Local and Global feature Descriptors Combination from RGB-Depth Videos for Human Action Recognition

Rawya Al-Akam and Dietrich Paulus
*Active Vision Group, Institute for Computational Visualistics, University of Koblenz-Landau,
Universitatsstr. 1, 56070 Koblenz, Germany*

Keywords: RGBD Videos, Global Features, Local Features, Action Recognition.

Abstract: This paper attempts to present human action recognition through the combination of local and global feature descriptors values, which are extracted from RGB and Depth videos. A video sequence is represented as a collection of spatio and spatio-temporal features. However, the challenging problems exist in both local and global descriptors for classifying human actions. We proposed a novel combination of the two descriptor methods, 3D trajectory and motion boundary histogram for the local feature and global Gist feature descriptor for the global feature (3DTrMBGG). To solve the problems of the structural information among the local descriptors, and clutter background and occlusion among the global descriptor, the combination of the local and global features descriptor is used. In this paper, there are three novel combination steps of video descriptors. First, combines motion and 3D trajectory shape descriptors. Second, extract the structural information using global gist descriptor. Third, combines these two descriptor steps to get the 3DTrMBGG feature vector from spatio-temporal domains. The results of the 3DTrMBGG features are used along with the K-mean clustering and multi-class support vector machine classifier. Our new method on several video actions improves performance on actions even with low movement rate and outperforms the competing state-of-the-art -temporal feature-based human action recognition methods.

1 INTRODUCTION

Human action recognition from the video has drawn attention from researchers in the field of computer vision, machine learning, and pattern recognition fields to characterize the behavior of persons. Also, it has been used in many applications fields like video surveillance, robotics human-computer interaction and a variety of systems that involve interactions between persons and computers (Chen et al., 2013). Therefore, the capability to design a machine that able to interact intelligently with a human-inhabited environment is playing important attention in recognizing activities of people from the video frames (Adem Karahoca, 2008). In the last few years, the human activity recognition researchers are essentially focused on recognizing human action from videos frames captured by traditionally visible light cameras (Yang and Tian, 2016). But recently, the action recognition research has entered a new phase technological advances and the appearance of the low-cost depth sensor like Microsoft Kinect (Soares Beleboni, 2014). These depth sensors have many advantages over RGB cameras, like provide useful extra information and 3D structural information to the complex problem of action recognition, as well as color image sequences in real time, also the depth information is invariant to lighting and color variations and can even work in total darkness that which makes it possible to find the solution for conventional problems in human action recognition (Kim et al., 2014) (Li et al., 2010). Of course, the depth camera also has an intense limitation which can be partially enhanced by fusion of RGB and depth. But all these advantages make it interesting to incarnate the RGBD cameras into a lot of challenging tasks.

Existing approaches for recognizing human action can be classified into two main categories: local feature approach and global approach. The essential steps of local feature approach are as follows: First, extracting local sub-regions or interest points from video or image sequences; Second, constructing descriptors to describe the local property such as SIFT (Bakheet and Al-Hamadi, 2016) (Azhar et al., 2015), optical flow (Nourani-Vatani et al., 2012); third, transforming these local feature vector descriptors to the

Al-Akam R. and Paulus D.
Local and Global Feature Descriptors Combination from RGB-Depth Videos for Human Action Recognition.
DOI: 10.5220/0006525002650272
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 265-272
ISBN: 978-989-758-276-9

form of words using a bag of words method and an action is represented by words combination; last, the final features are input to the classifier to perform action recognition. While global approaches generally extract the whole human body information, like MHI (motion history images) (Bobick and Davis, 2001), silhouette-based feature (Dedeoğlu et al., 2006) and MHV (motion history volume) (Weinland et al., 2006). The performance may have an influence on partial occlusion and background clutter.

