Science & Technology Development Journal, 21(2):52- 63

Original Research

Multiple modal features and multiple kernel learning for human daily activity recognition

Vo Hoai Viet*, Pham Minh Hoang

ABSTRACT

Introduction: Recognizing human activity in a daily environment has attracted much research in computer vision and recognition in recent years. It is a difficult and challenging topic not only inasmuch as the variations of background clutter, occlusion or intra-class variation in image sequences but also inasmuch as complex patterns of activity are created by interactions among people-people or people-objects. In addition, it also is very valuable for many practical applications, such as smart home, gaming, health care, human-computer interaction and robotics. Now, we are living in the beginning age of the industrial revolution 4.0 where intelligent systems have become the most important subject, as reflected in the research and industrial communities. There has been emerging advances in 3D cameras, such as Microsoft's Kinect and Intel's RealSense, which can capture RGB, depth and skeleton in real time. This creates a new opportunity to increase the capabilities of recognizing the human activity in the daily environment. In this research, we propose a novel approach of daily activity recognition and hypothesize that the performance of the system can be promoted by combining multimodal features. **Methods**: We extract spatial-temporal feature for the human body with representation of parts based on skeleton data from RGB-D data. Then, we combine multiple features from the two sources to yield the robust features for activity representation. Finally, we use the Multiple Kernel Learning algorithm to fuse multiple features to identify the activity label for each video. To show generalizability, the proposed framework has been tested on two challenging datasets by cross-validation scheme. **Results**: The experimental results show a good outcome on both CAD120 and MSR-Daily Activity 3D datasets with 94.16% and 95.31% in accuracy, respectively. **Conclusion**: These results prove our proposed methods are effective and feasible for activity recognition system in the daily environment.

Key words: HCI, HOF2, HOG2, MKL

University of Science, VNUHCMC, 227 Nguyen Van Cu Street, Ho Chi Minh, Viet Nam

Correspondence

Vo Hoai Viet, University of Science, VNUHCMC, 227 Nguyen Van Cu Street, Ho Chi Minh, Viet Nam

Email: vhviet@fit.hcmus.edu.vn

History

- Received: 28 August 2018
- Accepted: 19 September 2018
- Published: 03 October 2018

DOI :

https://doi.org/10.32508/stdj.v21i2.441

INTRODUCTION

Recognizing human activity is a challenging and engaging task in the community of computer vision research. It is one of the valuable research areas in computer vision with many applications in real-world, such as surveillance system, HCI system, smart city, smart home, gaming, health care and robotics. The literature reviews of human activity recognition may be found in some previous publications [1–4]. In general, the methods to the problem of human daily activity recognition contain four major steps:

i) feature detection,

ii) descriptor extraction,

iii) activity representation, and

iv) pattern classification.

In traditional approaches, researchers have focused on the descriptors that are extracted from image sequences that extend the spatial information in the 2D image to the spatial-temporal information. The studies have demonstrated positive results for human activity recognition.

In recent years, emerging 3D cameras such as Microsoft's Kinect and Intel's RealSense, show they can capture RGB, depth and skeleton in real time. This confers a unique opportunity to increase the capabilities of recognizing human activity in the daily environment. Many authors have exploited 3D spatial-temporal descriptors for depicting and classifying human daily activity [5–12]. In addition, Kinect can capture skeleton data that contain joints on the human body in real time. This helps to detect the bounding box for the individual human body and body parts easily, as well as remove the noise when extracting features. These approaches based on 3D cameras could be divided into four types:

i) RGB-representation approaches,

ii) depth-representation approaches,

iii) skeleton-representation approaches, and

iv) hybrid- representation approaches.

Cite this article : Hoai Viet V, Minh Hoang P. **Multiple modal features and multiple kernel learning for human daily activity recognition**. *Sci. Tech. Dev. J.;* 21(2):52-63.

Recognizing human activity from RGB image sequences

The approaches in this group can be divided into two kinds of categories: global features and local features. The early global features were introduced by Bobick and Davis [13]. They proposed two motion patterns: MEI and MHI. These templates were computed into Hu Moments for human activity representation. Similarly, Xinhua Sun [14] used Zernike moments for activity representation. These approaches, based on the global features, encode much information about the activity. However, they were sensitive to viewpoint, complex background, and occlusion. In order to overcome these above problems, the local features were proposed for activity representation. Many authors have introduced local spatial-temporal descriptors, such as HOG3D, HOF [15,16], SURF 3D [17], and SIFT 3D [18], for temporal information to obtain activity representation. These descriptors were the extended versions of HOG [19], SURF [20], and SIFT [15] that were very successful in solving image classification. The most successful method based on local features was dense trajectories [21,22] that extracted HOG/HOF/MBH for each interest point. However, these dense trajectories or 3D gradient features have a large computational cost in feature extraction.

Recognizing human activity from depth sequences

Approaches such as methods based on extending from the color image have been used [23,24]. For their similarly to MHI [13], Yang *et al.* [25] proposed DMM features that used the depth images projected on three orthogonal planes. Then, HOG [19] operation was applied to have a final vector for activity representation. Instead of accumulating the whole depth images, Li *et al.* [26] sampled around the 3D points of the boundaries that were projected on three orthogonal planes. In order to represent 4D information from depth images, Wang *et al.* [27] proposed the random occupancy pattern and Vieira *et al.* [23] introduced STOP descriptor. The descriptors were based on the idea of local features in RGB image. Many holistic features were similar to RGB image, such as HON4D [28] and SNV [29]. However, these algorithms have a high computational cost and high dimensionality.

Recognizing human activity from skeleton sequences

In addition to RGB and depth channels that are captured from 3D cameras, it is possible to capture 3D positions of skeleton joints with high precision in real time. This opens a new opportunity for recognizing activity in real time because the skeleton data is small and easy to extract features for representation. Xia *et al.* [30] proposed a HOJ3D descriptor to represent shape for each frame. The joints of skeletal data were projected into a spherical axis so that the descriptor is robust to the changes of view. Then, they used HMM model to encode the temporal information from serial feature sequences. Xiaodong *et al.* [31] introduced an EigenJoints descriptor which fuses activity information containing the static features of shape and the dynamic features of movement, based on differences of joints in positions, and Principal Component Analysis (PCA) to reduce the dimensionality of data. They used Naïve-Bayes-Nearest-Neighbor (NBNN) to classify activity using informative frame selection.

