Human Action Recognition using Meta Learning for RGB and Depth Information

S. Mohsen Amiri
Dept. of Electrical & Computer Eng.
University of British Columbia
Canada
mohsena@ece.ubc.ca

Mahsa T. Pourazad
TELUS Communications Inc. &
University of British Columbia
Canada
pourazad@ece.ubc.ca

Panos Nasiopoulos, Victor C.M. Leung
Dept. of Electrical & Computer Eng.
University of British Columbia
Canada
{panos, vleung}@ece.ubc.ca

***Abstract*—In this paper, we propose an efficient human action recognition technique, which utilizes Depth and RGB information of the scene. Our proposed technique, first builds a pair of classifiers based on RGB and depth information to independently predict the actions within a scene. Then, the obtained results from these classifiers are combined to achieve high accuracies in human action recognition. Our experimental results show that using an efficient amalgamation of depth-based and RGB-based classifiers improves human action recognition in smart home applications.**

Keywords—Kinect, Depth Camera, Smart home and human action recognition

I. INTRODUCTION

There is an increasing demand for human action recognition in the past few years and many researchers have focused on solving this problem using video information. One important bottleneck in vision-based human action recognition systems is that the commodity RGB cameras typically have a limitation in capturing 3D structure of the scene and human motions in a 3D space. To address this issue researchers in [1] have proposed to use several cameras to capture the scene (MoCap). However the high cost and difficulty of using such systems make them prohibitively impractical for many applications such as occupant monitoring in smart homes. Recently Microsoft has introduced a low-cost device called Kinect, which captures RGB and depth (RGB-D) information of the scene. The Microsoft Kinect driver for Microsoft Windows also provides a skeleton tracking system, which models human activities using a skeleton model with 20 joints. These features motivated several researchers in the field of human action recognition to use Kinect as a capturing device for 3D information.

To utilize depth information in action recognition, specific algorithms are required. This is due to the existing differences between the characteristics of RGB and depth information. In general the action recognition algorithms utilizing RGB-D information are categorized into two different groups: 1) skeleton-based techniques and 2) Depth map-based techniques.

The techniques under the skeleton-based category utilize the human skeleton information extracted by the Kinect driver to model human motions in a 3D space. The skeleton-based technique in [2] fits a two-layered maximum-entropy Markov model (MEMM) to the skeleton information. In this two-layer model, the top-layer represents activities and the mid-layer represents sub-activities [2]. J. Wang et al. in [3] propose to extract Local Occupancy Pattern (LOP) features from the skeleton and RGB-D information to discriminatively describe human motion and its interaction with objects in the scene. Moreover, they introduce Fourier Temporal Pyramid features (extracted from skeleton information) as the temporal descriptors of human motion dynamics. Other examples of skeleton-based features for human action recognition are Eigen-Joints [1], SMIJ (Sequence of Most Informative Joints) [4], and HOJ3D (Histogram of 3D Joint locations) [5]. Although skeleton information is a valuable piece of information, it is not guaranteed that the Kinect device always localizes body joints and provides accurate skeleton information (due to self-occlusion, object occlusion, or side-view). Therefore skeleton-based techniques with high dependency to skeleton information have limited applications and cannot be used in complex scenarios such as smart homes.

To overcome the issues with skeleton-based human action recognition techniques, recent approaches such as [6] and [7] have focused on designing new features to be extracted from RGB and Depth map streams. To this end new types of local spatiotemporal features have been designed for RGB-D data [6] and [7]. Note that local spatiotemporal features are less sensitive to occlusion and have shown high performance in RGB-based human action recognition [8-10]. Recently, Lu *et. al.* have proposed a new set of feature extractor and descriptor for RGB-D streams [6]. The proposed feature extractor in [6] is an extension of the Space-Time Interest Points (*STIP*) [11]. Unlike the original *STIP* proposed in [11], the modified STIP (*DSTIP*) extractor is less sensitive to high volume of noise and flickers of Kinect depth map data and avoids detecting false feature points. The proposed feature descriptors in [6] is called depth cuboid similarity feature (*DCSF*), which is calculated based on the histogram of depth values in a cuboid around each extracted *DSTIP* point. Another type of recently proposed features for human action recognition from RGB-D video sequences is "*Histogram of Oriented 4D surface Normals"* (*HON4D*) [7]. *HON4D* features are the extension of the "Histogram Of Normals" (*HON*) features [12] and designed to capture changes in the scene structure. The HON4D feature