The contribution of this paper is to address the problem of recognizing human behavior by using a novel method based on combining 3D Trajectory shape features from RGB-D video frames and both Motion Boundary Histogram (MBH) and global Gist features from depth video frames, which mean that the combination of the local and global descriptors is proposed. Local descriptors present a video as features extracted from a collection of patches, ideally invariant to clutter, occlusion, appearance change, and possibly to the rotation and scale change as well. While the main reason for using the global Gist descriptor is illustrated in three steps: First, Gist feature captures global structural information by filtering an image with different scales and orientations. In the case of realistic scenarios, it can be extracted more reliably than silhouettes feature proposed in (Wang and Suter, 2007). Second, the computational time of Gist feature is much less than optical flow features used in (Wang and Mori, 2010). Third, Gist feature can be represented as the concatenation of several local grids with implicit location information. In Figure 1 illustrate the general steps of action recognition progress.

2 RELATED WORK

Human action recognition has been a very active research topic over the recent years, there are different approaches have been introduced to address the action recognition problem. Depending on feature representations, action recognition systems can be categorized into ones based on shape and appearance-based representations (Niebles and Fei-Fei, 2007)(Ji et al., 2013), which appearance and shape-based approaches build models to represent actions and use these models in recognition tasks. Optical-flow based representations, that depends on calculating the optical flow to encode the energy of the action and represent actions as histograms of optical flow (Dalal et al., 2006)(Chaudhry et al., 2009). Dense-trajectory-based methods (Wang and Schmid, 2013), in this method, each dense trajectory is often represented as a vector of coordinates, this means that it is consequently missing the structural relationship between different trajectories. The researchers in (Shi et al., 2015) solved this problem by using deep trajectory descriptor to improve action recognition on RGB dataset. In another hand, some research depends on global descriptors, which is used for computing the structural information (shape, texture, and contour representation) from the images (Somasundaram et al., 2014). As in (Wang et al., 2013b), the global features was computed by applying a bank of 3D spatio-temporal filters on the frequency spectrum of video frames to extract the information about the motion and scene structure. The global Gist descriptor (Wang et al., 2014b) (Wang et al., 2013b) was used to extract the global grid-based features from image sequences to represent the human action recognition. The combination of local and global descriptor was improved in (Lisin et al., 2005) (Wang et al., 2015a) by using different combination strategy to improve the action classification tasks from RGB video frames. To address the problem of action recognition, some researcher used depth information related to RGB videos like a Kinect dataset to incorporate the RGBD cameras into more tasks. (Koperski et al., 2014) and (Xiao et al., 2014) which extended the 2D trajectory method to 3D trajectory by using RGB-D dataset. whereas we use the 3D Trajectory combined with global gist descriptor from RGBD video dataset.

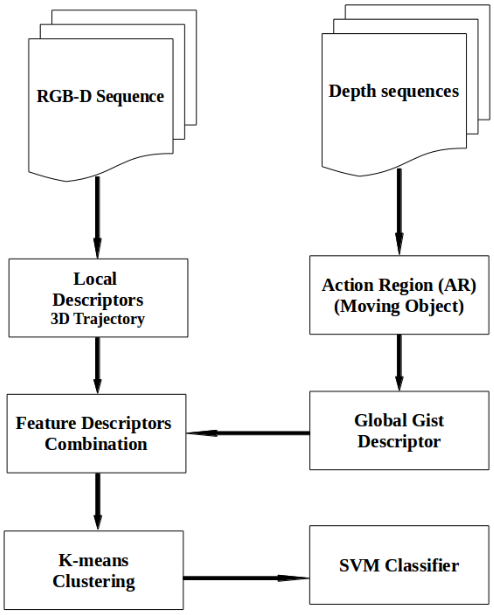
RGB-D Sequence
Depth sequences
Local Descriptors 3D Trajectory
Action Region (AR) (Moving Object)
Feature Descriptors Combination
Global Gist Descriptor
K-means Clustering
SVM Classifier

Figure 1: General structure of our proposed method.

