A COMPACT OPTICAL FLOW BASED MOTION REPRESENTATION FOR REAL-TIME ACTION RECOGNITION IN SURVEILLANCE SCENES

Shiquan Wang, Kaiqi Huang, Tieniu Tan

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{sqwang, kqhuang, tnt}@nlpr.ia.ac.cn

ABSTRACT

We address the problem of action recognition. Our aim is to recognize single person activities in surveillance scenes. To meet the requirements of real scene action recognition, we present a compact motion representation for human activity recognition. With the employment of efficient features extracted from optical flow as the main part, together with global information, our motion representation is compact and discriminative. We also build a novel human action dataset(CASIA) in surveillance scene with three vertically different viewpoints and distant people. Experiments on CASIA dataset and WEIZMANN dataset show that our method can achieve satisfying recognition performance with low computational cost as well as robustness against both horizontal(panning) and vertical(tilting) viewpoint changes.

***Index Terms*—** Surveillance, Pattern classification, Action recognition, Video signal processing, Motion detection, Action retrieval

1. INTRODUCTION

Human motion analysis is of great scientific interest in the field of computer vision. In particular, human activity classification plays an important role in a wide range of promising applications, $e.g.$, intelligent video surveillance and video content index and retrieval. Especially in intelligent video surveillance, recognition of human action allows the system to send an alarm when an action of interest has been recognized such as a person is fainting or to query a specified action in a long video.

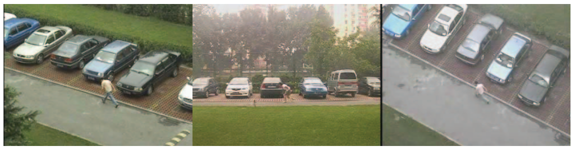


Fig. 1. Example frames from CASIA dataset. $angle$, $horizontal$ and $bird's\ eye$ are shown from left to right. A person is only **30 pixels** tall under $bird's\ eye$ viewpoint.

On the condition of acceptable correct recognition rate, much work has been done in the field of action recognition aiming at issues led by real scene applications. In [1], by modeling a Hidden Markov Model for each viewpoint respectively, categorizing human activity under different viewpoints is carried out. However, both the employment of Hidden Markov Model and the demand of one model for one viewpoint require large training data. Moreover, the viewpoint change is limited to horizontal change(panning). Though the desired view invariance should be against all directions of view changes, for real applications like intelligent video surveillance, there are several issues to be considered. First, since the vertical viewpoint is fixed for a static camera, several models can be built for corresponding vertical viewpoints to accomplish the task of action recognition. Second, though the vertical viewpoint is fixed for a static camera, rotation of target will bring into different horizontal viewpoints, thus the method should be horizontally view invariant. In [2], blurred optical flow is separated into four channels to suppress noise and form a template based method so that action at a distance is recognized. However, action types are limited to walking and running in different directions. Also, optical flow is used in a rather direct way and a best matching scheme is employed. Thus a large dataset is needed and only mid-level decision is available. Alternating optical flow into other forms will strengthen its discriminative power and bringing into classifiers like SVM and AdaBoost will result in better performance and a frame-to-frame recognition output [3]. Though satisfying action classification performance has been achieved in [3], high computational cost needs to be reduced. In [4], high performance in both correct recognition rate and speed is achieved by using an optical flow based volumetric feature. However, issues of distant objects and viewpoint change have not been addressed. In conclusion, a desired action recognition method should be efficient, robust and with instantaneous results.

Most current methods have one or two of the mentioned properties and it is desired to design a method that satisfies all these properties. In this paper, we first build a novel action dataset which posses both diverse action types at a distance and three vertically different viewpoints, then we propose a compact motion representation that can be used to recognize action at a distance and under different viewpoints at real-time. With the rich information contained in optical flow, our method can achieve high recognition performance; with the focus on using direction information of optical flow, our method can be robust; with a set of statistical values extracted from optical flow, our method can form a compact motion representation to be real-time. Fig.1 shows some example frames of CASIA dataset.

The remainder of the paper is organized as follows: In Section 2, we introduce the formation of motion representation. Experimental results and analysis on WEIZMANN dataset and CASIA dataset are presented in Section 3. Finally we draw our conclusion in Section 4.

