Robust Feature Extraction for Shift and Direction Invariant Action Recognition

Younghan Jeon, Tushar Sandhan, and Jin Young Choi(✉)

Perception and Intelligence Laboratory, Department of Electrical and Computer Engineering, ASRI, Seoul National University, Seoul, Republic of Korea
{yh1992,tushar,jychoi}@snu.ac.kr

Abstract. We propose a novel feature based on optical flow for action recognition. The feature is quite simple and has much lower computational load than the existing features for action recognition algorithms. It has invariance to scale, different time duration and direction of an action. Since raw optical flow is noisy on the background, several methods for noise reduction are presented. Firstly, we bundle up the fixed number of frames as a block and take the median value of optical flow (median flow). Secondly, we take normalization of histogram depending on the total magnitude. Lastly, we do low-pass filtering in frequency domain. Converting the time domain to frequency domain based on Fourier transform makes the feature invariant to shifted time duration of action. In constructing the histogram of optical flow, we align the direction of an action so that we can get direction invariant action representation. Experiments on benchmark action dataset (KTH) and our own dataset for smart class show that the proposed method gives a good performance comparable to the state-of-the-art approaches and has applicability to actual environments with smart class dataset.

Keywords: Action recognition · Activity recognition · Optical flow histogram · Behavior understanding · Video analysis

1 Introduction

Human action recognition (HAR) is an important research area of computer vision. Since figuring out the purpose of human in video is very important and every data is hard to be analyzed by human, automatic action recognition algorithm is required.

In the last decade, various approaches have been proposed for HAR. Previous works are roughly divided into two categories, appearance based [1–5] and motion based approaches [6–18]. Appearance based method captures the visual information in order to construct features. Main drawback of this model is ignoring the inherent temporal information of the action. In other words, it is hard to reflect sequential information when using the feature based on appearance. For this reason, the current researches mostly utilize a motion based algorithm. In particular, [7, 8] have shown outstanding performance in this area, however they require high computational load. In this paper, we aim to develop a motion based feature requiring low computation as well as giving good performance.

Y.-S. Ho et al. (Eds.): PCM 2015, Part II, LNCS 9315, pp. 321–329, 2015.
DOI: 10.1007/978-3-319-24078-7_32

In the motion based approaches, features should include the information on action dynamics which takes an important role in characterizing the actions. The motion features could be categorized into local and global features. Local features such as silhouette, dense trajectories [9], space-time descriptors [10, 11] only use the information of observed region. Computing the saliency is also needed, which requires much computational time. Also they have weakness for background clutter or illumination changes i.e. seriously affected by noise. Moreover, inherent data or constraints are discarded except salient part. Global features such as motion energy images [12] and motion history images [13] use whole sequence but ignore the temporal information. Moreover these features have different representation for same action with different directions and are hard to capture the action of different time duration. Frequencygram [14] discards the steady frames which lead to loss of data and also ignores the direction of action.

In this paper, we propose a new motion based feature resolving above issues with relatively low computational load. The proposed feature has invariance on scale, direction and time duration. It is based on optical flow and several techniques for reducing the noise are applied. To get direction-invariant action representation and recognition, direction is aligned in constructing the histogram of optical flow by considering the total magnitude of each direction (left bins or right bins) in histogram. In addition, frequency domain analysis is done to show the shift invariance property of the proposed feature through a similar framework with Frequencygram [14].

2 Proposed Method

Our feature is based on optical flow which is a sequence of motion patterns of objects, surfaces and edges in a video. Since various methods for computing optical flow has been developed, we use one of these algorithms. Lots of features and algorithms for action recognition use optical flow as it encapsulates information and dynamics of motion. Since its raw form is hard to be used because of noise from background, regulation strategy for the noise is required. Inspired by Dalal et al. [19], we use histogram approach for reducing the noise so as to obtain the representative features of actions.

2.1 Optical Flow Histogram

Each video divided into N_B blocks and each block consists of fixed number of video frames (N_f). Block analysis could help reducing the noise and give other side effects disturbing the representative feature extraction. Optical flow histogram is constructed in each block as follows.

2.1.1 Median Flow

As mentioned above, raw optical flow is hard to be applied intactly, we first reduce the background noise by combining N_f number of frames as a block. For every frames in each block, optical flow is computed. Then in every pixel the median value of optical flow among the included frames in each axis is picked. Median flow vector $\bar{v} = [u, v]^T$ for each pixel of each block is obtained.

Figure 1 shows the alternation between raw optical flow and median flow.

