Pseudo-3D Trajectories: An Effective Approach for Motion Representation in Depth Data

Chien-Quang LE

The Graduate University for Advanced Studies

Duy-Dinh LE

National Institute of Informatics

Shin'ichi Satoh

National Institute of Informatics

Abstract

Leveraging the motion information of trajectories shows the effectiveness to the human action recognition in 2D video. However, the issue is that this approach direction is effective or not when represents motions in 3D video is not still answered. In this paper, we will deal with this issue by conducting experiments based on 2D trajectory features to present motion information from one 3D video representation. Beside, in order to ensure including depth information, we propose a method based on compensating motion information from other representations. Evaluated on the benchmark datasets, our method significantly outperforms the 3D SoA methods.

Keywords: `Trajectory`, action recognition, depth, feature representation

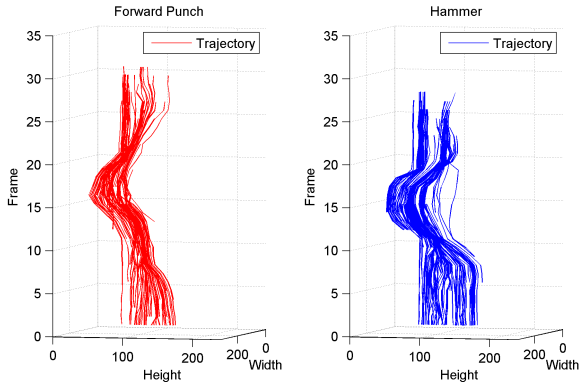
1. Introduction

Background and Challenges. Gần đây, với sự phát triển của RGB-D camera như Kinect, depth data đã mở ra nhiều hướng nghiên cứu tiềm năng cho bài toán Human Action Recognition. So sánh với intensity images thông thường, depth maps hỗ trợ nhiều advantages hơn. Ví dụ, depth maps cung cấp các thông tin về shape rõ ràng hơn so với intensity images. Hơn thế nữa, depth data ít bị ảnh hưởng bởi những thay đổi của ánh sáng. Tuy nhiên, các phương pháp dựa trên intensity liệu có hiệu quả trên depth data hay không vẫn chưa được quan tâm nhiều.

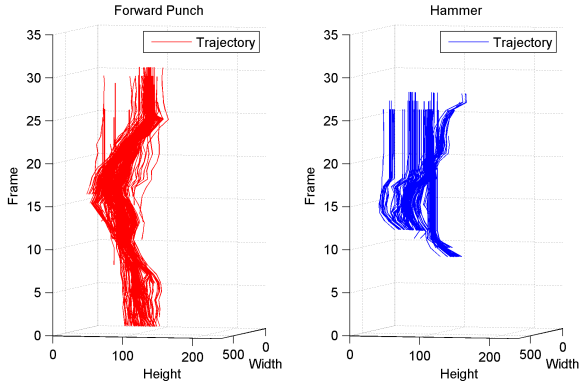
Existing approaches and drawbacks. Trong bài toán action recognition, để adapt các phương pháp dựa trên intensity cho depth data có 2 yếu tố chính. Thứ nhất, để capture motion information hiệu quả việc chọn lựa a robust feature representation là rất quan trọng. Thứ hai, để đảm bảo motion là đầy đủ thông tin trong depth video, việc bổ sung thông tin depth vào feature representation là yêu cầu không thể thiếu. Tuy nhiên, các phương pháp được đề xuất gần đây vẫn chưa hội tụ đủ 2 yếu tố này. Một số phương pháp như [1, 2] xem xét depth value như là intensity value và adapt các intensity-based techniques. Mặc dù, chúng có thể đạt được những kết quả hợp lý, nhưng tất cả chúng đều phải đối mặt với nhiều hạn chế. [DMM-HOG] có thể tận dụng thông tin depth từ các phép chiếu của depth maps. Nhưng its feature representation dựa trên global motion như HOG sẽ dễ gây nhầm lẫn bởi những similar postures. [2] có thể đảm bảo depth information trong việc tính toán features. Nhưng cách tiếp cận này không đảm bảo được sự tin cậy khi extract các local points, do bởi textureless data and depth noise. Ngoài hướng tiếp cận trên, các phương pháp như [3, 4] chỉ tập trung khai thác depth information nên không tận dụng được sức mạnh của các intensity-based features. Do đó, hướng nghiên cứu của chúng tôi là propose một phương pháp có thể đáp ứng đầy đủ cả 2 yếu tố nêu trên.

