Human action recognition with pose similarity

Shiquan Wang, Kaiqi Huang, Tieniu Tan

Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
E-mail:{sqwang, kqhuang, tnt}@nlpr.ia.ac.cn

Abstract: This paper presents a method for representing and recognizing human actions based on pose similarity. For pose representation, we extend Histogram of Oriented Gradients (HOG) with directional statistics to obtain a HOG based descriptor with a smaller dimension. Then a directional similarity measurement for the proposed descriptor is put forward to provide a measure consistent with human perception. To recognize human actions, each testing frame is classified with Nearest Neighbor classifier using the similarity measurement, and each testing sequence of frames is classified with an equal weight voting scheme. Detailed illustration and analysis on HOG with directional statistics are given to show that the proposed descriptor and similarity measurement are reasonable. Experiments on the WEIZMANN dataset demonstrate that with proper similarity measurement, very simple and direct method of human action recognition can achieve desirable performance.

Key Words: Action Recognition, computer vision, pattern recognition, HOG

1 INTRODUCTION

Human motion analysis is of great scientific interest in the field of computer vision. In particular, representation and recognition of human actions plays an important role in a wide range of promising applications, e.g. intelligent video surveillance, human computer interaction and video content retrieval. In this paper we focus on representing and recognizing single person actions.

This work is inspired by recent works in the following two directions: In [1], Schindler and Gool studied the problem of how many frames human action recognition requires. They used both visual cues and dynamic cues. Their conclusion is that "basic action can be recognized well even with very short snippets of 1-7 frames (at frame rate 25 Hertz)"; In [2], Weinland and Boyer worked on human action recognition with pure visual cues and they used a whole sequence to get a classification result. Their conclusion is that it is possible to recognize human actions with pure visual cues except for some special cases like an action and its reversal. Both their works endeavored to find the least information needed for human action recognition and to build a compact and simple human action recognition system. This encouraged us to find if we can achieve two goals at the same time.

To achieve this aim, we need to find a good representation of visual cues. Constraints of using shape/silhouette have been addressed for long time, e.g. it's usually hard to extract satisfying shape/silhouette due to noise, cluttered background. We turn our attention to HOG because recent works have shown that HOG is capable of representing poses well. HOG was initially used in human detection in [3], and is now still one of the most promising feature for human detection [4]. Study on HOG further extends to pose estimation with HOG in both 2D [5] and 3D [6], simultaneous tracking and action recognition with HOG [7], action recognition from pose primitives via HOG [8].

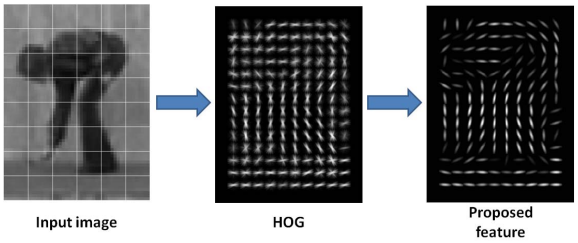

Fig. 1: Illustration of the proposed feature: The proposed feature is consisted of the mean direction and circular standard deviation of each sub-cell.

We notice that HOG is essentially directional data since histogram of oriented gradients is the approximated distribution of directions. This allows us to extend HOG with directional statistics to see what does HOG essentially capture to represent poses. We further propose a directional similarity measurement of HOG to embed poses in a manifold.

The remainder of this paper is organized as follows. In Section 2, some related knowledge of directional statistics is introduced; in Section 3, procedures of how to extract the extended HOG based feature and computation of the similarity measurement are described; in Section 4, analysis on the proposed feature and similarity measurement are given to demonstrate that they are reasonable and consistent with human perception; in Section 5, experimental results and analysis of human action recognition on a publicly available dataset, WEIZMANN dataset, are described; conclusion and future work are given in Section 6.