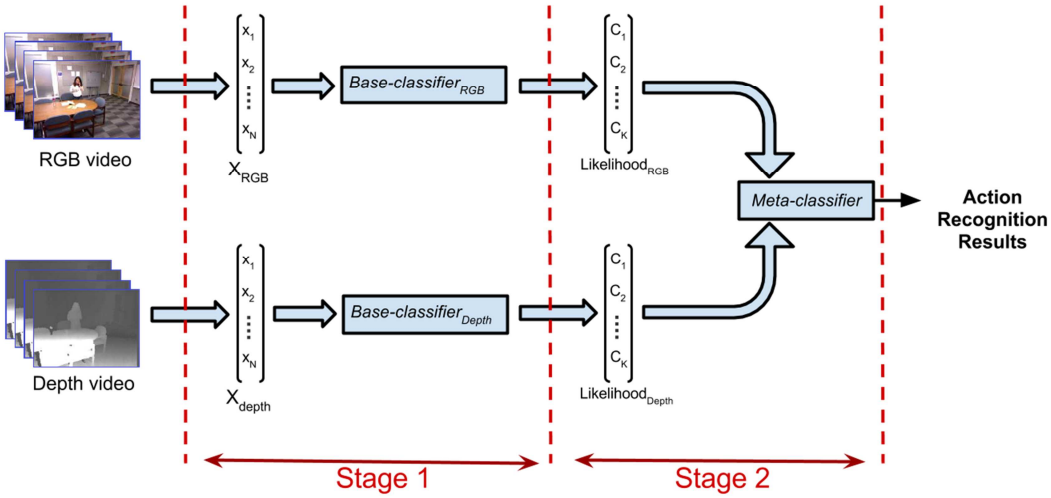

Fig 1. Block diagram of the proposed human action recognition using Depth and RGB video sequences.

extractor constructs features by quantizing the 4D space using a regular 4D extension of a 2D polygon, namely, a "600-cell Polyhedron" [13].

In this paper, we propose a new depth map-based human recognition approach (it does not require skeleton information. Our algorithm suggests extracting STIP features from depth and RGB streams and encoding these features using Non-negative sparse coding (NNSC) [10]. In our study we show that an efficient encoding of extracted STIP features can eliminate the effect of falsely detected features and attain competitive results. The proposed method applies a pair of base-classifiers to depth and RGB features, and then utilizes the mutual information between the base-classifier outputs to construct a meta classifier and improve the action recognition results. We test our approach on two publicly available human activity recognition datasets, MSR-Daily Activity 3D dataset [3] and *DMLSmartActions* dataset [14].

The rest of this paper is organized as follows: Section II provides a description of our proposed method, Section III discusses our experiments that evaluate the performance our proposed approach, and conclusion is drawn in Section IV.

II. PROPOSED METHOD

In this section, we describe our proposed method for human action recognition using RGB-D data. Depth map sequences and RGB streams are completely two different modalities. RGB data represent the perceived light by the sensor from the objects and depth map data represent the distance of objects from the depth sensor. In addition, due to the difference between sensor technologies, these two streams have different noise patterns. Due to the differences in the characteristics of RGB and depth sequences, some forms of actions and motions can be perceived better from RGB data, and some others can be detected better from depth data. Unfortunately most of the existing RGB-D human action recognition schemes are falling short of utilizing the differences between the characteristics of depth and RGB data to improve their performance in describing human actions.

To address this issue, we develop a two-stage learning system for action recognition as demonstrated in Figure 1. In the first stage two separate classifiers (Base-classifier) are used for RGB and Depth map information. These base-classifiers work in parallel to solve the same human action recognition problem using two different sources of information (i.e., RGB and depth data). Thus the classifiers potentially have different prediction results for each specific observation (for each observation, one depth sequence and one RGB sequence is captured), which can be used to reduce the variance of the prediction error. It is essential to note that prediction errors of these classifiers are not completely independent; hence they have some mutual information. In the second stage of our method, some of the errors in the action prediction are concealed by carefully observing and amalgamation of the outputs (action class likelihoods or probabilities of action classes) of the first stage and achieving more accurate action recognition (see Figure 1). The details of the model used as the base-classifier, and the proposed method for amalgamation of the outputs of two base-classifiers are provided in the following subsections.

