3D Human Posture Estimation
Using the HOG Features from Monocular Image

Katsunori ONISHI[†] Tetsuya TAKIGUCHI[‡] Yasuo ARIKI[‡]
[†] *Graduate School of Engineering, Kobe University*
[‡] *Organization of Advanced Science and Technology ,KobeUniversity*
katsu0920@me.cs.scitec.kobe-u.ac.jp {*ariki,takigu*}*@kobe-u.ac.jp*

Abstract

In this paper, we propose a method to estimate the 3D human posture from monocular image without using the markers. A 3D human body is expressed by a multi-joint model, and a set of the joint angles describes a posture. The proposed method estimates the posture using Histograms of Oriented Gradients(HOG) feature vectors that can express the shape of the object in the input image obtained from monocular camera. In addition, the feature dimension of the background region is reduced for reliability by principal component analysis (PCA) computed at every block of HOG. The joint angles in Human multi-joint model are estimated by linear regression analysis applied to its feature vector extracted from the input image. As a result of comparison experiment with the Shape Contexts features, the RMS error was reduced by about 5.35 degrees.

1. Introduction

It is widely studied to estimate the 3D configurations of complex articulated objects accurately from monocular images. The applications are expected in many fields of the human posture and kinematic information such as computer interface with a gesture input, an interaction with the robots, video surveillance and entertainment.

Various methods are dedicated to human posture estimation. There are methods to extract features from images, based on the structure of the human body for example, using skin color or the face position [1]. However, they impose restrictions on features such as clothes and orientation. There are other methods to extract silhouettes and edges from the image as features [2] [3] [4]. However, they rely on the stable extraction of the silhouettes and edges, and also they are weak in self-occlusion. To solve these problems, it is necessary to extract features inside the silhouettes, being independent of skin color or orientation. HOG [5] was originally proposed as the feature to express the shape of the object, but is effective for human posture estimation from the above viewpoint.

In this paper, we propose a method to estimate human posture using HOG feature which can describe the shape of the object as appearance-based approaches. The method doesn't depend on clothes and orientation under noisy condition so that 3D human posture can be estimated stably. However, the dimension of the extracted HOG feature vector is usually high in the background region, because the HOG feature is computed over the entire image. To solve this problem, we also propose a method to reduce the feature dimension in the background regions by PCA at every HOG block. Using the proposed methods, 3D human posture is estimated by linear regression of HOG feature.

2. Features

This section describes the HOG feature extracted from an image and the structure to represent the 3D human model. Moreover, this section describes the method to reduce the dimension of the HOG feature vector in the background region by PCA at every block.

2.1 Histograms of Oriented Gradients

HOG [5] and SIFT [6] were proposed for the gradient based features for general objects recognition. HOG describes similar features to SIFT. The difference is that SIFT describes the feature at the candidate location (keypoint), while HOG describes the feature over the given region. This means that HOG can represent the rough shape of the object.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

2.1.1 Gradient computation

Before extracting the HOG feature, the human region has to be detected by using the background subtraction method on the input image. The image size is normalized at this time, and the human region is located in the central position on the image. Then the image gradient is computed as follows.

$$\begin{cases} f_x(x,y) = I(x+1,y) - I(x-1,y) & \forall x,y \\ f_y(x,y) = I(x,y+1) - I(x,y-1) & \forall x,y \end{cases} \tag{1}$$

where f_x and f_y denote x and y components of the image gradient, respectively. $I(x,y)$ denotes the pixel intensity at the position (x,y). The magnitude $m(x,y)$ and orientation $\theta(x,y)$ are computed by

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \tag{2}$$

$$\theta(x,y) = \tan^{-1}(f_y(x,y)/f_x(x,y)) \tag{3}$$

In order to make the HOG feature insensitive to the clothes and the facial expression, we use the unsigned orientation of the image gradient computed as follows.

$$\tilde{\theta}(x,y) = \begin{cases} \theta(x,y)+\pi, & \text{if} \quad \theta(x,y) < 0 \\ \theta(x,y), & \text{otherwise} \end{cases} \tag{4}$$

2.1.2 Orientation histograms

The gradient image is divided into cells $c_w \times c_h$ pixels as shown in Fig.1. At each cell, the orientation $\tilde{\theta}(x,y)$ is quantized into c_b orientation bins, weighted by its magnitude $m(x,y)$ to make histogram. That is, the histogram with the c_b orientations is computed for each cell.

2.1.3 Block normalization

Fig.1 shows the orientation histogram extracted at every cell and the larger spatial blocks with $b_w \times b_h$ cells. Since a cell has c_b orientations, the feature dimension of each block is $d_b = b_w \times b_h \times c_b$ for each block. Let $\mathbf{v}$ denote a feature vector in a block, h_{ij} denote the unnormalized histogram of the cell in the position $(i,j), \{1 \leq i \leq b_w, 1 \leq j \leq b_h\}$ in a block. The feature vector of a certain block is normalized as follows.

$$h'_{ij} = \frac{h_{ij}}{\sqrt{\| \mathbf{v} \|^2 + \epsilon}} \qquad (\epsilon = 1) \tag{5}$$

Since the normalization is done by overlapping the block, the histograms h_{ij} are repeatedly normalized by the different block.

