Human Action Recognition using Depth Motion Map and Resnet

Thanh-Hai Tran*, Quoc-Toan Tran

Hanoi University of Science and Technology – No. 1, Dai Co Viet Str., Hai Ba Trung, Ha Noi, Viet Nam

Received: November 28, 2018; Accepted: June 24, 2019

Abstract

Human action recognition is an active research topic in recent years due to its wide application in reality. This paper presents a new method for human action recognition from depth maps which are nowadays highly available thanks to the popularity of depth sensors. The proposed method composes of three components: video representation; feature extraction and action classification. In video representation, we adopt a technique of motion depth map (DMM) which is simple and efficient and more importantly it could capture long-term movement of the action. We then deploy a deep learning based technique, Resnet in particular, for extracting features and doing action classification. We have conducted extensively experiments on a benchmark dataset of 20 activities (CMDFall) and compared with some state of the art techniques. The experimental results show competitive performance of the proposed method. The proposed method could achieve about 98.8% of accuracy for fall and non-fall detection. This is a promising result for application of monitoring elderly people.

Keywords: Human action recognition, depth motion map, deep neural network, support Vector Machine

1. Introduction

Human action recognition is becoming one of the most active research fields of computer vision. There are many applications of human action recognition in home / public security, human robot interaction or entertainment. Approaches for human action recognition could be divided in two main categories: hand crafted features based and deep learning based [1]. While hand crafted features based approach depends of expertise of feature designers and are only suitable for small dataset, deep learning based approach has been shown to be very successful on many big and challenging benchmarks [2]. Besides, with the rapid development of sensor technology, depth sensors are becoming very popular in the markets. Depth sensors have an attractive characteristic that is its independence of lighting condition, so they could avoid most challenges compared to conventional RGB cameras.

The work presented in this paper will deal with depth data for action recognition. The studied method belongs to the second approach which inherits the success of convolutional neural networks (CNN). Despite of this success there still exists many issues to be resolved. On the one hand, direct application of 2D CNNs totally ignores the temporal connection among frames [3]. On the other hand, some 3D CNNs tends to capture spatial-temporal features of the action but not long-term movement [4]. Both cases could lead to degrade the performance of action recognition.

We are motivated by the fact that a video could be compactly represented by a motion map. We could list some related popular techniques such Motion History Image (MHI) [5], Depth Motion Map [6], Gait Energy Images [7]. In these techniques, a sequence of consecutive images is represented by only one image. As a result, a conventional 2D neural network could be directly deployed to predict the action label.

In this paper, we propose a method for human action from depth maps by combining both techniques. Firstly, a motion depth map will be computed from consecutive frames of a video. We then deploy a 2D convolutional neural network for feature extraction and classification of actions. We experiment extensively this method and compare it with existing techniques, showing better results.

The remaining of this paper is organized as follows. In section II, we present related works on human action recognition and focus only to review depth based methods. In section III, we describe our proposed method with the use of depth motion map and convolutional neural network Resnet for action recognition. We will evaluate this method on a benchmark dataset. Section V concludes and gives ideas for future works.

*Corresponding-author: (+84)976.560.526

Email: thanh-hai.tran@mica.edu.vn

2. Related works

Action recognition techniques are broadly divided into two categories: methods using hand-crafted features, and deep learning based methods. In this section, we will focus on the state of the art works that are closely related to our works: action recognition from depth sensors.

The methods belonging to the first approach extract features from depth map. In [8], the authors computed 4D normal vector from each depth frame. They then created spatial-temporal cells and computed histogram of normal orientation vectors for each cell and concatenated them to produce the final vector for action representation (called HON4D). This method is simple and easy to implement. However, it is quite sensitive to noise of depth sensors. Other group of researches try to represent a sequence of depth frames by a depth motion map (DMM). Then different types of features have been extracted for example histogram of oriented gradient (HOG) in [9], local binary pattern (LBP) in [10], kernel descriptor (KDES) [11], [12]. The most advantage of DMM is its efficient computation. However, as DMM captures long-term movement of the human, some local movement could be omitted.

