Advances in Computer Science and Engineering
Volume 3, Number 3, 2009, Pages 175-186
Published Online: December 28, 2009
This paper is available online at http://www.pphmj.com

3D HUMAN POSTURE ESTIMATION BASED ON LINEAR REGRESSION OF HOG FEATURES FROM MONOCULAR IMAGES

KATSUNORI ONISHI, TETSUYA TAKIGUCHI and YASUO ARIKI

Graduate School of Engineering
Kobe University
1-1 Rokkodai, Nada, Kobe, 657-8501, Japan
e-mail: katsu0920@me.cs.scitec.kobe-u.ac.jp

Organization of Advanced Science and Technology
Kobe University
1-1 Rokkodai, Nada, Kobe, 657-8501, Japan
e-mail: {takigu, ariki}@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.

Keywords and phrases: posture estimation, histograms of oriented gradients, principal analysis, linear regression.

Received July 9, 2009

1. Introduction

The accurate estimation of the 3D configurations of complex articulated objects from monocular images [5] has widely been studied. Once the technology is perfected, there will be potential applications in many fields related to human posture and kinematic information, such as computer interfaces with gesture input, interaction with robots, video surveillance, and entertainment. However, this problem is extremely challenging due to the complicated nature of human motion and information limitations associated with 2D images.

Various methods focus on 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 facial position [3]. However, they impose restrictions on features, such as clothes and orientation. There are other methods to extract silhouettes and edges from images as features [1], [6] and [7]. Many methods represent human images using body silhouettes. This representation has the advantage of containing strong cues for posture estimation while being unaffected by changes in appearance and lighting. However, they rely on the stable extraction of the silhouettes and edges, and they are weak in regard to self-occlusion. To solve these problems, it is necessary to extract features inside the silhouettes, being independent of skin color or orientation. HOG [2] was originally proposed as features to express the shape of an object, but it is also effective for human posture estimation from the above viewpoint.

In this paper, we propose an appearance-based approach to estimate human posture using HOG features, which can describe the shape of the object. The method does not depend on clothes and orientation under noisy conditions, so 3D human posture can be estimated stably. However, the dimension of the extracted HOG features vector is usually high in the background region because the HOG features are computed over the entire image. To solve this problem, we also propose a method to reduce feature dimension in the background regions using principal component analysis (PCA) on every HOG block. Using the proposed methods, 3D human posture can be estimated by linear regression of HOG features.

It was confirmed that our method worked effectively for real images, and the experimental results show that our method reduces the RMS estimation error compared to the conventional method (shape contexts).

2. Features

This section describes the HOG features extracted from an image and the structure for representing the 3D human model. Moreover, this section describes the method to reduce the dimension of the HOG features vector in the background region using PCA on every block. Figure 1 shows the flow of HOG features extraction.

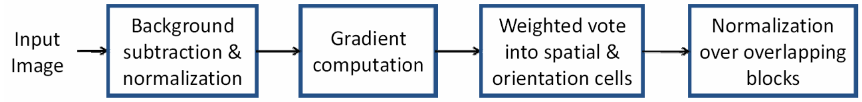
Input Image
Background subtraction & normalization
Gradient computation
Weighted vote into spatial & orientation cells
Normalization over overlapping blocks

Figure 1. The flow of feature extraction.

2.1. Histograms of oriented gradients

HOG [2] and SIFT [4] were proposed for gradient-based features for general object recognition. HOG and SIFT describe similar features. The difference is that SIFT describes the features at the candidate location (keypoint), while HOG describes the features over the given region. This means that HOG can represent the rough shape of the object as shown in Figure 2.