2. HUMAN MOTION REPRESENTATION

There are many efficient methods to get human blobs. We calculate optical flow within the union of two successive blobs using the Lucas-Kanade method [5]. Then we convert optical flow from the Cartesian coordinates to the Polar coordinates for the following reasons: Motion is more natural for human perception in the form of speed and direction than in the form of horizontal speed and vertical speed; direction is more robust against illumination variations and noise than the other components; moreover, direction pattern is robust to viewpoint changes.

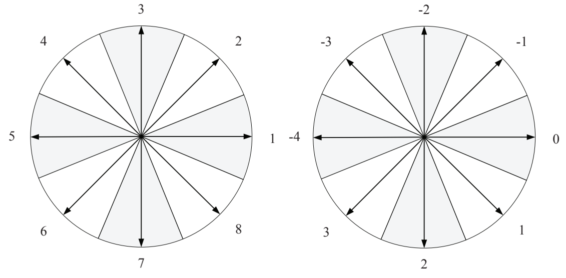
2.1. Motion representation formation

Then we partition optical flow into $N \times N$ blocks so that local to global information will be integrated in our motion representation for better performance. Each block is numbered from $1st$ to $(N \times N)th$ and the whole optical flow of the blob is numbered as the $0th$ block. We then histogram direction of motion pixels in each region into eight bins as shown in Fig.2(a) and normalize the histogram to get $NDi_j, i = 0, ..., N \times N, j = 1, ..., 8$ and $\sum_{j=1}^{8} NDi_j = 1$. Unlike meaningful image patterns whose HOG is usually complex and unique, direction of meaningful motion patterns are usually centralized thus its histogram of direction is simple. So instead of using histogram of direction to represent motion patterns, we further extract the following statistical values, which reflect certain physical meanings, in every block:

Motion Pixel Portion (MPP): The motion pixel indicates pixels belonging to moving objects. MPP represents the active level within the corresponding block and is calculated as

$$MPP_i = num\{MP_i\}/num\{PA_i\} \tag{1}$$

where $num\{MP_i\}$ is the number of motion pixels in ith block and $num\{PA_i\}$ is the number of all pixels in ith block.



(a) Direction histogram bin (b) Relative majority direction

Fig. 2. Direction assignment

Average Speed (AS): Instead of average optical flow we make use of average speed. This is because the average optical flow is not accurate in representing the speed level of a block and is sometimes erroneous. A special case is that the sum of all optical flows within a block is zero. The average speed is calculated as

$$AS_i = \frac{1}{num\{MP_i\}} \sum_{P(u,v) \in MP_i} \rho(u, v) \tag{2}$$

We are concerned with the motion pixels and it is the same when we extract other components of motion representation.

Relative Majority Direction (RMD): Majority Direction is the direction with most motion pixels and is assigned as

$$MD_i = \arg\max_{j=1,...8}\{NDi_j\} \tag{3}$$

where MDi is the majority direction of ith block.

Relative majority direction of each block is assigned relative to the whole optical flow majority direction as shown in Fig.2(b). This feature represents the block motion direction relative to the whole body motion direction and is calculated as

$$\begin{aligned} RMD_i = &\operatorname{mod}(MD_i - MD_0, 8) - \\ &8 \times (\operatorname{mod}(MD_i - MD_0, 8) \geq 4) \end{aligned} \tag{4}$$

Majority Direction Portion (MDP): This feature shows how centralized in direction the motion in the corresponding block is. It is calculated as

$$MDP_i = \max_{j=1,...8}\{NDi_j\} \tag{5}$$

Variance of Direction Distribution (VDD): The direction histogram can be seen as the direction distribution and its variance represents how complex the motion pattern is in the corresponding block. Usually small VDD refers complex motion, which is often caused by noise like swinging grass. VDD can distinguish from meaningful motion of people and meaningless motion of cluttered background. It is calculated as

$$VDD_i = \frac{1}{8}\sum_{j=1}^{8}(NDi_j - \overline{ND_i})^2 \tag{6}$$

Divergence of Direction Distribution (DDD): The divergence of direction distribution is an auxiliary feature for MDP and is calculated as

$$DDD_i = \sum_{j=1}^{8} NDi_j \times RMD\{(j - \arg\max_{l=1,\ldots,8}\{NDi_l\})\}^2 \tag{7}$$

where the $RMD\{x\}$ indicates the same mapping method as mentioned in calculating RMD.

