Human Action Recognition Based on RGB-D and Local Interactive Regions Detection

Suolan Liu and Lizhi Kong
Changzhou University, Jiangsu, China, 213164
lan-liu@163.com

Keywords: Action; RGB-D; Interactive regions; Recognition

Abstract. We propose a novel method to recognize human actions by fusing information from RGB-D sensors. Human action recognition is a challenging task because of the complexity movements, the variety of actions performed by different subjects and the changes of view and illumination. We propose to detect human motion from body parts by extracting sets of spatial-temporal interest points from RGB sequence and projecting them into depth map. Then, extract local interactive regions as supplementary information for action recognition. An improve classifier based on linear SVM coupled with dynamic time warping is developed for classification. We evaluate our method on two public datasets, including MSRDailyActivity3D dataset and ReadingAct RGB-D action dataset.

Introduction

Human action recognition is an active research topic in computer vision for its wild application in human-computer interaction and intelligent surveillance [1-3]. In the past few decades, many methods have been developed to recognize different actions from video sequences captured by visible light cameras and have been applied into some practical application. However, there are inherent limitations of this type sensor because it is sensitive to the change in illumination and color, dynamitic and cluttered background [4-7]. Despite great efforts have been made and great advance have been achieved, human action recognition still remains as a challenging task. Recently, with the wide spread of Microsoft Kinect sensor, it has become possible and convenient for us to capture RGB and depth information simultaneously, which can not only supply 3D information for distinguishing actions but also shows insensitive to changes in lighting conditions for depth maps [8,10]. Therefore, in recent years many researchers are paying closer attention to action recognition by fusing RGB and depth information. For examples, to address the issues appearing in home monitoring, e.g. cluttered backgrounds and loose clothes, authors in [10] put forward to applying spatio-temporal local features and a Bag-of-Words (BoW) model to identify single-person actions by fusing RGB and depth images. Tests on the self-collected dataset and two public datasets demonstrate its effect and better performance than only using RGB sequence for recognition. Kong et al. [11] proposed to use 3D kernel descriptors to project the low-level features on 3D patches from RGB and depth images into a more discriminative structure to describe different scenes and recognition different actions. Since view effects and camera motions have been viewed as difficult problems in action recognition, a view-invariant recognition method has been produced in [12]. Instead of extracting spatio-temporal features from every frame and using the feature vectors directly, they proposed to develop a spatio-temporal matrix (STM) by calculating the Euclidean distance between vectors and then the pyramid-structural BoW coupled with SVM is trained to recognize different actions. Due to the fusion of depth information, the recognition accuracy has been significantly improved in these methods. However, by analysis one may find that in some actions such as drinking, the recognition of interactive region between human and the cup can also provide help for action identification. Since the two-dimensional RGB image is the projection of three-dimensional objects [9], interaction detection in the traditional color images cannot effectively reflect the spatial relationship between human and the object, which will result in error extraction [13-14]. At this point, the depth image has a significant advantage [15]. The imaging principle of depth image reflects the distance between

objects and the sensor. Therefore, in this paper a human action recognition method based on depth image and local interactive regions detection is proposed.

The existing action recognition methods can be roughly divided into the following categories: target tracking based method, optical flow based method, template matching method and spatio-temporal interest points (STIPs) method. In the methods based on target tracking and template matching, extraction of human contour is usually needed before global features are obtained, which can be easily interfered by the noise and background changes. Using the optical flow information between two pixels to recognize human actions, method based on the optical flow is fragile to the change of light intensity. In spatio-temporal interest points' method, the extremal points are obtained by filtering and coefficient transformation and are labeled as the STIPs for action recognition. Recently, the use of STIPs has received increasing interest. STIPs method was firstly proposed for action recognition by Laptev and Lindeberg [16] who introduced a space-time extension of the popular Harris detector [17]. They detect regions having high intensity variation in both space and time as spatio-temporal corners. In [18], authors formulated the task of human action recognition as a learning problem penalized by a graph structure based on spatio-temporal features. Chakraborty et al. [19] proposed to use anisotropic filter to suppress interference and detect STIPs selectively. They improved the performance by gaining more repeatable, stable and distinctive STIPs for action recognition. Although advanced results have been published. STIPs based method remains vulnerable in tracking motive target for ignoring the spatial and temporal organization information. In practical applications, interaction information between human body and targets can effectively reflect human activities. Therefore, if interactive regions could be detected and features from these special regions be extracted as supplementary， which will provide great help for accurately recognizing of human actions.