978-1-4244-7210-9/10/$26.00 ©2010 IEEE

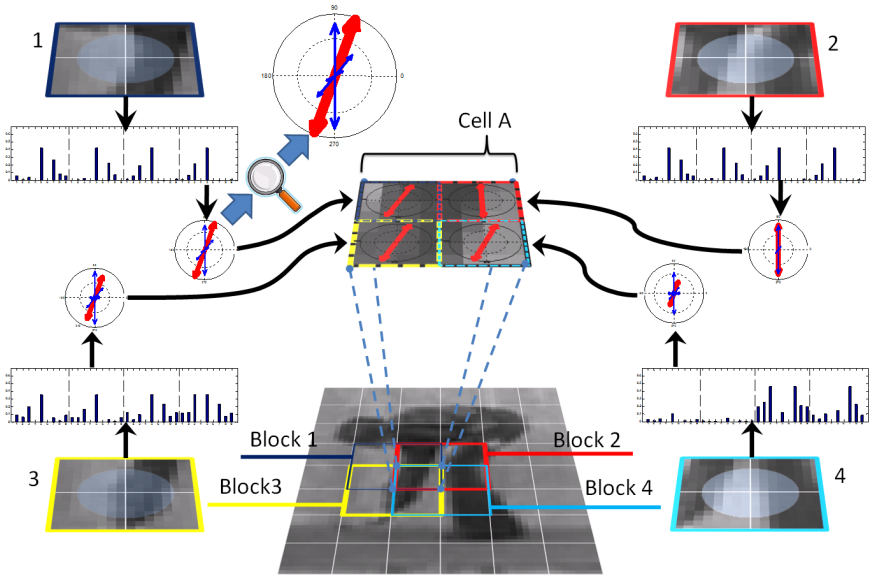

Fig. 2: This figure is best viewed in color. This figure illustrates how to extract the proposed feature from HOG. Details please refer to Section 3.

2 INTRODUCTION TO DIRECTIONAL STATISTICS

According to the definition in wikipedia [9], "directional statistics is the subdiscipline of statistics that deals with directions (unit vectors in R^n), axes (lines through the origin in R^n or rotations in R^n. More generally, directional statistics deals with observations on compact Riemannian manifolds".

The feature of directional data is that there is no specific start point nor end point. This leads to incorrect mean and standard deviation using regular statistics. For example, we get the average of 1° and 359° to be 180° with regular statistics. With directional statistics, rotation invariant statistics can be obtained. We will describe how to compute these statistics later.

We encounter directional data in the field of computer vision quite often, e.g., directions of gradients, hue component in HSV color space and directions of motion. Various applications in computer vision considering directional data are introduced in [10, 11, 12, 13, 14] respectively.

Two basic descriptive statistics in directional statistics are mean direction (MD) and circular standard deviation (CSTD). The mean direction $\overline{\theta}$ and the circular standard deviation σ_0 are calculated with Eq. 3 and Eq. 5 respectively.

$$\overline{C} = \frac{1}{n}\sum_{j=1}^{n}\cos\theta_j \tag{1}$$

$$\overline{S} = \frac{1}{n}\sum_{j=1}^{n}\sin\theta_j \tag{2}$$

$$\overline{\theta} = arc\tan(\overline{S}/\overline{C}) \tag{3}$$

$$\overline{R} = (\overline{C}^2 + \overline{S}^2)^{1/2} \tag{4}$$

$$\sigma_0 = \{-2\log(\overline{R})\}^{1/2} \tag{5}$$

When we apply directional statistics on HOG, two issues should be noted: 1. directions used in HOG is usually in the range of 0 to π; 2. HOG is grouped(binned) data. These issues have been discussed thoroughly in [15], for the first issue, we can simply multiply the unsigned direction by 2 before using directional statistics and divide resultant direction by 2 after; for the second issue, using bin centers as directions and bin value as weight of each direction will give results with ignorable difference as long as the bin width is smaller than 45°.

3 THE FEATURE AND SIMILARITY MEASUREMENT

First we compute HOG with an given image. Please refer to [3] for details about HOG. The parameter setting we choose to compute HOG is given in Table 1. As illustrated in Fig. 2, we can see that four blocks (bounding box colored in dark blue, red, yellow and light blue respectively) share one cell (*Cell A*). Lets look at *Block 1* (with dark blue bounding box), the histogram is built by concatenating histograms of its four cells together and the $4th$ histogram corresponds to *Cell A*. Due to the Gaussian weighting process (illustrated as a lighter part in each block) during histogramming, the histogram of each cell tend to describe the sub-cell near the center of its corresponding block. Thus the $4th$ part of histogram corresponds to the top-left part

Table 1: Parameter setting of HOG computation

Parameters of HOG			
$Height\_win$	64	$Width\_win$	48
$Height\_block$	16	$Width\_block$	16
$Height\_cell$	8	$Width\_cell$	8
num_bin	9	range of $direction$	0 to π

of *Cell A*. It is the same with other blocks so that four histograms correspond to four parts of *Cell A*. As a final result, we have four histograms for each cell (histograms for non-overlapping parts of cells on the boundary are set to be zeros).