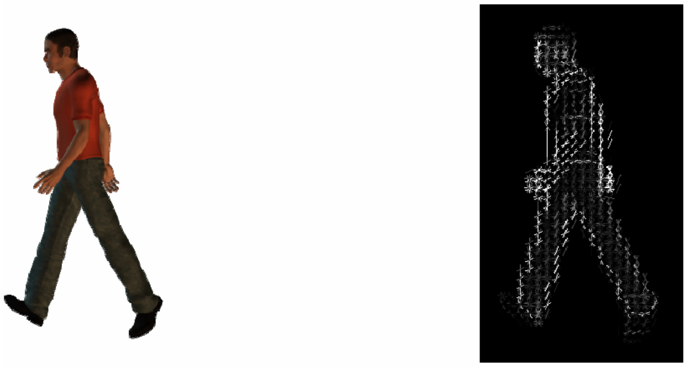

Figure 2. Input image (left) and image represented by HOG features (right).

2.1.1. Gradient computation

Before extracting the HOG features, the human region has to be detected 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 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 features insensitive to the clothes and facial expressions, we use the unsigned orientation of the image gradient computed as follows:

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

2.1.2. Orientation histograms

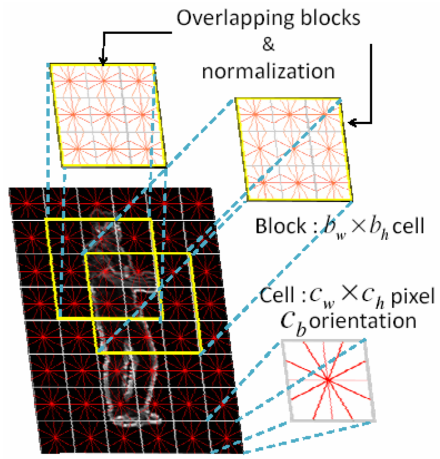

Figure 3. Block normalization.

The gradient image is divided into cells $(c_w \times c_h \text{ pixels})$ as shown in Figure 3. For 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

Figure 3 shows the orientation histogram extracted for 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 **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:

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

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

2.2. Dimension reduction using block-based PCA

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

For this purpose, PCA is carried out for every block using training data. The gray value in the background region is almost constant, although it includes noises, because background subtraction is already performed as preprocessing. Therefore, a lot of feature dimensions in the background region can be reduced by PCA. Conversely, number of features in the human region cannot be reduced too much because their values change in various ways. Therefore, the human region has a lot of feature dimensions, and in the background region is reduced as shown in Figure 4.

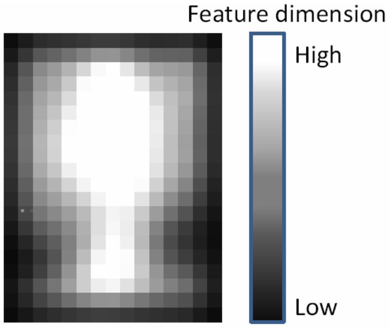

Figure 4. Dimension reduction from block-based PCA.

2.3. Structure of 3D human model

Humans are regarded as multi-joint objects that transform 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 **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 **y** has 24 $(3 \times 6 + 1 \times 6)$ dimensions for the joint angles (except for joints like a finger) as shown in Figure 5.

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

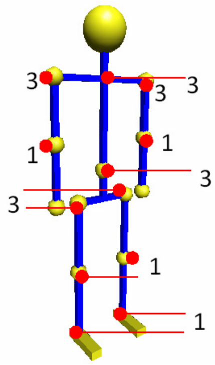

Figure 5. Structure of 3D human model.

3. Regression-based Approach

This section describes the method to estimate 3D human posture from image features. Regression analysis is employed to estimate the posture as used in [1]. 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:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \varepsilon. \qquad (6)$$

A is the $m \times d$ matrix, and ε is the residual error vector. The 3D human posture is estimated by converting the input image feature **x** to the 3D human model vector

y. In training the model (estimate A), n training pairs $\{(\mathbf{y}_i, \mathbf{x}_i) | i = 1 \cdots n\}$ is given (in our case, 3D postures and the corresponding image HOG features). The conversion matrix **A** is estimated by minimizing the least mean square error. Packing the training data into an $m \times n$ 3D posture matrix $\mathbf{Y} \equiv (\mathbf{y}_1 \mathbf{y}_2 \cdots \mathbf{y}_\mathbf{n})$ and $d \times n$ image feature matrix $\mathbf{X} \equiv (\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_\mathbf{n})$, the training is performed as follows:

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

In testing, 3D human posture vector **y** is estimated by converting HOG features vector **x** using the computed conversion matrix **A**. Figure 6 shows the regression-based estimation method.