Recognizing human activity from multiple modals

The approaches based on combining multiple descriptors are extracted from RGB, depth and skeletal data [8–10,12,32,33]. Zhao Yang [24] proposed the method that extends from the RGB approach by using local features. Firstly, the STIP method was applied to detect salient points. Then, HOG and HOF descriptors were used for RGB channel, and LDP descriptor was extracted from the depth channel. These features were used to yield visual words for activity representation. Wang *et al.* [34] combined skeletal data and depth channel to build ROP around each joint of the skeleton on 3D point cloud. Similarly, Sung *et al.* [35] used the joints of the skeleton to represent the individual person's body, with shaped parts as well as movement. In order to represent the characteristics of appearance, the authors extracted HOG from RGB and depth channel for the individual's body and parts at each frame. Then, maximum entropy Markov model (MEMM) was adapted to recognize a daily activity based on time series of sub-activities. L. Liu [36] proposed the GBGP approach based on evolution programming with the set of filters to extract the descriptors from RGB-D sequences automatically. The feature vectors were concatenated into a final vector for activity representation. Then, a support vector machine (SVM) classier was adapted to the activity classification phase. Pichao Wang [37] used deep learning to fuse RGBD sequences as on an entity to represent human activity from CNN. However, the deep learning methodologies have high computational cost, require high configuration in hardware, and require a lot of data that do not suit in some real-world applications. From the above review, we conclude that the feature extraction is a crucial step for obtaining a system

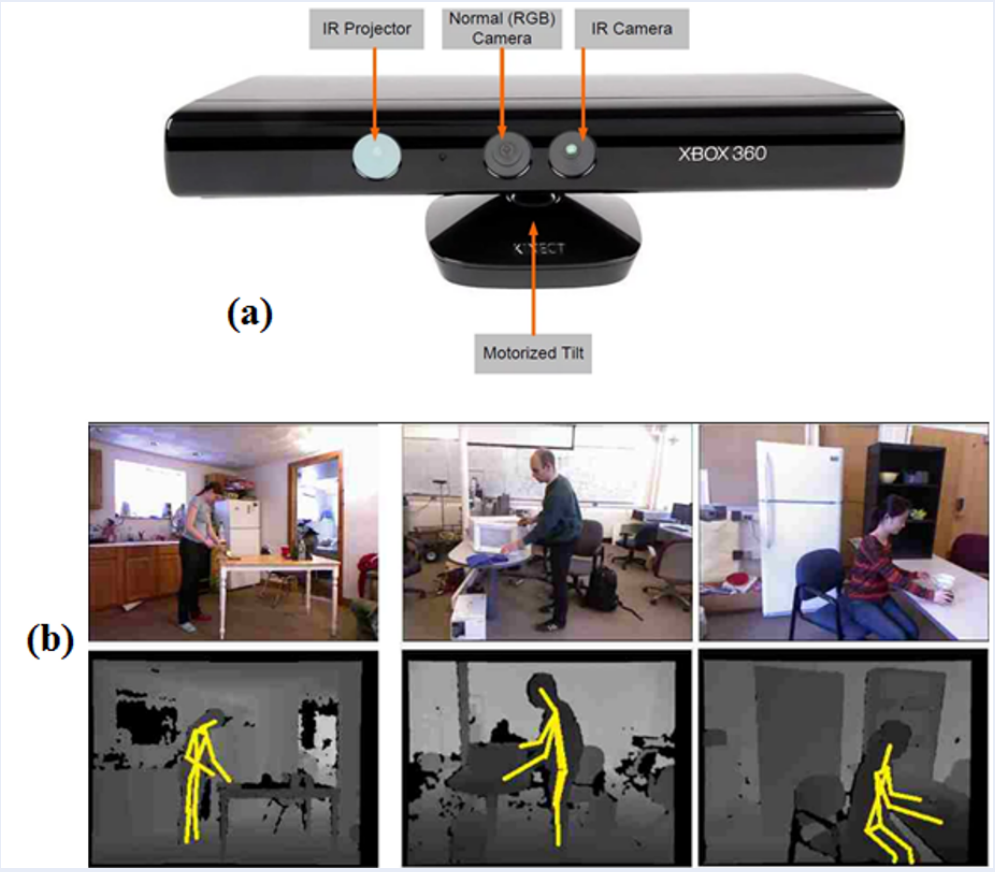

Figure 1: **Some sample frames are extracted from Kinect.** (**a**) Microsoft Kinect, (**b**) RGB, depth, and skeleton data.

that recognizes human daily activity with high performance. It is necessary to choose a set of appropriate descriptors that depict the discriminative characteristics for each activity. In our research, we concentrate on recognizing human daily activities which are captured from Microsoft Kinect (some samples frames can be seen in **Figure 1**). We propose the methodology for daily activity framework and hypothesize that the performance of the system can be promoted by combining multimodal features.

Firstly, we use skeleton data to detect the bounding boxes of the human body and parts, such as head, hands and feet. Then, we extract their shape, appearance, and motion feature to describe the human at each frame from RGB and depth channels. Next, we model the change of shape, appearance, and motion by pooling the frame descriptors in a matrix feature for each channel. After that, we apply HOG operation the second time on the matrix to obtain final vector feature for RGB and depth for activity representation. Both set of features are fused using the Multiple Kernel Learning technique at the kernel levels for human activity classification.

To sum up, the major contributions of our work are recapitulated as follows:

- A novel methodology for daily human activity recognition using the utility of multiple data sources from Microsoft's Kinect.
- A new spatial-temporal feature for motion descriptor named HOF2 that is inspired HOG2.
- Multiple kernel learning for activity classification of RGB-D and skeleton sequences.
- Evaluation of our proposed framework by performing experiments on two challenging daily activity datasets, namely CAD-120 and MSR-Daily Activity 3D.