Then we compute MD and CSTD of each histogram in HOG to obtain the proposed feature F with Eq. 6 and Eq. 7. Fig. 1 shows an example of computed feature.

$$F_{ij} = (\overline{\theta}_{ij}, \sigma_{0\_ij}) \tag{6}$$

$$F = \begin{bmatrix} F_{11} & F_{12} & ... & F_{1N_2} \\ F_{21} & F_{22} & ... & F_{2N_2} \\ ... & ... & ... & ... \\ F_{N_1 1} & F_{N_1 2} & ... & F_{N_1 N_2} \end{bmatrix} \tag{7}$$

Here ij refers to the ijth sub-cell. N_1 is the number of rows of sub-cells and N_2 is the number of columns of sub-cells. $\overline{\theta}_{ij}$ is MD of ijth sub-cell and σ_{0_ij} is the CSTD of ijth sub-cell.

The directional distance of two proposed features are computed with Eq. 8 and the directional similarity is computed with Eq. 9.

$$dist(F_m, F_n) = \sum_{i}^{N1} \sum_{j}^{N2} [(\pi - abs(\pi - abs(\overline{\theta}_{m\_ij} - \overline{\theta}_{n\_ij}))) + (\sigma_{0m\_ij} > 0)/2 + (\sigma_{0n\_ij} > 0)/2]/(N1 \times N2 \times \pi) \tag{8}$$

$$sim(F_m, F_n) = \exp(-dist(F_m, F_n)^2) \tag{9}$$

Here F_m and F_n refer to proposed features of two images.

4 ANALYSIS ON THE FEATURE AND SIMILARITY MEASUREMENT

We tested this method on Weizmann action dataset [16], which contains 5687 frames of nine people doing ten actions with variant poses. We add an index number for every frame consecutively following the original order of the dataset from 1 to 5687 so that we can refer to a single frame uniquely with its index. After computing HOG with parameter setting in Table 1, we extract the proposed features and calculate directional similarity matrix for each pair of all 5687 frames.

First we look into the nature of MD and CSTD in the proposed feature. Since each histogram of HOG corresponds to a sub-cell which is only $4pixel \times 4pixel$, its inner spatial structure is simple. We can only tell if there is a line in the sub-cell. Figure 3 shows some sub-cells with specific MD and CSTD. We can see that MD correspond to the direction of the line in the sub-cell and CSTD correspond to the existence of the line. Figure 1 shows an example of the proposed feature and its corresponding image and HOG.

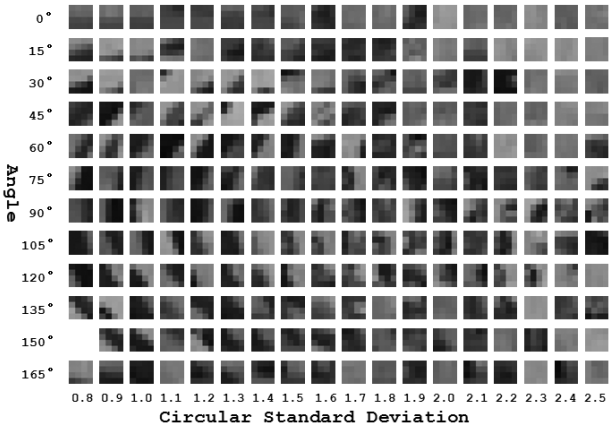

Fig. 3: Sub-cells with specific MD and CSTD: Smaller CSTD refers to smaller dispersion of gradient directions thus more probability that there is a line in the sub-cell. MD corresponds to direction of the line if there is one. There is no sub-cell with MD = 150° and CSTD = 0.8 in the dataset so it is left blank here.

Here we can see that the proposed feature refers to the existence and direction of a line in each sub-cell. It also demonstrates that HOG essentially captures the silhouette to represent poses as shown in 1.

Then we look into the property of the directional similarity. For layout convenience we only show the similarity matrix from frame 1 to frame 84, which belong to the sequence of first person performing bending action. Fig. 4 A gives out some example frames for comparison. Fig. 4 B is the similarity matrix with the directional similarity. Fig. 4 C is the similarity matrix with Euclidean distance from [17]. we can see that the similarity matrix with the directional similarity corresponds to the frames better than the similarity matrix with Euclidean distance, because adjacent frames(the frames are consecutive in a video sequence) look similar thus with smaller distance. The reason of this difference lies under the directional nature of HOG. For example, four histograms of direction:

$$a_1 = [1\ 0\ 0\ 0]; \quad a_2 = [0\ 1\ 0\ 0];$$
$$a_3 = [0\ 0\ 1\ 0]; \quad a_4 = [0\ 0\ 0\ 1];$$

we can tell that a_2 is more similar with a_1 than a_3 with a_1 in the sense of direction, however, with Euclidean distance, the pair-wise distance of four histograms are all 1.

