Human Body Pose Estimation Based on Histograms of Oriented Gradients and Relevance Vector Machine

Lin Deng[1], Min Jiang[1], J. Tang*[1,2]
[1] College of Computer Science and Technology,
Wuhan University of Science and Technology, 430081, Wuhan, P. R. China,
[2]School of Technology, Michigan Technical University, 1400 Townsend Drive
Houghton, Michigan 49931-1295, USA
{daniel.deng321@gmail.com, jiangminwust@gmail.com, *dadaotang@yahoo.com}

Abstract— **In this paper, a new method for the estimation of 3D human body poses from monocular images is proposed. Histograms of oriented gradients are used as the features for modeling human body poses. Human body poses are represented as 3D limb angles, which can remove the structure information from pose vector. Relevance Vector Machine is used to infer the mapping from image features to body poses. Experiments show that the proposed method is robust to camera views and can lead more accurate results than other pose estimation methods.**

Keyword: pose estimation; Histograms of oriented gradients; 3D limb angles; Relevance Vector Machine

I. INTRODUCTION

Human body pose estimation has important applications in computer vision filed, such as action recognition, motion animation, and human-computer interaction. Due to the variations of appearance, pose, clothes, and light conditions, 3D body pose estimation is still a challenging work. Especially, the estimation of 3D pose estimation from monocular images is more difficult as the influence of self-occlusion than pose estimation under multi-cameras condition

In this paper, we propose a new method to estimate 3D human body pose from monocular images. In the proposed method, image features are represented by the histograms of oriented gradients (HOG) and 3D body poses are represented by 3D limb angles. Relevance Vector Machine is used to infer body poses from the image features.

HOG was originally developed by Navneet Dalal and Bill Triggs[1] as a feature descriptor for head pose estimation[2], body pose detection[1][3] and human body pose estimation[4][5]. HOG is one of feature descriptors developed in the past. Besides HOG, there are also other image feature descriptors, such as shape context, SIFT, Hu moments. Generally speaking, shape context is the most often used descriptor for human body pose estimation [5][9].For example,Ryuzo Okada and Stefano Soatto[5] used shape context to represent image features of a human body and to estimate human body poses under monocular camera. However, shape context based human body pose estimation has many shortcomings. Because shape context is a shape descriptor based on the silhouette of a human body and one silhouette possibly corresponds to several probable poses due to self-occlusion, thus ambiguity will arise when shape context is adopted for human body pose estimation. In order to overcome the shortcomings of shape context based pose estimation, we adopt HOG as the image feature descriptor for pose estimation in the proposed method. HOG is based on statistical information of an image of a human body and it can capture the edge information inside the silhouette of a human body. Based on the past work, it is shown that HOG can obtain higher accuracy than pose estimation based on shape context.

In [4], Ronald Poppe proposed a method using HOG as the image feature descriptor for human pose estimation. There is a difference between his method and ours. The difference lies in the representation of human body poses and the inference algorithm. In Poppe's algorithm, 3D coordinates of human joints are used to represent a human body pose. This representation was also adopted by Ryuzo Okada [8]. The disadvantages of 3D joints based pose representation lie in that the dimension is large and thus increases the computational cost. Instead, in the proposed algorithm, we use 3D limb angles to represent a human body pose which can remove the structure information from pose vector and the computational time is greatly reduced. Besides, Poppe used a simple matching method based on the distance of the training data and testing data to estimate the poses and it only performs well in the multi-cameras condition. However, his method is

978-1-4577-0653-0/11/$26.00 ©2011 IEEE

too simple to handle nonlinear map problem under monocular camera situation. In the proposed algorithm, human pose estimation is treated as a regression problem. In order to solve the regression problem, Relevance vector machine (RVM) [6] is employed. RVM is one of the machine learning methods for non-linear problems. The basic idea of RVM is to convert the nonlinear problem to a linear problem in higher dimension space. Compared with Support vector machine (SVM), RVM needs less relevant vectors than SVM and thus it has been used in many applications including head pose estimation and hand pose estimation [7]. Experiments on HumanEva-I have shown that RVM performs better than simple matching methods. In out experiments, the performance of the proposed method will also be compared with the algorithm proposed by Ankur Agarwal in [5], which used RVM and shape context for pose tracking.

