A Robust Approach for Action Recognition Based on Spatio-Temporal Features in RGB-D Sequences

Ly Quoc Ngoc, Vo Hoai Viet, Tran Thai Son, Pham Minh Hoang
Department of Computer Vision and Robotics, University Of Science, VNU-HCM, Viet Nam

Abstract*—Recognizing human action is attractive research topic in computer vision since it plays an important role on the applications such as human-computer interaction, intelligent surveillance, human actions retrieval system, health care, smart home, robotics and so on. The availability the low-cost Microsoft Kinect sensor, which can capture real-time high-resolution RGB and visual depth information, has opened an opportunity to significantly increase the capabilities of many automated vision based recognition tasks. In this paper, we propose new framework for action recognition in RGB-D video. We extract spatiotemporal features from RGB-D data that capture both visual, shape and motion information. Moreover, the segmentation technique is applied to present the temporal structure of action. Firstly, we use STIP to detect interest points both of RGB and depth channels. Secondly, we apply HOG3D descriptor for RGB channel and 3DS-HONV descriptor for depth channel. In addition, we also extract HOF2.5D from fusing RGB and Depth to capture human's motion. Thirdly, we divide the video into segments and apply GMM to create feature vectors for each segment. So, we have three feature vectors (HOG3D, 3DS-HONV, and HOF2.5D) that represent for each segment. Next, the max pooling technique is applied to create a final vector for each descriptor. Then, we concatenate the feature vectors from the previous step into the final vector for action representation. Lastly, we use SVM method for classification step. We evaluated our proposed method on three benchmark datasets to demonstrate generalizability. And, the experimental results shown to be more accurate for action recognition compared to the previous works. We obtain overall accuracies of 93.5%, 99.16% and 89.38% with our proposed method on the UTKinect-Action, 3D Action Pairs and MSR-Daily Activity 3D dataset, respectively. These results show that our method is feasible and superior performance over the-state-of-the-art methods on these datasets.

***Keywords*—Action Recognition; Depth Sequences; GMM; SVM; Multiple Features; Spatio-Temporal Features**

I. Introduction

Automatic human action recognition is attractive research topic in the fields of computer vision and machine learning since it plays an important role in the applications such as human-computer interaction, intelligent surveillance, human action retrieval system, health care, smart home, and robotics. Due to its wide range of applications, automatic human action recognition has attracted much attention in recent years [11, 19, 20, 31, 37]. The goal of human action recognition is to automatically analyze ongoing action from an unknown video (i.e. a sequence of image frames).

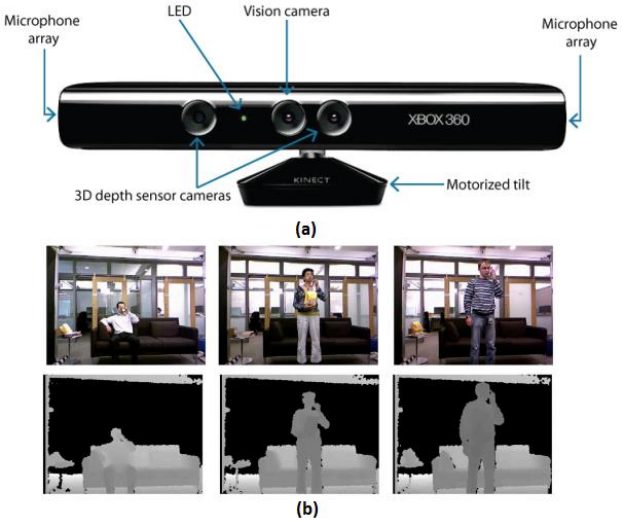


Fig. 1. Illustration of 3D camera and RGB-D data: a) Microsoft Kinect Device; b) Some examples of RGB-D data is captured by Kinect

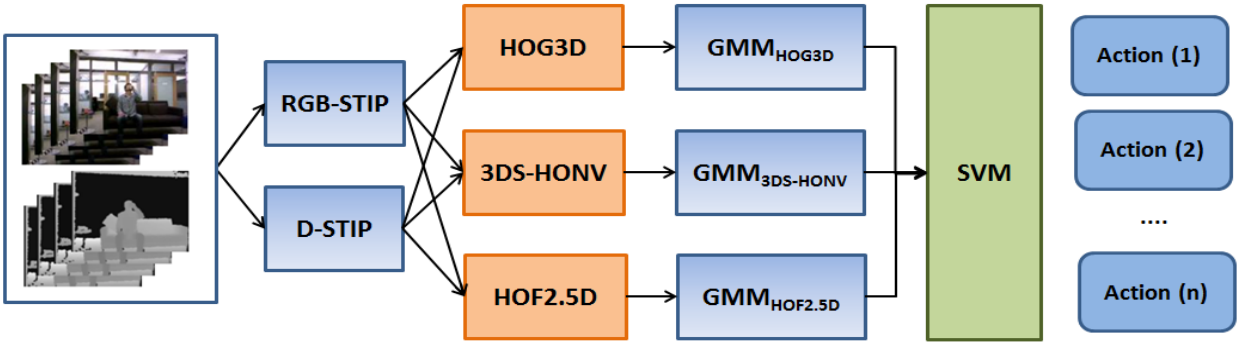


Fig. 2. Our proposal framework for action recognition in RGB-D video

Generally speaking, action recognition framework contains three main steps namely feature extraction, action representation, and pattern classification. Though much progress has been made [1, 6, 9, 11, 13, 19, 20, 31, 37], the problem of classifying action is currently of the most difficult challenges, especially in the presence of within-class variation, occlusion, background clutter, pose and lighting condition. These challenges address that the combination of different kinds of features action because action representation based on single feature is not enough to capture the imaging variations (view-point, illumination, etc…) and attributes of individuals (appearance, shape, motion, etc…).

Following the previous researches [1, 3, 6, 13, 15, 16, 18, 20], human action could be defined by structured patterns of the human's movements and poses. With the perspective, a robust feature extraction and description must capture shape and motion properties in action representation. As such, human action can be modeled by spatiotemporal features, where encode shapes and movements of the whole body or body parts, for instance temporal progression, e.g., one human action as the whole can decompose into local shapes and movements of parts. In the past two decades, a significant amount of research has been done in the area of human action recognition using a sequence of 2D images [1, 6, 13, 14, 15, 16, 18, 46]. A single spatiotemporal structure, however, is unlikely to be sufficient to represent a class of action in all but the simplest scenarios. Firstly, the execution of the action may differ from subject to subject, involving different body parts or different space-time progressions of body part movements. Secondly, the video capture process introduces intra-class variations due to occlusions or variations in camera viewpoint. Thus, the resulting space-time and appearance variations necessitate using a collection of spatiotemporal structures that can best represent the action at large. In addition, another property be also considered in action representation is evolution of action by time. It indicates that action also contains temporal structure for each action class. In this work, we apply video segmentation and max-poling technique which help to model temporal structure of action.