3 PROPOSED APPROACH

In this section, we describe our proposed approach for action recognition. This approach includes four steps. First, improved 2D dense trajectory method on RGB videos (Wang and Schmid, 2013) to 3D trajectory on RGB-D videos by mapping the 2D positions of the dense trajectories from RGB video frames to the corresponding positions in the depth video frames. That can recover the 3D trajectory of the tracked interest points, which captures important motion information along the depth direction. To represent the 3D trajectories, we apply motion boundary histogram (MBH) to depth direction and propose 3D trajectory shape descriptors. Second, Apply background subtraction on depth frames, only the area where human located is extracted from each frame of a video. Which called Action-Region (AR). Third, A Gist features are computed to each AR. The global long feature is composed of $m \times n$ Gist vectors located in non-overlapping local grids, where $m \times n$ is the number of the grids in an AR. finally, local and global features are then extracted and combined to encode video information. In the training step, features extracted from the training set are clustered using K-means (Xie et al., 2011) to generate a bag of words. Histograms based on occurrences of bag words in the training set are used as features to train classifiers. Finally, a multi-class SVM classifier is used to achieve action recognition. The following subsections explain each step in detail.

3.1 Preprocessing Input Data

The input videos which represented by color and depth data are analyzed as a frame sequence to extract the important features that presented in each video sequences. In order to reduce the computational complexity of the system in our work, we are used a lower resolution of 320×240. The depth maps data captured by the Kinect camera are often noisy due to imperfections related to the Kinect infrared light reflections. For reducing noise and to eliminate the unmatched edges from the depth images, two different denoising methods are used, the depth inpainting approaches (Xue et al., 2017) to paint the missing regions of specific kinds (e.g. occlusions, missing caused by sensor defects or holes caused by object-removal) and a spatio-temporal bilateral filtering to smooth depth images. The joint-bilateral filtering proposed in (Zhao et al., 2012) is formulated as in formula (1):

$$\hat{D}_{(P)} = \frac{1}{K_{(p)}} \sum_{q \in \Omega_p} f(p,q) g(\| \hat{D}_m(p) - \hat{D}_m(q) \|) \, h(\| I_{(p)} - I_{(q)} \|) \qquad (1)$$

where $f(p,q)$ denotes a domain term to measure the closeness of the pixels p and q. The function $g(.)$ refers to the depth range term that computes the pixel similarity of the modeled depth map. The function $h(.)$ perform an intensity term for measuring the intensity similarity. Moreover, Ω_p represents the spatio neighborhood of position p.

3.2 Feature Description

For the feature extraction step, we used two types of descriptor methods: local and global descriptors, which are represented by the 3D trajectory, MBH, and global Gist descriptor. The next subsection explains these methods in detail:

3.2.1 Local Feature Extraction

In this step, the 3D trajectory features are extracted from RGB-D video frames by extending the 2D trajectory in (Wang and Schmid, 2013)[1]. However, the dense trajectories achieve promising results on the 2D image plane, but, the motion information along the depth direction is missing in this work. To relieve this problem, we estimate the depth motion, by assume for each point P on the 2D RGB image plane, let $\overrightarrow{x} = (x_t, y_t)$ be its position at frame t by mapping $\overrightarrow{x} = (x_t, y_t)$ to the depth frame I_t^z, and the scene flow estimation which represent the 3D motion field of a scene (w_t) as illustrated in Algorithm 1 (Xiao et al., 2014).

However, the MBH descriptors (Wang et al., 2013a) in the directions of (MBH_x and MBH_y) gave a good performance in RGB. We extracted MBH from the depth direction (MBH_z). For the RGB-D data, besides these two MBH descriptors direction above. One major advantage of MBH_z is robustness to camera motion along the depth direction. We follow the model of MBH_x and MBH_y in (Wang and Schmid, 2013) to extract MBH_z.

[1] http://lear.inrialpes.fr/people/wang/improved _trajectories

Algorithm 1: 3D trajectories and feature extraction.

```
input  : 2D dense trajectories
Extend 2D dense trajectories to 3D trajectories
for each 2D trajectory $Tr^i_{2D}$ do
    for each 2D trajectory point $P^j_{2D}$ do
        • Map $P^j_{2D}$ to the depth frame $I^z_t$
        • Estimate the scene flow $w_t$ of $P^j_{2D}$
    end
    Extend $Tr^i_{2D}$ to 3D trajectory $Tr^i_{3D}$ with
    availability check
    If $Tr^i_{3D}$ is available, extract activity features
    along it
end
output: 3D trajectories and RGB-D activity features
```

3.2.2 Global Feature Extraction

In addition to the local descriptors, the global-based descriptors encode more spatio and temporal information within video sequences, which is representing the actions based on holistic information about the action and scene in each video frames. This descriptor method often requires the localization of the human body through alignment, background subtraction or tracking (Solmaz et al., 2013), and directly extract and describe the whole properties of human silhouettes or contours (Wang et al., 2015a).