Proposal, Idea and Steps. Trong bài báo này, chúng tôi sử dụng một feature representation dựa trên dense trajectories của [5], do bởi hiệu quả của cách tiếp cận này trong nhiều bài toán, including activity recognition and multimedia event detection. Các trajectories thu được bằng cách tracking các sampled points densely sử dụng optical flow fields. Sau khi extract trajectories, các trajectory-aligned descriptors sẽ được adopted. Sau đó, features tính toán được từ các descriptors này sẽ được sử dụng cho việc biểu diễn motion information trong video.

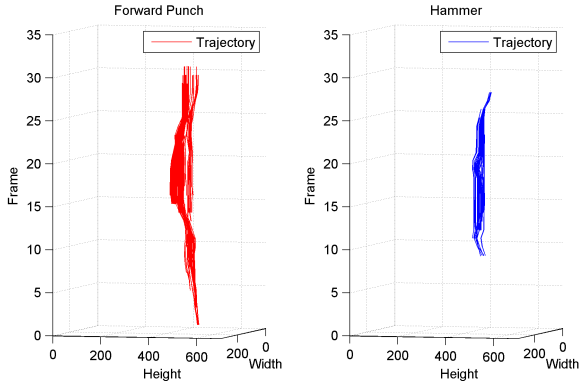
Tuy nhiên, việc thiếu sót depth information trong feature representation có thể gây ra các trường hợp bị confused, như được chỉ ra trong Figure 1a. Do đó, để đảm bảo việc không bỏ sót thông tin depth, ý tưởng cơ bản là combine thông tin chuyển động từ nhiều góc nhìn khác nhau. Các biểu diễn từ nhiều góc nhìn có thể đạt được bằng cách chiếu depth maps lên trên các mặt phẳng tương ứng. Việc chiếu này dễ dàng thực hiện được bởi những thuận lợi mà depth data mang lại.



(a) From front view



(b) From side view



(c) From top view

Hình 1: Minh họa sự tương tự giữa phần lớn các Trajectories của 2 actions: Forward Punch & Hammer.

Experiments and Results. Chúng tôi tiến hành các experiments trên challenging benchmark datasets, các kết quả thí nghiệm chỉ ra rằng phương pháp của chúng tôi đánh bại the SoA methods trên depth data. Các kết quả này đã cho thấy những contributions của our method: (1) We propose an adaptive method for 3D video representation by using 2D features. (2) We thực hiện comprehensive experiments on the state-of-the-art MSR Action 3D dataset and show that our method is the best when compared with the state-of-the-art 3D methods.

Paper structure. After a brief review of the related work in Section 2, the proposed method is described in Section 3. Section 4 presents the experimental results and their concerned discussions. The summaries of our work are given in Section 5.

2. Related Works

Tìm hiểu các thành phần của một hệ thống HAR hiện là một trong những hướng nghiên cứu quan trọng của CV. Feature representation là 1 trong số các thành phần thu hút được sự chú ý của cộng đồng nghiên cứu.

Works trích chọn features từ depth data. - Hướng xem depth value như intensity value. - Hướng sử dụng real depth value và skeleton information.

Điểm khác biệt của phương pháp hiện tại với các phương pháp trước. - Hướng sử dụng 2d trajectories cho 3d data.

Works in combining many types of features and sự khác biệt với our work. - Works gần đây sử dụng early fusion sheme. - Our method sử dụng late fusion sheme.

3. Proposed Method

This paper presents a effective depth video representation by adapting intensity trajectories-based motion features. First, chúng tôi sẽ cung cấp một brief review of the dense trajectories-based feature proposed by Heng Wang et al. [5]. Những phần liên quan như: dense sampling, tracking and feature descriptors is referred to. Our trajectories-based approach for depth data is mentioned at the end of this section.

