection of the best geometry: For a given threshold T, the direction d which generates less projection error is to be chosen. R_b is the number of bits to encode the quantized coefficients $\{Q(t)\}$, $\tilde{f}_\theta$ is the quantized f_θ. We use a uniformed quantize term.

$$Q(t) = \begin{cases} 0 & |x| \le 0 \\ sign(x) * (q + 0.5) * T & qT \le |x| \le (q+1)T \end{cases} \quad (3)$$

To select the optimal geometry structure, we must choose the direction d that minimizes function (1), where R_g is the number of bits needed to code the geometric parameter d with an entropy coder, and λ is the empirical penalty factor 3/28.

Step 6: The resulting 1D wavelet coefficients corresponding to the best geometry d can be restored inrto a 2D matrix of the same size as S, which is called Bandelet2 coefficients.

Step 7：Calculate the maximum statistical geometric flow of sparse Bandelet2 coefficients with assigned block size (namely 4*4, 8*8) into the final feature.

Brief illustration is given in Fig 2.The image structure is sparse in high level details. High frequency image regions indicate brief contour variation and shape pattern of the moving human subject. For an image size of 192*64 with 4*4 block-wise extractions, the feature representation is 48*16=768 in dimension. While with the 8*8 size block, the dimension reduces to 192 per image frame.

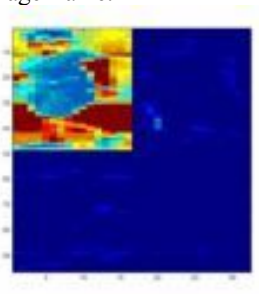


Fig 2. spectrum illustration after discrete reordering of wavelets. Four regions given above correspond to Level=1 wavelet transformation result.

The Bandelet2 coefficients are surely viable features. However, it's of expensive computational requirement due to its original high dimensional character. The feature matrix dimension equals to the pixel number of a frame, which will no doubt increase searching time for correspondence. Besides Bandelet2 coefficients, due to the original robustness of geometric flow, it's entirely possible and a necessary to adopt statistical features such as energy sums (within each block), entropy and max geometric flow instead. Statistical features, if managed well altogether would form a powerful image representation. However, correlation of weights and parameters for sub-features is difficult to learn, and some of these features alone is not intuitive enough to give specific physical explanation and discriminate subtle change between adjacent human images. Through experiments of feature extraction using each of the above mentioned features, we decide to take max geometric statistics as the final image representation in Step7, noted as *sparse geometry-stmax* (*stmax* for short). This *stmax* feature can well track the significant curve and shape change within a small block of image region where the max geometric flow represents the most dominant geometric regularity direction, thus preserves both interior and exterior information along image region that closely related to body shapes. Subtle human motion change will be discriminated through match of kernel dependency among frames of image features.

III. Experimental evaluation for Bandelet2 parameter selection

A Testing dataset for pedestrian detection

To verify the effectiveness of the speed up sparse geometric feature, we have to first utilize effective datasets to train parameters and use partial datasets to establish effective standards. As all kinds of approaches for human detection need comparable datasets to prove the capability of detection to the real world problems, we have to resort to datasets with comprehensive considerations, such as Caltech Pedestrian Dataset [27], MIT LabelMe dataset [26], and INRIA, the public available benchmark [29]; The Caltech 101 object categorization data [25], which is used to prove the performance of matching objects. In this paper, we use the INRIA dataset to train, select feature parameters and MIT Pedestrian Image Dataset to test, respectively.

The INRIA dataset was firstly used to verify the effectiveness of HOG feature by Dalal and Triggs etc., which contained both training and testing set [20]. All pictures are cut out of natural scene photos. We selected positive samples that each contains a human subject from training set with 2416 images and from testing set with 1132 images, both of size 128×64 as the positive samples set. We can obtain the negative samples (without human subjects in scene) in the same manner from training set that contains 6050 images and from testing set with 821 images as the negative sample set, all pictures are of the same pixel size as the positive set. The two sets above in pair forms the human body sample set. Fig 3 shows samples in both negative and positive sets.

Fig 3 (a)

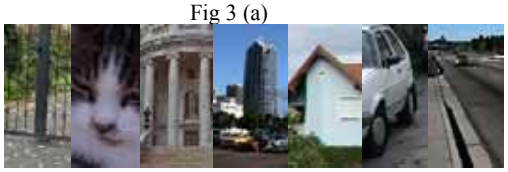


Fig 3 (b)

Fig 3 (a) positive samples with human subjects. (b) positive samples with non-human natural scenes, from INRIA [29].

For testing final results and compared evaluation, we use the MIT still human database of 460 static human body images. We have especially selected 300 challenging scenes to perform experiments with still images of human. Each image includes at least one fully front standing or back posing upright human body, with both horizontal and vertical size roughly about 500 pixels. Figure 4 shows examples of the dataset.

Here we use standard processing pipeline with the method proposed in [20] and only replace the stage of feature extraction with our proposed method. By doing this, we can eliminate other factors and easily find out the contribution of our descriptor.

Fig 4 partial examples of the MIT human dataset

To use sparse geometric features for visual detection, the input testing images are firstly rescaled adaptively with different ratio, like 0.4, 0.7, 1.0 etc with respect to the given image size. We perform an integration step to make the window scanning for features easier, with X=|X/8|×8, and Y=|Y/16|×16, X, Y is the number of pixels of columns and rows respectively. We use window scanning with every 8 pixels in the horizontal direction and 16 pixels in the vertical direction, and subdivide the blocks into 4 partitions. All the 4 quarters would form an integrated pixel grid to record the sparse geometric flow coefficients. See Fig 5 for illustration. After block-wised scanning, the statistical information within each cell can be calculated using *stmax* to capture the dominant geometric pattern in accordance with the training set.

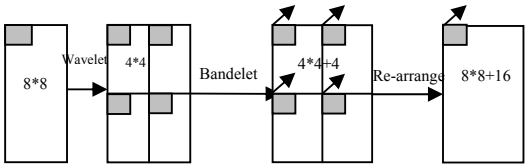


Fig 5 Window scanning subdivision and integration

B Optimal parameter selection

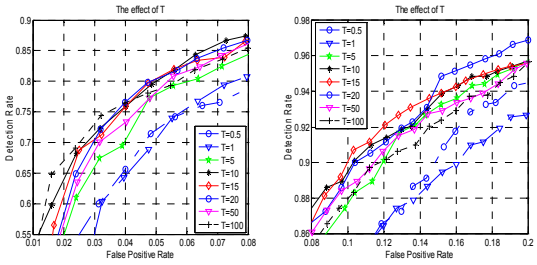
Parameters used specifically in our experiment to determine optimal features are T , L , $j_{\min}$ and $j_{\max}$. T is the quantized threshold, L is the level of wavelet, $j_{\min}$ is the smallest scale division size, while $j_{\max}$ is the largest integration scale of bottom-up quadtree fusion. By experiments, we can determine T to be 15, L be 1, $j_{\min}$ be 2, and $j_{\max}$ be 2. Details as given below.

We use standard ROC(Receiver operating characteristic) curve to track the sensitivity of the correlation between false positive rate and detection

rate. The upper trending line, or curve that bends more closely towards the upper-left corner within a plot ensures a better capacity in testing character. We use off-the-shelf SVM as the classification tool and

1) Quantized threshold T

The parameter T came with the mission to give higher image compression rate [35], where the higher T value will lead to a decrease in non-zero Bandelet coefficients. With a higher value of T, compressive rate rises while authenticity of image details declines, and in human detection and pose estimation work this largely affect the efficiency of Bandelet2 image feature in describing body contours, as well as interior and exterior flow vector counting. Previous experiments and published works [16] concerning Bandelet transform have suggested, neither a too high nor a too low value is preferable when attempting to locate the best direction of particular geometric flow. The selection of threshold varies upon different applications. For the basic need to encode image descriptor, we then adopt the hard thresh (HT) to build quantizing gate. Fig 5 shows the ROC curve on INRIA training set to determine T. Noted that Fig 9 (a) and (b) showed continuous false positive rate. T is thus set through experiment with training sets, with a uniform value of 15 over all frames.



(a) (b)

Fig 5 the ROC curve on training set to select threshold T. (a)false positive <0.08 (b)false positive >0.08

2) Quadtree division scale

In Geometric Bandelets' theory [23], the selection of minimum devised scale size is preferably to be set as small as possible and the maximum division scale is set as large as possible. However in our application to visual detection and estimation works, we noticed that there exist some tolerance. Based on the assumption of L=1 and T=15, ROC of j_max and j_min are given in Fig 6. We can further came to the conclusion that even both parameters are set in a small value and do not differ much, we can still achieve a fair result. The selection of both j_max and j_min is 2. Probally because the original image size is limited and do not need a cross level fusion of geometric flows. So it's natural to use a block size of 4*4 (2^j =4), which would ensure more appealing representativeness.