In this work, The Gist descriptor is used in the spatio-temporal domain to extract the structural feature from depth frames. GIST descriptor was proposed by (Wang et al., 2015a) and has been widely used in object classification and image retrieval (Ikizler-cinbis and Sclaroff, 2010). The global feature is computed as follow (Wang et al., 2015b):

- Given a video with n frames, the human region in each frame is separated from the background, which is called Action-Region (AR). i.e extraction to the foreground from all frame sequences (moving object).
- Gist descriptor is computed using a cluster of Gabor filters as shown below, which the AR Convolve with 32 Gabor filters at 4 scales, 8 orientations, producing 32 feature maps.

$$G(x,y) = exp(\frac{-(x_1^2+y_1^2)}{2\sigma^{2(l-1)}})cos(2\pi(F_x x_1 + F_y y_1)), \tag{2}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} cos\ \theta_l & sin\ \theta_l \\ -sin\ \theta_l & cos\ \theta_l \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \tag{3}$$

where (F_x, F_y) represent the frequency of the sinusoidal component. θ_l and l are the values of the orientations and scales respectively.
- Divide each feature map into 16 regions (by a 4×4 grid), and then average the feature values within each region.
- Concatenate the 16 averaged values of all 32 feature maps, resulting in a $16 \times 32 = 512$ GIST descriptor dimensions.

3.3 Action Recognition

In order to represent performance comparison for our action recognition system, we use a Support Vector Machines (SVM). It is one of the classification method, that is used hyper-planes in high dimensional space to divide the training data with the largest margin of the points and it is powerful and excessively used in the Bag-of-Words (BoW) state (Uijlings et al., 2015). In this work, to generate the Bag-of-Words (BoWs), each feature description of the video frame is compared to each centroid of the cluster in the dictionary using Euclidean distance measure e as formulated in equation (4) (Kundu et al., 2014):

$$e = \sum_{j=1}^{m}\sum_{i=1}^{n} \|X_i{}^{(j)} - C_j\|^2 \tag{4}$$

where $\|X_i{}^{(j)} - C_j\|^2$ is the selected distance measure between the feature vector point and the clustering center C_j. m is the clustering center length and n is the feature vector size. To classify video actions, we used multi-class SVM with radial basis function (RBF) kernel, An RBF is mapping the data into an infinite dimensional Hilbert space. The RBF is a Gaussian distribution, calculated as in formula (Mahesh et al., 2015):

$$k_{RBF}(\vec{x}, \vec{z}) = e^{-(\vec{x}-\vec{z})/2\sigma^2} \tag{5}$$

where $k_{RBF}(.)$ is the kernel function, $\vec{x}$ and $\vec{z}$ are the input vectors. The Bag of words vectors for all the videos are computed in training stage and labels are appended according to the class. This bag of words vectors are fed into the multi-class SVM in order to train the model that is further used in the testing stage for human action recognition as shown in Figure 2. In this figure, we show the steps of the dictionary generation after the feature vector extraction steps and this is the important step of BoW method. To generate the dictionary, the K-means clustering algorithm was used. The size of the dictionary is important for the recognition process, when the size of the dictionary is set too small then the BoW model cannot express all the keypoints and if it is set too high then it might

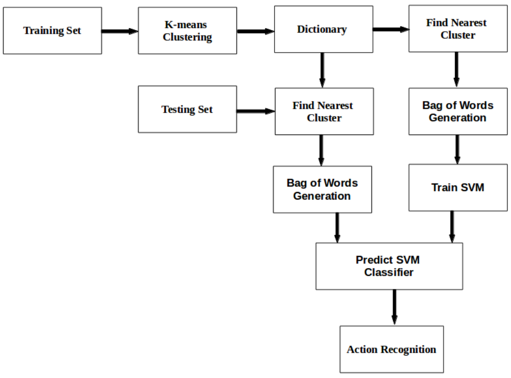

Figure 2: Action Classification Framework.