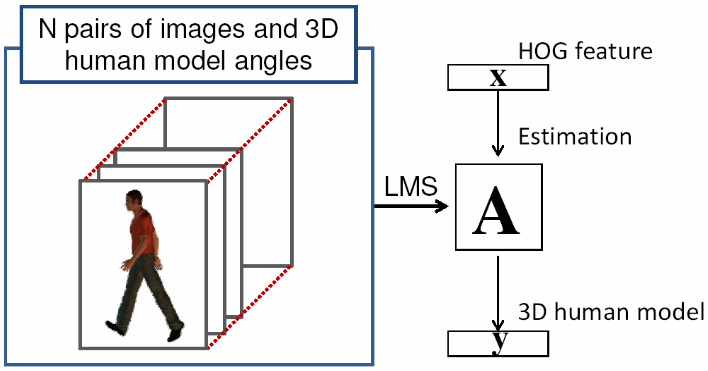

Figure 6. Regression-based estimation method.

4. Experiment

In this section, we show the results of our proposed method by comparison to the conventional method, which utilizes shape contexts descriptors [6] extracted from silhouettes.

4.1. Data and ground truth

Images were taken with a monocular camera with a resolution of 640 × 480 pixels, as shown in Figure 7. A standing human body was rotated horizontally. Images were taken from 8 directions at intervals of 45 degrees using a fixed camera. Five actions (standing, hands raised, arms open, walking, running) were taken continuously in each direction. We manually assigned joint angles to each posture beforehand, and the estimation result was evaluated by RMS error.

Figure 7. A sample image that was taken for the experiment.

For training data, 30 images were used in each direction, for 5 postures in total. For test data, 123 images were used; image (a) under the same condition as the training data, image (b) under various conditions, and image (c) downloaded from http://www.nada.kth.se/hedvig/data.html. The used images are summarized in Table 1.

Table 1. The number of images

Posture	The number of images			
	Training data	Test data		
		(a)	(b)	(c)
Standing	16	8	8	0
Hands raised	40	8	8	0
Arms open	24	8	8	0
Walking	80	16	16	11
Running	80	16	16	0

The image size was normalized to 150×200 using the background subtraction method. 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 features vector, the block was moved cell by cell. Because 234 blocks were made from an image, the dimension of the HOG features was 18,954. PCA was carried out for every block to reduce the 81 dimensions until the 98% cumulative proportion of the HOG features was achieved. The dimension of the HOG features was reduced to 8,998 as a result of computing block-based PCA.

4.2. Experimental result

It was confirmed that our method worked effectively for a real image. The results of the comparison experiment are shown in Figure 8.

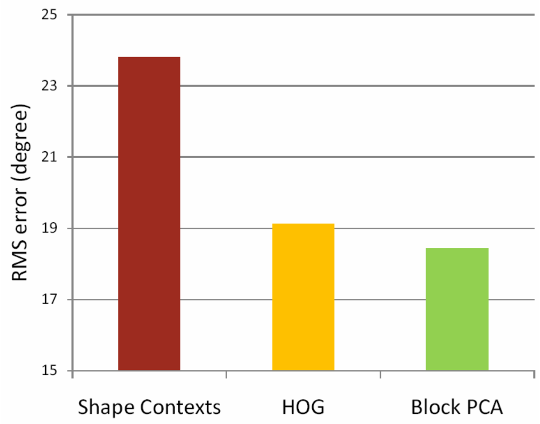

Figure 8. Comparison experimental results.