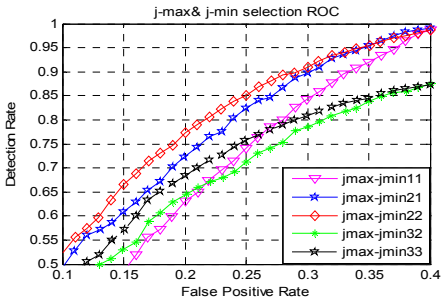


Fig 6 the ROC curve of j_min and j_max selection.

3) Wavelet transformation level

The 2D wavelet transformation was processed with different levels from L=0 to L=5 during previous stage of feature selection, and L=0 here means no 2D wavelet transformation. Fig 7 shows results on changing levels L.

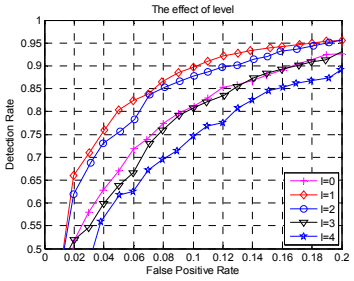


Fig 7 the ROC curve of wavelet transform level.

In accordance with our expectation that complexity in details does not necessarily ensure better performance, it turned out to be that only the level of wavelet L=1 is the most proper one. Explanation for the outcomes through further analysis would be: as the level of decomposition goes higher, the loss of feature representation in low frequency approximation outweighs far more than the gain of feature representation in high frequency detailed coefficients.

C Testing results on MIT dataset

And following are the results of testing on MIT data set. We have compared with HOG and show competitive performance. Fig 8 gives the curve of detection with both SG and HOG. See Fig 9-11 for comparison of our proposed method with HOG feature. Both methods are trained with SVM and with 10-fold cross validation.

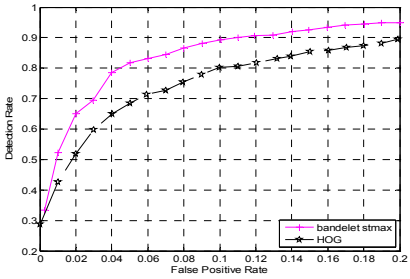
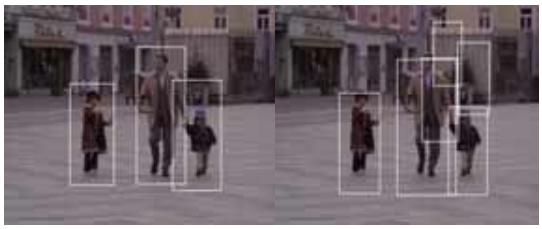


Fig 8. the ROC curve on training set of max-statistical feature, compared with HOG proposed in [20].

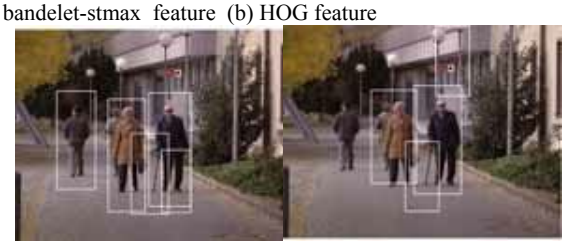


(a) (b)

Fig 9. Testing results on a frame with from (a) bandelet-stmax feature (b) HOG feature

(a) (b)

Fig 10. Testing results on a frame with from (a) bandelet-stmax feature (b) HOG feature



(a) (b)

Fig 11. Testing results with false alarm on a frame with from (a) bandelet-stmax feature (b) HOG feature

Table 1 Testing results on MIT for human detection. HOG features are extracted by Dalal and Tirggs' method [20] for comparison. Rounds indicates the training rounds of training with SVM.

Term / Feature	Rounds	Training Error	Accuracy	False Positive	s/frame
HOG	200	0.0936	86.585%	7.99%	0.119s
SG	108	0.0924	91.324%	6.54%	0.148s

We have further trying to find out whether the low dimensional bandelet-stmax feature is capable of discriminating types of objects under Caltech-101 object recognition dataset [25]. Using 15 training images per category while leaving 30 random images within the rest of the part to test, and the learning results are 40.2% with libSVM and 43.4% with kernel SVM, respectively. Since the feature is designed to match shape change with upstanding people in most frames, and do not acquire much robustness towards rotation and translation, the rotation that exists in most pictures would lead to a problem for correct matching with our block-wised descriptor. Such change would lead to lose in tracking of finding same binary patterns within corresponding position across frames. We are thinking of adding near neighboring characters to further enhance the weakness in recognizing objects with obvious rotation variance. Wang et al [34] came to an interesting finding that there is no best feature suit universally to all datasets, so the learning should better be designed directly from a particular dataset, which would partly explain our problem with Caltech-101 dataset as we have learned our feature from human datasets.

The main reason for our comparison building upon HOG is that both proposed feature and HOG are hand-engineered global features extracted from an unsupervised manner, and the latter is especially classical for its success in pedestrian detection and human recognition tasks. While others may have measurable impact such as SIFT feature, and BOW (bag of words), such local features yet need further time in preliminary learning for code books to get meaningful visual clusters which is extremely time consuming for large datasets. Local features and global features have different focus and merit, so this part of discussion and comparison are left to further work.

IV. Experimental Evaluation for visual estimation

A Gaussian Process Regression Review

Due to the inherently ambiguity in estimation from a single view point, a superior predictor is urgently in need to solve issue between high dimensional image feature inputs and high dimensional structured outputs of human poses. Different methods have been proposed to cope with pose estimation problems [2,3,4,5,6].In this section, we introduce the classical method for prediction (or regression). We denote the image feature vector as $x \in R_k$, pose vector as $y \in R_d$.

Being a collection of random variables, any finite

numbers of these have consistent joint Gaussian distributions, a Gaussian Process (GP) regression is a fully probabilistic method to model non-linear input-output dependencies [17]. It is a Bayesian approach which assumes a GP prior over functions:

$$p(f \mid X) = \mathrm{N}(0, K) \tag{4}$$

Where $f = [f_1, ..., f_n]^T$ is the vector with function values $f_i = f(x_i)$, $X = [x_1, ..., x_N]^T$, and K is a covariance matrix whose entries are given by a covariance function, $K_{i,j} = k(x_i, x_j)$. In practice, the function is commonly chosen as an RBF or Gaussian kernel. Because of the joint distribution on the prior, the posterior of prediction with a new test input is also a Gaussian conditioned on observed output, with mean and its covariance given by:

$$m(y^{(d)}) = Y^{(d)} K_X^{-1} K_X^x \tag{5}$$

$$\sigma^2(y^{(d)}) = K_X(x,x) - (K_X^x)^T K_X^{-1} K_X^x \tag{6}$$

Once we want to predict an output y , we can easily use (5) to directly infer from the distribution.

We validate the effectiveness of our approach mainly using real motion images from the Humaneva-I benchmark dataset from Brown University [9]. It contains four different human subjects (S1-S4) that mainly perform 6 actions (Walking, Jog, Box, Gesture, Throw/Catch, Combo) in a specific circumstance under calibrated camera sets. Video sequences are partitioned into training, validation, and test sets. Also, in test part of each subject, Combo sequences that contain walking, jogging and additional previously not appeared motions are included to evaluate estimation performance. Motion captured poses contain 10 parts: head, torso, upper arms, lower arms, upper legs, and lower legs, 20 joints each represents a 3D camera free location (x, y, z) , forming a 60d pose vector. During the experiments we use only one color camera information for monocular pose inference, be the case camera 'C1'. Pre-calculation for normalization in image serials is performed on each single human bounding box to overcome scale change in body size.

During experiments, we firstly use the training frames of each action to train parameters of corresponding subjects and tested on validation set. We compared with two types of state-of-the art available image features for human pose estimation: HMAX [12,13] (hierarchical descriptor on MAX operation) and HOG [13,14] (histogram of oriented gradients). Results are shown in S1 sequence in table 2, also 3 different descriptors are validated on the same predictor of TGPKNN in Fig 12. Nearest neighbors are set to WKNN k=15 for small training sets. In our TGP model, we set the nearest neighbors for k =100 for quantity evaluation.

Table 2 Validation results on S1 between different models. HOG descriptors are extracted by R. Poppe's for comparison. Mean relative 3D error in *mm* per joint are given.