With the recent advent of the cost-effective Kinect, depth cameras have received a great deal of attention from researchers. It is excited to promote interest within the vision and robotics community for its broad applications [27]. The depth sensor has several advantages over the visible light camera. Firstly, the range sensor provides 3D structural information of the scene, which offers more discerning information to recover postures and recognize actions. The common low-level difficulties in RGB imagery are significantly alleviated. Secondly, the depth camera can work in total darkness. There is a benefit for applications such as patient/animal monitoring systems which run 24/7. With these benefits, the Kinect has been opened a new opportunity to improve the performance of human action recognition significantly. Recently, researchers have paid more attention to using 3D spatiotemporal features for describing and recognizing human actions [3, 4, 29, 35, 36, 44, 46] based on depth information from Kinect. Compared with conventional color data, depth maps provide several advantages, such as the ability to reflect pure geometry and shape cues, or insensitive to changes in lighting conditions. Moreover, the range sensor provides 3D structural information of the scene, which offers more discerning information to recover postures and recognize actions. These properties help depth data provide more natural and discriminative vision cues than color or texture. Furthermore, the depth images provide natural surfaces which can be extracted to capture the geometrical structure of the observed scene in a rich descriptor. However, depth sensors cannot differentiate between objects of the same depth but different color, which is trivial for color cameras. Clearly the color and depth information are correlated but also complementary to a large extent, so it would be expected to have considerable benefits by fusing them appropriately together aiming at more robust pervasive action recognition systems.

In all case, furthermore, it is commonly believed that in order to obtain high recognition rate, it is important to select an appropriate set of visual features that usually have to capture the particular properties of a specific domain and the distinctive characteristics of each action class. The most important aspect of any action recognition system is to seek an efficient action representation. The target of the feature extraction is to find an efficient and effective representation of the action which would provide robustness during recognition process. Besides, in case action representation from multiple feature vectors will need a robust method to combine feature vectors in the right way so that the system achieves good performance. In this work, we use the average-pooling technique to aggregating visual words in BOW model and the max pooling technique to aggregating the segment feature vectors into the final feature vector for action representation.

In this manuscript, we build a new framework for action recognition upon our previous works in [35] and [36]. The proposed action recognition system consisted of a flowchart is shown in Fig. 2. The main contributions of this paper are summarized as follows: Firstly, we propose a new framework for action recognition, which takes profits of multi-modal RGB-D data by fusing information from both RGB images and depth maps. The spatiotemporal features are applied to capture shape and motion. A new action presentation method is proposed by using segmentation and max pooling technique in order to capture temporal structure of human action. In addition, we use GMM instead of k-means in BOW model in order to be more distinctive for action representation. Secondly, we systematically evaluate our frameworks on three challenging datasets. Moreover, we also evaluate the impact of video segmentation technique and spatiotemporal descriptors on the performance of the system in overall accuracy.

The rest of this paper is organized as follows: Section II gives a concise review of existing works on feature extraction from a sequence of images and depth. Section III presents feature extraction and description. Section IV introduces a scheme of action representation. Section V presents action classification. Section VI shows the experiment results on relevant benchmarks. Finally, section VII draws conclusions of our work and indicates future studies.

II. Related Works

Comprehensive reviews of the previous studies can be found in [19, 20, 31, 37]. Our discussion in this section is restricted to a few influential and relevant parts of literature, with a focus on RGB, depth and RGB-D for feature extraction and representation.

There has been a lot of works on human action recognition from images in recent decades, that could be divided into two types of approaches: global-based and part-based method. Global features is temporal templates is introduced by Bobick and Davis [6]. They use the two components of motion template (MEI and MHI) and Hu Moments for representation and recognition of human movement. Xinghua Sun [42] use Zernike moments instead of Hu moments for action representation. Beside global feature approaches, the local features methods such as: histogram of 3D oriented gradients (HOG3D) [1], histogram of optical flow (HOF) [17], 3D speeded up robust features (SURF3D) [13] extends from SURF [5], 3D scale invariant feature transforms (3D-SIFT) [34] extends from SIFT [10], local trinary patterns [26] and dense trajectories with HOG/HOG/MHB[15, 16] are used to extract the most salient features (edges, corners, orientation, and motion), the choice of which would greatly influence the performance of high-level vision tasks such as recognition. Viet Vo and Ngoc Ly .al [40] also proposed hybrid features that combine local and global features for action representation. In addition, soft-weighting scheme was used to achieve more descriptive in BOW representation.

Recently, with the availability of low-cost RGB-D sensors, The similarly to recognizing human action from 2D video, the depth map-based methods rely mainly on features, either local or global, extracted from the space time volume. Lu Xia at [29] proposed DSTIP based on STIP's idea in RGB images Liet al [28] sample representative 3D points extracting the points on the contours of the projections of the 3D depth map onto the three orthogonal Cartesian planes. To reduce the size of the feature vector, the method selects a specified number of points at equal distance along the contours of the projections. Wang al. [22] fuses the skeleton information and a local occupancy pattern based on the 3D point cloud around each joint. In a different approach, J.Wang al.[23] treat an action sequence as a 4D shape and propose random occupancy pattern features, which are extracted from randomly sampled 4D sub-volumes with different sizes and at different locations. These features are robust to noise and less sensitive to occlusions. Furthermore, holistic approaches for action recognition from depth sequences are recently becoming popular. Vieira al. [3] proposed Space-Time Occupancy Patterns. The depth sequence is represented in a 4d space-time grid. Then, a scheme is used to enhance the roles of the sparse cells which typically consist of points on the silhouettes or moving parts of the body. Oreifej and Liu [33] describe the depth sequence using a histogram that captures the distribution of the surface normal orientation in the 4D space of time, depth, and spatial coordinates. The similar Oreifej's idea, Quang D. Tran and Ngoc Q. Ly [35] proposed 3DS-HONV descriptor that uses Euler angles-based quantization to create 3D histogram for action representation. This approach is simpler than the Oreifej's approach in angle quantization step. In addition, optical flows are extracted from depth channel to obtain more descriptive. Xiaodong Yang .al [41] also proposed SNV based on the surface normal orientation with adaptive spatiotemporal pyramid. Yang al. [43] project the depth maps onto three orthogonal planes and accumulate the whole sequence generating a depth motion map (DMM), the similar idea to the motion history images [6]. Histograms of oriented gradients [32] are obtained for each DMM. The concatenation of the three HOG represents an action. These features encode more information about shape, motion and context.

Nearly, some researches focus on combining both color and depth data for action recognition. Zhao Yang [47] used STIP to detect interest point and descriptor is described by combining HOG/HOF from RGB and LDP from depth data. These descriptors are used to build codebook for action representation. Quang D. Tran and Ngoc Q. Ly [35, 36] also used STIP to detect interest points but the descriptor is combined by 3DS-HONV and HOG-HOF2.5D. And, sparse coding is applied on these descriptors for representation. In [24], L. Liu proposed graph-based genetic programming by applying filters into RGB and depth data to automatically extract discriminative spatiotemporal features for action representation and SVM was used to classify actions. However, feature learning approaches have complexity in computing.

In this work, we propose a new framework for human action recognition that combines both RGB images and depth maps. This approach falls in the part-based method category. More details, we use spatiotemporal features based on the interest points that are detected by STIP in both RGB and depth channels. These interest points are represented by HOG3D, 3DS-HONV and HOF2.5D that capture shape, appearance and motion of action. Moreover, we also apply video segmentation and max pooling techniques to capture the temporal structure for action representation.

III. Feature Extraction and Description

The key to the success of part-based methods is that the interest points are distinctive and descriptive. Following the approach commonly used for local interest points in images and video, the detection and description of spatiotemporal interest points are separated in two different steps. This section describes local feature detector and descriptor used in our approach. For spatiotemporal interest points detector, we apply STIP detector [18] as a space-time extension of the Harris detector [8]. For spatiotemporal interest points descriptors, we use three descriptors such as HOG3D [1], 3DS-HONV [35], and HOF2.5D [36].