Our method reduces the RMS estimating error by 5.35 degrees compared to the conventional method (shape contexts). Concerning the silhouette images, the limbs were sometimes ambiguous due to self-occlusion. However, in the HOG features, since it takes the internal edge into consideration, the posture differences can be distinguished so that the error decreased, as shown in Figure 9. In addition, HOG after PCA at each block can improve the RMS error by 0.68 degrees compared to the original HOG.

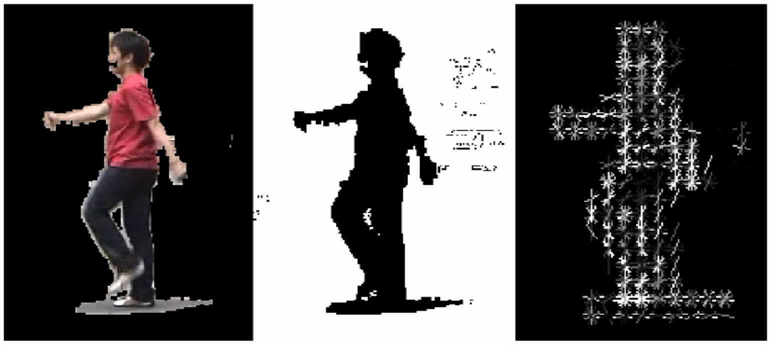

Figure 9. Images walking leftward. The left image is the input image, the middle image is the silhouette image, and the right image is the HOG features image.

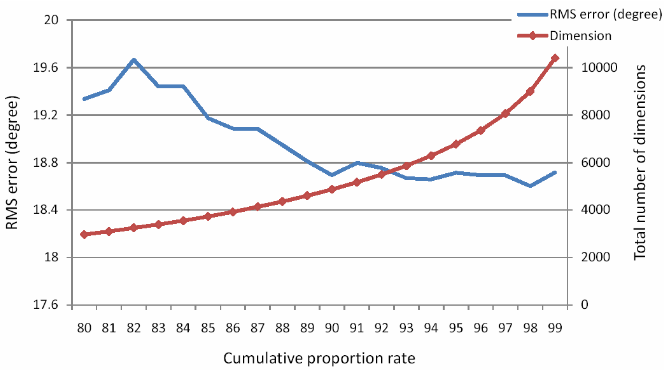

Figure 10. Results of dimension reduction using block-based PCA. The RMS error decreased most when the cumulative proportion rate was 98%.

The dimension reduction of PCA was decided according to the cumulative proportion rate. In Figure 10, the RMS error vs. the cumulative proportion rate (the number of feature dimensions) is plotted. As shown in Figure 10, the optimum cumulative rate (98%) was selected in our experiments. In addition, the block size (3×3) was also selected in our experiments. Figure 11 shows the experiment results, where the block size was changed. Furthermore, we compared PCA and our method, block-based PCA, in Figure 12. It can be confirmed that the result of 3×3 block-based PCA is good.

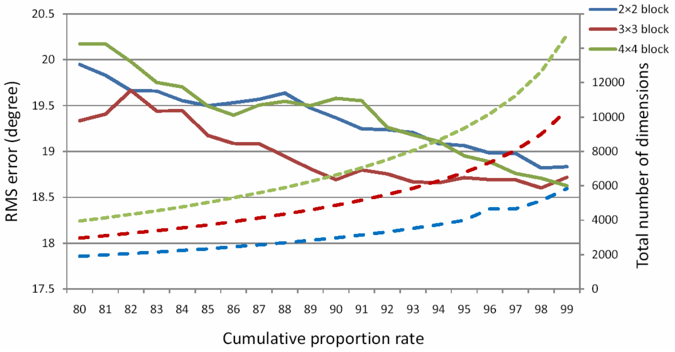

Figure 11. Results of experiments of each block size. Lines denote RMS error, and dash lines denote dimensions. The result of the 3×3 is the best one.