Model	*walking*	*jog*	*throw-catch*	*box*	*gesture*
GP+ *SGstmax*	29.55	48.15	**63.16**	26.44	**5.39**
GP+ HOG	38.92	48.37	68.32	49.79	11.38
WKNN+ SGstmax	25.86	48.48	63.83	26.72	6.69
WKNN+ HOG	38.36	46.38	69.79	48.31	13.17
TGP+ SGstmax	**24.66**	46.85	66.20	**24.92**	5.79
TGP+ HOG	28.71	**42.13**	64.19	43.25	10.38

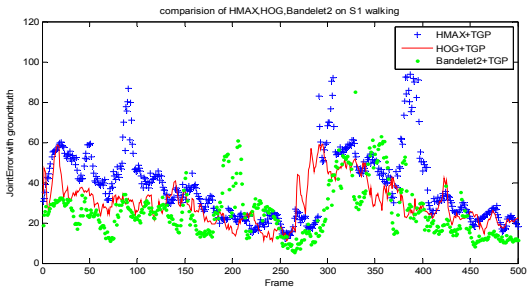


Figure 12 Results of mean joint error on validation part of continuous first 500 frames using different descriptors.

On the validation sets proposed feature shows better capacity than HOG in either model framework. It's easy to find out that in most types of action, SG cooperates better with given learning model. Comparing the three different model pipelines, TGP [13] shows a more promising result than the other two. It can be seen from Table 2 and Fig. 12, with the same pipeline model, our sparse geometric(SG) descriptor outperforms HMAX descriptor and is more competitive than HOG with a lower estimation error in most of the frames. Tough outliers are unavoidable, the maxim error of pose estimation with SG is much lower than that of HMAX. A few outliers in SG may be due to invalid feature extraction of geometric flow. In Table 2, compare the performance with either GP,WKNN or TGP learning model, the proposed SG feature is 5-10 mm lower in quantitive error to HOG feature extracted with Poppe's method of 270 dimensions, where the letter designed the feature blockwise with 5*6 blocks and 9 directions voting. With S1_Jog under TGP pipeline, HOG with 42.13mm

error is better than SG of 46.85mm. However, on other sequences especially walking, the validation error is about 8mm lower, a large improvement. For specific actions, Box and Gestures fit well to the training set when on validation, but generalize normally worse on test sets.

On the test part, we firstly train separately on a small set e.g.: train on S1 training part for walking and test on S1 test walking part. Noted that for the test sets, ground truth information is held back and results can only be gained through an online Humaneva evaluation system [9]. Evaluation results are shown in Table 3. The difference in machine learning model matters much to a successful estimation, but it's not the primary concern in here. The purpose of comparison with various models is to identify the merit of the proposed feature in such tasks. Within most of these pipeline models, the proposed SG feature outperforms its counterparts. Besides basic parameters, the SG-stmax feature is not particularly hand-crafted like HOG and HMAX to fit the human dataset. Indeed, features are best learned through particular dataset [34], so it's entirely possible to further optimize the SG with other recent multiscale analysis progress and improve the discriminative ability of the current version.

Table 3 Humaneva-I online evaluation results for test sequences, with our proposed method on GP, compared on HOG, HMAX models of Bo's work [13]. Both average error and SD in results are given. For space reason, only S1 is shown.

Model	walking	jog	box	gesture	Mean
Stmax	**59.05**	72.41	78.59	**24.93**	58.74
SD	***19.33***	*24.91*	*19.65*	***4.64***	\
HOG	62.1	**64.4**	**78.3**	26.8	57.9
SD	*25.9*	*21.1*	*21.7*	*11.2*	\
HMAX	65.3	69.6	90.8	30.8	64.13
SD	*22.3*	***19.8***	***15.6***	*6.7*	\

Though appearances and poses in the training and testing data have been relatively different, our method could still accurately infer the 3D poses (see table 3 and Fig. 13). Noted that results reported in table 3 of Bo's work are trained on all subjects and we have only trained on one single action sequence for each subject, still results are very competitive. It's clear that if training data is expanded better results would be achieved; we are going to explore more data during the next stage of our work. First row in figure 13 are original images from different subjects on test sets, second row are recovered 3D poses of the corresponding images.

Figure 13 Results on test sets and recovered poses. Different colors represent different limb parts.

In Fig 14 we show the recovered poses from test dataset of S3-Jog with GP model and our proposed feature. Visual results are good except for some miner estimation error in the legs. Pictures and poses are taken with 10 frame interval.

Figure 14 Results on test sets of S3_Jog action, 3d poses are projected to the frames without calibration.

V. Conclusion

In this paper we presented an effective feature as image representation to detect human in natural images, as well as inference high-dimensional, complex human motion poses. It's a novelty of our attempt to utilize sparse geometric representation to construct image features. Utilizing its representative character of geometric regularity from bandelization other than traditional contours or edges, there is no need of background subtraction or silhouette extraction beyond regions of bounding box, which is computationally more relaxed.

Firstly, a low dimensional geometric feature from the essence of sparse geometric is proposed. Features are learned from INRIA pedestrian dataset with multiscale analysis in parameter selection for valid feature identification. Testing results on human detection compared with classical HOG method is promising as we achieved good visual and quantity results with relatively low dimensional features of 512 per frame (image size 128*64) , whereas original HOG is about a 3800 highly dimensional representation. Training time is reduced using the proposed features to a half while being more effective with about 5%. However, the proposed feature is not well devised for

object recognition tasks as it turned out to be less robust to rotation variance as most of the images appear in Caltech 101 dataset. Here, we hope characters from local features and recent multiscale analysis methods would be helpful to blend in the current ones.

And here we also aim to tackle with the problem of 3D human pose estimation from monocular images by using a discriminative approach. Unlike classical image descriptors such as HOG etc. that based heavily on contour or edge information of the human body, which more than often needs background subtraction and cannot strictly depict image context through only edge information, geometric flow enables Bandelet2 features to preserves both interior image information and exterior information that closely related to body shapes. Experiments reveal that Bandelet2 can preserve both invariance as well as selectivity in image. Results of Humaneva-I reported are trained only on each particular action, there is no doubt that by incorporating a large training set of variant actions and poses, the performance will be much better and robust [11,13]. Also, a trend in combining discriminative and generative methods into a pipeline is preferable for more accurate pose estimation [10]. Since our work focus on the image representation and feature part, we will leave this to future work.

In future work we aim to further exploration of sparse geometric descriptors by fusion of other valid sub-features and effective exploration of training data to make better generalization on visual detection and pose estimation.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61075041, 61001206 and 61001206).

We thank Roland Poppe for giving great advice and generous support for making available materials from his work for testing and comparison.

References

[1] A. Agarwal, and B. Triggs. “Recovering 3d human pose from monocular images.” *IEEE transactions on PAMI, Number 1*, Vol.28, Jan 2006, pp. 44-58.

[2] H. Han, M. Tong*, J. Gou, R. Wang, G. Feng, “Discriminative Human Pose Estimation Based on the Bandelet2 Image Descriptor”, Proceedings of the ICIG 2011, pp. 679-684, 2011.

[3] C. Sminchisescu and B. Triggs , “Hyperdynamics importance sampling.” In *ECCV Vol. 1* , Copenhagen ,2002, pp.769–783.

[4] J. Deutscher , A. Blake , and I. Reid, “Articulated body motioncapture by annealed particle filtering.” In *IEEE CVPR’00*, 2000, pp. 126-133.

[5] G. Shakhnarovich, P. Viola, and T. Darrell. “Fast pose estimation with parameter-sensitive hashing.” In *ICCV 03 Vol 2*,Nice, France, October 2003, pp.750-759.

[6] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. “Discriminative Density Propagation for 3D Human Motion Estimation.” In *CVPR 2005*,San Diego,CA, 2005.pp.390-397.

[7] R. Urtasun and T. Darrell. “Sparse probabilistic regression for activity independent human pose inference.” In *CVPR’08*, Anchorage, AK, June 2008.pp.1-8.

[8] C. Sminchisescu, A. Kanaujia, and D. Metaxas. “BM3E: Discriminative density propagation for visual tracking.” In *PAMI Vol 104*, 11,2007, pp.2030-2044.

[9] L. Sigal, A. Blan, and M. J. Black, “HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion.” In *IJCV Vol87*, 2010, pp.1-2

[10] L. Sigal, A. Blan, and M. J. Black, “Combined discriminative and generative articulated pose and non-rigid shape estimation.” In *Advances in NIPS 20th*,Vancouver, Canada, December, 2008. pp. 1337–1344.

[11] L. Bo, C.Sminchisescu, A. Kanaujia, and D. Metaxas, “Fast algorithms for large scale conditional 3D prediction.” In *IEEE conferenceon CVPR’08*, Anchorage, AK, June, 2008, pp.1-8.

[12] A. Kanaujia, C. Sminchisescu, and D. Metaxas. “Semisupervised hierarchical models for 3d human pose reconstruction.” In *CVPR’07*, Minneapolis, MN, June, 2007. pp.1-8.