A. Base- classifier for human action recognition

In this part, we describe the base-classifier that is used for predicting the observed actions in RGB and depth streams (stage 1 in Figure 1). To build the base-classifiers, we customize the suggested framework in [10] for RGB and depth streams. This framework includes spatiotemporal feature extraction, building visual words, max-pooling and classification. For extracting spatiotemporal features from video sequences (both Depth and RGB), Harrise3D feature detector [17] and STIP feature extractor [11] are used. Note that neither STIP nor Hariss3D are designed to work with Depth video frames (Hariss3D may detect many false features due to the sensitivity to the noise of the depth map stream [7]). To address this issue, similar to the proposed frame work in [5], sparse coding is used for building visual words based on the extracted features. One advantage of this approach is

maintaining high recognition accuracy at the presence of high volume of noise [8, 10, 15]. For both base-classifiers, we used None-Negative Sparse coding (NNSC) for learning dictionaries with 8000 words, and then max-pooling for building global video descriptor as subscribed by [10]. For classification, the Linear Support Vector Machine (SVM) proposed in [9] and [10] was replaced by a Logistic Regression (Logit). Although this change results in lower accuracy in the output of the base-classifiers, it enhances the performance of the meta-classifier in the second stage of our method (see Figure 1). The reason is that the output of the Logit classifier is the probabilities of different action classes (likelihood of action classes), which are easier to be interpreted by the meta-classifier.

B. Multiple Source Amalgamation

To combine the output results from the pair base-classifiers and produce final results, we propose two different approaches as follows:

1) Naïve Bayssian Classifer (NBC): Naïve Bayssian Classifier (NBC) is a simple calssfier, which has a competetive performance in many tasks and is frequently suggested for amalgamation of the results from multiple base-classifiers [16]. To use NBC for action recongnition, we are interested to calculate the joint conditional probability of actions (likelihood) as follows:

$$P(c_i|X_{depth,}X_{RGB}) \propto P(X_{depth,}X_{RGB}|c_i) = P(X_{depth}|c_i).\ P(X_{RGB}|c_i) \quad (1)$$

where c_i represents action class i, X_{depth} and X_{RGB} represents depth and RGB descriptors respectively. Equation (1) is correct only under two assumptions: 1) $P(c_i)$ should be constant for all different actions (balances class distribution) and 2) given c_i, X_{RGB} and X_{depth} should be conditionally independent of each other. NBC has shown high performance as a classifier even if these two conditions are not satisfied [16], thus we use NBC as one of the options for amalgamation of the results from two base-classifiers.

2) Meta-classifier (MC): Albeit NBC provides a simple and efficient solution for source amagamation, it is unable to discover the existing mutual information between meta-features. To address this issue, we use the outputs (predicted probabilities of each action class) of the depth and RGB base-classifiers (meta-features) as the input of a meta-classifier to construct a stronger classifier. If we have two base-classifiers and K action classes, we have meta-features with 2.K dimenssions and a supervised classifier can be used as the meta-classifier. In this paper, we chose a SVM with an intersection kernel as the meta-classifier (MC).

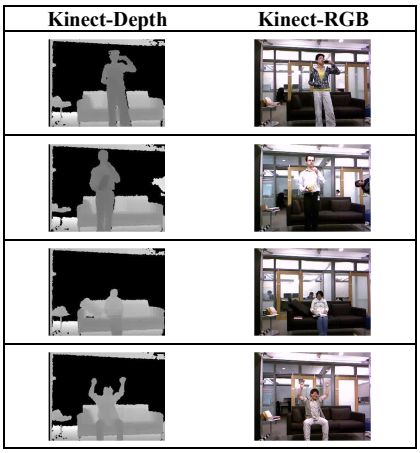

Fig 2. The snapshots of some of the activity samples from MSRDailyActivity3D dataset. Kinect-Depth is the depth stream from the Kinect device, and Kinect-RGB is the RGB stream from the Kinect device.

III. Implementation and Evaluation

In our study, we investigate the performance of our proposed action recognition algorithm and compare it with those of the RGB-D-based state-of-the-art human action recognition algorithms. The following sub-sections provide more details on the dataset, implementation aspects of our experiments, and our experiment results.