METHODS

In this section, we show our proposed framework architecture for human activity recognition system in

the daily environment. To be able to recognize what activities a person is doing, we rely on the shape, appearance, and the series of movements that he/she is performing during the course of the activity. The flowchart of our framework for recognizing human daily activity is shown as follows in **Figure 2.**

Shape and Appearance Features

The first characteristic often used in activity representation is the shape and appearance of the human body when performing the activity. In this work, we extract HOG2[30] to represent the changes in shape and appearance of hand activity in the spatial and temporal term.

Let *I*(*x,y*) as a *m* x *n* depth image, the gradient G_x, and G_y are calculated on I(*x,y*) by 1D mark [-1,0,1] to achieve a matrix *G* (i.e. computed magnitude of G_x, and G_y,), matrix θ is quantized orientations from G*x*, and G_y, and *B* denotes the number of bins by extracted histograms.

I(*x,y*) is divided into *M x N* blocks which overlap 50% each other. At each block, we compute an orientation histogram h^s with *B* bins. Let G^s and θ^s be magnitude matrix and orientation matrix at *s*th block with $s \in \{1, ..., M \cdot N\}$, so *q*th bin of histogram h^s is denoted as:

$$h^s(q) = \sum_{x,y} G^S_{x,y} \,.\, 1[\theta^S(x,y) = \theta]$$

Where:

. $\theta \in \{-\pi + \frac{2\pi}{B} : \frac{2\pi}{B} : \pi\}$

. 1 is the indicator function

. q $\in \{1, ..., B\}$

After that, the local histogram h^S of *s*th block is normalized by L2-norm:

$$h^S \rightarrow h^S / \sqrt{IIh^S II_2^2 + \in}$$

By 50% overlapping, we can obtain completely local spatial information of each block and express correlation of blocks. Finally, HOG histograms of bocks are concatenated to form the HOG descriptor h_t at frame $t \in \{1, ..., T\}$. In this work, we extract HOG for 7 bounding boxes, in which, 1 is for the whole body, 6 for 6 joints (left arm, left hand, right arm, right hand, head, and torso of each frame) (**Figure 3**).

Similar, we collect HOG histograms h_t over images to form a 2D matrix called S . Changes of the descriptors according to rows in S represent the changes of the shape and appearance of the activity.

On HOG matrix S, we apply pooling techniques to summarize spatial feature of a depth video. Pooling techniques can help avoid over-fitting in the next recognition step. One of two kinds of pooling techniques (max pooling and average pooling) is used to get the first spatial component hS of the final feature. In this work, we adopted the max pooling technique in our experiments.

Each row in S matrix is HOG feature in each frame, so when calculating derivative along row vectors of S, the result represents the change of body shape in the temporal term. Therefore, HOG algorithm is applied one more time on S matrix to extract the second temporal component hT of the final feature histogram.

$$hT = HOG(S) = HOG\left(\begin{bmatrix} h_1 \\ \vdots \\ h_T \end{bmatrix}\right)$$

The final feature h is formed by concatenating hS and hT and is normalized by L2-norm.

$$h = [hT; hS]$$

The final feature h is called HOG2 because HOG algorithm is applied twice as Figure 4. In our case, the size of HOG block M, N, and the B bins of the histogram features are fixed in two times HOG applying, so the size of HOG2 feature is $1 \; x \; (2 \cdot M \cdot N \cdot B)$. Therefore, HOG2 feature describes the two important elements in activity representation, which are the shape and temporal shape when performing the activity.

HOF2

Since motion is an important source of information for activity representation, we introduce a descriptor, which is extracted from Optical Flow and HOF, to represent the changes of motion flow of activity in the spatial and temporal term.

Let $I(x, y)$as a frame of depth sequence with the size of $m \; x \; n$. Farneback dense optical flow estimation algorithm[16] is applied on two continuous frames to extract optical flow image $I_{OF}(x, y)$.

$$I_{OF}(x, y) = OF(I_{t-1}(x, y), \; I_t(x, y)) \qquad (7)$$

where:

* $I_{OF}(x, y)$ is an optical flow image

* $t \in \{2, \; ... \; , \; T \}$

* OF is the Optical flow estimation function.

After that, $I_{OF}(x, y)$ is split into $M \; x \; N$ blocks with 50% overlapping. At each block, a magnitude matrix G^s and an orientation matrix θ^S with $S \in \{1, \; ... \; , \; M \cdot N \}$ are calculated to build a B-bin orientation histogram h^S_{OF}. Finally, an orientation matrix h_t is created by concatenation local orientation histograms. In this work, we extract HOF for 7

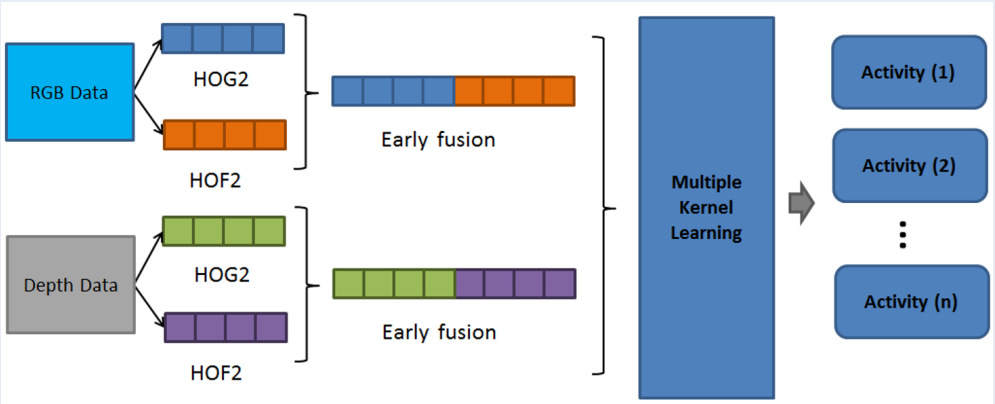

Figure 2: **Flowchart of our methodology for human daily activity from Microsoft's Kinect.**