[13] L. Bo, and C. Sminchisescu. “Twin Gaussian Processes for Structured Prediction.” *In IJCV Vol. 87*, *no. 1-2*, 2010, pp. 28-52.

[14] R. Poppe. “Evaluating Example-based Pose Estimation: Experiments on the HumanEva Sets,” In *CVPR 2nd Workshop on EHuM*, Minneapolis, MN, June 2007.

[15] C. Sminchisescu, A. Kanaujia, and D.Metaxas. “Learning joint top-down and bottom-up processes for 3d visual inferenc.” In *CVPR’06 Vol 2*, New York, June 2006, pp. 1743–1752.

[16] E. L. Pennec and S. Mallat. “Sparse Geometric Image Representations With Bandelets”. In *IEEE transactions on IP Vol. 14 no. 4*, April 2005.

[17] J. Qui.nonero-Candela and C. E. Rasmussen. “A unifying view of sparse approximate gaussian process regression.” In *Journal of Machine Learning Research*, 2005.

[18] D.G. Lowe, Distinctive image features from scale-invariamt keypoints, IJCV, 60(2):91-110, 2004.

[19] S. Belongie, J. Malik, and J. Puzicha, Matching shapes, The 8th *ICCV*, Vancouver, Canada, pages 454-461, 2001.

[20] N. Dalal and B. Triggs, Historgram of oriented gradient for human detection, In *CVPR*, 2005

[21] Bo Wu, Ram Nevtia, Optimizing Discrimination-Efficiency Tradeoff in Integrating Heterogeneous Local Features for Object Detection, CVPR, Pp:1-8, 2008.

[22] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio, Example-Based object Detection in images by Components, IEEE Transactions on Pattern Analysis And Machine Intelligence, 2001(23)4.

[23] G. Peyré and S. Mallat, Surface compression with geometric bandelets, ACM Transactions on Graphics (TOG), 2005, 7, 24(3). Pp: 601- 608.
[24] Pedro F. Felzenswalb, Ross B. Girshick, David McAllester, and Deva Ramanan, object Detection with Discriminatively Trained Part Based Models, IEEE transactions on *PAMI* 2010,pp:1627-1645.
[25] Pedro F. Felzenswalb, Daniel P. Huttenlocher, Pictorial structures for object recognition, *IJCV*, 61(6):55-79, 2005.
[26] P. Dollár, C. Wojek, B. Schiele, Pedestrain Detection:A Benchmark, *CVPR*, Pp: 304-311,2009. [10] B. Russela,A. Torralba, K.P.Murphy, and W.T Freeman, LabelMe: A database and web-based tool for image annotation, *IJCV*, 2008.77(1-3):157-173.
[27] D. Martain, C. Fowlkes, and J. Malik. Learning to detect nature image boundaries using local brightness, color, and texture cues, *PAMI*, 26(5):530-549,2004.
[28] P.F. Felzenszwalb, D. McAllester, D. Ramanan, A Discriminatively Trained, Multiscale, Deformable Part Model, *CVPR*, pp:1-8,2008.
[29]INRIA Static Person Data Set. http://lear.inrialpes.fr/data.
[30] Pedro F. Felzenswalb, Daniel P. Huttenlocher, Efficient matching of pictorial structures, In Proc. Of the *CVPR*, Hilton Head Island, USA, pp.66-75, 2000.
[31] N. Dalal, B. Triggs, and C. Schmid. Human Detection Using Oriented Histograms of Flow and Appearance, Proc. European Conference on Computer Vision, 2006.
[32] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng, Fast Human Detection Using a Cascade of Histograms of Oriented Gradients, Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2006.
[33] C. Harris and M. Stephens. "A combined corner and edge detector". Proceedings of the 4th Alvey Vision Conference. 1988. pp. 147–151
[34] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In BMVC, 2010.

SCIENCE CHINA

Variable Structure Multiple Model for Articulated Human Motion Tracking from Monocular Video Sequences

HAN Hong[1*], TONG MingLei[2], CHEN ZhiChao[1],FAN YouJian[1]

[1]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Institute of Intelligent Information Processing, Xidian University, Xi'an 710071, China,
[2]School of computer and information engineering, Shanghai University of Electric Power, Shanghai 200090, China,

Received: ; accepted:

Abstract：A new model-based human body tracking framework with learned-based theory is introduced in this paper. We introduce a Variable Structure Multiple Model (VSMM) framework to the challenging problem such as uncertainty of motion styles, imprecise detection of feature points and ambiguity of joints location. Key Human joint points are detected automatically and the undetected points are estimated with Kalman filters, multiple motion models are learned from motion capture data with ridge regression method. The model set which covers the total motion set is designed on the basis of topological and compatibility relationship among them, and VSMM algorithm is used to estimate the quaternion vectors of joints rotation. Experiments using real images sequences and simulation videos demonstrate the highly efficiency of our proposed human tracking framework.

Keywords：Human motion tracking, ridge regression, Model groups design, Variable Structure Multiple Model (VSMM), Motion model group adaptation

Citation Han H et al. Variable Structure Multiple Model for Articulated Human Motion Tracking from Monocular Video Sequences.

1 Introduction

Human body motion tracking is an important branch of computer vision, which is of interest in a broad set of applications as visual surveillance, human-computer interface, motion capture, animation, etc. And now it is attracting ever increasing attention. Due to the huge variability of human body shapes, large number of DOF (Degree of Freedom), multitude of local minima solutions, occlusion in images and the ambiguous locations of joint points, human tracking is still a challenging research topic in human motion analysis.

Mikic et al.[1] introduced an integrated system for automatic acquisition of the human body model and motion tracking using input from multiple synchronized video streams. Xu and Li.[2] learned a certain correlation between the left-side and the right-side body motion in specific human activities from modest amount of HumanEva training data by Partial Least Square (PLS), They performed tracking using Rao-Blackwellised particle filter (RBPF) tracking framework. Sigal et al.[3] introduced HumanEva data which contain synchronized and ground-truth 3D motion in detail and gave a baseline algorithm to test the performance of new methods. Their algorithm uses a relatively standard

*Corresponding author (email: hanh@mail.xidian.edu.cn)

Bayesian framework with optimization in the form of Sequential Importance Re-sampling and Annealed Particle Filtering. Tong[4] learned three motion models from motion capture data and recovered 3D human pose by a two-layer filter framework, and the human pose were recovered in the second layer where the interactive multiple model framework was in. Bo [5] proposed a new structured learning methodStructured Output-Associative Regression (SOAR)that models not only the input-dependency but also the self-dependency of outputs, in order to provide an output re-correlation mechanism that complements the (more standard) input-based regressive prediction.Many methods employ the motion model to guide human tracking,though the learned motion model can reduce motion ambiguities and enhance tracking accuracy as well as stability, smaller motion model group can't achieve good pose estimation results due to the complexity and uncertainty of human motion. If only by increasing the types of motion model, computational complexity would increase considerably as the number of model types increases, what's more the efficiency of tracking will be reduced because of excessive competition from the unnecessary models. Therefore we introduce VSMM framework for tracking articulated human motion, with the aim to improve tracking performance , and at the same time without increase of computational complexity.

We propose a model-based human tracking framework with learned-based theory, mainly aiming to deal with multi-modality in action variations of distinguished motion patterns. Here we agree with a basic fact: the set of typical human poses is far smaller than the set of kinematically possible ones [7]. Therefore, due to this fact, we directly learn lots of motion models from motion capture data instead of learning model mappings from image features and capture data with regression methods which is another potential branch namely discriminative approach and then 3D human body pose can be estimated from image sequence with these individually independent yet exchangeable models, which could well alleviate motion ambiguity and self-occlusion etc. Here the image features are human key joint points which can be detected automatically from within each image frame. Since poses can be of different parameters under various models, it is technically not preferable when actions changes a lot while prediction methods remains the same towards different motions，unless the generalization capacity is adequate-which is currently still a problem to be solved. Thus sorts of strategies are progressively proposed to deal with images that covers a wide range of actions and discriminate each single action under a specific condition. An observation Driven GPLVM model is proposed to learn and estimate poses from different parameterized actions/gestures, which exhibit large systematic variation in joint angle space for different instances due to difference in contextual variables [15]. Ryuzo Okada and Stefano Soatto trained on several pose clusters which are subsets of the training samples with similar poses and discriminated the pose clusters by Support Vector Machines (SVM) with pose-dependent feature selection, and then they depend on several correspondent linear regressors to cover poses within the predicted cluster, where each regressor responds for a class of action type[16]. Urtasun and Darrel [17] used local Gaussian Processes to learn poses relating to actions of local cluster, so this manner of dealing with multiple sub-models is also active-dependent and fits well when correctly specified the current action type. In this work, we follow the conceptually similar but a general approach: within the framework of Variable Structure Multiple Model (VSMM), several styles of motion models including walking, jumping and squat etc. are trained with ridge regression. A total human motion model set is established firstly, so the models could be grouped in terms of the physical meanings, topological and compatibility relationship among them[7,8], in each subset actions follow the same pattern and motions are easier to track within the same category. Then the activation and termination of model-groups are performed according to the changes of human body segments projection angle. At last, the present pose can be reconstructed with hybrid estimations of VSMM.