II. Algorithm

A. *Image Feature Extraction*

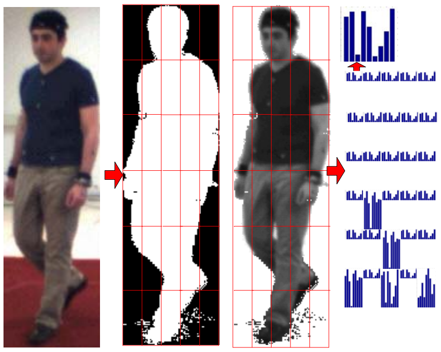
HOG is a feature descriptor based on local gray level information. Two steps will be used in the proposed method to compute the HOG of a human body. We first use background subtraction to obtain the foreground of the image and then we perform the computation of HOG.

Background subtraction: In the proposed method, background subtraction is used to obtain the foreground (human body) in the image. In order to perform background subtraction, a background model is needed. In this paper, we adopt Gaussian mixture model to model the background. Beside, to remove the shadow of human body, the image is transformed to HSV color space because saturation component is not sensitive to shadows. The results of image binaryzation performed on the original RGB color space and HSV color space are combined. Because in our experiments, there is only one active person in the video, so we set the largest connected domain in the combined binary image as the silhouette of target human body.

HOG Computation: After the silhouette is obtained, we will compute the gradient magnitude and the gradient direction angle of the gray image inside the silhouette. The computation of HOG is based on the method of R-HOG [1]. For HOG computation, the silhouette is first divided into 5×6 blocks, and then the magnitudes and the angles of gradient of pixels inside the bounding box of silhouette are computed. In each block, the angles of gradient direction is divided into 9 bins uniformly ranging from -90° to 90°. Each pixel at (x,y) is assigned to a bin based on its' gradient angle $\alpha(x,y)$, which can be represented by the following equation[1].

$$V_k(x,y)=\begin{cases} G(x,y) & \alpha(x,y)\in bin_k,(x,y)\in S \\ 0 & \end{cases} \qquad (1)$$

where $G(x,y)$ is the gradient magnitude of the pixel at location (x,y) and $V_k(x,y)$ represents the gradient magnitude of the pixel at location (x,y) in the kth bin($k\in[1,9]$), S is the pixel set of the silhouette area. Obviously, only the pixels inside the silhouette of a body are involved in HOG computation.



(a) (b) (c) (d)

Fig.1 An Example of HOG Computation

From equation (1), we obtain the sum of gradient magnitude of the kth bin in block C_i using $BV_k^i=\sum_{(x,y)\in C_i} V_k(x,y)$. The feature of an image is a 270 dimension vector H, which is computed by

$$H=\{BV_k^i \mid k\in[1,9], i\in[1,30]\} \qquad (2)$$

where BV_k^i represents the sum of the kth bin in the C_i. Fig. 1 shows the original image, the silhouette of the original image, the division of the silhouette, and the histograms of HOG features.