A. Prepocessing Stage

The 3D sensors such as Kinect based on structured light to estimate depth information, it is prone to be affected by noises due to reflection issues. These effects of noise could significantly decrease the overall performance of RGBD-based action recognition framework. Therefore, we firstly relieve the missing data and oultliers from the depth channel. As a result at [16], we adopted the bilateral filter for smoothing the depth channel. The bilateral filter [30] is a combination of a domain kernel, which gives priority to pixels that are close to the target

pixel in the image plane, with a range kernel, which gives priority to the pixels which have similar labels as the target pixel. This filter is often useful to preserve edge information based on the range kernel advantages. The edge is important information to represent shape of action. The bilateral filter is defined as follows:

$$I^f(x) = \frac{1}{W_p}\sum_{x_i\in\Omega} I(x_i) f_r(\|I(x_i) - I(x)\|) g_s(\|I(x_i) - I(x)\|)$$

$$W_p = \sum_{x_i\in\Omega} f_r(\|I(x_i) - I(x)\|) g_s(\|I(x_i) - I(x)\|)$$

Where I^f is the filtered image, I is the original input image , x are the coordinates of the current pixel to be filtered, Ω is the window centered in x, f_r is the range kernel for smoothing differences in intensities and g_s is the spatial kernel for smoothing differences in coordinates. In this research, f_r and g_s are supposed as Gaussian functions.

B. Interest Point Detection

The STIP or Harris3D detector was proposed by Laptev and Lindeberg in [18], which is an extension of the well-known Harris detector in the temporal dimension. The STIP detector first computes the second-moment 3 × 3 matrix μ of first order spatial and temporal derivatives. Then, the detector searches regions in the video with significant eigenvalues λ_1, λ_2, λ_3 of μ, combining the determinant and the trace of μ:\

$$H = |\mu| - k \cdot T_r(\mu)^3$$

where $|\cdot|$ corresponds to the determinant, $Tr(\cdot)$ computes the trace, and k stands for a relative importance constant factor. A commonly used value of k in the literature is $k \approx 0.005$ As we have RGB-D data, we apply the STIP detector separately on the RGB and depth channels, so we get two sets of interest points for description step.

C. HOG3D Descriptor

The HOG3D descriptor was proposed by Kläser et al. [1]. It is based on histograms of 3D gradient orientations and can be seen as an extension of the well-known SIFT descriptor [10] to video sequences. Gradients are computed using an integral video representation. Regular polyhedrons are used to uniformly quantize the orientation of spatiotemporal gradients. The descriptor, therefore, combines shape and motion information at the same time. A given 3D patch is divided into $n_x \times n_y \times n_t$ cells. The corresponding descriptor concatenates gradient histograms of all cells and is then normalized. The process of computing the HOG3D descriptor for a patch in an action depth sequence is described in Fig. 3.

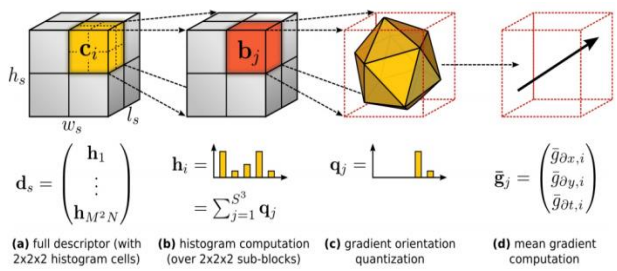


Fig. 3. Process of extracting HOG3D descriptor [1]

D. 3D Spherical Histogram of Oriented Normal Vectors (3DS-HONV) Descriptor

The 3DS-HONV descriptor was proposed by Quang D. Tran, Ngoc Q. Ly in [35], which based on HONV in [38]. The process of computing the 3DS-HONV descriptor for a patch in an action depth sequence is described in Fig. 4. For each patch, the orientation of the normal vector at each depth point is first computed, quantized in spherical coordinate by using 3 angles θ, φ ψ, and voted into a 3D histogram $q_i \in R^{b_\theta \times b_\varphi \times b_\psi}$, where b_i is the relevant bin size. Those 3D histograms at all interest points are then accumulated to create a histogram of normal occurrences distribution. Implementation of this computing process is described as follows:

1) Spatio-Temporal Surface Oriented Normal Vectors

The depth sequence can be considered as a function $R^3 \rightarrow R^1 : z = d(x, y, t)$ ($d(.)$ (is a function of depth sequence) which constitutes a surface in the 4D space represented as the set of points $\{p = (x, y, t, z)\}$ satisfying $S(p) = d(x, y, t) - z = 0$. The normal to the surface S is computed as:

$$n = \nabla S = (z_x, z_y, z_t, -1) = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, \frac{\partial z}{\partial t}, -1\right)$$

where z_x, z_y, z_t are first derivatives of the depth map z over x, y, t, which can be computed by using the finite difference approximation respectively. Since only the orientation of the normal could describe the shape of the 4D surface, the computed normal vectors are then normalized to a unit length as follows:

$$\hat{n} = \left(\hat{z}_x, \hat{z}_y, \hat{z}_t, -1/\left\|(z_x, z_y, z_t, 1)\right\|_2\right)$$

2) Spherical Quantization and 3D Histogram Representation:

In our work, the orientation of spatiotemporal surface normal is characterized by three Euler angles { $\theta, \phi,$ and ψ } ∈ [0; π] computed in spherical coordinate. The Euler angles are a classical way to specify the orientation of an object in space with respect to a fixed set of coordinate axes [21]. According to Euler's rotation theorem[21], any rotations may be described using three angles; therefore, we clarify that by just using 3 Euler angles θ, ϕ, ψ for quantization, the resulting histogram can encode any kinds of surface normal orientation in a rich representation. Euler angles-based quantization is simple, intuitive, but also more efficient than quaternions-based quantization. The approximate computation of Euler angles θ, ϕ, and ψ [21] are summarized as follows:

$$\theta = \tan^{-1}\left(\frac{\partial \hat{z}}{\partial \hat{y}} \Big/ \frac{\partial \hat{z}}{\partial \hat{x}}\right)$$

$$\phi = \tan^{-1}\left[\left(\left(\frac{\partial \hat{z}}{\partial \hat{x}}\right)^2 + \left(\frac{\partial \hat{z}}{\partial \hat{y}}\right)^2\right)^{1/2} \Big/ \frac{\partial \hat{z}}{\partial \hat{t}}\right]$$

$$\psi = \left(\left(\frac{\partial \hat{z}}{\partial \hat{x}}\right)^2 + \left(\frac{\partial \hat{z}}{\partial \hat{y}}\right)^2 + \left(\frac{\partial \hat{z}}{\partial \hat{t}}\right)^2\right)^{1/2}$$

In order to create 3D histogram representation for each depth point, the [0; π] interval is subdivided in b_θ, b_ϕ ,b_ψ bins,

so that the histogram has a total of $b_\theta \times b_\phi \times b_\psi$ bins, and is then normalized to compute the proportion of normals falling into each bin. In this work, we use the tuple of bins' size which are { $b_\theta = 5$, $b_\phi = 5$, $b_\psi = 6$}, this means a 150-dimensions 3DS-HONV for each interest point.