带格式的：字体：（默认）Times New Roman，（中文）宋体，小四

带格式的：字体：（默认）Times New Roman，（中文）宋体，小四

带格式的：字体：（默认）Times New Roman，（中文）宋体，小四

带格式的：字体：（默认）Times New Roman，（中文）宋体，小四

带格式的：字体：（默认）Times New Roman，（中文）宋体，小四

带格式的：字体：（默认）Times New Roman，（中文）宋体，小四

2 Human skeleton model

In this paper, the human skeleton model we employed resembles the one in literature [4] in order to take advantage of kinematic chains framework. The 3D skeletal model is shown in Fig.1 that contains 17 joints and 16 segments. Referring to paper [4], we employ unit quaternion to express the orientation of joint because quaternion can avoid the problems encountered when in the usage of Euler angles, such as Gimbal lock. On the other hand, quaternion can improve the stability of rotation matrix.

In Fig.1, the global coordinate system denoted with T_0 originates at the centroid of the human body, including 3 global translations and a unit quaternion rotation vector, each local coordinate of joints is located at the end of the father segments and contains only a unit quaternion rotation vector Q_i.

3 Human body key joint points location

It's a rather challenging problem that automatic initialize human body key joint point location from uncali-brated monocular image sequences because of self-occlusion and unknown foreground. At present, most of initialization methods are performed by manual start or learned from example. In this paper, we find the key joint points automatically with the method learned from paper [1,5], Fig.2 is the schematic.

We locate the head point first because it is the most stable shape of the human body and tend to avoid occlusion by other human parts. We utilize a center of homocentric circle gliding along every pixel in the skeleton line of human region. The location of the lower neck is found as an average over the silhouette points.

The limbs are easily detected from the skeleton line when there is no occlusion on the human body.The four ends of the limbs are aligned with the end points of silhouette skeleton line.

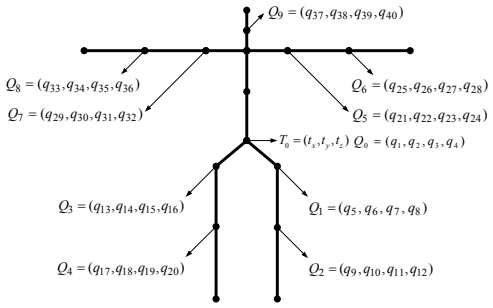


Fig. 1 3D human body skeletal model represent byunitquaternion

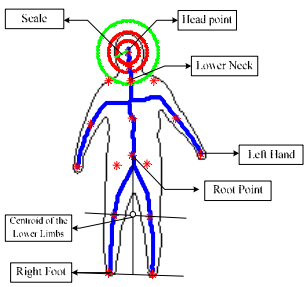


Fig. 2 Key joint points location and Scale metering

The points on the skeleton line which have the minimum distance difference between the distance from one point to the shoulder point and the distance from this point to the corresponding hand point are chosen as Elbow points. And we calculate the average point of lower half body (below centroid) first and draw a line through the point parallel to the connected line between two foot points. The points on the segmented legs that minimize the distance to the line are taken as knee points.

It is noteworthy that alignment may be failed due to the occlusion of limbs, discontinuity of skeleton of image, or isolated point of noise. For instance, if one hand is occluded by torso, it is difficult for us to find a proper point in image as a measurement of hand point on model. Here we perform a Kalman filter to predict the points that are not detected. Detailed process refers to reference[11].

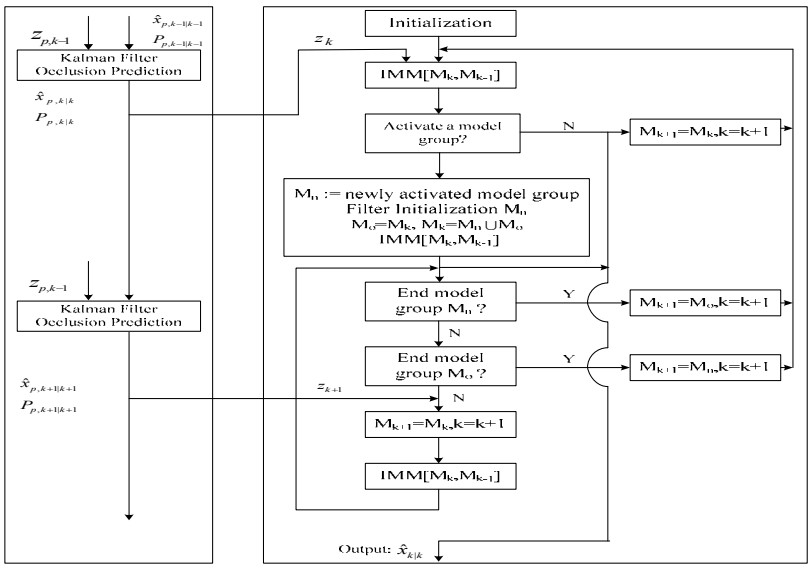


Fig. 3 Human body tracking based on VSMM

During the process of detecting key points, if situations occur when there is background cluster or ~~turbulance~~turbulence in video context, detection will be compromised indirectly. However, the LMedS method adopted in our proposed framework exhibit strong robustness, while Extended Kalman Filter designed to tackle with occlusion problems would also help to smooth images and eliminate aboveeffects.

带格式的：两端对齐，右侧： 1.08 厘米，段落间距段前：10.8 磅，行距：固定值 12.95 磅

4Human body tracking framework based on VSMM

As we know, VSMM framework can solve the incompleteness problem of model set. Working under the multiple hypothesis frameworks, paper [10] proposed three VSMM schemes: active digraph (active model-set), digraph switching (model set switching), and adaptive grid schemes. In this section, we introduce model set switching scheme to human body motion tracking. The VSMM is especially preferable when dealing with digraph structured patterns, and that is why we tend to believe models can be adeptly changed according to our switching rules and well represent changing conditions of human motions as well as internal relationships among each model. The VSMM framework consists of two layers: model set adaptation (MSA) and model set sequence conditioned estimation. Before we perform motion tracking, various motion models should be trained and grouped. The tracking framework based on VSMM is show in Fig.3.

4.1 Motion model learned and model groups design

The EKF filter we used in the VSMM framework is defined by:

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{F}\mathbf{x}(k) + \mathbf{w}(k) \\ \mathbf{z}(k) &= \mathbf{H}(\mathbf{x}(k)) + \mathbf{v}(k) \end{aligned} \tag{1}$$

where $X(k+1)$ and $Z(k)$ represent state vector and measurement vector respectively, $W(k)$ and $V(k)$ are sequences of zero-mean, white Gaussian noise with covariance matrices $Q(k)$ and $R(k)$. The state and measurement vectors are defined by:

$$\mathbf{x}(k) = (t_x, t_y, t_z, \mathbf{Q}_0, \mathbf{Q}_1, \cdots, \mathbf{Q}_8, \mathbf{Q}_9) = (t_x, t_y, t_z, q_1, q_2, \cdots, q_{39}, q_{40}) \quad (2)$$

$$\mathbf{z}(k) = (\mathbf{p}_0^x, \mathbf{p}_1^x, \cdots, \mathbf{p}_{15}^x, \mathbf{p}_{16}^x, \mathbf{p}_0^y, \mathbf{p}_1^y, \cdots, \mathbf{p}_{15}^y, \mathbf{p}_{16}^y) \quad (3)$$

There are 43 parameters(Table 1) in the state vector, including 3 translation parameters and 10 unit quaternion vectors to be estimated. The state transition matrix is learned by ridge regression from CMU motion capture data. In the measurement vector, total 34 parameters of the human model including x and y coordinates of 17 points constitute the Kalman filter measurement in Table 2. And we project human model to an image plane with perspective projection.