In addition to the above features, we also employ features from shape and trajectory: blob size as $H_k \times W_k$, blob W/H ration as W_k/H_k, acceleration of trajectory in vertical direction as $\nabla^2 Y_k$ to compensate the lack of global information of optical flow.

Then by cascading all values we have the final motion representation of $(5 + 6N^2 + 3)$ dimensions, which is much more compact than optical flow.

3. EXPERIMENTS

In this paper we build a weak classifier with every dimension of our motion representation and use the method in [6] to construct a multi-class AdaBoost classifier. In this way we are able to produce a classification decision for every frame of the video. For a sequence of video frames, we use a voting scheme with equal weight for every frame that the sequence is classified as the category with the most votes from its frames. We test our method on a PC with a Pentium IV 3.0GHz CPU and 1.5 GB RAM at real-time.

3.1. Datasets

We test our method on CASIA dataset and WEIZMANN dataset [7] for discriminative power and robustness.

CASIA dataset contains 710 video clips of 6 single person action types performed by 24 people under 3 horizontally different viewpoints. The videos are $320 \times 240@25fps$. We obtain human blobs through a Gaussian Mixture Model (GMM) background modeling and Nearest Neighbor (NN) tracking in this dataset.

Human silhouettes are given in WEIZMANN dataset. To test our method against blob detection noise, we add Gaussian noise of $\mu = 0$ and $\sigma = 5$ to the width and height of silhouettes' bounding boxes as human blobs.

3.2. Experiments on CASIA dataset

The dataset is divided into training set and testing set half and half. A person is about $50pixels$ tall under $angle$ viewpoint, about $50pixels$ tall under $horizontal$ viewpoint and about $30pixels$ tall under $bird's\ eye$ viewpoint.

We test our method on CASIA dataset with $N = 2, 3, 4, 5$ and the test result is shown in Fig.3. From the result we can see that the results under $angle$ viewpoint and $horizontal$ viewpoint are better. This is reasonable that under $bird's\ eye$ viewpoint people are smaller due to larger distance and many action types look similar due to self-occlusion caused by viewpoint. The CCR(Correct Classification Rate) increases along with N at first, later for some larger N the increase of CCR slows down.

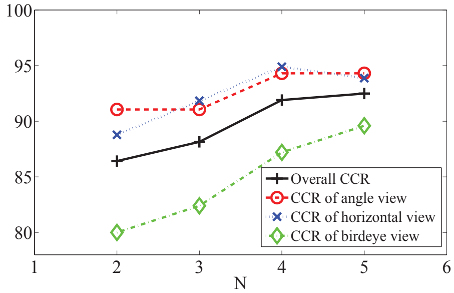


Fig. 3. **Test result on CASIA dataset:** Four sets of test results are given for $N = 2, 3, 4, 5$ respectively. The test results are given by sequences.

Table 1 gives the recognition result of $N = 5$ in detail. We achieve an average CCR of 92.49% by sequence. The test results show that our method can effectively work under the three different viewpoints for the six activities.

angle	bend	crouch	faint	jump	run	walk	CCR
bend	22	0	0	0	0	1	95.65%
crouch	0	24	0	0	0	0	100.00%
faint	0	2	4	0	0	0	66.67%
jump	0	1	0	20	1	1	86.96%
run	0	0	0	0	24	0	100.00%
walk	0	0	0	1	0	22	95.65%

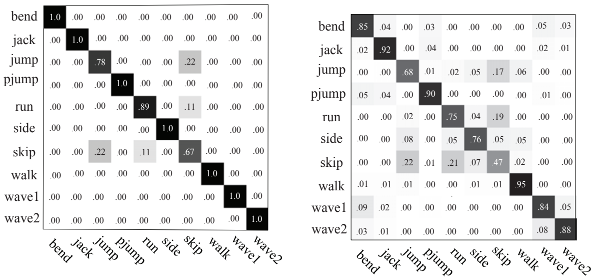
horizontal	bend	crouch	faint	jump	run	walk	CCR
bend	20	0	0	0	0	0	100.00%
crouch	0	19	1	0	0	0	95.00%
faint	0	0	6	0	0	0	100.00%
jump	1	0	0	17	1	1	85.00%
run	0	0	0	2	16	0	88.89%
walk	0	0	0	0	0	14	100.00%

bird's eye	bend	crouch	faint	jump	run	walk	CCR
bend	22	0	0	2	0	0	91.67%
crouch	0	22	2	0	0	0	91.67%
faint	0	1	5	0	0	0	83.33%
jump	0	0	0	19	1	4	79.17%
run	2	0	0	1	20	0	86.96%
walk	0	0	0	0	0	24	100.00%