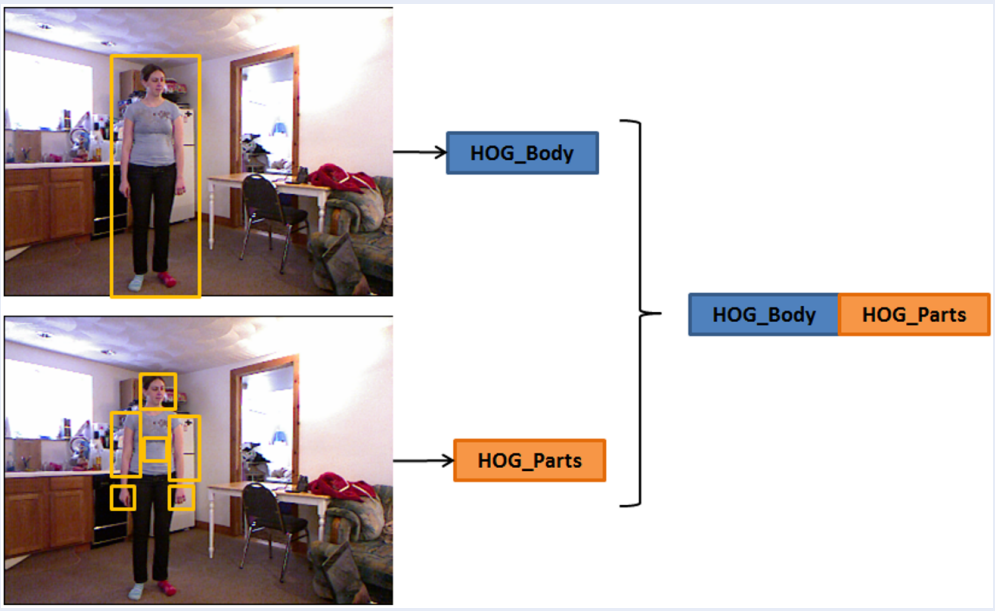

Figure 3: **The HOG extraction at each frame.**

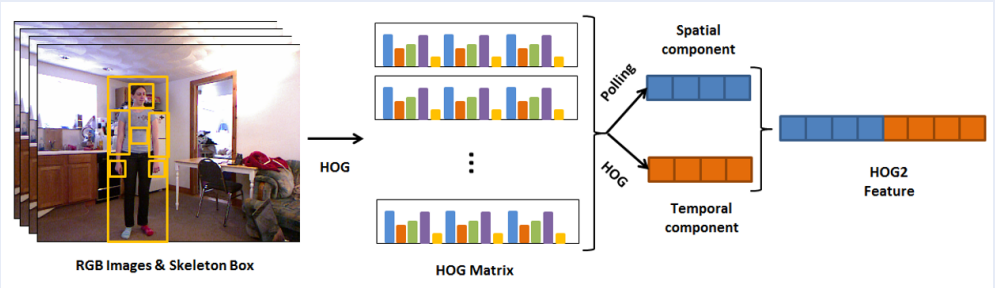

Figure 4: **Illustration of HOG2 extraction for the person's bodyand parts for each video.**

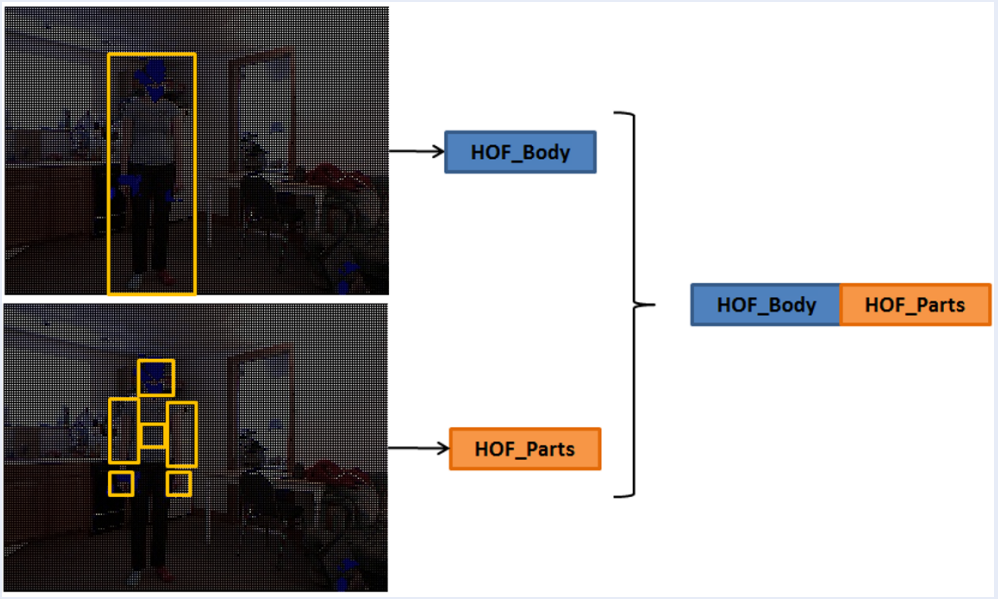

Figure 5: **The HOF extraction at each frame.**

bounding boxes in which are 1 is for the whole body and 6 for 6 joints (left arm, left hand, right arm, right hand, head, and torso for each frame) (**Figure 5**).

An orientation histogram matrix S_{OF} is formed by collecting orientation histograms over frames. Changes of the horizontal vector of S_{OF} represent the changes in the movement of the activity.

On matrix S_{OF}, pooling techniques (which are mentioned in the previous section) are applied to obtain the first spatial component $h_{OF}S$ of the HOF2 feature. Then, HOG operator is used one more time onS_{OF} to represent the second temporal component $h_{OF}T$ of the HOF2 feature.

$$h_{OFT=HOG(S_{OF})} = HOG\left(\begin{bmatrix} h_1 \\ \vdots \\ h_{T-1} \end{bmatrix}\right)$$

The final HOF2 feature h is formed when we concatenate $h_{OF}S$ and $h_{OF}T$ and is normalized by L2-norm, L1-sqrt or L2-Hys[24]. The HOF2 extraction process is similar to the HOG2 extraction method, so the final extracted feature is named HOF2 as in **Figure 6**. In this case, the size of block M, N, and the B bin of histogram feature are fixed, so the size of h is $1\ x\ (2 \cdot M \cdot N \cdot B)$. Thus, HOF2 feature describes the two important elements in activity representation are motion and temporal dynamics when performing the activity.