Table 1 Variables in state vector

T_0	Global Translation	Q_5	LeftShoulder Rotation
Q_0	Global Rotation	Q_6	Left Elbow Rotation
Q_1	Left Hip Rotation	Q_7	RightShoulder Rotation
Q_2	Left Knee Rotation	Q_8	Right Elbow Rotation
Q_3	Right Hip Rotation	Q_9	Neck Rotation
Q_4	RightKneeRotation		

Table 2 Variables in measurement vector

p_0	Root Point	p_9	Neck Point
p_1	Left Hip Point	p_{10}	Head Point
p_2	Left Knee Point	p_{11}	Left Shoulder point
p_3	Left Foot Point	p_{12}	Left Elbow Point
p_4	Right Hip Point	p_{13}	Left Hand Point
p_5	Right Knee Point	p_{14}	Right Shoulder Point
p_6	Right Foot Point	p_{15}	Right Elbow Point
p_7	Chest Point	p_{16}	Right Hand Point
p_8	Lower Neck Point		

Motion models learning The state transition matrix is learned by ridge regression from CMU motion capture data. For the motion vectors using CMU database are formed by Euler angle, we transform the motion vector to unit quaternion, the formula of ridge regression is defined by:

$$X^i(k+1) = F^i X^i(k) + \varepsilon \tag{4}$$

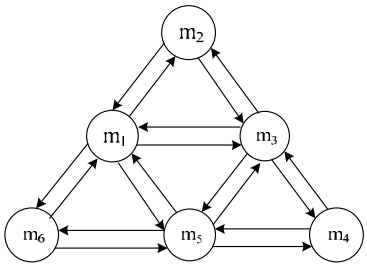
Where i represents the i-th model to be trained; F^i is a 43 x 43 state transition matrix to be trained. To train these models, we use a set of training pairs $\{(X^i(k+1), X^i(k)) \mid k = 1, 2, \cdots, n-1\}$. The matrix F^i can be represented as below:

$$F^i = \arg\min_{F^i} \{\sum_{i=1}^{n} \left\| F^i X^i(k) - X^i(k+1) \right\|^2 + R(F^i)\} \tag{5}$$

The so called ridge regression use $R(\mathbf{F}) \equiv \lambda \|\mathbf{F}\|^2$ instead of $R(\mathbf{F}) \equiv 0$ to reduce the severe over-fitting which produced by simple least squares estimation, because high dimensional regression is an intrinsically ill-conditioned problem. And λ is the normalization factor.Model groups design The basic idea of model set switching is that the model set used for an multiple model (MM) estimator can be made adaptive by switching among a number of predetermined model groups according to certain rule[7]. We designed a set cover of the total model set firstly and the total model set in the following experiment contains 6 motion models, such as walking stiff (m_1), walking (m_2), walking with arm out to keep balance (m_3), jumping jack (m_4), jumping (m_5) and squat (m_6). The topology and transition probability among models are given in Fig. 4 and Table 3.

Table 3 Transition probability among models

带格式的：允许文字在单词中间换行



	m_1	m_2	m_3	m_4	m_5	m_6
m_1	0.86	0.04	0.03	0	0.04	0.03
m_2	0.04	0.92	0.04	0	0	0
m_3	0.03	0.04	0.9	0.03	0	0
m_4	0	0	0.03	0.92	0.05	0
m_5	0.04	0	0	0.05	0.87	0.04
m_6	0.03	0	0	0	0.04	0.93

Fig. 4 Topology of motion model set

Where m_i $(i = 1, \cdots, 6)$ is i-th trained motion model, and the different motion model groups share a common model, as shown in Table 4. Now, we design motion model set cover according to the topology and transition probability of model set, as shown in Table 3.

Table 4 Motion model groups design

1st motion model group	m_1, m_2, m_3
2nd motion model group	m_3, m_4, m_5
3rd motion model group	m_5, m_6, m_1

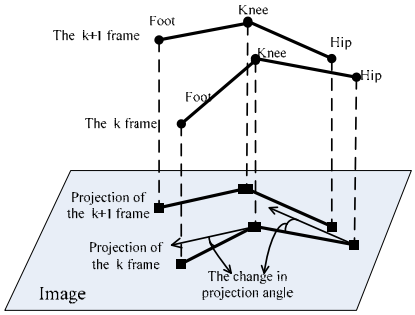


Fig. 5 Change in projection angle of kinematic chain

4.2Motion model group adaptation

The model group currently in effect is called current model group M_k, the others are called candidate model group. The model set adaptation consists of model group activation and termination.

Candidate motion model group activation

The human body model is an articulated 3D rigid body, as shown in Fig. 1, rotation angle of each joint depends on the motion mode currently in effect. We designed the model group activation rules according to the change in projection angle of kinematic chain, as shown in Fig. 5.

In Fig. 5, there is a lower limb and its projection of different frames. from that, we can see that differences of projection angle can reflect transition of motion mode: (1)In 1^{st} model group, the changes in projection angle of limbs are very small, but the contrary is the case in the 2^{nd} model group; (2)In 3^{rd} model group, the change in projection angle of hip equals to that of group 2, but projection positions of upper limbs are discrepancy. In order to eliminate the detection error of single frame, an average projection angle change of 10 frames is used in this paper.

The rules of model groups activation are designed as follows:(1) If the changes in projection angle of hips are twice as much as the previous ones, 3^{rd} model group will be in activation; (2) If the changes in projection angle of the most lower limbs are twice as much as the previous ones, 2^{nd} model group will be in activation; (3) If the changes in projection angle of most lower limbs or the hips are half as much as the previous one, we will activate 1^{st} model group.

Motion model group termination

In order to terminate the candidate or current model group, we consider the probability ratio and likelihood ratio of the model set sequence. The sum of all probabilities and the probabilistically weighted sum of all likelihoods of the motion model in group M_l are defined as:

$$\begin{aligned}\hat{\mu}_k^{M_l} &= P(M_k^l \mid M_k^{n+o}, z^k) = \sum_{m_i \in M_l} \mu(m_k^i \mid M_k^{n+o}, z^k) \\ L_k^{M_l} &= p[z_k \mid M_k^l, z^{k-1}] = \sum_{m_i \in M_l} p[z_k \mid m_k^i, M_k^l, z^{k-1}]\mu(m_k^i \mid M_k^l, z^{k-1})\end{aligned} \tag{6}$$

where $M_k^{n+o} = M_n \cup M_o$. $\mu(m_k^i \mid M_k^{n+o}, z^k)$ is the probability of the model m_i at time k. $p[z_k \mid m_k^i, M_k^l, z^{k-1}]$ is the likelihoods of the model m_i .The value of l is n or o.

The rules of model group termination are designed as:If the condition $\mu_k^{M_n} / \mu_k^{M_o} \prec t_1^p$ or $\prod_{k=k_0}^{k} \left(L_k^{M_n} / L_k^{M_o} \right) \prec t_1^L$ is satisfied, the candidate model group M_n is terminated at the following frames. k_0 is the time at which the model group M_n activated. If both of the conditions $\mu_k^{M_n} / \mu_k^{M_o} \prec t_2^p$ and $\prod_{k=k_0}^{k} \left(L_k^{M_n} / L_k^{M_o} \right) \prec t_2^L$ are satisfied, the current model group M_o is terminated at next frame.

4.3 Algorithm of human body tracking based on VSMM

The detailed solving algorithm is given as follows:

Human Body tracking using VSMM
step 1: Initialization current model group M_1;
step 2: Run a cycle of IMM[M_1];
step 3: Check if a candidate model group is activatedaccording to the activation rules. If no one is activated, output state vector estimation $\hat{\mathbf{x}}_{k\mid k}$ and state error covariance Pk|k ,update model probability set $\{\mu_k^i\}_{m_i \in M_k}$ .Let $M_{k+1} = M_k$, $k = k+1$,and then

带格式的：字体：倾斜

带格式的：字体：倾斜

return to step 2;

step 4: Record newly activated model group as M_n ,and then let $k_0 = k$, $M_0 = M_k$, $M_k = M_n \cup M_o$ .The probability of model mi in group M_n is initialized as $\hat{\mu}(m_k^i \mid M_k^n, z^k) = \arg\max(\hat{\mu}(m_k^j \mid M_k^o, z^k))$, and then normalized $\hat{\mu}(m_k^i \mid M_k^{n+o}, z^k)$; RunIMM[M_1];

step 5: output state vector estimation $\hat{\mathbf{x}}_{k|k}$ and state error covariance $P_{k|k}$,update model probability set $\{\mu_k^i\}_{m_i \in M_k}$;

step 6: For model groups $M_l = M_n$, M_0:

$$\mu_k^{M_l} = \sum_{m_i \in M_l} \mu_k^i$$

$$L_k^{M_l} = \frac{1}{\hat{\mu}_{k|k-1}^{M_l}} \sum_{m_i \in M_l} L_k^i \hat{\mu}_{k|k-1}^i$$

if $\mu_k^{M_n} / \mu_k^{M_o} \prec t_1^p$ or $\prod_{k=k_0}^{k} \left(L_k^{M_n} / L_k^{M_o} \right) \prec t_1^L$ is satisfied,then terminate model group M_n; let $M_{k+1} = M_0$, $k = k+1$; return to step 2;

if $\mu_k^{M_n} / \mu_k^{M_o} \succ t_2^p$ and $\prod_{k=k_0}^{k} \left(L_k^{M_n} / L_k^{M_o} \right) \succ t_2^L$ is satisfied, then terminate model group Mo; let $M_{k+1} = M_n$, $k = k+1$; return to step 2;

step 7: Let $M_{k+1} = M_k$, $k = k+1$; Run IMM[M_1]; Return to step 5.