lead to over-fitting and increasing the complexity of the system . The K-means clustering was applied on all BoW from training videos, the K has represented the dictionary size. The centroids of each cluster are combined to make a dictionary. In our method, we got the best result with a value of $K = 100$ as a dictionary size (Al-akam and Paulus, 2017).

4 EXPERIMENTATION AND EVALUATION

In this section, we illustrate the two types of datasets that used in our approach and the experimental results.

4.1 Dataset

To assess the performance of our approach, the public dataset like MSR DailyActivity 3D Dataset[2] and Online RGBD Action dataset (ORGBD)[3] are used. This dataset is widely used in the evaluation of action recognition algorithm.

4.1.1 MSR-DailyActivity3D Dataset

The MSR DailyActivity 3D Dataset is designed to cover human's daily activities in the living room and it is captured by a Kinect device and it is represent a 3D daily activity dataset (Wang et al., 2014a). This dataset contains 16 action and 10 subjects; each subject performs each activity in two different poses: *drinking, eating, read a book, call cell phone, writing on a paper, using laptop, using vacuum cleaner, cheer up, sitting, still, tossing paper, playing game, laying down on sofa, walking, playing guitar, stand up and sit down.* See Figure 3 (Al-akam and Paulus, 2017).

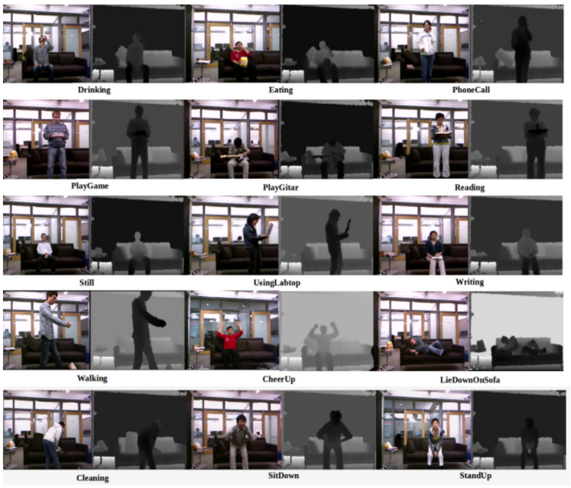

Figure 3: Sample frames of MSR-Daily Action 3D Dataset.

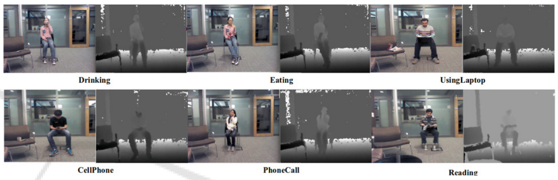

Figure 4: Sample frames of Online RGBD Action Dataset.

4.1.2 Online RGBD Action Dataset

The Online RGBD Action dataset (ORGBD) (Yu et al., 2015) are captured by the Kinect device. Each action is represented by 16 subjects and each action is performed two times. In this dataset, the seven types of actions are recorded in the living room: *drinking, eating, using a laptop, picking up a phone, reading phone (sending SMS), reading a book and using a remote.* As shown in Figure 4 (Al-akam and Paulus, 2017). We compare our approach with the state-of-the-art methods on the same environment test setting, where half of the subjects are used as training data and the rest of the subjects are used as test data.

4.2 Experiments and Results

To improve the performance of the proposed approach (3DTrMBGG), We test the two types of the dataset as presented in sec 4.1.1 and sec 4.1.2, and try to experiment the results in three different models:

- Local descriptors, which means that the feature vectors are extracting using 3D Trajectory from RGB-D combined with 3D motion boundary histogram (3DTrMB) from depth information. The feature descriptor dimensions from 3DTrMB are 628 for each video action.