Activity Representation

In the previous step, we have presented HOG2 and HOF2 descriptors that are used to depict activities. These descriptors are spatial-temporal histograms to show changes in shape and movement when performing activities. In this work, we extract HOG2 and HOF2 for both RGB and depth channels. As the results, we have 4 feature vectors: h^{RGB}_{HOG2}, h^{RGB}_{HOF2}, h^{D}_{HOG2} and h^{D}_{HOF2} for each activity. Thus, we use the feature set for activity representation instead of a fixed length vector like traditional methods. In order to classify the daily activities, we can use early or late fusion techniques.

Activity Classification

In the previous section, our proposed method for daily activity was represented. Here, a set of feature vectors are used instead of a fixed length vector of features as in the previous approaches. Almost all classification algorithms accept input vectors that have the same fixed length in order to train and test the model for activity classifiers. Therefore, we can concatenate the set of vectors into a final vector to build the model. This approach may run into the problem of dimensionality, causing the performance of the system to fall. To overcome this problem, we use the multiple kernel learning (MKL)[38] methods to fuse the multiple features based on building weights that encode the relation of features from multiple sources. The main

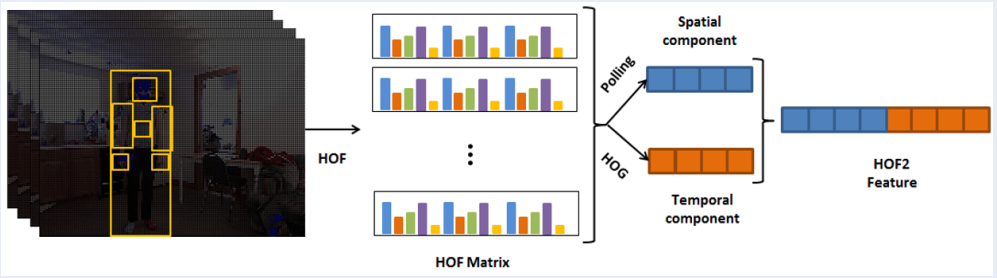

Figure 6: Illustration of HOF2 extraction for a person's body and parts from each video.

idea of MKL algorithm is to use many kernel functions so that multiple feature sources are fused into a nonlinear manner instead of linear combination in late fusion technique. Moreover, the MKL method builds the model by using training data to create good weights to select useful information pieces of the feature vectors from multiple sources. This helps to improve the accuracy rate of the daily activity recognition system and utilizes the advantages from multiple data sources.

In the study, we adopt the SimpleMKL method[38] which is deployed by using SVM with the different kernel functions. The algorithm yields the distinguishing weights to combine the multiple feature vectors from multiple SVM classifiers with multiple kernel functions. Traditional SVM classier only applies for binary classification problems. To be applicable for multiclass problems, we use the one-against-one technique that is proposed in the published study[39].

EXPERIMENTS

Our approach is experienced on two benchmark daily activity datasets, such as CAD 120 and MSR-Daily Activity 3D. The datasets are recorded by Microsoft Kinect to capture human activity in the daily environment.

For the classification phase, we adopt the SimpleMKL two kernel functions to combine the multiple features, extracted from RGBD data. The two kernel functions are defined as follows:

$$K_{Gaussian}(x,y) = e^{-\|x-y\|^2/\delta^2}$$

$$K_{Poly}(x,y) =< x,y >^d,\ d \in N$$

Where $K_{Gaussian}$ is a Gaussian function and σ is the kernel parameter; we adopted $\sigma^2 \in \{0.1, 1, 2\}$; where $K_{Poly}(x,y)$ is a polynomial function and d is the kernel parameter, we adopted $d \in \{1,2,3\}$. Thus, we have 6 parameters for two kernels. We adopt the Leave-One-Out-Subject scheme of evaluation so that we achieve fair comparisons with the other approaches.

We also evaluate these different approaches on the datasets to investigate in greater detail the nature of the problem of daily human recognition. Lastly, we trust that our evaluations in this research study will encourage innovative development and increase studies on daily activity recognition algorithms.

CAD 120 Dataset

The CAD 120 dataset[33] consists of 10 different activity classes where each activity is executed three times. It has a total of 120 samples and is performed by four people. These peoples have to perform an activity many times and interact with different things. Some frames of the dataset are seen in **Figure 7.**

MSR Daily Activity 3D Dataset

MSR-Daily Activity 3D dataset[27] consists of 16 different daily activities in the living room environment. Each activity class has 160 samples and is performed by two different contexts:

i) a standing human and

ii) a sitting human (on a sofa).

This dataset is more difficult than the others due to human-object interaction in frequency when people are performing the activities. Some frames of the dataset are seen in **Figure 10.**

DISCUSSION

In this work, we use the utility of multimodal systems for daily activity representation. From the analysis of the previous approaches, we observed that no one kind of feature can recognize all activity datasets. Therefore, we should combine different kinds of features to improve the performance of the daily activity system. Specifically, we extract HOG2 and HOF2 spatial-temporal features for RGB-Depth data based skeleton joints that can depict shape, appearance, and motion of humans when performing the activities. As