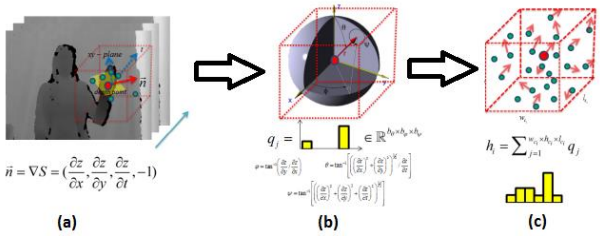


Fig. 4. Process of extracting 3DS-HONV descriptor from an interest point [36]: (a) Surface normal is computed at each point, (b) 3D histogram of normal distribution in spherical coordinate is constructed, (c) 3D histograms at all points are accumulated

E. HOF2.5D Descriptor

According to many previous researches [20, 31], the motion plays important role in human action analysis. In order to have a good representation for human action is feature descriptor must capture this property. The 3DS-HONV descriptor was proposed by Quang D. Tran and Ngoc Q. Ly in [36] which contains the human motion. This descriptor is not generated from a unified image sequence function f(x, y, z, t), but instead of capturing separately the xy-motion from pairs of RGB images and the z movements from pairs of depth channel. With assuming that is the position of each pixel in RGB images can be mapped to the related cloud point in depth maps. In specific, each pixel ($p_t^{RGB} = \{x_t^{RGB}, x_t^{RGB}\}$) in RGB-D frame F_t can be easily projected to its corresponding position ($p_t^D = \{x_t^D, x_t^D\}$) in the depth map. The process of computing HOF2.5D descriptor is described as follows: each RGB frame F_t^{RGB}, the $\{V_x, V_y\}$ components of the optical flow fields (OF) at every pixels are computed using algorithm that was proposed by G. Farnebäck algorithm [12]. In order to create OF2.5D at each calibrated pixel ($p_t = \{p_t^{RGB}, p_t^D\}$), we utilize the information of available depth maps to compute the V_z component of the OF vector as this formulation:

$$V_z = F_{t+1}^D(p_{t+1}^D) - F_t^D(p_t^D)$$

As results, each RGB-D frame F_t, we obtain a feature descriptor $D = \{D_1, V_2, \ldots, V_n\}$, where each element $D_i = \{V_x, V_y, V_z\}$ is a 3D vector that captures satisfactorily 3D motion information of a particular pixel. As a final representation for each interest point, we perform a histogram quantization using three orthogonal planes xy, xz, yz as shown in Fig. 6. The orientations of each OF2.5D are computed on three projected planes as follows:

$$\alpha_1 = \tan^{-1}\left(\frac{V_y}{V_x}\right)$$

$$\alpha_2 = \tan^{-1}\left(\frac{V_z}{V_x}\right)$$

$$\alpha_3 = \tan^{-1}\left(\frac{V_z}{V_y}\right)$$

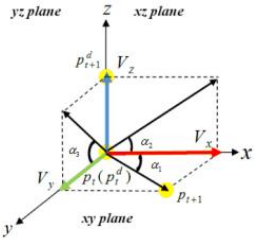


Fig. 5. Quantization scheme for computing HOF2.5D [36]

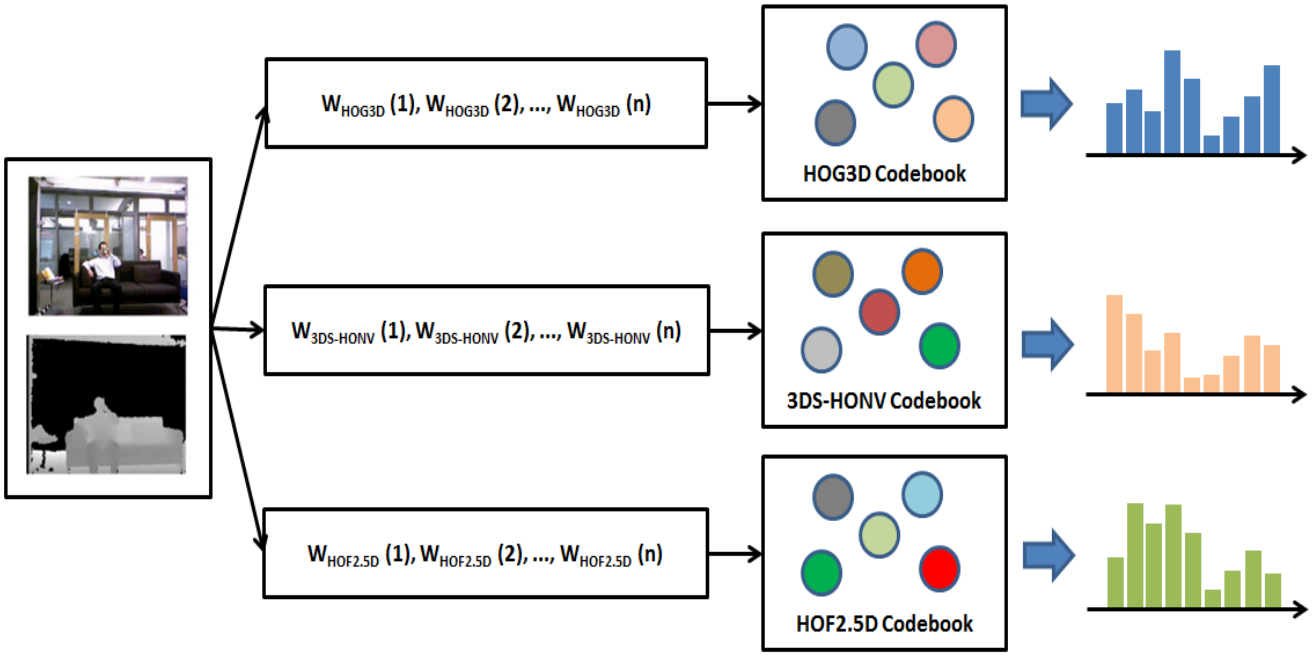


Fig. 6. Illustration of BOW for action representation in RGB-D data

We then evenly deploy $b_{\alpha 1}$, $b_{\alpha 2}$, $b_{\alpha 3}$ orientations binning on three orthogonal planes to finally generate a histogram representation of each semi-scene flow vector, namely as HOF2.5D. In all experiments, we set $b_{\alpha 1} = b_{\alpha 2} = b_{\alpha 3} = 8$. As a consequence, for each interest point descriptor, by accumulating all HOF2.5D descriptors at all pixels, we achieve a 24-bins histogram that captures the distribution of motion flows.

IV. Action Representation

A. Bag of Word

In part-based methods, a video is modeled by the bag of words (BOW) model which is the way of constructing a feature vector based on the number of occurrences of word. Each visual word is just a feature vector of patch. The major issue of BOW is vector quantization algorithms to create effective clusters. The original BOW used k-means algorithm to quantize feature vectors. Although k-means is used widely in clustering, its accuracy is not good in some cases. In addition, binary weighting for histogram of word occurrences which indicates the presence and absence of a visual word with values 1 and 0 respectively, was used. Generally speaking, all the weighting schemes perform the nearest neighbor search in the vocabulary in the sense that each interest point is mapped to the most similar visual word. Many researches argue that, for visual words, directly assigning an interest point to its nearest neighbor is not an optimal choice, given the fact that two similar points may be clustered into different clusters when increasing the size of visual vocabulary. On the other hand, simply counting the votes is not optimal as well. For instance, two interest points assigned to the same visual word are not necessarily equally similar to that visual word, meaning that their distances to the cluster centroid are different. Ignoring their similarity with the visual word during weight assignment causes the contribution of two interest points equal, and thus more difficult to assess the importance of a visual word in video.