5 Experiment results and discussion

The framework presented in this paper is evaluated in experiments with video sequences and simulated images. All the programs are coded by Matlab and performed in Window XP with a HP work station. We compare our algorithm with Extended Kalman Filter(EKF) based on Random Walking Model [1] and Interacting Multiple Model (IMM) [4]. EKF was employed in VSMM framework.

5.1Experiments of real video sequences

The experimental video is made with monocular camera with video images size of 320 × 240. The video includes a variety of human motion, such as step(1-120th frame), hand wave & striding step(121st-250th frame), jumping jack & jumping up and down(251st-390th frame),and squat(391st-450th frame) etc. The results are shown in Fig.6, Fig.7 and Fig.8.

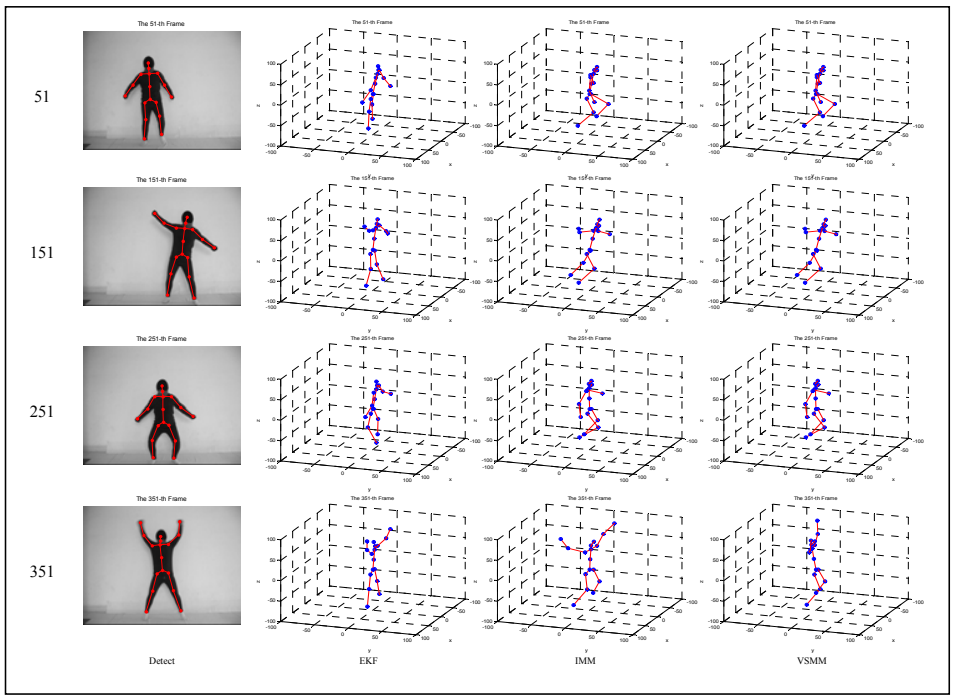
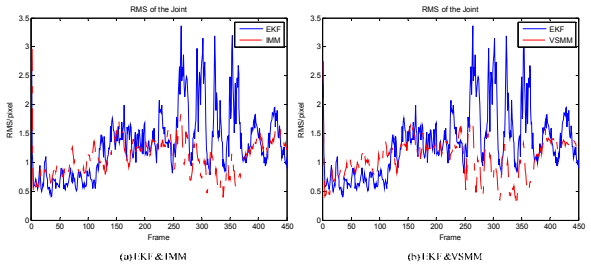
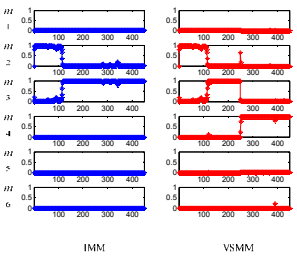


Fig. 6 Tracking a human motion from a sequence of real video images



(a) EKF &IMM (b) EKF &VSMM

Fig.7 Comparison of RMS between the projections of feature points and estimated points



IMM VSMM

Fig.8 The probability of motion models

From Fig.6, detected joints results is shown in the first columns, and 3D results are shown in the second(EKF),third (IMM) and fourth column (VSMM).We can see the result using VSMM is much better than the result using IMM from 3D view, and EKF is the worst. Fig. 7 gives the RMS between the projections of feature points and estimated points. The result is shown that EKF obtains worst RMS. In the Fig. 8, probabilities of motion models in the two frameworks are compared. We can see that there are only two effective models in IMM method while the VSMM method has four models in effect. Model group currently in effect shifts with the motion mode changing in the video. Selected models are more accordant with the actual conditions.

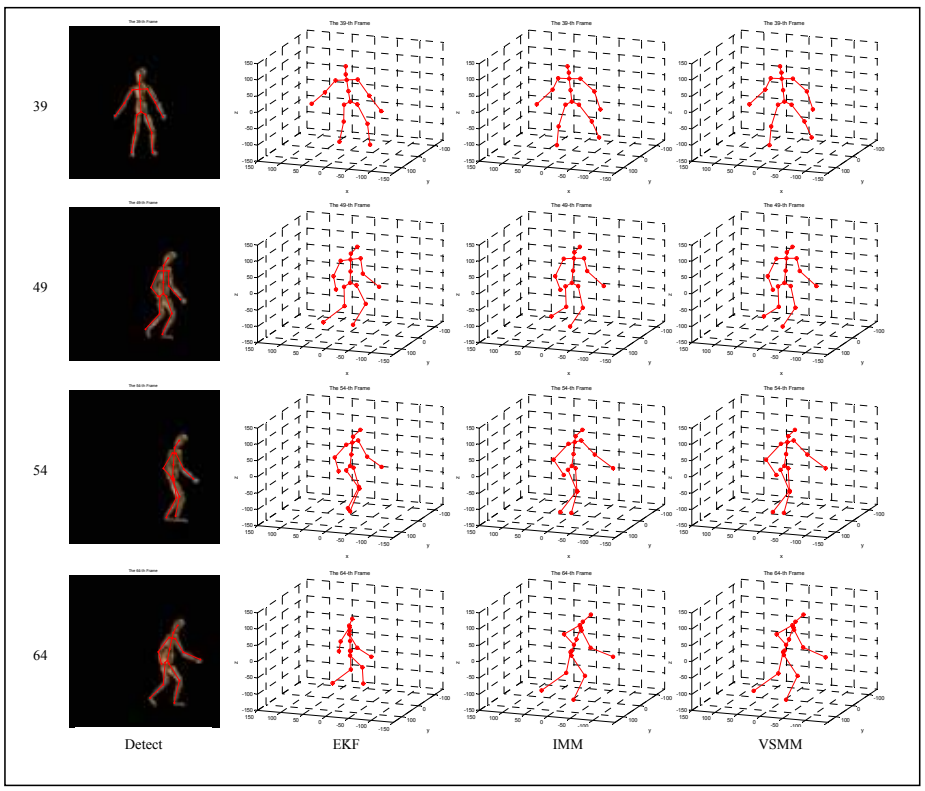
Through qualitativeanalysisin our work of the three methods in computational complexity,EKF only need to establish one filter, IMM algorithm need to establish 6 filters, and VSMM only need to establish 3 filters, therefore, the computational requirement of VSMM is only half of the IMM, andVSMM achieves a far more less average RMS compared to the latter. In addition, in the respect of promotion and generalization capabilty, as the whole model set extends, computational cost of IMM will be further mounted, while VSMM only need to change the topology,and the basic structure of computational cost remain unchanged.

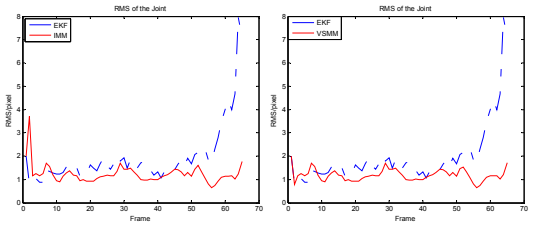
Through qualitative analysis in our work of the three methods in computational complexity, EKF only need to establish one filter, IMM algorithm need to establish 6 filters, and VSMM only need to establish 3 filters, therefore, the computational requirement of VSMM is only half of the IMM, and VSMM achieves a far more less average RMS compared to the latter. In addition, in the respect of promotion and generalization capability, as the whole model set extends, computational cost of IMM will be further mounted, while VSMM only need to change the topology, and the basic structure of computational cost remain unchanged.