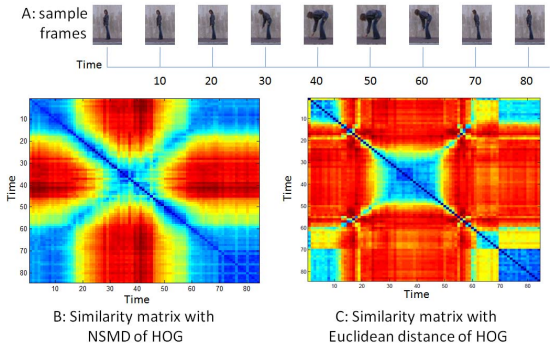

Fig. 4: Similarity matrix of frames 1-84(belong to the video sequence of person 1 doing action bend): This figure is best viewed in color. Adjacent frames are similar to each other, however, similarity matrix of HOG with Euclidean distance is not in accordance with this fact, especially for frame 10-20. Similarity matrix of HOG with the directional similarity corresponds the frames well. Red color refers to higher dissimilarity and blue color refers to lower dissimilarity.

5 EXPERIMENTS ON HUMAN ACTION RECOGNITION

Finally we carried out an experiment of action recognition on Weizmann dataset using the proposed feature and naive Nearest Neighbor search.

The experiment is carried out with following settings: Sequences of n people are used as query models and other sequences are used as test samples. We test all combinations of C_9^n with $n = 1$ to $n = 8$. Results are given on a frame basis that a frame is classified as the same type of its nearest frame in the query model, and on a sequence basis with an equal weight voting scheme from classification results of frames belonging to the sequence.

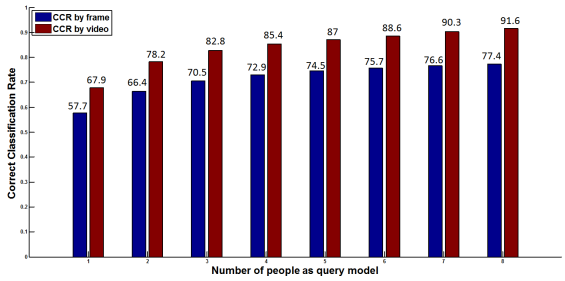

Fig. 5: Overall CCR versus number of people as query model from 1 to 8.

Fig. 5 gives overall results from $n = 1$ to $n = 8$, the highest Correct Classification Rate(CCR) is obtained when $n = 8$ that $CCR = 77.4\%$ on a frame basis and $CCR = 91.6\%$ on a sequence basis. The result is good in the sense that we use purely visual cues in one frame based method. We can achieve comparable performance with fewer model samples. We further present the confusion matrix of $n = 1$ and $n = 8$ in Fig. 6, from which we can see that major false classifications occur among actions of skip, run, and jump since most of their frames look quite similar with each other. Only static appearance based method, like the one we use, is not effective enough for such actions. Incorporating motion information would result better performance. From above results and analysis we can infer that the proposed feature and directional similarity measurement perform fairly well as we expected.

We also conduct an experiment using the same setting as in [2]. The result is $CCR = 82.07\%$ by frame and $CCR = 100\%$ by sequence. But this setting uses one frame in every 20 frames as query model. In this way, frames of every person are used. This is not like regular experiment setting in action recognition so we will not compare this result with others.

We then compare our results on this dataset with other methods. Results on this dataset have been proposed in many papers [7, 8, 16, 17, 18, 19]. We list some methods with classification results both by sequence and by frame for the ease of comparison in Table 2. Performance of our method is comparable with others. Noting that a very simple classifier is employed, this method requires fewer sample data to achieve acceptable results while other methods give result of a leave-one-out cross validation, and our aim is to do human action recognition with very few frames and visual cues only.

Table 2: Comparison of different methods

Methods	By seq.(%)	By frame(%)
This paper	91.6	77.4
Niebles et al. [18]	72.8	55.0
Thurau2007 [19]	86.66	57.45
Thurau2008 [8]	94.40	70.4

6 CONCLUSION

In this paper, we point out that it is important to consider the nature of HOG as directional data and extend HOG with directional statistics and propose a directional similarity for human action recognition. For pose representation, we extend HOG with directional statistics to obtain a HOG based descriptor with a smaller dimension. Then a directional similarity measurement for the proposed descriptor is put forward to provide a measure consistent with human perception. Experiments on WEIZMANN dataset demonstrate that with proper similarity measurement, we achieve the goal of human action recognition with very few frames and visual cues only.