In this work, we propose GMM instead of k-means in BOW model. We denote the parameters of the K-component GMM by $\lambda = \{w_k, \mu_k, \Sigma_k, k = 1, \dots, K\}$, where w_k, μ_k and Σ_k are respectively the mixture weight, mean vector and covariance matrix of Gaussian k and subject to $\sum_k w_k = 1$. In this work, we set K = 512 that are used in many researches. We estimate the GMM parameters on a large X training set of local spatiotemporal descriptors using the Expectation-Maximization (EM) [2] algorithm to optimize a Maximum Likelihood (ML) criterion. For GMM, soft quantization corresponds to assigning features partially to each of the GMM clusters, according to their posterior probabilities:

$$v_i = \left[P_{K|X}(1, x_i), P_{K|X}(2, x_i), \dots, P_{K|X}(K, x_i)\right]$$

$$P_{K|X}(k, x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right\}$$

where v_i is the vector of soft-counts associated with feature x_i. The soft-weights of each visual word, contributed by all features in the video, are then pooled into a histogram:

$$H(V) = F(v_1, v_2, \dots, v_n)$$

which is the final video representation (n is the number of descriptors). The standard average pooling operator aggregates word counts into bins of $H(V)$ and normalizes as follows:

$$F_{av}(v_1, v_2, \dots, v_n) = \frac{1}{n}\sum_i v_i$$

$H(V)$ represents a histogram for the video V.

B. Video Segmentation

Video segmentation is the method that divides video into fixed length segments. These approaches can be divided into two types: non-overlapping and overlapping segments. For non-overlapping segments, a video is divided into continuous and equal length segments. The method does not take account information about the semantic boundary of a segment. However, this information is important because it keeps semantic meaning of each segment.

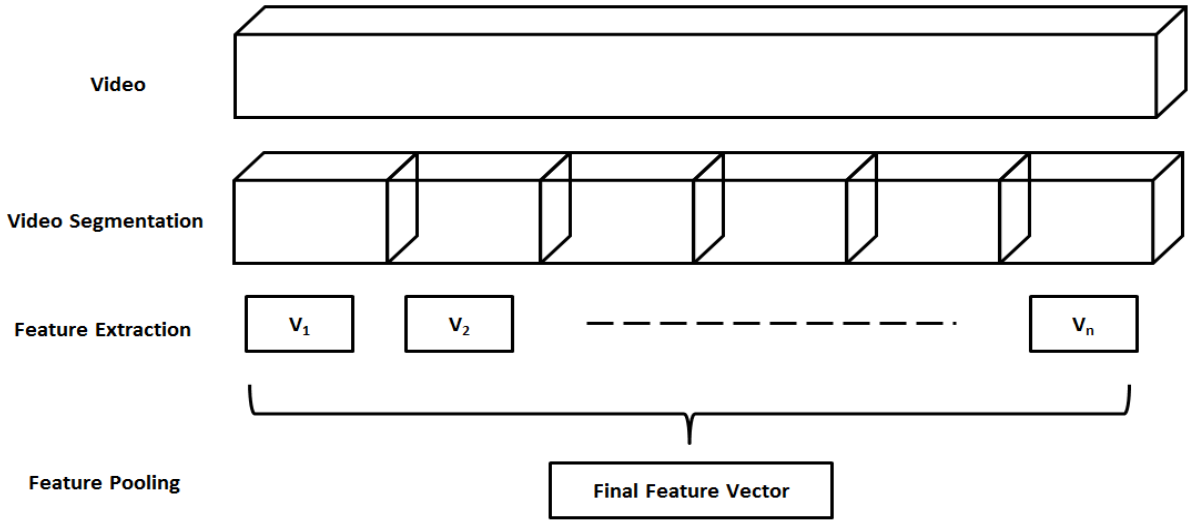


Fig. 7. Illustration of video segmentation method for action representation

This method also has the advantage that the subsequent ranking algorithm does not have to deal with problems arising from length differences. A variant of this fixed length method uses overlapping segments. In this method, a video is divided into overlapping and equal length segments. This approaches can be used that try to identify lexically and semantically coherent segments.

For all used methods we have to determine the length of the segments or the number of segments for a video. For the action recognition task as described above long segments clearly have two disadvantages: longer segments have a higher risk of covering several subtopics and thus give a lower score on each of the included subtopics. In the second place, long segments run the risk that they include the relevant fragment but that the beginning of the segment is nevertheless too far away from the jump-in point that should be found. Short segments, on the other hand, might get high rankings based on just a view words. Furthermore, short segments make the recognizing process more costly. In our approach, we choose different length segments to select the optimal one.

C. Action Represention with Feature Pooling

In video segmentation stage, we divide the video into the set of segments. Each segment is represented by three feature vectors (HOG3D, 3DS-HONV, and HOF2.5D) that are computed by BOW. We use the following temporal aggregation pools feature values for each feature dimension over time as Fig. 7. Pooling features over time means that the temporal structure of action will be modeled. With three descriptors, we have three feature vectors for action representation. Finally, we concatenate them into a final feature vector that presents for action. The vector feature will be provided to classifier to identify the label of action class which performed in video. In this research, the max pooling technique are proposed for aggregating feature vectors.

V. Action Classification

SVM is the most popular discriminative classifier and was proposed by Vladimir Vapnik [39]. It provide the state-of-art performance in many real applications such as text categorization, image classification etc… It is known as the maximum margin classifier. Consider the given training data set $\{(x_i, y_i)\}$ where i = 1, 2,…n and x_i, is N-dimensional feature vector with label y_i = +1 or −1 denoting the class it belongs to. The feature vectors are assumed to be normalized between [−1, 1] or [0, 1] to obviate the undesirable domination of any particular dimension(s) in deciding the decision boundary. SVM strives to find the hyper-plane $w \cdot x - b = 0$ that best separates the training data with regards to the distance from this hyper-plane. The optimal values for w and b can be found by solving a constrained minimization problem, using Lagrange multipliers α_i (i =1,…, n).

$$f(x) = \sum_{i=1}^{n} \alpha_i \, y_i \, K(x_i, x) + b$$

Where α_i and b are found by using an SVC learning algorithm. And $K(x_i, x)$ is a kernel function for the training sample x_i and the test sample x.

The multi-class classification problem is commonly solved by a decomposition to several binary problems for which the standard binary SVM can be used. The one-against-rest decomposition is often applied. In this case, the classification problem to k classes is countered by training k different classifiers, each one trained to distinguish the examples in a single class from the examples in all remaining classes. When it is desired to classify a new example, the k classifiers are run, and the classifier which outputs the largest (most positive) value is chosen. We use non-linear SVM with a RBF kernel which have shown good performance in many researches. In this work, we use LibSVM [7] for SVM classifier implementation. The penalty parameter is set as C = 100.

VI. Exprimental Results

We firstly evaluate the performance of the proposed approach on the three challenging 3D action datasets such as UTKinect-Action, 3D Action Pairs, and MSR-Daily Activity dataset. Then we compare our results to the state-of-the-art methods to demonstrate the superiority of the proposed approach.

Secondly, we evaluate the performance of separation of descriptors and combination of descriptors that are used BOW with k-means and GMM to yield the histogram represents for actions. Furthermore, in order study the effect of the size of the video segmentation on the final classification performance, we choose segment lengths of 10, 15, 20, and 25 frames on non-overlapping and overlapping segmentation. And we use uniform segment sampling with 50% of overlapping. Therefore, the number of segments will be doubled for each overlapping experiment.

A. UTKinect-Action dataset

UTKinect-Action dataset [25] contains 10 different action classes performed by 10 subjects, collected by a stationary Kinect sensor. The 10 action classes are: walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, clap hands. Each action was collected from 10 different persons for 2 times: 9 males and 1 female. Depth sequences are provided with resolution 320 × 240, and skeleton joint locations are also provided in this dataset. In our experiment, we used the same setting is the leave-one-out scheme in [25].