B. *Body Pose Representation*

Body pose representation is important for pose estimation. The most popular body representation is to use 3D coordinates of the body joint. In this representation, a body pose y, it can be represented as $y=\{p_i \mid i\in[1,N_{joints}]\}$. Here p_i is the 3D coordinate vector of body joint i, and N_{joints} is the total number of body joints. However, this representation has some disadvantages. For example, the dimension of 3D joints based pose representation is too big and it includes body structure parameters, such as limb length. In the proposed method, we adopt a shape-invariant representation. The representation is a set of 3D limb angles, $y = \{\theta_i \mid i \in [1, N_{\lim b}]\}$. Here θ_i is the degree of freedoms of body limb i. This representation removes the body structure information from pose vector and the dimension of pose is reduced. This representation can reduce the computational time greatly. In this representation, the pose configuration is composed of pelvis orientation θ_1 in global coordinates, degrees of freedoms of body limbs $\{\theta_i\}_2^{N_{\lim b}}$ in local coordinates (such as upper arm, lower arm, upper leg , lower leg, and so on.) Global orientation θ_1 can be obtained by the position of hip, pelvis and the orientation of torso. Local degrees of freedoms θ_i from 3D body joint can be obtained as follows by Eq(3) and Eq(5).

Given the affine transformation matrices of head and thorax as M_{head} and M_{throax}, we can get the local transformation matrix of head relative to thorax $M_{headThroax}$ as

$$M_{headThroax} = inv(M_{head}) \times M_{throax} \qquad (3)$$

where

$$M_{headThroax} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & 0 \\ h_{21} & h_{22} & h_{23} & 0 \\ h_{31} & h_{32} & h_{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (4)$$

Let the local degrees of freedoms of the head be $\theta = \{\alpha, \beta, \lambda\}$. Here α, β, γ represent the local angles of head relative to its' parent node thorax in the X-axis, Y-axis and Z-axis, respectively. Then, α, β, γ can be derived by

$$\alpha = \tan^{-1}\frac{h_{23}}{h_{33}},\ \beta = \sin^{-1} h_{13},\ \gamma = \tan^{-1}\frac{h_{12}}{h_{11}} \qquad (5)$$

C. Pose Estimation

Human body pose estimation can be described as follows: given an image HOG feature vector H, $H \in R^d, d = 270$, try to obtain the human body pose configuration vector y ($y \in R^c, c = 28$). The estimation problem can be treated as a nonlinear regression problem to find the nonlinear mapping function ψ, $y = \psi(H)$. In the past, SVM has been used to solve the problem. However, SVM is prone to overfitting. In order avoid overfitting, we adopt RVM to solve the problem. By using RVM, the problem can be described as follows: Given the image feature H, pose vector y can be evaluated by Eq. (6)[6].

$$y = \psi(H) = \sum_{i=1}^{N} w_i K(H, x_i) + w_0 \qquad (6)$$

where $K(\bullet)$ is a kernel function, w is a weighting matrix, $w = \{w_i\}_{i=0}^N$, x_i is the relevance examples in training data. N represents the number of pose vectors.

III. Experiments

A. RVM Setting

In the experiments, we selected HumanEva-I as training and test datasets. 1110 frame image of three videos of three different peoples were chosen as training data, and the remaining 393 frames were chosen test data. The pose vector corresponding to $\{X_i\}_{i=1}^N$ is denoted by $t = \{t_i\}_{i=1}^N$ .In RVM, the choice of the kernel function is very important. For simplicity, we chose Gaussian function as the kernel function. Besides, the performance of RVM is also affected by the width parameter of Gaussian kernel function. For example, small value of width will lead to overfitting. We tested the width in the range from 0 to 300 to find the suitable parameters. The experimental results of using different width parameters are shown in Fig. 2. The error is the Euclidean distance between the estimated pose and true pose.