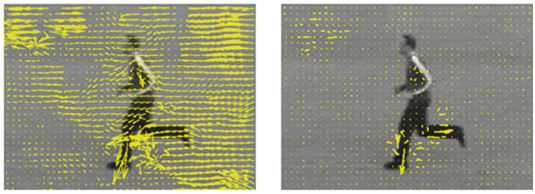

Fig. 1. Comparison between raw optical flow (left) and median flow (right)

2.1.2 Sum of Magnitude in Each Range

Each pixel in the block has median flow vector $\bar{v}$. Initial histograms of blocks are constructed by the following process. Angle from $-\pi$ to π is divided into a number of bins N_{bin}. N_{bin} must be multiple number of 4 in order to be symmetric as shown in Fig. 2. For each range, magnitude of the corresponding median flow vector is summed up to be the weight of each bin as follows:

$$W_i = \sum\nolimits_{\bar{v}\in\Theta_i} \|\bar{v}\|, \tag{1}$$

where,

$$\Theta_i = \left\{ \bar{v} \middle| -\pi + \frac{2\pi\,(i-1)}{N_{bin}} \le \theta\,(\bar{v}) \le -\pi + \frac{2\pi i}{N_{bin}} \right\}, \tag{2}$$

$$\theta\,(\bar{v}) = \frac{(1 - sgn\,(u))\,sgn\,(v)}{2}\pi + \tan^{-1}\frac{v}{u}. \tag{3}$$

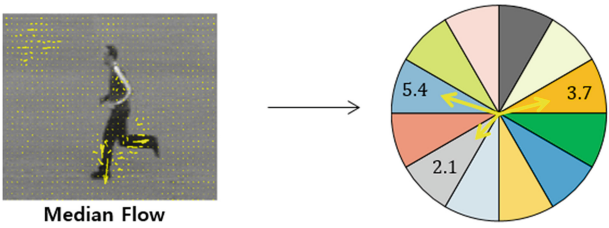

Fig. 2. Weight computation for each bin from median flow

W_i is the weight of each bin for $1 \le i \le N_{bin}$ and $\theta\,(\bar{v})$ is the direction angle of vector $\bar{v}$ from $-\pi$ to π. Using these weights, initial weighted histograms for blocks $H^I = \left\{ h_1^I, h_2^I, \cdots, h_{N_B}^I \right\}$ are constructed.

2.1.3 Aligning Direction

N_B number of initial histograms were generated in the previous steps. However these histograms have different distributions for the same actions with different direction. For

example, stretching left arm to the left side and right arm to the right side have different (but symmetric) histograms. To avoid this issue, direction should be aligned so that we can get the same histograms for the same actions with different directions. Predicting direction of actions could be resolved by comparing total magnitude of each half side in the histogram. Right half is from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ and left half side is the rest. If sum of magnitude in the right half side is bigger than the left, weight of each bin is switched with the symmetric bin. Then we get the aligned histograms for blocks $H^d = \left\{h_1^d, h_2^d, \cdots, h_{N_B}^d\right\}$. Figure 3 shows overview of direction aligning.

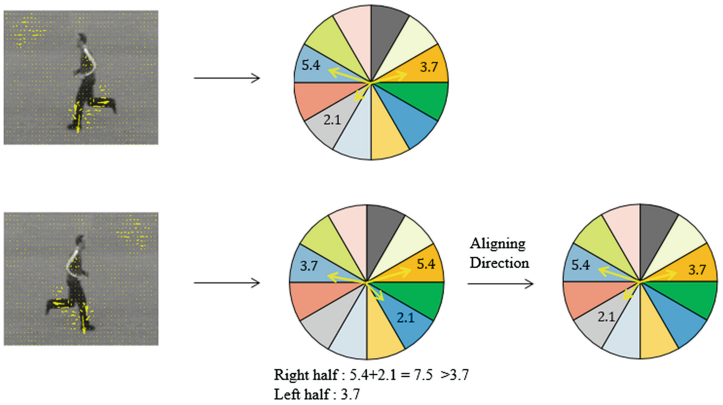

Fig. 3. After aligning, same action with different direction has same distribution

2.2 Histogram Normalization and Concatenation

Normalization of histogram is necessary for scale invariant feature generation since the weight of histogram varies with the person's size in the scene even if the action is same. However, normalizing every histogram to have total weight of 1 could enlarge the noise effect in the actionless block. Introducing a threshold τ in total weight of histogram could help us to sort out this problem. We only normalize the high energy histogram, namely $\|h_i^d\| \geq \tau$, to be $\|h_i^d\| = 1$. That means weights of histograms are divided by their total weight of histogram. The rest low energy histograms ($\|h_i^d\| < \tau$) are not divided by their total weight but by τ to weaker the effect while maintaining the information. Then we get the final histogram $H = \left\{h_1, h_2, \cdots, h_{N_B}\right\}$. Concatenating these histograms as columns, histogram pattern matrix $\mathbf{H} = \left[h_1 h_2 \cdots h_{N_B}\right]$ for each video which have $N_B \times N_{bin}$ dimension is constructed. Figure 4 shows whole processes in this step. Although $\mathbf{H}$ is a representation of an action, it is hard to work as a feature directly since each actor starts action at different time. Also it has a large amount of noise even if somewhat reduced by median. Section 2.3 deals with this problem.