[2] http://www.uow.edu.au/w̃anqing/#Datasets

[3] https://sites.google.com/site/skicyyu/orgbd

- Global descriptor, in this step the feature vectors are extracting using Gist descriptor from depth information(3DGIST). The feature descriptor dimensions from 3DTrMB are 512 for each video action.
- Local and Global descriptor combination. The feature vectors are extracting by the combination between 3DTrMB and 3DGIST feature values in one vector 3DTrMBGG for each video action on RGB-D videos. The feature vector dimensions from 3DTrMBGG are 1140 for each video action.

Table 1 shows the comparison results of the three steps illustrated above. For each step, we perform K-means clustering to the feature descriptors values, which yields codebooks with size K. In our experiments, we set K as 100. Finally, the multi-class SVM classifier is applied to compute the accuracy values from the proposed method.

From our test results, we notice that the comparison between this three steps results show that the combination between local and global give the accuracy as 95.62% and 95.62% using MSR-DailyActivity3D and Online RGBD action dataset, while the results from using local or global descriptors separately is less than the combined results on the same dataset.

In Table 2 and Table 3 compares our results with the existing state of the art using the same dataset with different action recognition methods.

5 CONCLUSIONS AND FUTURE WORKS

This paper proposed a novel action representation by combining 3D trajectory, motion boundary histogram and global Gist feature (3DTrMBGG) using the bag of word (BoW) model. And also, this work improved the benefit of combining a set of local features vector with a single global feature vector in a suitable manner. The advantage is that not an only main global attribute of human action are kept but also the impact of occlusion and noise is reduced. Evaluations on two challenging realistic scenario's action datasets, 3D MSR-Daily action, and Online-RGBD datasets, prove that our proposed method has the capability of recognizing diverse actions in a large variety of RGB-D videos. From the Experiment results showed that the proposed scheme can effectively recognize the similar action with high movement rate as walking, cleaning, etc., and improves performance on actions with low movement rate like: reading, using laptop, etc. It gives a 95.62% on 3D MSR Daily action dataset and 97.62% on ORGBD dataset as a recognition rates.

As the future work, we will focus on salient object detection method to detect salient objects in video frames and only extract features for such objects and will combine a new feature vector values like local binary pattern (LBP). Also for the classification task, we will use convolution neural networks (CNN), and K-nearest neighbor (KNN).

ACKNOWLEDGEMENTS

This work was partially supported by Ministry of Higher Education and Scientific Research (MHESR), Iraq, and University of Koblenz Landau, Germany.

REFERENCES

Adem Karahoca, M. N. (2008). Human motion analysis and action recognition. In *1st WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering*, pages 156–161.

Al-akam, R. and Paulus, D. (2017). RGBD Human Action Recognition using Multi-Features Combination and K-Nearest Neighbors Classification. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(10):383–389.

Azhar, R., Tuwohingide, D., Kamudi, D., Sarimuddin, and Suciati, N. (2015). Batik Image Classification Using SIFT Feature Extraction, Bag of Features and Support Vector Machine. *Procedia Computer Science*, 72:24–30.

Bakheet, S. and Al-Hamadi, A. (2016). A Discriminative Framework for Action Recognition Using f-HOL Features. *Information*, 7(4).

Bobick, A. F. and Davis, J. W. (2001). The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267.

Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. (2009). Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pages 1932–1939.

Chen, G., Zhang, F., Giuliani, M., Buckl, C., and Knoll, A. (2013). Unsupervised Learning Spatio-temporal Features for Human Activity Recognition from RGB-D Video Data. In *Social Robotics*, pages 341–350, Cham. Springer International Publishing.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Computer Vision – ECCV*, pages 428–441, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dedeoğlu, Y., Töreyin, B. U., Güdükbay, U., and Çetin, A. E. (2006). Silhouette-Based Method for Object

Table 1: Our Experiment Results on RGB-D dataset.