The methods belonging to the second approach learn features from training data. Many techniques using deep learning have been proposed for human action recognition from RGB video [13], [14]. However, less methods have been studied on depth data. One reason could be that the deep learning requires big data for training. 2D or 3D CNNs for action recognition inherit from very big dataset of RGB images or videos. However, the depth datasets of human action are still limited. In this paper, we would like to investigate how to combine the two techniques (DMM and deep learning) in a unified framework. Instead of using conventional handcrafted features extracted on depth map, we will use deep learning to learn features. The studied neural architecture is Resnet which have been the best deep network for images based task [15]. We will investigate if Resnet is convenient on depth motion map for action recognition task.

3. Proposed method

3.1 The proposed framework

We propose a framework for action recognition from depth map illustrated in Fig. 1. It composes of three main steps:

- *Step 1: Computation of a compact representation of video by a unique image:* In the first step, given a sequence of consecutive images, we compute a depth motion map (DMM) that is a compact and efficient representation of a video.

- *Step 2: Extraction of features:* We extract the descriptor for the DMM computed from previous step. At this step, we deploy a 2D convolutional neural network (Resnet-101) which has been shown to be very efficient for many image based tasks.

- *Step 3: Action classification:* We could use scores produced from softmax layer of Resnet-101 to make final decision of action classification or we could learn a SVM models from training data and use for predicting action label at testing phase.

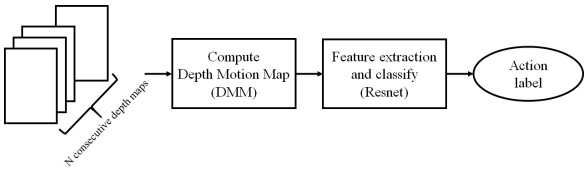


Fig. 1. General framework of proposed method for action recognition.

In the following, we will explain in more detail each step of the proposed framework.

3.2 Depth Motion Map (DMM)

Depth Motion Map technique tries to represent a sequence of frames by summing all movements of pixels between two consecutive frames. This representation was shown to be computationally very fast and compact. It captures historical movements of all pixels in the sequence. Thanks to its valuable properties, in this work, we deploy DMM technique for action representation from depth maps.

The computation of DMM is following. Given a sequence of N depth maps $\{D^1, D^2, ..., D^N\}$, the depth motion map is defined as follows:

$$DMM = \sum_{i=1}^{N-1} | \mathrm{D}^{i+1} - \mathrm{D}^{i} |$$

Fig. 3 illustrates a DMM computed from a falling action sequence of fig. 2. We notice this image represents well the long-term movement of human. Note that the original resolutions of RGB and depth are of the same resolution but for better illustrating the DMM we have zoomed in the DMM in Fig.3.

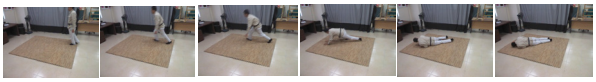


Fig. 2. A sequence of consecutive frames (shown in RGB for better understanding)