3.1. Dense trajectories

In order to obtain trajectories, there are two important steps: sampling and tracking. [5] propose sampling on a dense grid with a step size of 5 pixels. The sampling is performed at multiple scales with a factor of $1/\sqrt{2}$. Then, tracking is the next step to form trajectories. At each scale, in frame t, each point $P_t = (x_t, y_t)$ is tracked to point $P_{t+1} = (x_{t+1}, y_{t+1})$ in next frame *t+1* by:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)}, \tag{1}$$

where $\omega = (u_t, v_t)$ denotes the dense optical flow field, M is the kernel of median filtering, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of P_t. The algorithm of [6] is adopted to compute the dense optical flow. And to avoid a drifting problem, a suitable value of trajectory length is set to 15 frames. Beside, trajectories with sudden changes are removed.

After extracting trajectories, two kinds of descriptors: a trajectory shape descriptor and a trajectory-aligned descriptor can be adopted.

Trajectory Shape Descriptor. This descriptor describes the shape of a trajectory in the simplest way. Given a trajectory of length L, its shape is concatenated by a sequence of displacement vectors $S = (\Delta P_t, ..., \Delta P_{t+L-1})$, where $\Delta P_t = P_{t+1} - P_t = (x_{t+1} - x_t, y_{t+1} - y_t)$. In order to make the descriptor invariant to

scale changes, the final result is then achieved by normalizing the shape vector by the overall magnitude of the displacement vectors:

$$\bar{S} = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{k=t}^{t+L-1} \|\Delta P_k\|}, \tag{2}$$

Trajectory-aligned Descriptor. The descriptors are much more complex than the trajectory shape descriptor. They are computed within a space-time volume ($N \times N$ spatial pixels and L temporal frames) around the trajectory. This volume is divided into a 3D grid (spatially $n_\sigma \times n_\sigma$ grid and temporally n_τ segments). The default settings of these parameters are $N = 32$ pixels, $L = 15$ frames, $n_\sigma = 2$, and $n_\tau = 3$.

In order to capture the local motion and appearance around a trajectory, three kinds of descriptors have been employed: the Histogram of Oriented Gradient (HOG) [7], the Histogram of Optical Flow (HOF) [8], and the Motion Boundary Histogram (MBH) [9]. For HOG, orientation information is quantized into 8-bin histogram. HOF is 9-bin histogram. Since the feature of a trajectory is calculated and concatenated from sub-volumes of a 3D volume, the final representation has 96 dimensions for HOG and 108 dimensions for HOF. MBH descriptor computes derivatives on both horizontal and vertical components of optical flow $I_\omega = (I_x.I_y)$. Similar to HOG descriptor, the orientation information is quantized into 8-bin histogram. Since the motion information is combined along two directions, the final representation is $96 \times 2 = 192$-bin histogram. By presenting gradient of optical flow, MBH descriptor is able to suppress global motion information and only keep local relative changes in pixels.

According to the authors [8, 5, 10, 11], all the three descriptors have shown the effectiveness for action recognition. The experimental settings for these descriptors are based on an empirical study showed in [5]. We also conduct our experiment on all the three descriptors when compared to the depth-based state-of-the-art methods.

3.2. Pseudo-3D trajectory-based Approach for Motion Feature in Depth Data

Our proposed trajectory-based approach for human action recognition in depth data is as follow. At first, intensity representations are formed from the sequence of depth maps, as illustrated in Figure 2. In particular, we choose number of the representations of 3. Number 3 represents 3 view directions: front, side, and top in 3D space. Forming the representations is necessary due to dimensional gap when we adapt 2D techniques for 3D data. After that, the dense trajectories are extracted from the intensity representations. And the feature descriptors are also computed in this step. At the next step, with each intensity representation, corresponding feature representation is quantized from raw trajectory features by apply a "bag-of-words" model. A "late fusion" scheme is used to generate the final feature representation for action in the sequence of depth maps (Fig. 3).