A. Datasets

To compare the performance of the proposed method in this paper, we use two different RGB-D datasets, which contain human actions in home environment.

1) MSRDailyActivity3D Dataset [3]: DailyActivity3D dataset is a chanllenging dataset, which is designed to cover human's daily activities in a living room. The dataset was captured by a Kinect device and contains sixteen activity types (*drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up,* and *sit down*) performed by 10 subjects. To add more inter-class variance for the actions in the dataset, each subject performs an activity in two different poses: "*sitting on sofa*" and "*standing*". The total number of the activity samples is 320. Figure 2 demonstrates the snapshots of some of the activity samples.

This dataset aslo includes the 3D joint positions, which are extracted by the skeleton tracker of the MS Kinect driver. Note that our proposed algorithm only uses Depth and RGB channels of this dataset and does not use 3D joint positions for action recognition.

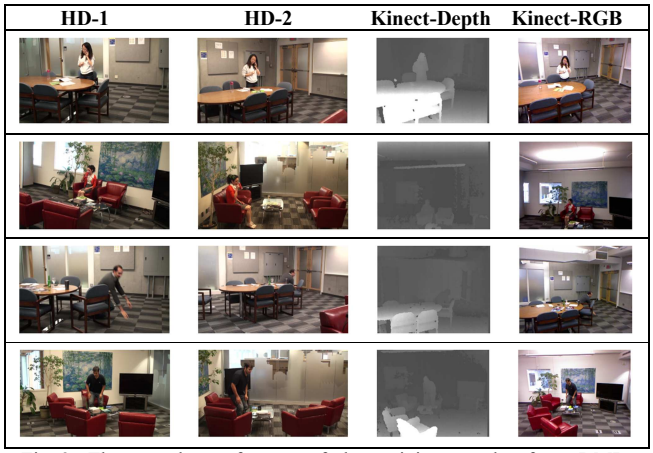

Fig 3. The snapshots of some of the activity samples from DML-SmartActions dataset. HD-1 and HD-2 are sample frames from two HD cameras. Kinect-Depth is the VGA depth stream from the Kinect device, and Kinect-RGB is the VGA RGB stream from the Kinect device.

2) DMLSmartActions [14]: The DMLSmartActions dataset contains twelve different actions performed by 17 subjects (6 females and 11 males), which are common in people's day-to-day life in home environments (*clean-table, drink, drop-and-pickup, fell-down, pick-something, put-something, read, sit-down, stand-up, use-cellphone, walk, and write*).

In total the DMLSmartActions dataset contains 932 annotated samples and contains about 3 times more samples compare to what DailyActivity3D [3] has (each sample was collected using two static High Definition RGB cameras and one Kinect device). Figure 3 shows sample frames from the DMLSmartAction dataset. In this work, we only use the Depth and RGB streams captured by the Kinect device to generate the results (RGB streams captured by HD cameras are not used).

B. Implementation and parameters setting

In this work, we use the implementation of Hariss3D [17] and STIP [11] from the binary provided by [7]. The SPAMS (SPArse Modeling Software) library is used for NNSC implementations [18]. The number of visual words in the dictionary is fixed to 8000. Other parameters in SPAMS are tuned using the provided instructions in [10].

An implementation of a Logistic regression (Logit) from the Scikit-learn library [19] is used as the base-classifiers. The meta-features are constructed by concatenation of the predicted probabilities of the pair Logit classifier trained with the depth and the RGB data. This meta-feature is used for training a SVM with an intersection kernel (meta-classifier).

To have a fair comparison with the performance of algorithms proposed in [3, 6, 7], similarly we use the Leave-One-Out (LOO) strategy. In our LOO, we iterate over different subjects and each time exclude all the samples belonging to one subject and train the system using the remaining samples. Then, the excluded samples are used to test the trained system and find its accuracy. Finally, the overall accuracy of each algorithm is calculated as the average of the accuracy values obtained over all iterations.