The Proposed Method

Human Body and Interactive Regions Detection. Inspired by the Successful Work in [19] and [16], we extract human body from RGB sequence. In our scheme, STIPs are firstly extracted for human body detection. STIPs can be roughly detected by method proposed in [16]. To obtain the most effective STIPs, anisotropic filter is applied to eliminate the influence of interference points. For every STIP labeled as (x,y), its neighborhood can be defined as $(x-a, y-b)_{o<|a|,|b|\le r}$, where r is the size of the local window. Then we define a gradient orientation discriminant as:

$$q_\sigma(x, y, x-a, y-b) = \left|\cos(\theta_\sigma(x, y) - \theta_\sigma(x-a, y-b))\right| \quad (1)$$

where σ is scale size and $\theta_\sigma(x,)$ is gradient orientation. If a neighboring point have the same orientation with its center (x,y), the suppression strength can be expressed as follows:

$$p_\sigma(x, y) = \sum_a \sum_b M_\sigma(x-a, y-b) q_\sigma(x, y, x-a, y-b) \quad (2)$$

Then the next function can be used to identify a STIP should be reserved or deleted.

$$S_\sigma(x, y) = S(M_\sigma(x, y) - \lambda p_\sigma(x, y)) \quad (3)$$

l is the weighting factor. M_σ is gradient magnitude $S(z) = z$ when $z \ge 0$. Zero for negative z.

Furthermore, by comparing the coordinates of the selective STIPs, we can localize human motion region in a rectangular box.

Interactive regions detection is mainly involved to detect the targets associated with the human motion parts. For an example, in the activity of drinking, in addition to the human hand movement is effective information, cup can also provide additional information for identifying the action. Therefore, the interactive regions with human body can also provide valuable information in action recognition. However, to RGB sequence, we can only detect interactive regions by setting specific space constraints. Because of the interference from complex background, it is difficult to accurately detect the regions. But in depth sequence, there is compact continuity between objects belonging to

the same object or touching each other. Based on these analyses, in this paper we propose to extract interactive regions in depth sequence by projecting human motion region from RGB to depth sequence and calculating the depth continuity value from motion parts.

An example in Figure.1 displays the extraction of human motion region and interactive region. For the action "Hoovering floor", the movement of the upper body is very remarkable, and at the same time the vacuum cleaner is obviously moved in the operating from the human. It can be seen that the relatedly sensitive regions can be effectively extracted by our proposed method.

Figure 1 Regions detection of the action "Hoovering floor"

Feature Descriptor and Classification. Once human motion regions and interactive regions are effectively detected, the next step is to extract feature vector for action recognition. For local motion feature description, an approach based on literatures [20] and [21] is developed here for the construction of feature vector. We firstly extract dense trajectories from the above regions. Each point $p_t = (x_t, y_t)$ at frame t is tracked to its next frame t+1 by using a median filter in the dense optical flow field $\psi = (u_t, v_t)$.

$$p_{t+1} = (x_t,\ y_t) + (K * \psi)|(\tilde{x}_t,\ \tilde{y}_t) \quad (4)$$

Where K is the median filtering kernel. $(\tilde{x}_t,\ \tilde{y}_t)$ denotes the mean position of (x_t, y_t) in a local window region. For each trajectory, the histogram of oriented gradient (HoG), histogram of optical flow (HoF) and motion boundary histogram (MBH) are extracted and concatenated to form a feature vector.