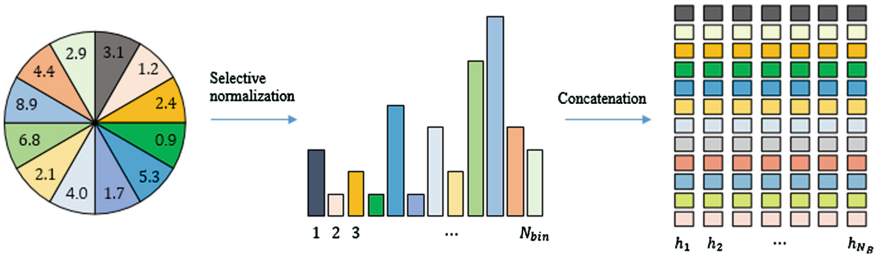

Fig. 4. Histogram normalization and concatenation

2.3 Low-Pass Filtering in Frequency Domain

Handling in frequency domain has two main advantages. First, we can get the shift-invariant property by using only magnitude after Fourier transform. Second, since noise commonly has high frequency, it could be removed by low-pass filtering in the frequency domain. Fourier transform converts the domain from time to frequency.

$$\hat{H}_w(k,l) = \sum_{n=0}^{N_B-1} \sum_{m=0}^{N_{bin}-1} e^{-i(\omega_k n + \omega_l m)} \cdot \mathbf{H}(n,m), \tag{4}$$

$$H_w = \begin{cases} \left|\hat{H}_w(k,l)\right| & if\ k \in \left[-\omega_B, \omega_B\right],\ l \in \left[-\omega_{bin}, \omega_{bin}\right], \\ 0 & otherwise \end{cases} \tag{5}$$

where $\omega_k = \frac{2\pi k}{N_B}$, $\omega_l = \frac{2\pi l}{N_{bin}}$ and ω_B, ω_b are fixed constants determining bandwidth of low-pass filter. H_w is the final feature of action representation.

3 Experimental Results

We evaluated the performance on two datasets, benchmark dataset (KTH, [20]) and our own dataset (Smart class). We used the same experimental setting in each dataset as $N_f = 5, N_{bin} = 32, \tau = 0.1, w_{bin} = 2\pi, w_B = 0.5\pi$. Classification is performed by multi-class SVM [21]. For the optical flow extraction we used the algorithm proposed by Liu. [22]. This could be replaced by any other methods.

3.1 KTH Dataset

Videos in KTH dataset have 6 actions (walking, jogging, running, boxing, hand waving, hand clapping) for 4 scenarios (indoor, outdoor, scale variation, different clothes) with 25 subjects, totally 600 scenes (see Fig. 5).

Fig. 5. KTH dataset

Table 1 shows recognition accuracy and confusion matrix in KTH dataset. We used original split, 16 subjects for training and 9 subjects for testing, suggested in [20]. Our method shows comparable performance with state-of-the-art algorithms in spite of simple process and relatively low computational load.

Table 1. Recognition accuracy and confusion matrix in KTH dataset

Method	Accuracy
Fathi et al. [16]	90.5
Wu et al. [5]	94.5
Laptev et al. [6]	91.8
Kovashka et al. [17]	94.5
Sadanand et al. [8]	98.2
Cai et al. [18]	94.2
Our method	96.3

3.2 Smart Class Dataset

Smart class dataset has 6 actions (standing, sitting, lying face down, stretching waist, hand raising, hand waving) with 30 people and 12 sets each, totally 2160 activity videos and additional multi-camera group scenes with 4 cameras.

3.2.1 Single-View

We applied leave-one-and-out method for measuring the accuracy. Smart class dataset and confusion matrix in single-view is shown in Fig. 6. Average precision is 93.17 % that has main confusion on stretching waist and hand raising. It seems that the feature has similar motion pattern in stretching waist or raising hand due to normalization and discarding geometric information.