In this dataset, Table I shows the experimental results of our different methods. From the results one can see that 3D-HONV is the best descriptor in case only one descriptor is used and the fusion of HOG3D, 3DS-HONV and HOF2.5D outperforms the single descriptor in using BOW with k-means and GMM. Fig. 8 presents a comparison of the accuracy of overlapping and non-overlapping segmentation with the difference on the length of segment. The overlapping method is better than the non-overlapping method in all cases. And, the length of 15 frames for each segment achieve the best performance. Table II compares our approach results with state-of-the-art results on UTKinect-Action dataset. We can see that our result of 93.5% in accuracy is better than all previous results using the same settings. Our recognition rate is more than the current best rate by 1.6%.

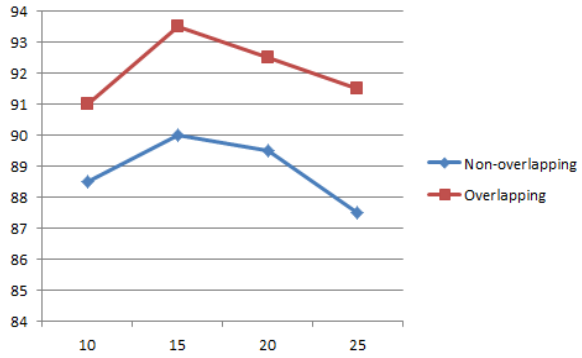


Fig. 8. Experimental results from non-overlapping and overlapping segmentation on UTKinect-Action

TABLE I. EXPERIMENTAL RESULTS OF OUR METHOD ON UTKINECT ACTION DATASET

Methods	Accuracy (%)	
	KM-BOW	GMM-BOW
HOG3D	83.5	85
3DS-HONV	88	91.5
HOF2.5D	86	87.5
Combined	**91**	**93.5**

TABLE II. COMPARISION OF THE PROPOSED METHOD WITH THE STATE OF THE ART METHODS ON UTKINECT ACTION DATASET

Methods	Accuracy (%)
HOJ3D [25]	90.92
STIPS + Joint [46]	91.9
Our approach	**93.5**

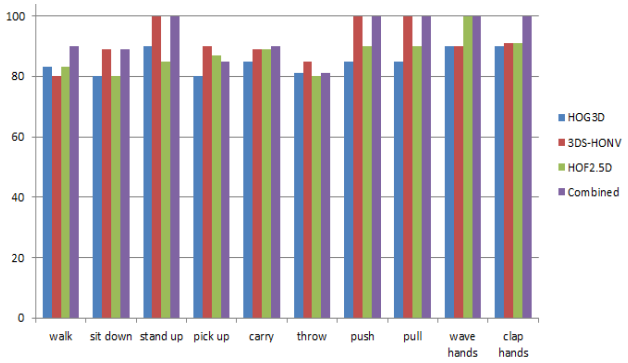


Fig. 9. Comparison of our proposed methods on UTKinect Action dataset

B. 3D Action-Pairs dataset

The 3D Action-Pairs dataset contains activities which are selected in pairs such that the two activities of each pair are similar in motion and shape. For example, "Pick up" and "Put down" actions have similar motion and shape. This dataset has six pairs of activities: "Pick up a box/Put down a box", "Lift a box/Place a box", "Push a chair/Pull a chair", "Wear a hat/Take off a hat", "Put on a backpack/Take off a backpack", and "Stick a poster/Remove a poster". The dataset includes 12 activities performed by 10 different subjects. Each action was performed three times by each subject. We used this dataset in order to emphasize two points: 1) to evaluate the performance of our proposed method in the case of actions that have similar trajectories and objects; 2) to show the advantage of using the feature fusion to enhance the recognition rate.

TABLE III. EXPERIMENTAL RESULTS OF OUR METHOD ON 3D ACTION PAIRS DATASET

Methods	Accuracy (%)	
	KM-BOW	GMM-BOW
HOG3D	92.22	94
3DS-HONV	92.27	91.38
HOF2.5D	93.61	95.28
Combined	**94.44**	**99.16**

TABLE IV. COMPARISION OF THE PROPOSED METHOD WITH THE STATE OF THE ART METHODS 3D ACTION PAIRS DATASET

Methods	Accuracy (%)
Skeleton+LOP [23]	63.33
Depth Motion Maps [43]	66.11
Skeleton + LOP + Pyramid [23]	82.22
HON4D [33]	96.67
SNV [41]	98.89
BHIM [45]	100
Our Approach	**99.16**

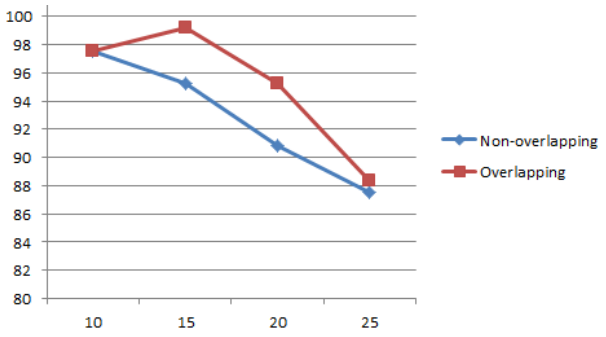


Fig. 10. Experimental results from non-overlapping and overlapping segmentation on 3D Action Pairs

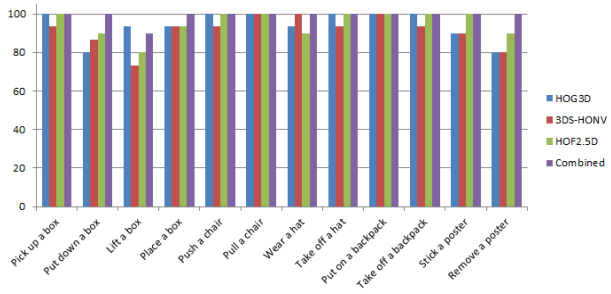


Fig. 11. Comparison of our proposed methods on 3D Action Pairs dataset

In this dataset, Table III shows the experimental results of our different methods. From the results one can see that HOF2.5D is the best descriptor in case only one descriptor is used and the fusion of HOG3D, 3DS-HONV and HOF2.5D outperforms the single descriptor in using BOW with k-means and GMM. Fig. 10 presents a comparison of the accuracy of overlapping and non-overlapping segmentation with the difference on the length of segment. The overlapping method is better than the non-overlapping method in most cases. And, the length of 15 frames for each segment obtain the best performance. Table IV compares our approach results with state-of-the-art results on 3D Action Pairs dataset. We can see that our result of 99.16% in accuracy is better than most previous results using the same settings. Our recognition rate is less than the current best rate is 100% by 0.84%.

C. MSR-Daily Activity 3D dataset

The MSR-Daily Activity 3D dataset contains 16 different human activities: drink, eat, read book, call cell phone, write on a paper, use laptop, vacuum cleaner use, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand-up, sit-down, and each subject performs an activity in two different poses: a standing pose and a sitting on sofa pose. Each pose has 160 total samples, with each subject is one sample per activity in each pose. This dataset is created to cover daily activities and human–object interactions in the living room. These tests are more challenging than the other datasets because of frequent human–object interactions.