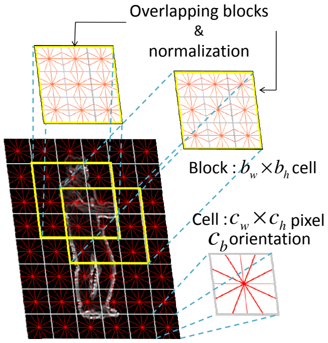

Figure 1. Block normalization.

2.2 Dimension reduction by PCA

HOG feature vector usually has high dimension even in the background region, because the gradients are computed over the entire image. Since the feature is required inside the human region, the feature in the background region should be removed for the 3D human posture estimation.

For this purpose, PCA is carried out at every block using training data. Gray value in the background region is almost constant, although it includes noises, because the background subtraction is already performed as preprocessing. Therefore, a lot of feature dimensions in the background region can be reduced by PCA. On the contrary, the feature in the human region can not be reduced too much because its value changes variously. Therefore, the human region has a lot of feature dimensions and that in the background region is reduced as shown in Fig.2.

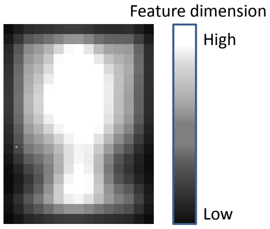

Figure 2. Dimension reduction by PCA.

2.3 Structure of 3D human model

Human is regarded as a multi joint object that transforms into various shapes. However, the segment part which connects two joints can be regarded as rigid. Therefore, it is possible to express a 3D human model by joint angles. That is, in order to express the posture of a 3D human model, the values of joint angles are important.

Let $\mathbf{y}$ denote the vector composed of the angles at joints (elbow, waist, knee, etc.) of the 3D human model. Various postures can be expressed by changing these joint angles. The $\mathbf{y}$ has 24 $(3 \times 6 + 1 \times 6)$ dimensions of the joint angles (except for joints like a finger) as shown in Fig.3.

The various postures are expressed by estimating these joint angles from the input image.

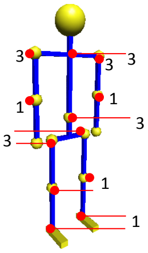

Figure 3. Structure of 3D human model.

3. Regression based approach

This section describes the method to estimate 3D human posture from an image feature. Regression analysis is employed to estimate the posture as used in [2]. The relation between the HOG feature vector $\mathbf{x} \in \mathbb{R}^d$ and 3D human model vector $\mathbf{y} \in \mathbb{R}^m$ is approximated by the following formula.

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \varepsilon \tag{6}$$

Here, $\mathbf{A}$ is $m \times d$ matrix, ε is residual error vector. The 3D human posture is estimated by converting the input image feature $\mathbf{x}$ into the 3D human model vector $\mathbf{y}$. In training a model (estimate A), n sets of training pairs $\{(\mathbf{y}_i, \mathbf{x}_i) \mid i = 1 \cdots n\}$ is given (in our case, 3D poses and the corresponding image HOG features). The conversion matrix $\mathbf{A}$ is estimated by minimizing the least mean square error. Packing the training data into $m \times n$ 3D pose output matrix $\mathbf{Y} \equiv (\mathbf{y_1}\ \mathbf{y_2}\ \cdots\ \mathbf{y_n})$ and $d \times n$ image feature matrix $\mathbf{X} \equiv (\mathbf{x_1}\ \mathbf{x_2}\ \cdots\ \mathbf{x_n})$, the training is performed as follows.

$$\mathbf{A} := \arg\min_{\mathbf{A}} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|^2 \tag{7}$$

In testing, 3D human posture vector $\mathbf{y}$ is estimated by converting HOG feature vector $\mathbf{x}$ using the computed conversion matrix $\mathbf{A}$.

4. Experiment

In this section, we show the result of our proposed method by comparing with the conventional method, shape contexts descriptor [4] extracted from a silhouette.

4.1 Data and ground truth

Images were taken by a monocular camera with the resolution of 640 × 480 pixel. A human body stood up and rotated horizontally. Images were taken from 8 directions by a fixed camera. Five actions (standing, hands up, opening arms, walking, running) were taken continuously in each direction. We manually gave the joint angles to each posture beforehand, and the estimation result was evaluated by RMS error.