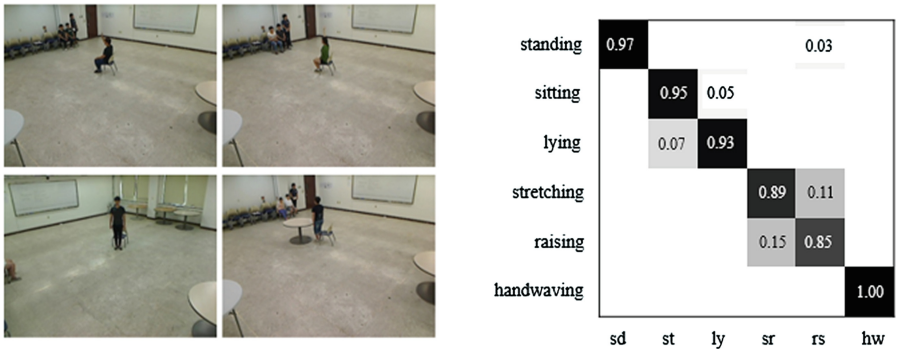

Fig. 6. Smart class dataset in single-view and confusion matrix

3.2.2 Multi-View

Multi-view video has 10 people in the same place and tracking box of each person is given. Figure 7 shows the multi-view group scenes in different views. Each group scene has 4 views and we decide the action by majority vote using SVM score. Specifically, training was performed in single-view videos and we used multi-view scene as testing. First we judged the action of each people in each view, then we had 4 decisions with SVM score for one action. After ignoring the decision which had lower SVM score than threshold, we conducted majority vote and got the final decision of action. Average precision of multi-view scene is 87.04 % which shows discriminative performance despite severe occlusions with objects or other person.

Fig. 7. Multi-view group scenes in different views

4 Conclusion

In this paper, we proposed a novel human action recognition algorithm by extracting a new feature based on optical flow. The proposed feature has property of direction invar-

iance and shift invariance, i.e. it has the same output for the same action with different direction or time duration. It is also resilient to scale variation. In the process, several strategies for generating action specified representation by reducing background noise of optical flow were also presented. Median flow, normalization of histogram and low-pass filtering in frequency domain are comprised in the proposed method. Experimental results show the efficiency of the proposed algorithm and applicability to real-life environment.

Acknowledgement. This work was supported by the Brain Korea 21 Plus Project in 2015 and also by the IT R&D program of MOTIE/KEIT. [10041610, The development of automatic user information(identification, behavior, location) extraction and recognition technology based on perception sensor network(PSN) under real environment for intelligent robot].

References

1. Rahman, M.M., Ishikawa, S.: Robust appearance-base human action recognition. In: ICPR (2004)
2. Wu, X., Xu, D., Duan, L., Luo, J.: Action recognition using context and appearance distribution features. In: CVPR (2011)
3. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV (2003)
4. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Workshop VS-PETS (2005)
5. Wu, X., Xu, D., Duan, L., Luo, J.: Action recognition using context and appearance distribution features. In: CVPR (2011)
6. Yamato, J., Ohya, J., Ishii, K.: Recognizing humanaction in time-sequential images using hidden markov model. In: CVPR (1992)
7. Luo, G., Hu, W.: Learning silhouette dynamics for human action recognition. In: ICIP (2013)
8. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: CVPR (2012)
9. Wang, H., Klaser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: CVPR (2011)
10. Niebles, J.C., Wang, H., Fei-fei, L.: Unsupervised learning of human action categories using spatialtemporal words. In: BMVC (2006)
11. Klser, A., Marszaek, M., Schmid, C.: A spatiotemporal descriptor based on 3d-gradients. In: BMVC (2008)
12. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. In: PAMI (2001)
13. Ahad, M., Tan, J., Kim, H., Ishikawa, S.: Motion history image: its variants and applications. In: MVA (2012)
14. Sandhan, T., Choi, J.Y.: Frequencygrams and multi-feature joint sparse representation for action and gesture recognition. In: ICIP (2014)
15. Sandhan, T., Yoo, Y., Yoo, H., Yun, S., Byeon, M.: Multi-task learning with over-sampled time-series representation of a trajectory for traffic motion pattern recognition. In: AVSS (2014)
16. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: CVPR (2008)

17. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: CVPR (2010)
18. Cai, Q., Yin, Y., Man, H.: Learning spatio-temporal dependencies for action recognition. In: ICIP (2013)
19. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
20. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR (2004)
21. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machine. In: ACM TIST (2011)
22. Liu, C.: Beyound pixels: exploring new representations and applications for motion analysis. In: Doctoral Thesis. Massachusetts Institute of Technology (2009)