Figure 7: **Examples of frames sampled from CAD 120 Dataset.**

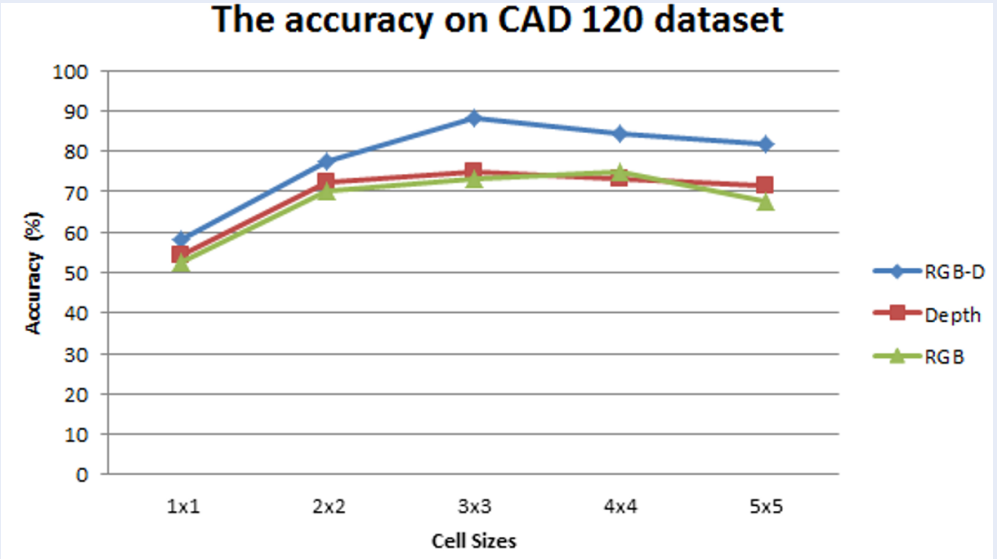

Figure 8: **The different cell sizes in the HOG2 and HOF2 features for daily activity classification with SVM on CAD 120 dataset.**

Table 1: **The experimental results of the different methods on CAD 120 dataset**

Method	Accuracy (%)
RGB + Skeleton + SVM	73.34
Depth + Skeleton + SVM	75
RGB-D + Skeleton + SVM	88.34
RGB-D + Skeleton + MKL	94.16

Table 2: Comparison of results with the state-of-the-art methods

Method	Accuracy (%)
Salient Proto-Objects [7]	78.2
Object Affordances [33]	84.7
Spatio-Temporal Structure [11]	93.5
RGB-D + Skeleton + SVM	88.34
RGB-D + Skeleton + MKL	94.16

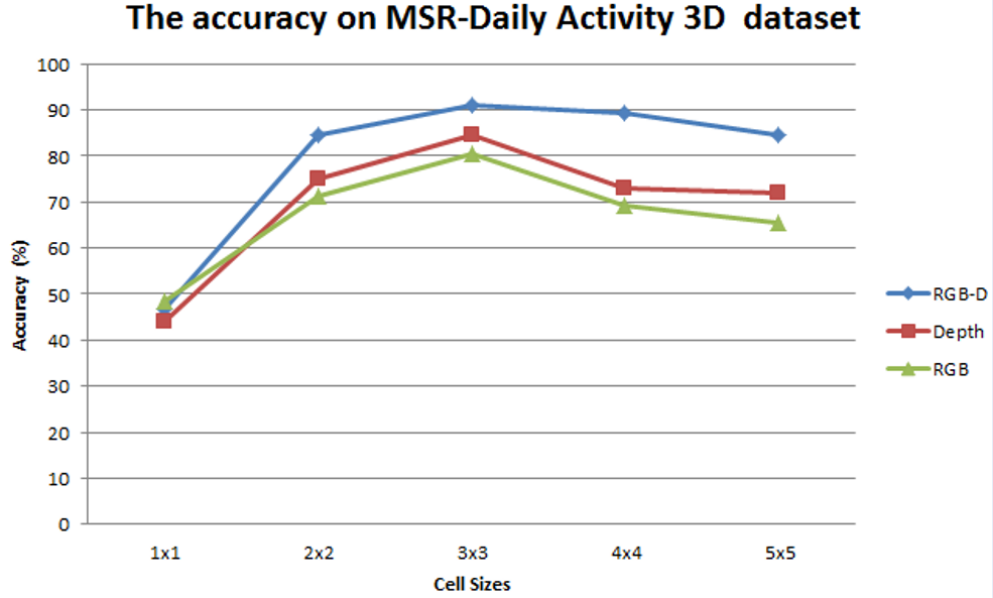

Figure 9: The different cell sizes in the HOG2 and HOF2 features for daily activity classification with SVM on MSR-Daily Activity 3D dataset.

Table 3: The experimental results ofthe different methods on MSR-Daily Activity 3D dataset

Method	Accuracy (%)
RGB + Skeleton + SVM	80.63
Depth + Skeleton + SVM	84.38
RGB-D + Skeleton + SVM	90.94
RGB-D + Skeleton + MKL	95.41

well, HOG2 and HOF2 also present the changes in shape and motion in the time axis. In order to classify a set of features for activity representation, multiple kernel learning approaches are used to find the best discriminative weights for each descriptor in the feature set. So, the performance of the activity recognition system is improved when using multiple modal features.

To evaluate the parameters of the proposed approach, we also perform different experiments on two datasets. The HOG2 and HOF2 descriptors (mentioned in Methods) with different sizes of the cell were compared in **Figures 8 and 9** when classifying with SVM. The results are shown as cell size of 4x4 in computing the HOG2 and HOF2; representation is shown the best results on the two datasets. We also note that different sizes have different results on the datasets. The fusion of RGB and depth descriptors shows best

Table 4: Comparison of results with the state-of-the-art approaches on MSR-Daily Activity 3D dataset

Method	Accuracy (%)
Actionlet Ensemble Model [27]	0.857
HON4D [28]	0.800
SVN [29]	0.862
RGGP [36]	0.904
RGB-D + Skeleton + SVM	90.94
RGB-D + Skeleton + MKL	95.31

results across datasets. We also evaluate the different proposed methods when classifying with SVM and MKL. As shown in **Tables 1 and 3**, the MKL method was shown to work well and improve recognition rate significantly.

Tables 2 and 4 compared our experimental results with the state-of-the-art methods of two challenging datasets. On the CAD120 dataset (**Table 2**), we only improve slightly on accuracy compared to Koppula's methods [33], which is 0.56%. As seen by the MSR Daily Activity 3D in Table 4, our accuracy rate is 95.31%, more than the best result [36] by 5.27%. This increase is due to our approaches to only extract the features from the interest regions based on skeleton data to remove a huge amount of redundant information. Moreover, our approach involves the addition of temporal adjacent features from matrix features. The experimental results show that our proposed methods are efficient, theoretically feasible, and practical.