Training the meta-classifier using meta-features is a tricky task and should be performed very carefully. The main advantage of meta-classifiers appears in the modeling of the "generalization error". "Generalization error" is the error of base-classifiers in predicting the action classes for un-observed samples (video samples of un-observed subjects in training set). If the whole dataset is divided into a training and testing datasets, the base-classifier is trained using a portion of training samples (base-set), and the un-used samples of training dataset (meta-set) should be used to train the meta classifier (i.e., the samples used for training the base-classifier cannot be used for training the meta-classifier). In our study, to produce more data for better training the meta-classifier, similar to what we do in LOO, each time all the samples belonging to one subject in the training dataset is excluded to build the base-set and meta-set. Then, we train the base-classifiers and meta-classifier using the base-set and meta-set respectively.

TABLE 1. Achieved accuracies of different algorithms on the MSRDailyActivity3D dataset.

Algorithm name	Accuracy
Actionlet Ensemble[3]	85.75%
Local Occupancy Pattern (LOP) [3]	42.50%
DCSF+Joint [6]	88.20%
DCSF[6]	83.60%
HON4D [7]	80.00%
Base-classifier (Only RGB)	74.37%
Base-classifier (Only Depth)	62.81%
Our Proposed NBC Method	75.18%
Our Proposed MC Method	**84.36%**

C. Performance evaluation

In this subsection, we evaluate the performance of our proposed method for human action recognition using RGB-D data in two publicly available datasets, MSRDailyActivity3D dataset and DMLSmartActions dataset.

The accuracy of different algorithms on the MSRDailyActivity3D dataset is reported in Table 1. As it can be observed, when only the base-classifier (modified version of [10]) is applied on the RGB data, the obtained accuracy is 74.37%, and the accuracy of the base-classifier on the Depth data is 74.37%. By amalgamation of the outputs of two base-classifiers using NBC, the accuracy improves by less than 1%, while the proposed meta-classifier boosts the results by 10% and 21.55% compare to the cases where only RGB-based base classifier and depth-based base classifier are respectively used (see Table 1). This large margin supports the idea that using the dependencies between the outputs of two base-classifiers is a suitable approach to improve the accuracy.

As Table 1 denotes, our MC method outperforms other algorithms which utilize only RGB-D data (LOP [3] and DCFS [6]). In addition the proposed MC algorithm can achieve higher accuracies compare to HON4D [7], which utilizes skeleton information in addition to RGB-D data. Although our MC method does not utilize the skeleton information, it manages to achieve comparable results with Actionlet Ensemble [3] and DCSF+Joint [6], which use skeleton information in addition to RGB-D data.Figure 4 demonstrates the confusion matrix of the proposed method for action recognition using the MSRDailyActivity3D dataset.

In a different experiment, we apply the proposed algorithm on the DML-SmartActions dataset. Table 2 reports the summary of the results for different algorithms. As it can be observed, the modified version of [10] achieves 37.11% accuracy when only depth information is used, and gains the accuracy of 54.13% when only RGB data is used. By combining the outputs of two base classifiers using NBC, the accuracy improves by 2.2%. Our proposed MC algorithm improves the accuracy of the RGB-based and depth-based base-classifiers by 23% and 40% respectively (see Table 2). The experimental results confirm the effectiveness of the meta-learning approach for the action recognition problem. We also compared the results of MC algorithm with the highest reported results in [10]. As Table 2 shows by efficiently using RGB-D data (MC algorithm), the accuracy of action recognition improves by 19% compared to the approach in [10]. Note that the best reported results in [10] are achieved

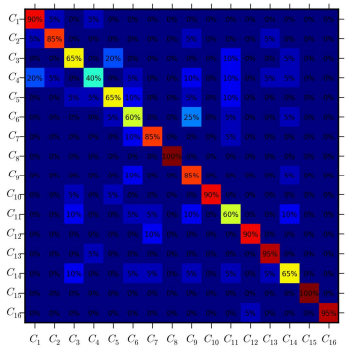

Fig 4. Confusion matrix of the proposed action recognition system for MSRDailyActivity3D dataset (C_1:drink, C_2:eat, C_3:read book, C_4:call cellphone, C_5:write on a paper, C_6:use laptop, C_7:use vacuum cleaner, C_8:cheer up, C_9:sit still, C_{10}:toss paper, C_{11}:play game, C_{12}:lay down on sofa, C_{13}:walk, C_{14}:play guitar, C_{15}:stand up and C_{16}:sit down)

using High Definition video sequences, STIP features and a χ^2-SVM. These results show that by efficiently using RGB-D data, the recognition accuracy can highly be improved. Figure 5 demonstrates the confusion matrix of the proposed method in for action recognition task using the DML-SmartActions dataset.