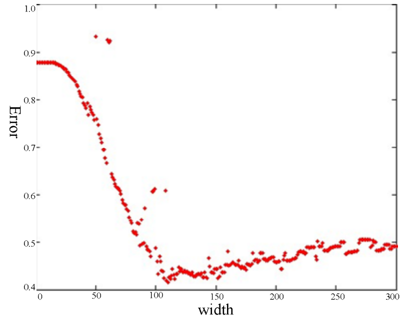


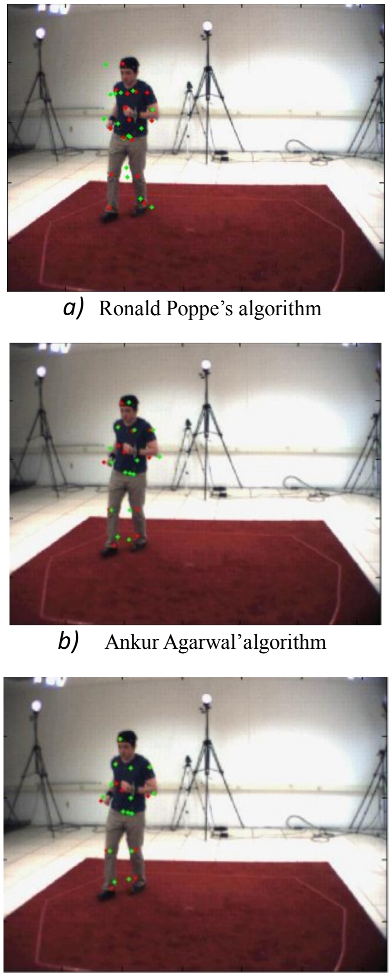
Fig.2 The results of using different width

As shown in Fig.2, the optimal value of width parameters lies in the range of [100,150] and we found when the width is set to be 133, the error is the smallest and thus the width is set to 133 in the experiments. In RVM algorithm, the performance will be greatly influenced by iteration number. The number of iterations affects the quality of the relevance vectors and the accuracy of the estimation. Generally speaking, greater number of iterations can lead to more accurate estimated results. However, the increasing number of iterations could increase the computational cost. In our experiments, we found that 100 is a reasonable number of iterations because the accuracy and the running time are acceptable. At last, we can estimate the mapping function ψ , which $\{X_i\}_{i=1}^N$ and t are the same as equation (6). We can evaluate the weight matrix $w = \{w_i\}_{i=0}^N$ and the relevant vector x_i by the algorithm in [7].

B. Experiment Results

Two other algorithms were compared with the proposed algorithm. The other two algorithms used for comparison were the algorithms developed by Ronald Poppe's[4] and Ankur Agarwal's[5]. The former algorithm used simple template matching and HOG feature for pose estimation. Experiments proven that it only had good performance under environment with multiple cameras or monocular camera with static view. The latter algorithm used shape context feature and RVM for pose estimation. Experiments also showed the performance of Ankur Agarwal's algorithm is good only if the silhouettes derived from background extraction are clear.

However, both of Ronald Poppe's[4] and Ankur Agarwal's algorithms don't work well under environments with monocular camera and arbitrary camera view. The experimental results of the three algorithms are shown in Fig.3, Fig.4 and Table 1. Table1 shows that the comparison results of different algorithms in different camera settings. In our experiments, we used three cameras with different view angles in the HumanEva-I. In Fig 3, red points represent the joints of true pose, green points represent the joints of estimated pose. Fig.3 shows the results of three algorithms with the same video. Because the original algorithm of Ankur Agarwal [5] is for pose tracking, we used a simplified version of his algorithm which omitted the dynamical model for comparison. Experiments show that Ronald Poppe's algorithm is not stable when the camera angles changed, the performance of the proposed algorithm is better than those two algorithms.


a) Ronald Poppe's algorithm

b) Ankur Agarwal'algorithm

c) The proposed method

Fig. 3 the results of human pose estimation with three alogrithms.

TABLE 1 COMPARISON RESULTS OF DIFFERENT METHOD IN DIFFERENT CAMERA SETTINGS

Train camera	**CAM1**		**CAM2**		**CAM3**	
Test camera	***CAM2***	***CAM3***	***CAM1***	***CAM3***	***CAM1***	***CAM2***
Poppe's method	242.96	238.87	239.98	345.20	247.46	347.89
Agarwal's method	100.64	98.11	100.67	102.07	104.76	130.20
Our method	75.86	77.58	68.85	66.88	69.13	66.32

IV. CONCLUSION

In this paper, we proposed a novel method to estimate 3D human body poses from monocular images. The image features are represented by HOG and 3D body pose is represented by 3D limb angles. Relevance Vector Machine was used to infer body pose from image features. Experiments on HumanEva-I show that the proposed method lead more accuracy results than Ronald Poppe and Ankur Agarwal 's methods.