CONCLUSIONS

In this work, we have discovered the utility of the multimodal approach for the problem of daily human activity recognition in RGB, depth, and skeleton sequences. The activity descriptors are extracted from shape, appearance, and motion in RGB, depth, and skeleton data. We use skeleton joints to capture regions of interest when performing activities; these regions include body, hands, foots, etc. Moreover, we yield a robust activity representation by fusion of HOG2 and a new motion feature (called HOF2), aggregated from RGB and depth features. These features help capture the appearance and motion of the human body and parts as spatial-temporal characteristics for activity representation in 2D and 3D. Thus, each activity is presented by a set of features instead of one fixed length vector as in traditional approaches. Finally, we adopt the MKL method in order to combine multiple features that are extracted from RGB and depth data in the classification step. We evaluate our methodology on two challenging public datasets, such as CAD120 and MSR Daily Activity 3D, with 94.16% and 95.31% in accuracy, respectively. We have shown the benefit of a multimodal approach in the addressing the challenge of daily activity recognition. The evaluations of the different parameters and methods allow us to better understand the nature and problems of daily activity recognition. The experimental results in this study are potentially useful in real-life applications, such as in healthcare, smart home technologies, and robotics.

COMPETING INTERESTS

The authors hereby declare there are no conflicts of interest associated with this work.

AUTHORS' CONTRIBUTIONS

Vo Hoai Viet has made significant contributions in making the literature review, coding algorithms, conducting experiments, interpreting the data, analyzing the experimental results, and composing the manuscript to be submitted.

Pham Minh Hoang has contributed in coding algorithms, conducting experiments, and composing the manuscript.

ACKNOWLEDGMENTS

This research is supported by University of Science, Vietnam National University — Ho Chi Minh City under grant number T2017-01.

ABBREVIATIONS

CNN: Convolution Neural Network
D: Depth
DMM: Depth Motion Map
GBGP: graph-based genetic programming
HCI: Human Computer Interactions
HMM: Hidden Markov Model

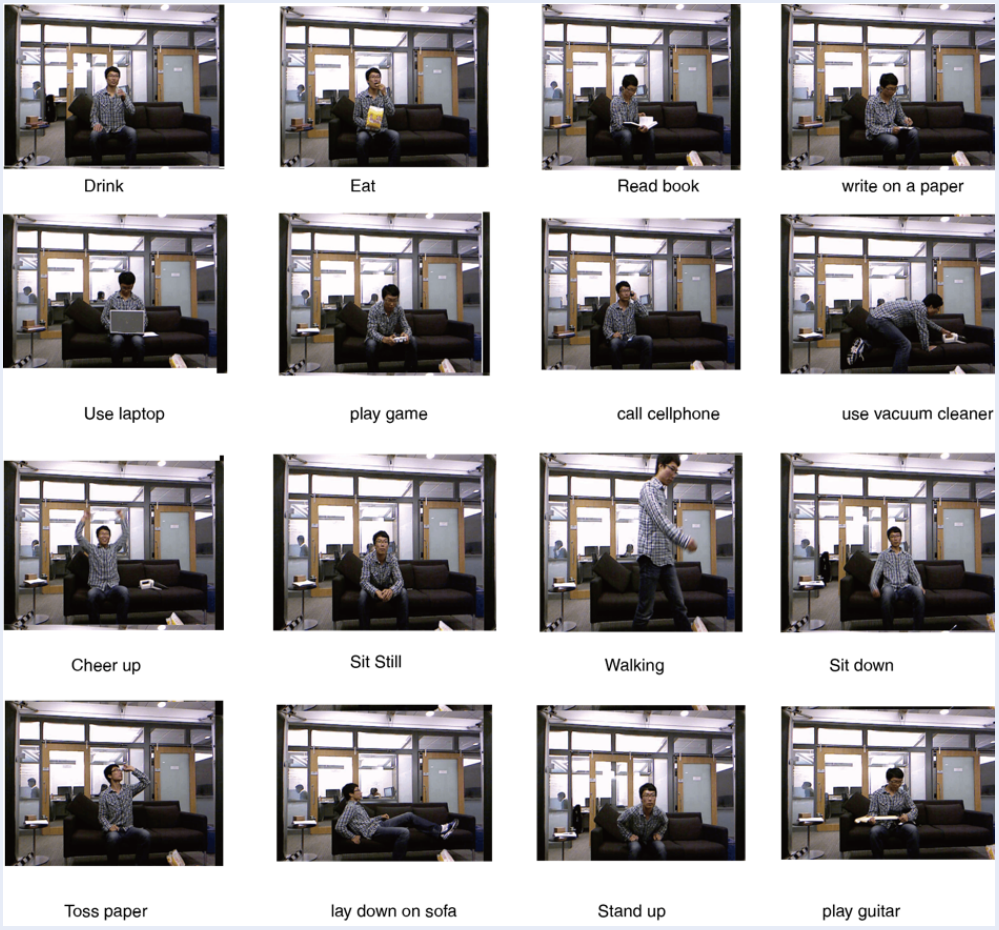

Figure 10: **Some frames are sampled from MSR-Daily Activity 3D dataset.**

HOF: Histogram of Optical Flow
HOG: Histogram of Gradient
HOG3D: Histogram of Gradient 3D
HOJ3D: Histograms of 3D Joints
HON4D: Histogram of oriented 4d normals
LDP: Local Depth Pattern
MBH: Motion Boundary Histogram
MEI: Motion Energy Image
MEMM: maximum-entropy Markov model
MHI: Motion History Image
MKL: Multiple Kernel Learning
NBNN: Naïve Bayes Nearest Neighbor
PCA: Principal Component Analysis
RGB: Red Green Blue
RGB-D: Red Green Blue Depth
ROP: random occupancy pattern
SIFT: Scale-Invariant Feature Transform
SNV: Super Normal Vector
STIP: Space Time Interest Points
STOP: Space-time occupancy patterns
SURF: Speeded Up Robust Features
SVM: Support Vector Machine