Future work may contain two parts, learning pose primitives with spectral clustering using the similarity matrix and incorporating some directional distribution knowledge in the analysis of HOG and similar features, e.g., SIFT and optical flow.

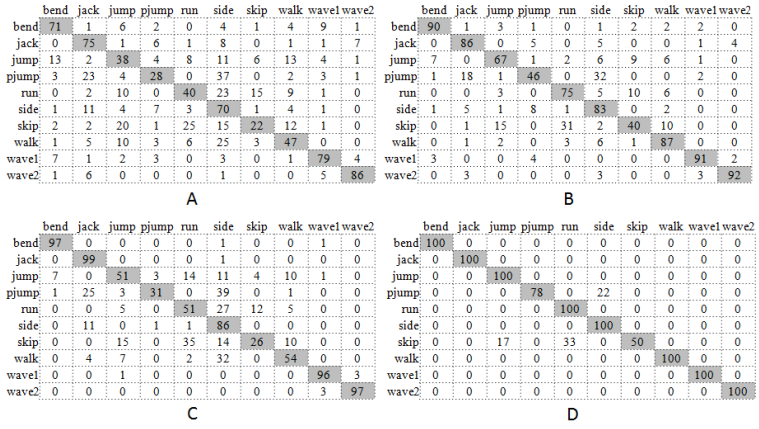

Fig. 6: Confusion matrix of our method: A: $n = 1$ by frame; B: $n = 8$ by frame; C: $n = 1$ by sequence; D: $n = 8$ by sequence. Note that most false classifications occur among action with similar poses like skip, jump and run.

ACKNOWLEDGEMENTS

This work is supported by National Basic Research Program (Grant No. 2004CB318100), National Natural Science Foundation of China (Grant No. 60736018, 60702024, 60723005), National Hi-Tech Research and Development Program of China (2009AA01Z318). The authors thank the anonymous reviewers for their valuable comments.

References

[1] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *Proc. CVPR 2008*, pages 1–8, 2008.

[2] Daniel Weinland and Edmond Boyer. Action recognition using exemplar-based embedding. In *Proc. CVPR 2008*, pages 1–7, 2008.

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR 2005*, volume 1, pages 886–893 vol. 1, 2005.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proc. CVPR 2009*, pages 304–311, 2009.

[5] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P.H.S. Torr. Randomized trees for human pose detection. In *Proc. CVPR 2008*, pages 1–8, 2008.

[6] K. Onishi, T. Takiguchi, and Y. Ariki. 3D human posture estimation using the HOG features from monocular image. In *Proc. ICPR 2008*, pages 1–4, 2008.

[7] Wei-Lwun Lu and J.J. Little. Simultaneous tracking and action recognition using the PCA-HOG descriptor. In *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, page 6, 2006.

[8] Christian Thurau and Valclav Hlavac. Pose primitive based human action recognition in videos or still images. In *Proc. CVPR 2008*, pages 1–8, 2008.

[9] `http://en.wikipedia.org/wiki/Directional_data`.

[10] Din-Chang Tseng, Yao-Fu Li, and Cheng-Tan Tung. Circular histogram thresholding for color image segmentation. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 2, pages 673–676 vol.2, 1995.

[11] Sung-Hyuk Cha and S.N. Srihari. Distance between histograms of angular measurements and its application to handwritten character similarity. In *Proc. ICPR 2000*, volume 2, pages 21–24 vol.2, 2000.

[12] Jinye Peng, Bianzhang Yu, and Dakai Wang. Images similarity detection based on directional gradient angular histogram. In *Proc. ICPR 2002*, volume 1, pages 147–150 vol.1, 2002.

[13] J. Rabin, J. Delon, and Y. Gousseau. Circular earth movers distance for the comparison of local features. In *Proc. ICPR 2008*, pages 1–4, 2008.

[14] Ko Nishino. Directional statistics BRDF model. In *Proc. ICCV 2009*, pages 476–483, 2009.

[15] Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, 2 sub edition, January 1999.

[16] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc. ICCV 2005*, volume 2, pages 1395–1402 Vol. 2, 2005.

[17] Imran Junejo, Emilie Dexter, Ivan Laptev, and Patrick Prez. Cross-View action recognition from temporal self-similarities. In *Proc. ECCV 2008*, pages 293–306. 2008.

[18] J.C. Niebles and Li Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proc. CVPR 2007*, pages 1–8, 2007.

[19] Christian Thurau. Behavior histograms for action recognition and human detection. In *Proceedings of the 2nd conference on Human motion: understanding, modeling, capture and animation*, pages 299–312, Rio de Janeiro, Brazil, 2007. Springer-Verlag.