As the training data, 30 images were used in each direction, for 5 postures in total. As the test data, 123 images were used; image (a) under the same condition as the training data, image (b) under various conditions, image (c) downloaded from `http://www.nada.kth.se/~hedvig/data.html`. The used images are summarized in Table1. The image size was normalized to 150 × 200. The values of HOG parameters were c_w=10,c_h=10,c_b=9,b_w=3,b_h=3. In computing the HOG feature vector, the block was moved by a cell. PCA was carried out at every block to reduce the 81 dimensions until the 90% cumulative proportion of the HOG feature is achieved.

Table 1. The number of image.

Posture	The number of image			
	Training data	Test data		
		(a)	(b)	(c)
Standing	16	8	8	0
Hands up	40	8	8	0
Opening arms	24	8	8	0
Walking	80	16	16	11
Running	80	16	16	0

4.2 Experimental result

It was confirmed that our method worked effectively for the real image. The result of the comparison experiment is shown in Fig.4. Our method reduces the RMS estimating error by 5.35 degrees compared with the conventional method (shape contexts). Concerning the silhouette images, the limbs were sometimes ambiguous by the self-occlusion. However, in the HOG

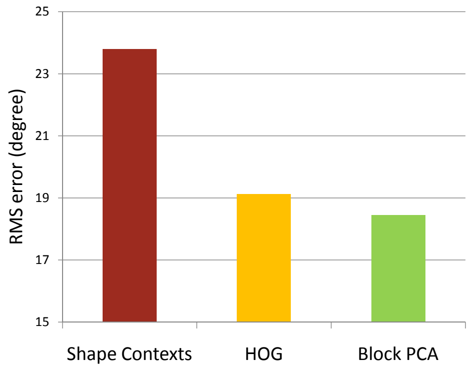

Figure 4. Comparison experiment result.

feature, since it takes the internal edge into consideration, the posture differences can be distinguished so that the error decreased. In addition, HOG after PCA at each block can improve the RMS error by 0.68 degrees compared with the original HOG.

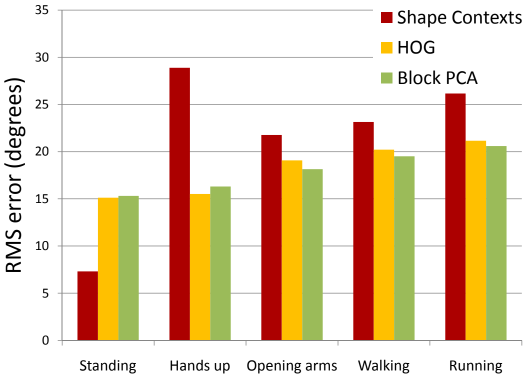

Figure 5. Estimation result of postures.

Next, the evaluation result of postures is shown in Fig.5. The conventional method (shape contexts) showed the small error in the standing posture. This is because noises occurred when the human moved quickly. In a case of stationary posture like standing, there were few noises in an image, so that it was stabilized and the human silhouette was extracted accurately.

However, the purpose is to estimate not the standing but various postures. In this view point, our method can be said to be effective as shown in Fig.4 when considered all postures.

Fig.6 shows examples of the result of 3D posture estimation.

5. Conclusion

We described a method to estimate 3D human posture from a monocular image. In this paper, we proposed to use the HOG feature, which can be extracted without depending on cloths and orientation, and to reduce the feature dimension in the background region by PCA at every block. In future work, without using background subtraction, human detection with HOG feature will be integrated with our method.

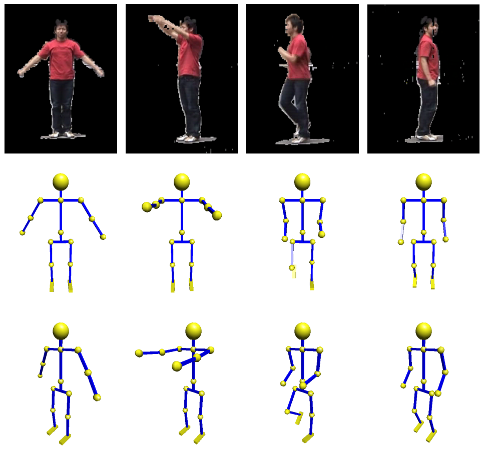

Figure 6. Sample poses reconstructed.

References

[1] M.Lee, I.Cohen. A Model-Based Approach for Estimating Human 3D Poses in Static Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL.28, No.6, 2006.

[2] A.Agarwal and B.Triggs. 3D Human Pose from Silhouettes by Relevance Vector Regression. *IEEE Conference on Computer Vision and Pattern Recognition*, 882–888, 2004.

[3] G.Mori and J.Malik. Recovering 3D Human Body Configurations using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL.28, No.7, 1052-1062, 2006.

[4] G.Mori and J.Malik. Estimating Human Body Configurations using Shape Context Matching. European Conference on Computer Vision, 150-180, 2002.

[5] N.Dalal and B.Triggs. Histograms of Oriented Gradients for Human Detection. *IEEE Computer Vision and Pattern Recognition*, 886–893, 2005.

[6] D.Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60(2), 91-110, 2004.