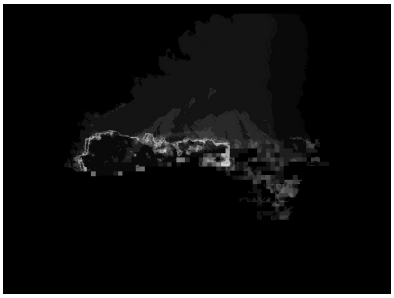
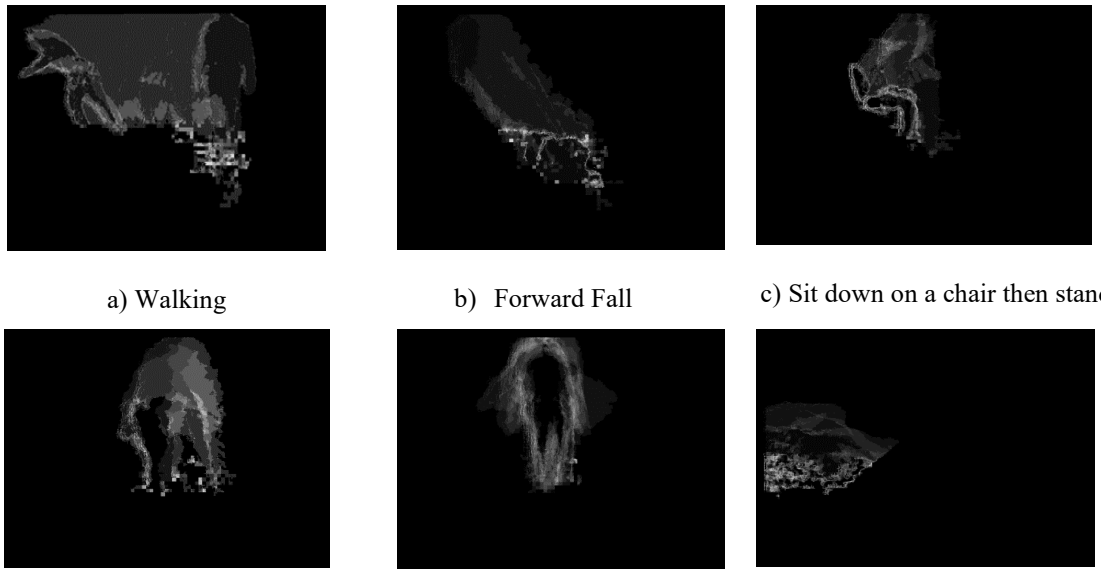


Fig. 3. The DMM computed from the corresponding depth sequences of falling action in Fig. 2

Fig. 4 illustrates different DMMs computed from different action sequences. We observe the difference among DMMs which could be a good indicator for classification.

3.3 Feature extraction using Resnet

Given a DMM computed of a video sequence, we extract features from this DMM for classification. In this work, we would like to try an advanced learning technique using deep neural network to automatically extract features from DMM. There are many deep neural architectures such as VGG16, Google Lenet, Alexnet, etc. One of problems of such deep neural networks is that when the deeper networks start converging, accuracy will get saturated then degrades rapidly. In 2015, Kaiming He and his colleagues tried to resolve this issue by deep residual learning framework (called Resnet) [15]. The idea of Resnet is instead of learning a direct mapping of x to y with a function *H(x)* (plain block composed of a few of stacked non-linear layers), Resnet learns a residual function $y = F(x) = H(x)-x$ (residual block composed of staked non-linear layers and an identity function) where *F(x)* is easier to be optimized than *H(x)*. *F(x)* is called Residual function. Resnet has been demonstrated to outperform in both ILSVCR'15 and COCO'15 challenges. Motivated by its performance, in this paper, we will deploy Resnet for action recognition. The original Resnet has been trained on RGB dataset and efficient for RGB still images based task. In our work, DMM is depth motion map, which has totally different characteristique than RGB images. Then one of contributions in this work is to investigate if Resnet is still efficient on DMM for action recognition. In the original paper [15], there are five architectures of Resnet (18 layers, 34 layers, 50 layers, 101 layers, 152 layers). Resnet-101 will be chosen for investigation due to its balances between accuracy and computational time. Resnet-101 has been trained and test on COCO'15 dataset. To be deployed on DMMs images, we have to fine-tune the network on our DMM dataset. We normalize all DMMs to 224x224x3. We use batch normalization after every convolutional layer. Stochastic Gradient Descent (SGD) with momentum 0.9. Learning rate is set to 0.001 with mini batch size 16, weight decay 1e-6, cross entropy is loss function. The training data is described in Section 4.



a) Walking b) Forward Fall c) Sit down on a chair then stand up

d) Crouch down to pick up things by left hand e) Run slowly f) Left fall while lying on a bed

Fig. 4. Illustration of different DMMs computed from different action sequences

3.4 Action classification

Once the network has been trained, we can use scores given by softmax layer for making decision. We can also extract features at the layer just before softmax and put into a SVM classifier. We will report classification result using softmax and SVM at experiment section.