IV. CONCLUSION

In this paper, we introduce a new approach for human action recognition using RGB-D data. The proposed method uses a pair of base-classifiers and the mutual information between their outputs to construct a meta-classifier. Our experimental results show that the proposed method outperforms the state-of-the-art human action recognition algorithms which utilize RGB-D data and no skeleton data. One important advantage of the proposed method is its independency to human body skeleton information. As the result our proposed method can be applied in the difficult circumstances, where the skeleton is noisy or cannot be detected due to occlusion.

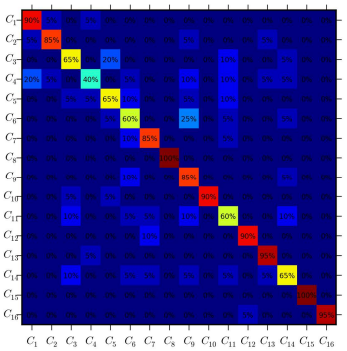

Fig 5. Confusion matrix of the proposed action recognition system for DML-SmartActions dataset (C_1:clean-table, C_2:drink, C_3:drop-and-pickup, C_4:fell-down, C_5:pick-something, C_6:put-something, C_7:read, C_8:sit-down, C_9:stand-up, C_{10}:use-cellphone, C_{11}:walk, and C_{12}:write)

ACKNOWLEDGMENT

This work was made possible by NPRP grant # NPRP 4-463-2-172 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors..

TABLE 2. Achieved accuracies of the proposed method the DMLSmarthome dataset

Algorithm name	Accuracy
STIP on High definition video [10]	58.20%
Base-classifier (Only RGB)	54.13%
Base- classifier (Only Depth)	37.11%
Our Proposed *NBC* Method	56.32%
Our Proposed MC Method	77.19 %

REFERENCES

[1] X. Yang, Y. Tian, "EigenJoints-based Acton Recognition Using Baive-Bayes-Nearest Neighbor," The 2nd International Workshop on Human Activity Understanding from 3D Data 2012 (HAU3D12)

[2] Jaeyong Sung, Colin Ponce, Bart Selman, Ashutosh Saxena, "Human Activity Detection from RGBD Images," In AAAI workshop on Pattern, Activity and Intent Recognition (PAIR), 2011.

[3] Jiang Wang, Zicheng Liu, Ying Wu, Junsong Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), Providence, Rhode Island, June 16-21, 2012.

[4] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, "SMIJ: A New Representation for Human Skeletal Action Recognition," The 2nd International Workshop on Human Activity Understanding from 3D Data, 2012 (HAU3D12)

[5] L. Xia, C.C. Chen, J. K. Aggarwal, View Invariant Human Action Recognition Using Histogram of 3D Joints, The 2nd International Workshop on Human Activity Understanding from 3D Data, 2012 (HAU3D12)

[6] Lu Xia and J.K. Aggarwal, "Spatio-Temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera," In Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, Oregon, June 2013

[7] Omar Oreifej and Zicheng Liu, "HON4D: Histogram of Oriented 4D N Normals for Activity Recognition from Depth Sequences," In P roceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, Oregon, June 2013

[8] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatialtemporal volumes for human action recognition," in ACCV, 2010.

[9] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in BMVC, 2009.

[10] S. M. Amiri, P. Nasiopoulos, and V. C. Leung, "Non-negative sparse coding for human action recognition," in ICIP, 2012.

[11] Ivan Laptev, Marcin Marszalek, Cordelia Schmid and Benjamin Rozenfeld, "Learning Realistic Human Actions from Movies," in Proc. CVPR'08.

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

[13] H. S. M. Coxeter, "Regular polytopes," In 3rd. ed., Dover Publications. ISBN 0-486-61480-8, 1973

[14] S. Mohsen Amiri, Mahsa T. Pourazad, Panos Nasiopoulos and Victor C.M. Leung, "Non-Intrusive Human Activity Monitoring in a Smart Home Environment," in IEEE HealthCom'13

[15] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in CVPR, 2009.

[16] Harry Zhang, "The Optimality of Naive Bayes, " in FLAIRS 2004

[17] I. Laptev and T. Lindeberg, "Space-time interest points," in ICCV, 2003.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," JMLR, 2010.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python ," JMLR 2011.