- Hình 2 - Illustration of proposed method - Depth maps -> 3 projections -> dense trajectories

- Hình 3 - Framework overview for our system - Depth data -> 3 feature extraction -> 3 BoW model -> 3 histogramintersection-SVM -> concatenated-score features -> cai-kernel SVM classifier

In order to generate intensity representations from the sequence of depth maps, we use the approach proposed in [12]. This technique is also used in [1]. Basically, this method projects depth maps onto three orthogonal planes in Casterian space to obtain corresponding intensity representations. However, motion representation for human action in the previous approaches is accumulated from global motion information. Therefore, these approaches must deal with the challenges from human segmentation problem in more complicated datasets. In contrast to the previous ones, we pay attention to capture local motion information for representing human actions. With the approach, we do not care the challenges for segmenting human body. To effectively use local motion in-

formation, we leverage the effectiveness of trajectory-based representation. In practice, we adopt the dense trajectory-based approach proposed in [5]. Thus, motion information in depth data can be reproduced by complementary motion information in different intensity representations.

Our proposed trajectory-based approach is compared with the state-of-the-art methods in human action recognition using depth data. Actually, our approach does not care skeleton extraction, which is used as an important factor in some works, such as [3, 13]. In fact, extracting skeleton exactly is still an unsolved problem, due to the challenges, such as: cluttered background, hardware quality, camera motion, ... Figure 4 illustrates an example case when extract skeleton information.

- Hình 4 - An example for skeleton extraction error.

4. Experimental Settings

4.1. Dataset

We test our method on MSR Action 3D dataset. This dataset contains 20 actions, as showed in Table 1. Actions are performed by ten subjects for two or three times in the context of game console interaction. In total, there are 567 sequences of depth maps. The depth maps are shot at frame rate of 15 fps. The size of the depth map is 640×480, we resize into 320×240 to ensure processing efficiency.

ID	Action Name	ID	Action Name
1	high arm wave	11	two hand wave
2	horizontal arm wave	12	side-boxing
3	hammer	13	bend
4	hand catch	14	forward kick
5	forward punch	15	side kick
6	high throw	16	jogging
7	draw x	17	tennis swing
8	draw tick	18	tennis serve
9	draw circle	19	golf swing
10	hand clap	20	pick up & throw

Bång 1: 20 actions in MSR Action 3D dataset

In order to conduct a fair comparison, we use the same experimental settings as [12, 13, 1, 3, 2, 4]. In the settings, the dataset is divided into three action subsets. Each subset has 8 actions (Table 2). The two subsets AS1 and AS2 present that grouped actions have similar movements. The subset AS3 groups complex actions together. For instance, action *hammer* seems to be confused with action $forwardpunch$ in AS1 or similar movements between action *handcatch* and action *side boxing* in AS2. As for each subset, we select half of the subjects as training and the rest as testing (i.e. cross subject test).

Action Subset 1 (AS1)	**Action Subset 2 (AS2)**	**Action Subset 3 (AS3)**
horizontal arm wave	high arm wave	high throw
hammer	hand catch	forward kick
forward punch	draw x	side kick
high throw	draw tick	jogging
hand clap	draw circle	tennis swing
bend	two hand wave	tennis serve
tennis serve	side-boxing	golf swing
pick up & throw	forward kick	pick up & throw

Bảng 2: The three action subsets used in the experiments

4.2. Evaluation Method

Figure 3 shows our evaluation framework for the trajectory-based features. We perform experiments using the proposed approach and compare with the state-of-the-art methods on depth data. We use the application available online[1] to extract dense trajectories and aligned-features. To quantize a large number of features obtained by densely sampling, the "bag-of-words" (BoW) model is applied. At first, in each intensity representation, we randomly get about 80,000 extracted trajectories for clustering with K-mean algorithm. Then, a codebook of 2000 visual codewords is formed for each. After that, the hard-assignment technique is used to compute histograms of the visual words on the corresponding intensity representations.

Once all the BoW histograms are generated, we adopt the late-fusion scheme with the popular Support Vector Machine (SVM) for classification. In practice, we use the precomputed-kernel technique with the histogram intersection mea-

[1] http://lear.inrialpes.fr/∼wang/dense_trajectories

surement for the first classification step and apply χ^2 kernel for the next step. In our implementation, we use the libSVM library published online by author[2] and perform the one-vs-all strategy for multi-class classification. We adopt the format requirements of the library to synchronize the annotation and the data. For testing, the scores of each intensity representation at the first classification step are concatenated to generate the final feature representation as the input of the final prediction. The predicted value is defined as the maximum score obtained from all the classifiers. This score shows that a human action is confused with another or not.