VSMM framework processes the test video in 44.5s, and EKF expends 12s, but the IMM method consumes 62.1s. It is clear that bigger model set will increase computational time and degrade the system performance, because the excessive models make the unnecessary competition.

5.2Experiments of simulation video sequences

In this experiment, the simulation video sequences are created with Poser® to render the real walk in circle motion capture data from CMU database. The tracking result using VSMM is given in Fig. 9. Fig. 10 shows the RMS between the projections of feature points and estimated points. The models' probability is shown in Fig. 11.





(a) EKF & IMM (b) EKF & VSMM

Fig.10 Comparison of RMS of simulation video between the projections of feature points and estimated points

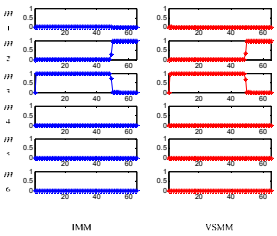


Fig.11 The probability of motion models

From the tracking results of above experiments, it shows that the VSMM framework can obtain good tracking results, and it is not concerned with the variation of depth or rotation of human body. Th+ough the probability of motion models are consistent, the runtime are totally at odds: EFK: 2.3s, IMM: 8.3s, VSMM: 7.6s.The time consumption and tracking results show that the VSMM method is substantially more cost-effective than the IMM.

Tracking results of EKF in the Fig.6 show that human body pose have distortion which does not agree with the human body kinematics rules, i.e. the left crus has bend ahead in 251-th; 3D tracking results of IMM show that the torso lean forward in the 351-th frame. The VSMM framework obtain the best results at all the views, this is because the motion model used in this paper are trained from motion capture data, which let a step predicted results are reasonable so as to reduce motion ambiguity, and the model group switching rules designed in this paper improve the correctness of model switching which has shown in Fig. 8. The model switching rules of IMM don't get the correct jump among different models, and this is caused by unfair competition of unnecessary motion models.

From Fig.9, human turn round from the 48-th frame to the 65-th frame, the RMS of VSMM is better than EKF, this is because random walking model have very great random and can't well match human motion when the movement direction changes.

We can further see from Fig6 ad Fig9 that, image observation of recovered poses do not show clear distortion and are basically congruent with the groundtruth. Form Fig7 and Fig10 we can see, the RMS error is much lower than that of IMM and also EKF, especially when motion patterns of the human subject changes, compared to EKF, both VSMM and IMM would precisely transform the action and do not lead to a sudden leap. From Fig8 and Fig11 it can be seen that form the transition of motion groups, compared with IMM, VSMM responds more instant and precise. When there is a change of motion pattern, VSMM would be most actively engaged into the current motion.

Though better results have been obtain by VSMM framework, sometimes it still unsatisfactory and can't get the best tracking results. We analysis the reason of the error as follows: (1) the motion models trained on the CMU database don't match the motion mode included in the test video. This is because human motion impacted by a variety of factors,

such as speed, step length, terrain variability etc.; (2) the unknown depth and the error of measure are the important factors which affect the tracking accuracy

6 Conclusions

In this paper, VSMM framework is introduced to track 3D human body motion from uncalibrated monocular video sequences. This manner of dealing with multiple sub-models based on model set of potentially possible actions is an active-dependent one, and fits well when correctly judges the current action type. The key joint points, which are taken as observations, are detected automatically. At the same time, the occluded joint points are predicted using Kalman filters, and various human motion models are learned from motion capture data with ridge regression method in order to access informative priors to track poses from even previously not appeared scene. The model set ~~cover of~~that covers the total motion models set is designed on the basis of topological and compatibility relationship among them. This specific design enables timely exchange of estimation information according to the change in projection angle of kinematic chain, model groups transited with the motion modes. We have performed VSMM tracking for human motion from monocular video images and simulation video sequences. From the experimental results, we can see that though working framework appears to be homogeneous, the tracking results using VSMM is better than that of using IMM, this is because the result of VSMM is a probabilistically weighted sum of all filters based on admissible mode set sequences rather than of all filters based on possible mode sequences as in the IMM, thus the unnecessary and redundant image information is filtered out to reduce estimation ambiguity. Also, the tracking results using VSMM is better than that using classical EKF, this is because we use motion models which are trained from motion capture data to enclose prior knowledge which would effectively direct the tracking process. From the experimental results of the proposed algorithm, we came to the conclusion that: If better results of human body motion tracking are aimed to obtain, we must firstly train enough human capture data of human body motion because of the variety and complexity nature of human motion itself. In the future work, we plan to advance the key points detection method, and improve the rules of motion model group activation, thereby we can hope to get more accurate and robust tracking results.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61075041, 61001202and 61001206), the Fundamental Research Funds for the Central Universities (No. K50510020009), the Key Project of Ministry of Education of China (No. 108115), and the National Research Foundation for the Doctoral Program of Higher Education of China (No. 2010020312005).

References

1 MikicI,Trivedi M, Hunter E, et al. Human Body Model Acquisition and Tracking Using Voxel Data, International Journal of Computer Vision(IJCV), 2003,53(3):199–223
2 Xu X Y, Li B X, Learning Motion Correlation for Tracking Articulated Human Body with a Rao-Blackwellised Particle Filter. In: ICCV, 2007, 1–8
~~2~~3] Liang Wang, Tieniu Tan, HuazhongNing, Weiming Hu, “Silhoutte analysis based gait recognition for human identification”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 12, pp. 1505-1518, 2003
~~3~~4SigalL,BalanA,Black M J, HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV, 2010,87(1): 4–27
~~4~~5Tong M L, Three Dimensional Human Motion Analysis from Uncalibrated Monocular Image Sequences. PhD thesis(in Chinese). Shanghai Jiao Tong University, ShangHai, China, 2007
6 Bo L F , Sminchisescu C, Structured Output-Associative Regression.In: Proc. of IEEE

Conf. on Computer Vision and Pattern Recognition,2009,2403–2410
5 Liefeng Bo, and Cristian Sminchisescu. “Twin Gaussian Processes forStructured Prediction.”INIJCV Vol. 87, no. 1-2, 2010, pp. 28-52.
8 GraumanK,ShakhnarovichG,DarrellT,Inferring 3D Structure with a Statistical Image Based Shape Model. In: ICCV , 2003, 1:641–647
6 Lee, C.-S., &Elgammal, A. (2010). Coupled visual and kinematicmanifold models for tracking. International Journal of ComputerVision, 87(1–2).
7 10 Li X R, Zhi X R , Zhang Y M,Multiple-model estimation with variable structure-part III:model-group switching algo-rithm.IEEE Transactions on Aerospace and Electronic Systems , 1999, 35 (1):225–241
8 11 Li X R,Zhang Y M,Zhi X R,Multiple-model estimation with variable structure-part IV: design and evaluation of model-group switching algorithm.IEEE Transactions on Aerospace and Electronic Systems , 1999, 35(1):242–254
12 Han H, Yue L C, Jiao L C, Wu X, Human Body Motion Tracking Based on Quantum-Inspired Immune Cloning Algorithm. In:MIPPR 2009: Pattern Recognition and Computer Vision, Proc. of SPIE, 2009,7496(05), 1-8
13 Sigal, L., Blan, A., & Black, M. J. (2010). HumanEva: Synchronizedvideo and motion capture dataset and baseline algorithmfor evaluation of articulated human motion. International Journalof Computer Vision, 87(1–2).
9 RomeilSandhu and Samuel Dambreville,etc. “A Nonrigid Kernel-Based Framework for 2D-3DPose Estimation and 2D Image Segmentation.”In IEEE transactions on PAMIVol. 33, no. 6,June 2011,pp.1098-1115.
14 Li X R,Bar-Shalom Y, Multiple-model estimation with variable structure.IEEE Transactions on Automatic Control, 1996,4: 478–493
15 Abhinav Gupta, Trista Chen, etc.Context and Observation Driven Latent Variable Model for Human Pose Estimation, CVPR 2008.
16 RyuzoOkada and Stefano Soatto.Relevant Feature Selection for Human Pose Estimationand Localization in Cluttered Images
10 17 Urtasun, R., and Darrell, T. Local probabilistic regression foractivity-independent human pose inference. In IEEE conference on computer vision and pattern recognition (CVPR) 2008.

带格式的: 字体: 10 磅
带格式的: 字体: (默认) Times New Roman
带格式的: 字体: (默认) Times New Roman, 小四
带格式的: 字体: (默认) Times New Roman, 小四, 非加粗
带格式的: 字体: (默认) Times New Roman, 小四, 非加粗
带格式的: 字体: (默认) Times New Roman, 小四
带格式的: 字体: (默认) Times New Roman, 小四, 非倾斜
带格式的: 字体: (默认) Times New Roman, 小四, 非倾斜

中华人民共和国国家知识产权局

[redacted]

[redacted]

[redacted]

此致

[redacted]

[redacted]