4. Experiments

4.1 Data set and performance measurement

To evaluate the performance of the proposed method, we use a benchmark dataset CMDFall [16]. This dataset contains 20 actions captured by Kinect sensors in simulated home environment with 50 subjects (30 males and 20 females) aging from 21-40. The depth sensor is set at resolution of 640x480, 16bit depth images and captures frames at 20fps. In this work, we will investigate only depth maps from one Kinect view (K3). 20 actions contain normal actions and abnormal actions. These actions are grouped in 6 groups and 2 classes. List of actions is presented in Tab. 1. Totally we have 1967 samples of 20 classes. We used the same data split as [16] for training and testing the method. 993 samples of all classes for training and 974 for testing. We use accuracy as performance measurement.

Table 1. List of actions and categorization

S_1	S_2: 6 groups	ID	S_3: 20 activities
Fall	Fall while walking	1	Front fall
		2	Back fall
		3	Left fall
		4	Right fall
	Fall while lying on the bed	5	Lie on bed then fall left
		6	Lie on bed then fall right
	Fall while sitting on the chair	7	Sit on chair then fall left
		8	Sit on chair then fall right
Non Fall	Horizontal movement of the whole body	9	Walk
		10	Run slowly
		11	Stagger
		12	Crawl
		13	Move chair
	Hands and legs movement	14	Move hand and leg
		15	Left hand pick up
		16	Right hand pick up
	Vertical movement of the whole body	17	Jump in place
		18	Sit on chair then stand up
		19	Sit on bed then stand up
		20	Lie on bed then sit up

4.2 Experimental results

4.2.1 Evaluation of the number of layers in Resnet

As we mentioned in the section 3.3, the original paper about Resnet has introduced different architectures which differ from the number of layers. We have tested Restnet with 34, 50, 101, 152 layers and obtained results as shown in Tab.2. We see that the accuracy increases gradually when the number of layers increases from 34 to 101 but it seems to be saturated when the number of layers reaches to 152. As a result, we will choose Resnet with 101 layers for further analysis.

We observe that the proposed method DMM-Resnet using softmax for classification achieved 66.1% of accuracy in case of classifying 20 actions. This accuracy is still low because of high variation of actions and intra-class similarity. However, when we group them into 6 groups, accuracy has increased to 94.6%. In addition, when we would like to distinguish only fall and non-fall, the method could produce very impressive results (98.5%). This shows a good performance of the method for fall detection from normal daily activities.

Table 2. Accuracy (%) of action classification with different layers of Resnet

Methods	20 actions	6 groups	Fall and Non-Fall
DMM-Resnet 34-softmax	52.0	87.4	94.1
DMM-Resnet 50-softmax	64.1	93.9	97.8
DMM-Resnet 101-softmax	66.1	94.6	98.5
DMM-Resnet 152-softmax	66.6	94.3	98.4

4.2.2 Comparison with existing methods

We compare the proposed method with other methods [11]. The method [11] used exactly DMM for action representation as this method, but Kernel descriptor (KDES) was extracted from DMM for action description. Another method proposed to characterize a sequence of frames by static Pose Map (SPM) [17]. We have computed SPM from action sequences then apply both KDES-SVM and Resnet-101 for comparison. In addition, beside using softmax of Resnet-101 for making classification decision, we extract features from layers before fully connected layer and train SVM for classification. We report the comparative results in Tab. 3. actions.

We found that DMM-Resnet101-SVM produced the best result comparing to existing methods. Using Resnets101-SVM, the accuracy increases more than 16.2% in case of 20 action classification, 11% in case of 6 groups classification and 5.5% in case of fall and non-fall classification.

Table 3. Comparison of different methods in term of accuracy (%)

Methods	20 action	6 groups	Fall and Non-Fall
DMM-KDES-SVM [11]	51.2	84.2	93.5
SPM [17] -KDES-SVM	51.6	85.5	93.0
DMM-Resnet 101-softmax	66.1	94.6	98.5
SPM [17] -Resnet 101-softmax	63.0	92.9	96.1
SPM [17] - Resnet 101- SVM	64.1	93.0	97.2
Our proposed DMM-Resnet 101-SVM	**67.4**	**95.2**	**98.8**

SPM gives similar or lightly lower accuracy than DMM when combining with KDES or Resnet. DMM-Resnet-SVM gives higher accuracy than DMM-Resnet-softmax and highest result among all methods. We have investigated in details failure cases generated by DMM-Resnet101-SVM. We found that

in case of 20 action classification, the most failure appears at front fall with back fall; left fall with right fall, lie on bed then fall left with lie on bed with fall right; left hand pick up with right hand pick up. In case of 6 groups classification, we observe once again fall in different directions are confused with fall from bed. The confusion is significantly reduced with the case of fall and non-fall classification.