Dataset	Methods	Accuracy
MSR-DailyActivity3D	**3DTrMB**	**92.0%**
	3DGist	**93.33%**
	3DTrMBGG	**95.62%**
ORGBD	**3DTrMB**	**81.52%**
	3DGist	**83.81%**
	3DTrMBGG	**97.62%**

Table 2: Comparison of Recognition Accuracy with other Methods on MSR Daily Activity 3D Dataset.

Methods	Accuracy
CHAR (Zhu et al., 2016)	54.7%
Discriminative Orderlet (Yu et al., 2015)	60.1%
Relative Trajectories 3D (Koperski et al., 2014)	72.0%
Moving Pose (Zanfir et al., 2013)	73.80%
CoDe4D+Adaptive MCOH (Zhang and Parker, 2016)	86.0 %
Unsupervised training (Luo et al., 2017)	86.9 %
3DTrMB	**92.0%**
3DGist	**93.33%**
3DTrMBGG	**95.62%**

Table 3: Comparison of recognition accuracy with other methods on Online RGBD (ORGBD) Dataset.

Methods	Accuracy
HOSM (Ding et al., 2016)	49.5%
Orderlet+SVM (Yu et al., 2015)	68.7%
Orderlet+ boosting (Yu et al., 2015)	71.4%
Human-Object Interaction(Meng Meng et al., 2015)	75.8%
3DTrMB	**81.52%**
3DGist	**83.81%**
3DTrMBGG	**97.62%**

Classification and Human Action Recognition in Video. In *Computer Vision in Human-Computer Interaction*, pages 64–77, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ding, W., Liu, K., Cheng, F., and Zhang, J. (2016). Learning Hierarchical Spatio-temporal Pattern for Human Activity Prediction. *Journal of Visual Communication and Image Representation*, 35:103–111.

Ikizler-cinbis, N. and Sclaroff, S. (2010). Object, Scene and Actions : Combining Multiple Features for Human Action Recognition. In *11th European Conference on Computer Vision (ECCV)*, pages 494–507.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

Kim, D., Yun, W.-h., Yoon, H.-s., and Kim, J. (2014). Action Recogntion with Depth Maps Using HOG Descriptors of Multi-view Motion Appearance and History. In *UBICOMM 2014 - 8th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, pages 126–130.

Koperski, M., Bilinski, P., and Bremond, F. (2014). 3D Trajectories for Action Recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 4176–4180.

Kundu, M. K., Mohapatra, D. P., Konar, A., and Chakraborty, A. (2014). *Advanced Computing - Volume 2, Wireless Networks and Security Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014), Springer*.

Li, W., Zhang, Z., and Liu, Z. (2010). Action Recognition Based on A Bag of 3D Points. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW)*, pages 9–14.

Lisin, D. A., Mattar, M. A., Blaschko, M. B., Learned-Miller, E. G., and Benfield, M. C. (2005). Combining Local and Global Image Features for Object Class Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, Washington, DC, USA. IEEE Computer Society.

Luo, Z., Peng, B., Huang, D.-A., Alahi, A., and Fei-Fei, L. (2017). Unsupervised Learning of Long-Term Motion Dynamics for Videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10.

Mahesh, Y., Shivakumar, D. M., and Mohana, D. H. S. (2015). Classification of human actions using non-linear svm by extracting spatio temporal hog features with custom dataset. *International Journal of Research In Science & Engineering (IJRISE)*, 1:1–6.

Meng Meng, Drira, H., Daoudi, M., and Boonaert, J. (2015). Human-object interaction recognition by learning the distances between the object and the skeleton joints. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6.

Niebles, J. C. and Fei-Fei, L. (2007). A Hierarchical Model of Shape and Appearance for Human Action Classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8.

Nourani-Vatani, N., Borges, P. V. K., and Roberts, J. M. (2012). A study of feature extraction algorithms for optical flow tracking. In *Proceedings of Australasian Conference on Robotics and Automation, Victoria University of Wellington, New Zealand.*, pages 1–7.

Shi, Y., Zeng, W., Huang, T., and Wang, Y. (2015). Learning Deep Trajectory Descriptor for action recognition in videos using deep neural networks. *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Soares Beleboni, M. G. (2014). A brief overview of Microsoft Kinect and its applications. *Interactive Multimedia Conference, University of Southampton, UK*, pages 1–6.