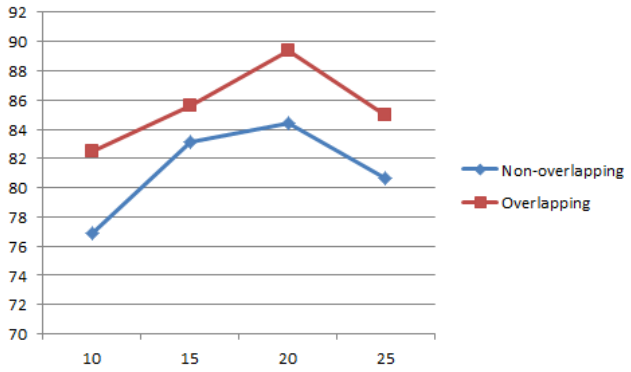


Fig. 12. Experimental results from non-overlapping and overlapping segmentation MSR-Daily Activity

TABLE V. EXPERIMENTAL RESULTS OF OUR METHOD ON MSR-DAILY ACTIVITY DATASET

Methods	**Accuracy (%)**	
	KM-BOW	**GMM-BOW**
HOG3D	78.75	80.63
3DS-HONV	84.36	87.5
HOF2.5D	81.25	82.5
Combined	**86.25**	**89.38**

TABLE VI. COMPARISION OF THE PROPOSED METHOD WITH THE STATE OF THE ART METHODS ON MSR-DAILY ACTIVITY DATASET

Methods	**Accuracy (%)**
LOP [23]	42.50
Depth Motion Maps [43]	43.13
Local HON4D [33]	80.00
Actionlet Ensemble [23]	85.75
SNV [41]	86.25
BHIM [45]	86.88
Our approach	**89.38**

In this dataset, Table V shows the experimental results of our different methods. From the results one can see that 3DS-HONV is the best descriptor in case only one descriptor is used and the fusion of HOG3D, 3DS-HONV and HOF2.5D outperforms the single descriptor in using BOW with k-means and GMM. Fig. 12 presents a comparison of the accuracy of overlapping and non-overlapping segmentation with the difference on the length of segment. The overlapping method is better than the non-overlapping method in all cases. And, the length of 20 frames for each segment achieve the best performance. Table VI compares our approach results with state-of-the-art results on MSR-Daily Activity dataset. We can see that our result of 89.38% in accuracy is better than all previous results using the same settings. Our recognition rate is higher than the current best rate is 86.88% by 2.5%.

VII. CONCLUSION

In this work, we present a new framework for action recognition in RGB-D video based on spatiotemporal features and segmentation technique. We use STIP detector to select interest points for both RGB and depth channels. Spatiotemporal descriptors consist of HOG3D, 3DS-HONV and HOF2.5D are extracted. These descriptors capture shape, appearance and motion information which are vital properties for action representation. We use GMM instead of k-means in BOW model to create more distinctive for action representation. Also, we apply segmentation and max pooling technique to capture the temporal structure of action. Our approach systematically is evaluated on several benchmark datasets such as UTKinect-Action, 3D Action Pairs, and MSR-Daily Activity 3D dataset with final recognition accuracies of 93.5%, 99.16% and 89.38% for fusion of descriptors, respectively. The experimental results have shown outcome performance compare to the-state-of-art methods in overall in most cases. For the spatiotemporal descriptors, 3DS-HONV has shown robust descriptor in most cases. However, HOG3D is better than 3DS-HONV in case that needs to distinguish these objects that have the similar shape as 3D Action Pairs dataset. And, HOF2.5D is better than HOG3D and 3DS-HONV in case that needs to differentiate these actions that have the similar motion. Thus, to improve the action recognition system, fusion of the descriptors is the best way. For the part-based model, from experimental results also show that GMM is more powerful than k-means when using to create visual words in BOW model. For segmentation method, in addition, we indicate that overlapping method performs the best in most cases. And, the length of segment also impacts to the performance of the system. However, the length is not fixed for all the dataset that it depends on the descriptors are used and the nature of dataset. In this work, the experimental results indicate that the length of segment is 15 and 20 frames are the best performances.

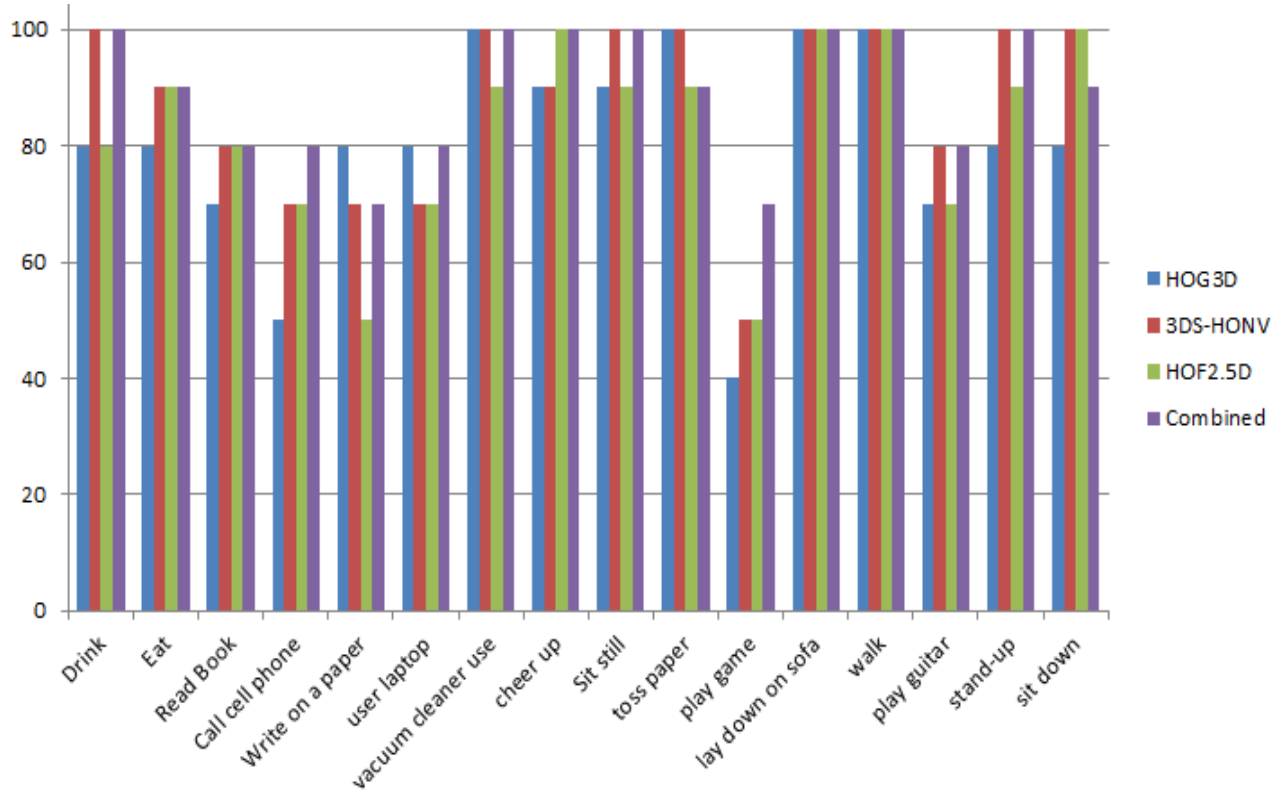


Fig. 13. Comparison of our proposed methods on MRS-Daily Activity dataset

In summary, the key problems of this research are summed up as follows: firstly, we have explored the utility of spatiotemporal features derived from RGB and depth information. These features are extracted to capture both shape and motion in action. Secondly, GMM used to instead of k-means in BOW model to have more distinctive and descriptive for action representation. Finally, we have modeled temporal structure of action based on video segmentation and max pooling technique.