ACKNOWLEDGMENT

Thanks Ronald Poppe for the matlab code of HOG. This work was supported in part by Project (2008TD04) from Science Foundation of Wuhan University of Science and Technology, and the Project (2009CDA034) from Hubei Provincial Natural Science Foundation, P. R. China, and the

Project(61003127) supported by National Natural Science Foundation of China.

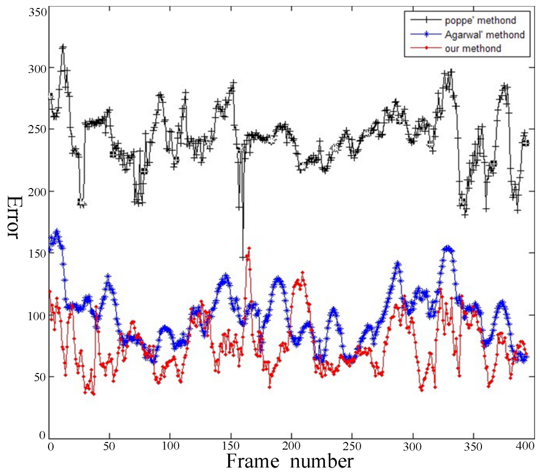


Fig.4 The results of human pose estimation with three algorithms .The red, blue, and black curves are the results obtained by the proposed method, Ankur Agarwal's and Ronald Poppe's methods respectively.

REFERENCES

[1]. Navneet Dalal,Bill Triggs, "Histograms of Oriented Gradients for Human Detection,"IEEE Conference on CVPR, 1:886–893 vol. 1, 2005.

[2]. Elisa Ricci, Jean-Marc Odobez, "Learning Large Margin Likelihoods for Realtime Head Pose Tracking,".IEEE Conference on Image Processing, Cairo, 2009.

[3]. Zhu, Q., Avidan, S., Yeh, M.C., Cheng, K.T. "Fast human detection using a cascade of histogramsof oriented gradients," In: Proc. of CVPR, vol. 2, pp. 1491–1498 (2006).

[4]. Ronald Poppe, "Evaluating example-based pose estimation: experiments on the HumanEva sets,"In: Computer Vision and Pattern Recognition (*CVPR 2007*) workshop on Evaluation of Articulated Human Motion and Pose Estimation (*EHuM2*) (*2007*).

[5]. Agarwal, A., Bill Triggs. "Recovering 3D human pose from monocular images," IEEE Trans.on PAMI 28(1), 44–58 (*2006*).

[6]. Michael. E. Tipping, " Sparse Bayesian learning and the relevance vector machine," J.Machine Learning Research, pages 211{244, 2001}.

[7]. Arasanathan Thayananthan, Ramanan Navaratnam, BjÄorn Stenger, "Multivariate Relevance Vector Machines forTracking," Computer Vision – ECCV, Volume 3953, 124-138, Graz,2006

[8]. Ryuzo Okada,Stefano Soatto, "Relevant Feature Selection for Human Pose Estimation and Localization in Cluttered Images,"Computer Vision – ECCV, Volume 5303, 434-445, Graz,2008.

[9]. Martin Urschler, Joachim Bauer, Hendrik Ditt, and Horst Bischof, "SIFT and Shape Context for Feature-Based Nonlinear Registration of Thoracic CT Images," Computer Vision Approaches to Medical Image Analysis Second International ECCV Workshop, Volume 4241, 73-84,Graz,2006.