REFERENCES

1. Ye M, Zhang Q, Wang L, Zhu J, Yang R, Gall J. A survey on human motion analysis from depth data; 2013.
2. Aggarwal J, Ryoo M. Human activity analysis: A review. ACM Computing Surveys. 2011;43:1–43. Available from: Doi:10.1145/1922649.1922653.
3. Aggarwal JK, Cai Q. Human motion analysis: a review. Proceedings IEEE Nonrigid and Articulated Motion Workshop. 1997;p. 90–102. Available from: 10.1109/namw.1997.609859.
4. Ali HH, Moftah HM, Youssif AA. Depth-based Human Activity Recognition: a comparative perspective study on feature extraction. Future Computing and Informatics Journal. 2018;3:51–67.
5. Zhao Y, Liu Z, Yang L, Cheng H. Combing RGB and Depth Map Features for human activity recognition. APSIPA. 2012;p. 1–4.
6. Jalal A, Sarif N, Kim JT, Kim TS. Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home. Indoor and Built Environment. 2013;22:271–9. Available from: Doi:10.1177/1420326x12469714.
7. Rybok L, Schauerte B, Al-Halah Z, Stiefelhagen R. "Important stuff, everywhere!" Activity recognition with salient proto-objects as context. 2014 IEEE Winter Conference on Applications of Computer Vision (WACV). 2014;p. 646–651.

8. Jalal A, Kim Y, Kamal S, Farooq A, Kim D. Human daily activity recognition with joints plus body features representation using Kinect sensor. Informatics, Electronics & Vision (ICIEV), 2015 International Conference on 2015 Jun 15. 2015;p. 1–6.
9. Zhang C, Tian Y. RGB-D camera-based daily living activity recognition. Journal of Computer Vision and Image Processing. 2012;2:12.
10. Farooq A, Jalal A, Kamal S. Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map. Transactions on Internet and Information Systems (Seoul). 2015;9:1856–69.
11. Koppula H, Saxena A. Learning Spatio-Temporal Structure from RGB-D Videos for Human Activity Detection and Anticipation. International Conference on Machine Learning. 2013;p. 792–800.
12. Koppula H, Saxena A. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. International Conference on Machine Learning. 2013;p. 792–800.
13. Bobick AF, Davis JW. The recognition of human movement using temporal templates. IEEE transactions on pattern analysis and machine intelligence. 2001;23:257–267.
14. Sun X, Chen M, Hauptmann A. Action recognition via local descriptors and holistic features. Computer Vision and Pattern Recognition Workshops, 2009 CVPR Workshops 2009 IEEE Computer Society Conference on 2009 Jun 20. 2009;p. 58–65.
15. Lowe DG. Distinctive image features from scale invariant keypoints. International Journal of Computer Vision. 2004;60:91–110. Available from: Doi:10.1023/b:visi.0000029664.99615.94.
16. Farnebäck G. Two-frame motion estimation based on polynomial expansion. Scandinavian conference on Image analysis. 2003;p. 363–370.
17. Willems G, Tuytelaars T, Gool LV. An efficient dense and scale-invariant spatio-temporal interest point detector. European conference on computer vision. 2008;p. 650–663. null.
18. Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition. Proceedings of the 15th ACM international conference on Multimedia. 2007;p. 357–360.
19. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, 2005 CVPR 2005 IEEE Computer Society Conference on. 2005;1:886–893.
20. Bay H, Tuytelaars T, Gool LV. Surf: Speeded up robust features. European conference on computer vision. 2006;p. 404–417.
21. Wang H, Kläser A, Schmid C, Liu CL. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision. 2013;103:60–79. null. Available from: DOI:10.1007/s11263-012-0594-8.
22. Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. Computer Vision and Pattern Recognition, 2008 CVPR 2008 IEEE Conference on 2008 Jun 23. 2008;p. 1–8.
23. Vieira AW, Nascimento ER, Oliveira GL, Liu Z, Campos MF. Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. Iberoamerican congress on pattern recognition. 2012;p. 252–259.
24. Yang X, Zhang C, Tian Y. Recognizing actions using depth motion maps-based histograms of oriented gradients; 2012.
25. Yang Z, Liu Z, Hong C. RGB-Depth Feature for 3D Human Activity Recognition, Communications, China, 2013. Hanbo Wu, Xin Ma, Zhimeng Zhang, Haibo Wang and Yibin Li, Collecting public RGB-D datasets for human daily activity recognition. International Journal of Advanced Robotic Systems. 2017;14.
26. Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3d points. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on 2010 Jun 13. 2010;p. 9–14. null.
27. Wang J, Liu Z, Wu Y, Yuan J. Learning actionlet ensemble for 3D human action recognition. IEEE transactions on pattern analysis and machine intelligence. 2014;36:914–927.
28. Oreifej O, Liu Z. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. Proceedings of the IEEE conference on computer vision and pattern recognition. 2013;p. 716–723. null.
29. Yang X, Tian Y. Super normal vector for activity recognition using depth sequences. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014;p. 804–811.
30. Xia L, Aggarwal JK. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013;p. 2834–2841.
31. Yang X, Tian YL. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on 2012 Jun 16 . 2012;p. 14–19.
32. Zhu Y, Chen W, Guo G. Fusing spatiotemporal features and joints for 3d action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2013;p. 486–491.
33. Koppula HS, Gupta R, Saxena A. Learning Human Activities and Object Affordances from RGB-D Videos. The International Journal of Robotics Research. 2013;32:951–70. Available from: Doi:10.1177/0278364913478446.
34. Wang J, Liu Z, Chorowski J, Chen Z, Wu Y. Robust 3d action recognition with random occupancy patterns; 2012.
35. Sung J, Ponce C, Selman B, Saxena A. Unstructured human activity detection from rgbd images. Robotics and Automation (ICRA), 2012 IEEE International Conference on 2012 May 14. 2012;p. 842–849.
36. Liu L, Shao L. Learning Discriminative Representations from RGB-D Video Data. IJCAI. 2013;1:3.
37. Wang P, Li W, Gao Z, Zhang Y, Tang C, Ogunbona P. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks; 2017.
38. Rakotomamonjy A, Bach F, Canu S, Grandvalet Y, Simple MK. Journal of Machine Learning Research. 2008;9:2491–521.
39. Ohn-Bar E, Trivedi M. Joint angles similarities and HOG2 for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2013. p. 465–470.