Next, the evaluation results of postures are shown in Figure 13. The conventional method (shape contexts) showed a small error in the standing posture. This is because noises occurred when the human moved quickly. In a case of stationary posture, such as standing, was little noise in an image, so it was stabilized, and the human silhouette was extracted accurately.

However, the purpose is to estimate not only standing but various other postures as well. With this in mind, our method can be said to be effective, as shown in Figure 8 when considered all postures.

Figure 14 shows examples of the results of 3D posture estimation.

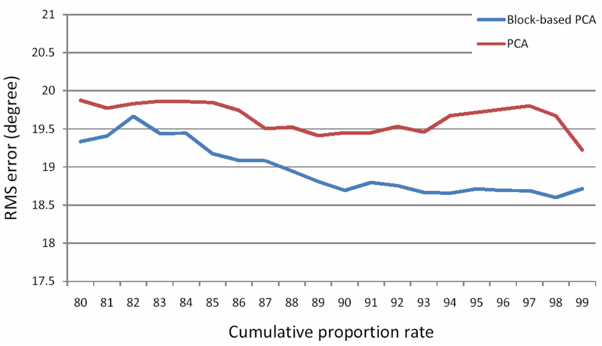

Figure 12. Comparison between PCA and block-based PCA.

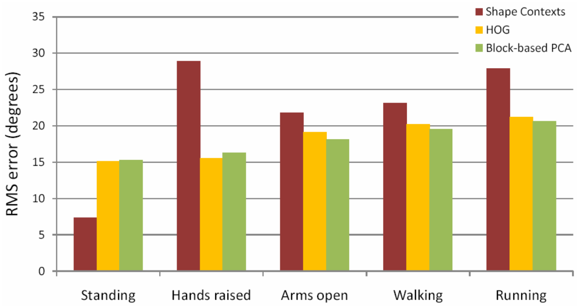

Figure 13. Estimation result of postures.

5. Conclusion

We described a method for estimating 3D human posture from a monocular image. In this paper, we proposed the use of HOG features (which can be extracted without dependence upon clothes and orientation) and reducing the feature dimension in the background region by PCA for every block. In future research, human detection with HOG features will be integrated with our method without using background subtraction.

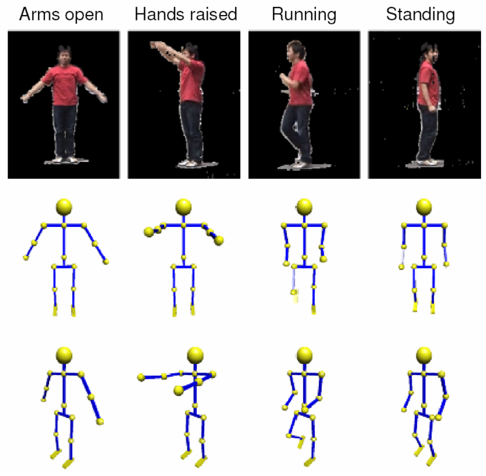

Figure 14. Reconstructed sample postures.

References

[1] A. Agarwal and B. Triggs, 3D human pose from silhouettes by relevance vector regression, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2004), 882-888.

[2] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, IEEE Computer Society International Conference on Computer Vision and Pattern Recognition (2005), 886-893.

[3] M. Lee and I. Cohen, A model-based approach for estimating human 3D poses in static images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28(6) (2006), 905-916.

[4] David G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60(2) (2004), 91-110.

[5] Thomas B. Moeslund and Erik Granum, A survey of computer vision-based human motion capture, International Journal of Computer Vision and Image Understanding 81 (2001), 231-268.

[6] G. Mori and J. Malik, Recovering 3D human body configurations using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 28(7) (2006), 1052-1062.

[7] C. Sminchisescu, A. Kanaujia and D. N. Metaxas, BM^{3}E : discriminative density propagation for visual tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 29(11) (2007), 2030-2044.