In the future, we will investigate new method to improve appearance, motion properties as well as consider the impact of context and evolution of human when performing the action.

ACKNOWLEDGMENT

This research is funded by Vietnam National University, Ho Chi Minh City (VNU-HCM) under grant number *B2014-18-02*.

REFERENCES

[1] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D- gradients. In BMVC , 2008.

[2] A. P. Dempster, N. M. Larid, and D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society Series B(Methodological), vol. 39(1), pp. 1–38, 1977.

[3] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. M. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in CIARP, 2012, pp. 252–259.

[4] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in ICCV, 2011.

[5] Bay, H., Tuytelaars, T., and Van Gool, L, "SURF:Speeded Up Robust Features", In Proceedings of the Ninth European Conference on Computer Vision, May, 2006.

[6] Bobick, A. and Davis, J.: The Recognition of Human Movement Using Temporal Templates. IEEE Trans. On Pattern Analysis and Machine Intelligence, 2001.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. ACM TIST, 2(27):1–27, 2011.

[8] C. Harris and M.J. Stephens. A combined corner and edge detector. In Alvey Vision Conference, 1988.

[9] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In ICPR , 2004.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–110, 2004.

[11] Daniel Weinland, Remi Ronfard, Edmond Boyer, A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition, INRIA, 2010.

[12] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in Proc. 13th Scandin. Conf. Image Anal. (SCIA), 2003, pp. 363–370.

[13] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In ECCV, 2008.

[14] Hao Zhang, Wenjun Zhou, Christopher Reardon, Lynne E. Parker Simplex-Based 3D Spatio-Temporal Feature Description for Action Recognition, CPVR2014.

[15] H. Wang, A. Kl ¨ aser, C. Schmid, and C.-L. Liu: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, Mar. 2013.

[16] Heng Wang and Cordelia Schmid, Action Recognition with Improved Trajectories. ICCV 2013.

[17] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In CVPR, 2008.

[18] I. Laptev and T. Lindeberg. Space-time interest points. In ICCV, 2003.

[19] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," ACM Computing Surveys, Apr. 2011.

[20] J. Aggarwal and Q. Cai. Human motion analysis: a review. In Nonrigid and Articulated Motion Workshop, pages 90–102. IEEE, 1997.

[21] J. Diebel, "Representing attitude: Euler angles, unit quaternions, and rotation vectors," 2006.

[22] J. Wang, Z. Liu, J. Chorowski, Z. Chen, , and Y. Wu. Robust 3d action recognition with random occupancy patterns. In ECCV, 2012.

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

[24] L. Liu, L. Shao, Learning discriminative representations from RGB-D video data, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2013.

[25] L. Xia, C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pages 20– 27. IEEE, 2012.

[26] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In ICCV, 2009.

[27] Leandro Cruz, Djalma Lucio, Luiz Velho: Kinect and RGBD Images: Challenges and Applications. SIBGRAPI Tutorials, pp 36-49, 2012.

[28] Li, W., Zhang, Z., and Liu, Z. Action Recognition based on A Bag of 3D Points. IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2010.

[29] Lu Xia and J.K. Aggarwal, Spatio-Temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera, CVPR 2013.

[30] M. Camplani and L. Salgado, Efficient spatio-temporal hole filling strategy for kinect depth maps, A. M. Baskurt and R. Sitnik, Eds., vol. 8290, no. 1. SPIE, 2012, p. 82900E.

[31] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, Juergen Gall, A Survey on Human Motion Analysis from Depth Data. Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications Lecture Notes in Computer Science Volume 8200, 2013, pp 149-187

[32] Navneet Dalal, Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR, 2005.

[33] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in CVPR, 2013.

[34] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In MULTIMEDIA, 2007.

[35] Quang D. Tran, Ngoc Q. Ly, Sparse Spatio-Temporal Representation of Joint Shape-Motion Cues for Human Action Recognition in Depth Sequences, 2013 IEEE RIVF International Conference on Computing & Communication Technologies -Research, Innovation, and Vision for the Future (RIVF), 2013.

[36] Quang D. Tran, Ngoc Q. Ly, An Effective Fusion Scheme of Spatio-Temporal Features for Human Action Recognition in RGB-D Video, IEEE-International Conference on Control, Automation and Information Sciences (ICCAIS), 2013.

[37] Ronald Poppe, A survey on vision-based human action recognition, Image and Vision Computing 28, 976–990, 2010.

[38] S. Tang, X. Wang, X. Lv, T. X. Han, J. M. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in ACCV, 2012.

[39] V.Vapnik, "Statistical learning theory", *John Wiley and Sons*, New York, 1998.

[40] Viet Vo, Ngoc Ly, Robust human action recognition using improved BOW and hybrid features, 2012 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2012

[41] Xiaodong Yang, YingLi Tian. Super Normal Vector for Activity Recognition Using Depth Sequences, CVPR, 2014.

[42] Xinghua Sun, Mingyu Chen, Alexander Hauptmann,Action Recognition via Local Descriptor and holistic features, Computer Vision and Pattern RecognitionWorkshop, IEEE, 2009.

[43] Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps based histograms of oriented gradients. In: ACM International Conference on Multimedia. (2012) 1057-1060.

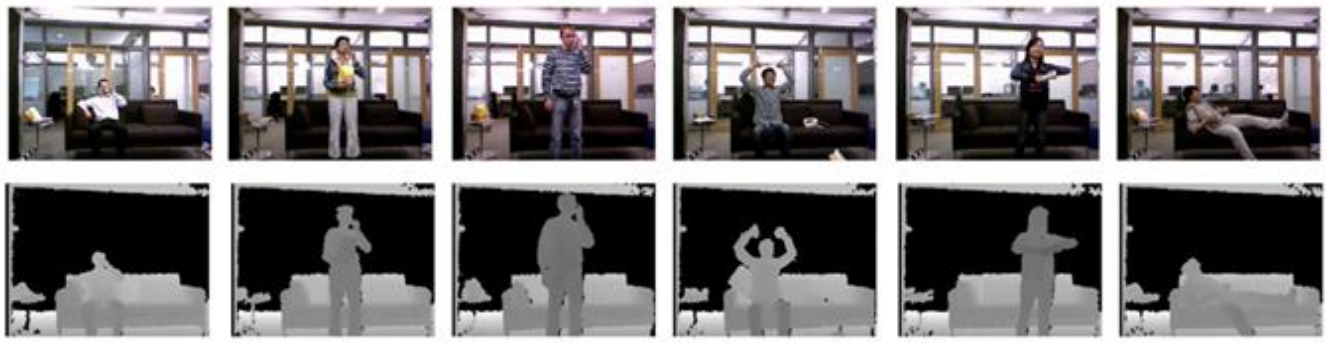
[44] Yen-Yu Lin, Ju-Hsuan Hua, Nick C. Tang, Min-Hung Chen, Hong-Yuan Mark Liao, Depth and Skeleton Associated Action Recognition without Online Accessible RGB-D Cameras, CVPR 2014.

[45] Yu Kong, Yun Fu, Bilinear Heterogeneous Information Machine for RGB-D Action Recognition, CVPR 2015.

[46] Yu Zhu, Wenbin Chen, and Guodong Guo, Fusing Spatiotemporal Features and Joints for 3D Action Recognition, 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops.

[47] Zhao Yang, Liu Zicheng, Cheng Hong. RGB-Depth Feature for 3D Human Activity Recognition, Communications, China, 2013.

(a)



(b)

Fig. 14. Some sample frames from UTKinect-Action (a) and MSR-Daily Activity 3D (b) dataset