5. Experimental Results

This section presents the experimental results from applying our proposed approach on MSR Action 3D dataset. We also report the results on the main intensity representation (i.e. front projection). Beside, an evaluation related to selecting compensation information from other representations will be also mentioned. All the results are compared in terms of the accuracy. The best performance is highlighted in bold.

5.1. Recognize Actions from An Intensity Representation

Table 3 - lists the results from our trajectory-based approach on front representation. Interestingly, the result table indicates that this approach beats all the state-of-the-art methods based on silhouette features [12, 1], skeletal joint features [13, 3], local occupancy patterns [14, 15], normal orientation features [4] and cuboid similarity features [2]. However, the results also show that there is significant difference of the performance among the used feature descriptors.

[2] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Method	Accuracy (%)
Bag of 3D Points [?]	74.70
STOP [?]	84.80
EigenJoints [?]	82.33
Random Occupancy Patterns [?]	86.50
Local Occupancy Patterns [?]	88.20
Depth Motion Maps-based HOG [?]	91.63
Histogram of Oriented 4D Normals [?]	88.89
Depth Cuboid Similarity Feature [?]	89.30
Ours	**94.53**

Bảng 3: Results on front representation using MBH descriptor.

Action Subsets		
AS1	**AS2**	**AS3**
92.45	92.04	99.11

Bảng 4: Results on three subsets.

Consider the results on action subsets, we found that 2 subsets AS1, AS2 contain many confused actions (Table 4). For example, action-pair "hammer" and "forward punch" in AS1, or "side-boxing" and "hand catch" in AS2, as showed in Table 5. When analyzing confused actions, we found that the main cause is due to similar movements. And, since depth data is textureless, it makes recognition more difficult. That is a reason why we need compensate information from other intensity representations.

	a02	**a03**	**a05**	**a06**	**a10**	**a13**	**a18**	**a20**
a02	**0.833**	0	**0.167**	0	0	0	0	0
a03	0	**0.917**	**0.083**	0	0	0	0	0
a05	0	**0.364**	**0.636**	0	0	0	0	0
a06	0	0	0	**1.0**	0	0	0	0
a10	0	0	0	0	**1.0**	0	0	0
a13	0	0	0	0	0	**1.0**	0	0
a18	0	0	0	0	0	0	**1.0**	0
a20	0	0	0	0	0	**0.067**	0	**0.933**

(a) Action Subset 1

	a01	**a04**	**a07**	**a08**	**a09**	**a11**	**a12**	**a14**
a01	**1.0**	0	0	0	0	0	0	0
a04	**0.083**	**0.833**	**0.083**	0	0	0	0	0
a07	0	0	**0.786**	**0.071**	**0.071**	0	**0.071**	0
a08	0	0	0	**1.0**		0	0	0
a09	0	0	0	**0.133**	**0.867**	0	0	0
a11	0	0	0	0	0	**1.0**	0	0
a12	0	**0.133**	0	0	0	0	**0.867**	0
a14	0	0	0	0	0	0	0	**1.0**