Next, we apply linear SVM as classifier. In SVM, the kernel function calculates a distance between feature vectors to recognize action. However, in our method since the amount of extracted regions may be different, the histograms' number is not equal correspondingly to each action sample. Therefore, the basic SVM approach cannot be used as classifier directly. Dynamic Time Warping (DTW) is an algorithm for measuring the difference between two sequences, which may have different speeds or positions in time. To a feature vector constructed by histogram time series, a DTW based kernel is produced to use in SVM for recognition as follows:

$$V(v_n, v_m) = -DTM_{nm} \quad (5)$$

Where v_n and v_m are feature vector from two action sequences. DTM_{nm} is DTW distance.

Experiments and Discussion

We evaluate the recognition accuracy of our proposed method on two available datasets: MSRDailyActivity3D dataset [22] and ReadingAct RGB-D action dataset [10]. The MSRDailyActivity3D is the most common dataset for 3D human action recognition according to Wang et al. and is composed by 10 subjects performing 16 actions, which includes: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar, stand up, sit down. There are background objects and persons appear at different distances to the front sensor. Most actions involve human-object interaction. ReadingAct RGB-D action dataset is composed by 20 subjects with ages ranging from 20s to 60s, performing 19 actions: coming in, going out, walking past, walking around, switching light, talking on phone, phone call (mobile), picking up from floor, putting on jacket, hoovering floor, sitting down, standing up, lying down, getting up, reading a book, typing on computer, having meal, drinking (sitting), drinking (standing). To conduct a fair comparison, we only use the sequences

recorded in front of the person from reading Act RGB-D action dataset. Some examples of this dataset are shown in Figure.1.

MSRDailyActivity3D dataset. In the experiments, the most popular approach of cross-subject splitting [22] is used. The subjects labeled as odd numbers are used for training and others for testing. Here, we compare to the approaches proposed in [15], [22] and [10]. We do the tests by using RGB information only, depth information only and the combined information proposed in our method. The accuracy of our proposed approach is 98.6%, which is the best result on the dataset as far as we know. The comparable results are listed in Table.1. Most actions are classified without any mistake while only two actions involving playing game and tossing paper present low recognition accuracies from 86.7% to 90.9%.

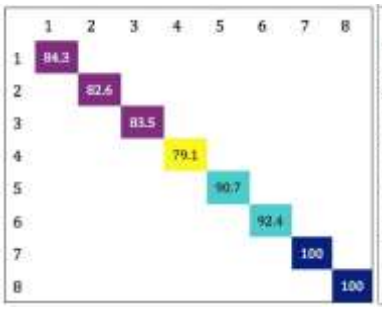
Table 1 Classification accuracy of different methods

MSRDailyActivity3D	RGB only(%)	Depth only(%)	Combined(%)
Method in [15]	81.5	75.2	83.4
Method in [22]	80.4	72.5	88.7
Method in [10]	79.1	65.3	84.9
Our method	87.6	77. 4	98.6

ReadingAct RGB-D Action Dataset. In this test we divide all actions into two subsets as listed in Table.2. For each action and each subject, the first two action sequences are used as training samples and the rest as test samples.

Table 2 Two activities subsets from ReadingAct RGB-D action dataset

Subset1	1. coming in ; 2. going out; 3. walking past; 4. walking around; 5. sitting down; 6. standing up; 7. lying down; 8. getting up;
Subset2	1. switching light; 2. talking on phone; 3. phone call(mobile); 4. picking up from floor; 5. putting on jacket; 6. hoovering floor; 7. reading a book; 8. typing on computer; 9.having meal; 10. drinking (sitting); 11. drinking (standing);



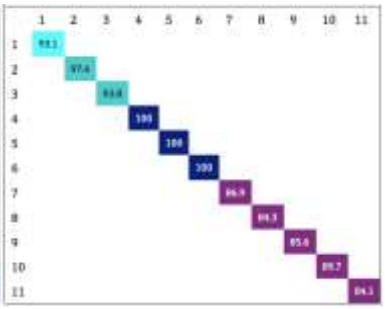


Figure 2 Confusion matrix of subset1 Figure.3 Confusion matrix of subset2

Figure.2 and Figure.2 are the confusion matrixes for the two subsets respectively. One can see the recognition accuracies obtained by our approach. For some actions such as lying down, getting up, picking up from floor, putting on jacket and hoovering floor, the local interactive regions can be detected observably and the motion regions from human body have obviously distinguishability from others. Therefore, these actions can be recognized with a high accuracy as 100%. On the other hand, since there is few interactive regions can be detected, actions including walking around and going out are classified with low accuracies about 80%.