Solmaz, B., Assari, S. M., and Shah, M. (2013). Classifying web videos using a global video descriptor. *Machine Vision and Applications*, 24(7):1473–1485.

Somasundaram, G., Cherian, A., Morellas, V., and Papanikolopoulos, N. (2014). Action recognition using global spatio-temporal features derived from sparse representations. *Computer Vision and Image Understanding*, 123:1–13.

Uijlings, J., Duta, I. C., Sangineto, E., and Sebe, N. (2015). Video classification with Densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off. *International Journal of Multimedia Information Retrieval*, 4(1):33–44.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013a). Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *International Journal of Computer Vision, Springer Verlag*, 103(1):60–79.

Wang, H. and Schmid, C. (2013). Action Recognition with Improved Trajectories. *IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2014a). Learning Actionlet Ensemble for 3D Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):914–927.

Wang, L. and Suter, D. (2007). Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. In *EEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Y., Li, Y., and Ji, X. (2013b). Recognizing Human Actions Based on Gist Descriptor and Word Phrase. *Proceedings of International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)*, pages 1104–1107.

Wang, Y., Li, Y., and Ji, X. (2014b). Human Action Recognition Using Compact Global Descriptors Derived from 2DPCA-2DLDA. In *Proceedings of IEEE International Conference on Computer and Information Technology (CIT)*, pages 182–186.

Wang, Y., Li, Y., and Ji, X. (2015a). Human Action Recognition Based on Global Gist Feature and Local Patch Coding. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(2):235–246.

Wang, Y., Li, Y., Ji, X., and Liu, Y. (2015b). Comparison of Grid-Based Dense Representations for Action Recognition. In *Intelligent Robotics and Applications, Springer International Publishing*, pages 435–444, Cham. Springer International Publishing.

Wang, Y. and Mori, G. (2010). Hidden Part Models for Human Action Recognition: Probabilistic vs. Max-Margin. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1310–1323.

Weinland, D., Ronfard, R., and Boyer, E. (2006). Free Viewpoint Action Recognition Using Motion History Volumes. *Computer Vision and Image Understanding*, 104(2):249–257.

Xiao, Y., Zhao, G., Yuan, J., and Thalmann, D. (2014). Activity Recognition in Unconstrained RGB-D Video Using 3D Trajectories. In *SIGGRAPH Asia Autonomous Virtual Humans and Social Robot for Telepresence*, pages 1–4, New York, NY, USA. ACM.

Xie, J., Jiang, S., Xie, W., and Gao, X. (2011). An efficient global K-means clustering algorithm. *JOURNAL OF COMPUTERS (JCP)*, 6(2):271–279.

Xue, H., Zhang, S., and Cai, D. (2017). Depth Image Inpainting: Improving Low Rank Matrix Completion With Low Gradient Regularization. *IEEE Transactions on Image Processing*, (9):4311–4320.

Yang, X. and Tian, Y. (2016). Super Normal Vector for Human Activity Recognition with Depth Cameras. *IEEE Transactions on Pattern Analysis and Machine Inteligence*, pages 1–12.

Yu, G., Liu, Z., and Yuan, J. (2015). Discriminative orderlet mining for real-time recognition of human-object interaction. *Lecture Notes in Computer Science (including subseries in Artificial Intelligence and Bioinformatics)*, 9007:50–65.

Zanfir, M., Leordeanu, M., and Sminchisescu, C. (2013). The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2752–2759.

Zhang, H. and Parker, L. E. (2016). oDe4D: Color-Depth Local Spatio-Temporal Features for Human Activity Recognition from RGB-D Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):541–555.

Zhao, Y., Liu, Z., Yang, L., and Cheng, H. (2012). Combing RGB and Depth Map Features for Human Activity Recognition. In *Proceedings of The Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4.

Zhu, G., Zhang, L., Shen, P., and Song, J. (2016). An Online Continuous Human Action Recognition Algorithm Based on the Kinect Sensor. *MDPI, Sensors (Basel)*, 16(2):1–18.