(b) Action Subset 2

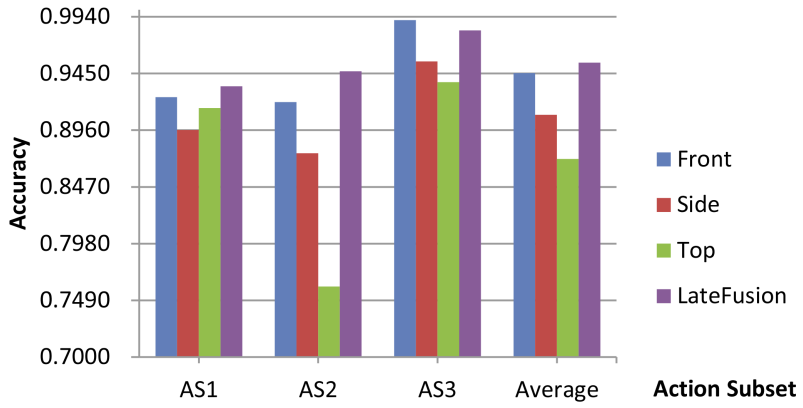
	a06	**a14**	**a15**	**a16**	**a17**	**a18**	**a19**	**a20**
a06	**1.0**	0	0	0	0	0	0	0
a14	0	**1.0**	0	0	0	0	0	0
a15	0	0	**1.0**	0	0	0	0	0
a16	0	0	0	**1.0**	0	0	0	0
a17	0	0	0	0	**1.0**	0	0	0
a18	0	0	0	0	0	**0.933**	**0.067**	0
a19	0	0	0	0	0	0	**1.0**	0
a20	0	0	0	0	0	0	0	**1.0**

(c) Action Subset 3

Bảng 5: Confusion matrices on three subsets.

5.2. Compensate Motion Information from Other Representations

Trong phần thí nghiệm này, chúng tôi bổ sung thông tin từ các representation còn lại cho representation ban đầu. Figure 2 chỉ ra a better view khi so sánh hiệu quả của Fusion với các separate representations. Expectedly, the fusion performance, which is 0.9543 accuracy, is better than all the separate ones on each representation in terms of average. Obviously, our proposed approach outperforms the mentioned state-of-the-art methods.



Hình 2: Results from using the late fusion scheme on representations

Ngoài ra, dựa vào kết quả được chỉ ra ở Figure 2, việc bổ sung thông tin từ các representations khẳng định 2 điều. The first one bảo đảm rằng kết quả nhận biết từ front representation là đóng vai trò quan trọng nhất. The second one chỉ ra rằng thông tin bổ sung từ các represenations khác có thể hỗ trợ hiệu quả cho final predictions.

6. Discussions

6.1. Evaluate the Role of Intensity Representations

6.2. The Impact of Our Method on Descriptors

- Mô tả thí nghiệm trên 3 descriptors: HOG, HOF, MBH - Nhận xét trước khi fusion - Nhận xét sau khi khi fusion: + Compensate information + Keep salient information

7. Conclusions

We proposed using the trajectory-based approach for human action recognition using depth data in this work. We evaluated our approach by using the dense trajectories motion feature on MSR Action 3D datasets. More interestingly, our proposed trajectory-based approach only applied for one representation beats all the recent state-of-the-art approaches in terms of depth data. Beside, in order to deal with confused actions due to similar movements, compensating information from other representations is proposed. Therefore, the effectiveness of our approach on depth datasets like MSR is confirmed.

A trajectory-based approach with compensating information from separate representations shows promising results. This opens a general approach to leverage intensity-based techniques for depth data. This also suggests the importance of trajectory-based motion information on human action recognition using depth data. Therefore, exploiting depth-trajectory-based motion information for human action can be beneficial for an action recognition system. This is also an interesting idea for our future work.

References

[1] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: Proceedings of the 20th ACM international conference on Multimedia, ACM, 2012, pp. 1057–1060.

[2] L. Xia, J. Aggarwal, Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 2834–2841.

[3] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 1290–1297.

[4] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 716–723.

[5] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Action Recognition by Dense Trajectories, in: IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs, United States, 2011, pp. 3169–3176.
URL http://hal.inria.fr/inria-00583818/en

[6] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Image Analysis, Springer, 2003, pp. 363–370.

[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 886–893.

[8] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[9] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: Computer Vision–ECCV 2006, Springer, 2006, pp. 428–441.

[10] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al., Evaluation of local spatio-temporal features for action recognition, in: BMVC 2009-British Machine Vision Conference, 2009.

[11] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos "in the wild", in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 1996–2003.

[12] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE, 2010, pp. 9–14.

[13] X. Yang, Y. Tian, Eigenjoints-based action recognition using naive-bayes-nearest-neighbor, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, IEEE, 2012, pp. 14–19.

[14] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3d action recognition with random occupancy patterns, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 872–885.

[15] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, M. F. Campos, Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer, 2012, pp. 252–259.