Table 1. Test result on CASIA dataset when $N = 5$

3.3. Experiments on WEIZMANN dataset

For the sake of comparison, we follow the leave-one-out strategy as others take. We evaluate the performance of our method in frame-by-frame classification as well as video sequence classification. The confusion tables are shown in Fig.4 with $N = 4$. Comparing with Niebles and Li's [8] result of 72.8% by sequence and 55.0% by frame, our method obtained better performance of 93.3% by sequence and 82.37% by frame. Our result is comparable with [9] while our method is computational light and real-time. Notice that confusions are mostly among $jump$, run and $skip$, which is reasonable because they are very similar with each other.



(a) Classification by Sequence

(b) Classification by Frame

Fig. 4. Classification results on WEIZMANN dataset

3.4. Test on horizontal view change

To evaluate the effectiveness of our method under horizontal view change, we conducted a set of robustness tests about horizontal view changes with the dataset for viewpoint robust in [7].

The dataset of horizontal viewpoint changes contains walking sequences with horizontal viewpoint change of $0°$, $5°$, $10°$, $15°$, $20°$, $25°$, $30°$, $40°$ and $45°$ respectively.

We use the classification model built in $Subsection 3.2$ under horizontal viewpoint and the classification model built in $Subsection 3.3$ to classify this dataset, respectively. Table 2 shows classification result when $N = 3$. The result demonstrates our method is robust against horizontal view changes. Model from WEIZMANN dataset achieves better result, which is reasonable since this robustness dataset is part of the WEIZMANN dataset.

Test sequence	CASIA 1st best		WEIZMANN 1st best	
Walking in 0	walk	89.71%	walk	98.53%
Walking in 5	walk	95.31%	walk	100.00%
Walking in 10	walk	98.46%	walk	100.00%
Walking in 15	walk	96.05%	walk	100.00%
Walking in 20	walk	94.81%	walk	100.00%
Walking in 25	walk	87.95%	walk	98.80%
Walking in 30	walk	73.53%	walk	98.53%
Walking in 40	walk	33.64%	walk	62.73%
Walking in 45	bend	48.86%	walk	79.55%

Table 2. Test against horizontal viewpoint change using model of CASIA(Left) and WEIZMANN(Right)

4. CONCLUSION

In this paper, we have presented a compact optical flow based approach to single person activity classification. With human blobs detected, statistical values are extracted from optical flow to form a compact motion representation. The method is real-time with good classification performance; is effective under vertically different viewpoints and robust against horizontal viewpoint changes. We also build a novel dataset concerning distant objects and vertical viewpoint changes.

Acknowledgment

This work is funded by research grants from the National Basic Research Program of China (Grant No. 2004Cb318100), National Natural Science Foundation of China (Grant No. 60736018, 60723005), NLPR 2008NLPRZY-2, Nature Science Foundation of China (60605014, 60875021), National 863 plans projects(2009AA01Z318). We thank for the anonymous reviewers for their hard work and good suggestions.

5. REFERENCES

[1] F. Niu and M. Abdel-Mottaleb, "View-invariant human activity recognition based on shape and motion features," *IEEE Sixth International Symposium on Multimedia Software Engineering, 2004. Proceedings.*, pp. 546–556, 2004.

[2] A.A. Efros, G. Berg, A.C. adn Mori, and J. Malik, "Recognizing action at a distance," *IEEE Conference on Computer Vision, 2003. ICCV 2003.*, vol. 2, pp. 726–733, 2003.

[3] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pp. 1–8, 2008.

[4] Yan Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," *IEEE Conference on Computer Vision, 2005. ICCV 2005.*, vol. 1, pp. 166–173, 2005.

[5] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *International Joint Conference on Artificial Intelligence, 1981. IJCAI '81.*, pp. 674–679, 1981.

[6] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," 1998.

[7] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Conference on Computer Vision, 2005. ICCV 2005.*, vol. 2, pp. 1395–1402, 2005.

[8] Juan Carlos Niebles and Fei-Fei Li, "A hierarchical model of shape and appearance for human action classification," *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR 2007.*, pp. 1–8, 2007.

[9] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?," *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pp. 1–8, 2008.