5. Conclusions

In this paper we have presented a method for human action recognition from depth map using combination of depth motion map and Resnet. Resnet has been shown to be very for RGB images based task. In this paper, we have demonstrated that Resnet is still very efficient on depth motion map. We have compared the proposed method with Kernel descriptors and found that the method outperformed it. The highest classification has been achieved in case of fall and non-fall classification with 98.8% of accuracy. This is a promising result because it could help for alarming falling of people as soon and accurate as possible in elderly or kid monitoring. In the future, we will explore other modalities such as RGB and skeletons for improving performance of the method.

References

[1] R. Poppe; A survey on vision-based human action recognition; Image Vis. Comput., vol. 28, no. 6, pp. 976–990, Jun. 2010.

[2] J. Carreira and A. Zisserman; Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, ArXiv170507750 Cs, May 2017.

[3] O. Russakovsky et al.; ImageNet Large Scale Visual Recognition Challenge, ArXiv14090575 Cs, Sep. 2014.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri; Learning Spatiotemporal Features with 3D Convolutional Networks; in Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Washington, DC, USA, 2015, pp. 4489–4497.

[5] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa; Motion history image: its variants and applications, Mach. Vis. Appl., vol. 23, no. 2, pp. 255–281, Mar. 2012.

[6] C. Chen, K. Liu, and N. Kehtarnavaz; Real-time human action recognition based on depth motion maps; J. Real-Time Image Process., vol. 12, no. 1, pp. 155–163, Jun. 2016.

[7] X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren; Gait Energy Response Functions for Gait Recognition against Various Clothing and Carrying Status, Appl. Sci., vol. 8, no. 8, p. 1380, Aug. 2018.

[8] O. Oreifej and Z. Liu; HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences, in 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 716–723.

[9] X. Yang, C. Zhang, and Y. Tian; Recognizing Actions Using Depth Motion Maps-based Histograms of Oriented Gradients; in Proceedings of the 20th ACM International Conference on Multimedia, New York, NY, USA, 2012, pp. 1057–1060.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz; Action Recognition from Depth Sequences Using Depth Motion Maps-Based Local Binary Patterns; in 2015 IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 1092–1099.

[11] T.-H. Tran and V.-T. Nguyen; How Good Is Kernel Descriptor on Depth Motion Map for Action Recognition; in Computer Vision Systems, 2015, pp. 137–146.

[12] T.-H. Tran, T.-L. Le, V.-N. Hoang, and H. Vu; Continuous detection of human fall using multimodal features from Kinect sensors in scalable environment; Comput. Methods Programs Biomed., vol. 146, pp. 151–165, Jul. 2017.

[13] K. Simonyan and A. Zisserman; Two-Stream Convolutional Networks for Action Recognition in Videos; ArXiv14062199 Cs, Jun. 2014.

[14] V. Khong and T. Tran; Improving Human Action Recognition with Two-Stream 3D Convolutional Neural Network; in 2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR), 2018, pp. 1–6.

[15] K. He, X. Zhang, S. Ren, and J. Sun; Deep Residual Learning for Image Recognition; ArXiv151203385 Cs, Dec. 2015.

[16] Thanh-Hai Tran et al.; A Multimodal multiview dataset for human fall analysis and preliminary investigation on modality; in The 20th International Conference on Pattern Recognition (ICPR'2018), Beijing, China.

[17] Z. Zhang, S. Wei, Y. Song, and Y. Zhang; Gesture Recognition Using Enhanced Depth Motion Map and Static Pose Map; in 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), 2017, pp. 238–244.