Conclusion

In this work, we presented a new framework for human action recognition by combining RGB information with depth maps. We proposed extracting sets of spatial -temporal interest points from RGB sequence and projecting the detected motion regions into depth map. Furthermore, local interactive regions are detected from depth sequence by using the 3D distance information. An improved linear SVM based on dynamic time warping is applied as classifier. Extensive tests with the produced framework demonstrate that all the developed steps contribute significantly to improve recognition accuracy. We concluded that spatial-temporal information, as well as local interactive regions, could be efficiently combined by forming feature vector, which results in a significant increase of performance.

References

[1] [S Megrhi, M Jmal and W Souidene. Spatio-temporal action localization and detection for human action recognition in big dataset. Journal of visual communication and image representation, 2016,41: 375-390
[2] H Liu. RGB-D action recognition using linear coding, Neurocomputing, 2015, 149: 79-85
[3] J Zhang. RGB-D-based action recognition datasets: A survey. Pattern recognition, 2016(60): 86-105
[4] C Chen, R Jafari, N Kehtarnavaz. Action recognition from depth sequences using depth motion maps-based local binary patterns. In: WACV, 2015: 1092–1099
[5] N Dalal, B Triggs. Histograms of oriented gradients for human detection. CVPR2015: 886–893
[6] C L Diogo, T Hedi, P David. Learning features combination for human action recognition from skeleton sequences, Pattern recognition letters, 2017
[7] I Lillo, J Niebles, A Soto. Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos. Image and vision computing, 59(2017): 63-75
[8] R. Slama, H. Wannous, M. Daoudi, A. Srivastava, Accurate 3d action recognition using learning on the grassmann manifold, Pattern Recognition, 2014, 48(2): 556-567
[9] D Maxime, W Hazem. Space-time pose representation for 3D human action recognition, Springer Berlin Heidelberg, 2013, 8158:456-464
[10] L Chen, H Wei, J Ferryman. ReadingACT RGB-D action dataset and human action recognition from local features. Pattern recognition letters, 2014, 50: 159-169
[11] Y Kong, B Satarboroujeni, Y Fu. Learning hierarchical 3D kernel descriptors of RGB-D action recognition. Computer vision and image understanding, 2016,14:14-23.
[12] Y Hsu, C Liu, T Chen. Online view-invariant human action recognition using rgb-d spatio-temporal matrix. Pattern recognition, 2016,60:215-226
[13] X Li, M Fang, J Zhang. Learning coupled classifiers with RGB images for RGB-D object recognition. Pattern recognition, 2017,61:433-446
[14] L Fademerecht, I Bulthoff. Action recognition is viewpoint-dependent in the visual periphery, Vision Research, 2017,135:10-15
[15] S Liu, R Lagonegro. Action recognition scheme based on RGB-D mixture model. Journal of engineering research and technology, 2016,5(12): 352-356
[16] I Laptev, T Lindeberg. Space-time interest points, ICCV Workshops, 2003.
[17] C Harris. A combined corner and edge detector, in: Alvey Vision Conference, 1988.
[18] W Guo, G Chen. Human action recognition via multi-task learning based on spatial-temporal feature. Information Sciences, 2015,65(1): 37-43.
[19] C Bhaskar, M Holte, T Moeslund. Selective spatio-temporal interest points. Computer vision and image understanding, 2011
[20] H Wang, A Klaser. Action recognition by dense trajectories. CVPR, 2011: 3169–3176
[21] H Wang, C Schmid. Action recognition with improved trajectories. In ICCV, 2013.
[22] J Wang, Z Liu. Mining action let ensemble for action recognition with depth cameras. In CVPR, 2012:1290–1297.