Human Activity Recognition Using Multi-modal Data Fusion

Andres Felipe Calvo[1(✉)], German Andres Holguin[1], and Henry Medeiros[2]

[1] Faculty of Engineering, Universidad Tecnológica de Pereira, Pereira, Colombia
{afcalvo,gahol}@utp.edu.co
[2] Department of Electrical and Computer Engineering, Marquette University, Milwaukee, USA
henry.medeiros@marquette.edu

Abstract. The automated recognition of human activity is an important computer vision task, and it has been the subject of an increasing number of interesting home, sports, security, and industrial applications. Approaches using a single sensor have generally shown unsatisfactory performance. Therefore, an approach that efficiently combines data from a heterogeneous set of sensors is required. In this paper, we propose a new method for human activity recognition fusing data obtained from inertial sensors (IMUs), surface electromyographic recording electrodes (EMGs), and visual depth sensors, such as the Microsoft Kinect®. A network of IMUs and EMGs is scattered on a human body and a depth sensor keeps the human in its field of view. From each sensor, we keep track of a succession of primitive movements over a time window, and combine them to uniquely describe the overall activity performed by the human. We show that the multi-modal fusion of the three sensors offers higher performance in activity recognition than the combination of two or a single sensor. Also, we show that our approach is highly robust against temporary occlusions, data losses due to communication failures, and other events that naturally occur in non-structured environments.

Keywords: Human activity recognition · Multimodal sensors · Data fusion · Support vector machine · Hidden Markov Model

1 Introduction

The analysis of human activity is a critical component for applications in fields such as health, security, sports, among others. Performing this task in an automatic manner is challenging and has prompted several researchers to attempt a multitude of approaches [3,5,19]. Among the most common devices used for this task are depth cameras (Kinect®). Some approaches use the spatial coordinates of human body joints and then compute feature vectors that can be used for classification. In [16], the authors use polar coordinates for the characterization of joints in order to achieve higher performance in activity classification. Other

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 946–953, 2019.
https://doi.org/10.1007/978-3-030-13469-3_109

methods use classifiers (K-means, SVM) to generate a codebook with key postures and subsequently employ a Hidden Markov Model (HMM) to recognize the different combinations of postures and thus identify the activity being performed. However, all these methods present limitations caused by partial occlusions of the target [4,15].

Sensors such as Inertial Measurement Units (IMUs) are also used for activity recognition [2,18]. However, these sensors require high processing capabilities [13], and most of the time a single sensor is not sufficient to perform satisfactory detection [1]. Electromyographic signal sensors (EMG) are also useful for activity recognition [6,14]. However, sophisticated signal processing and multiple sensors are also required for adequate detection accuracy. Kang et al. use Mel-Frequency Cepstral Coefficients (MFCC), obtaining an activity recognition accuracy of 85% [11]. Korbinian et al. use an HMM for activity recognition and neural networks for motion segmentation, reporting high accuracy rates between 93% and 100% [12]. There is a consensus that fusing data from different sensors improves human activity recognition systems [5,16]. Also, a single sensor modality is generally not capable of identifying the wide range of human activities. Although several methods for human activity recognition that use multi-modal fusion approaches have been proposed [7,9], few techniques take more than two sensing modalities into account at the same time. Among the few recent works that do use multi-modal approaches, Zhand et al. employ a model that is based on primitive motions to classify movements using Bag of Features (BOF) techniques with histograms of primitive symbols [17]. In particular, to the best of our knowledge, no existing method fuses the information of IMUs, EMGs, and depth sensors simultaneously. Therefore, we propose a fusion method that combines the strengths of each sensor to provide better performance.

2 Proposed Method

This paper proposes an activity recognition method based on primitive motion detection. Our method is comprised of two main steps. First, we analyze the sensor data over a small time window to perform primitive motion classification, creating a motion sequence from each sensor. Second, this sequence of primitives is fed into a Hidden Markov Model that classifies the overall activity. An overview of the prediction and training methods is shown in Fig. 1. To validate our method, we built an annotated database containing 5 different human activities. Each activity was performed 3 times by 16 different individuals. For each subject, we captured raw data from 4 IMUs, 4 EMG sensors, and a Kinect® device. Our dataset is publicly available at https://goo.gl/6F82wd.

2.1 Primitive Motions Recognition

Models based on primitive motions are inspired on techniques from human speech analysis [8]. In speech recognition, phrases are generally divided into isolated phonemes. These phoneme models are used as basic blocks in order to build

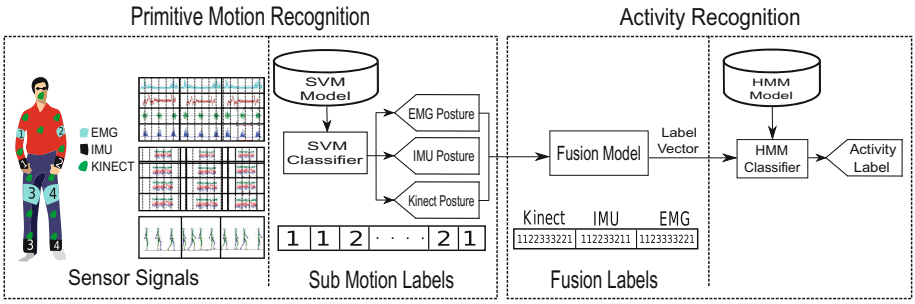

Fig. 1. Overview of the training and classification processes of our proposed approach.

words and phrases in a hierarchical way [10]. Our motion detection model follows a similar idea to that of Zhand et al. [17] in the sense that each activity is represented as a sequence of sub-movements, or primitive motions, generating a unique signature that will be used for classification of the overall activity.

Primitive Motions Encoding. In this work, we propose eight primitive motions to train the HMM system. These motions are: (1) Repose, (2) Partially crouched, (3) Fully crouched, (4) In midair, (5) Quarter rise arm, (6) Three-quarters arm rise, (7) Step forward with right foot and (8) Step forward with left foot.

Feature Extraction. From each sensor modality, a set of features is extracted from the video sequence during a time observation window, which was set experimentally to 3 s. For the Kinect®, the descriptor vector is obtained from the 14 human pivot points. The sensor is able to provide data at 30 samples per second. However, our feature vector is composed of groups of 3 samples, corresponding to an overall rate of 10 samples per second. Given the set of body joints in Cartesian coordinates, all these points are converted to polar coordinates vis-a-vis the center of mass:

$$P_i = [r_1\, \theta_1\, r_2\, \theta_2 \ldots r_{14}\, \theta_{14}]\,, \tag{1}$$

where i is the i-th sample window, with $i = \{1, 2, 3\}$. In addition, the mean m and standard deviation v are computed over all the coordinates. The final feature vector for the Kinect sensor is then defined according to

$$\text{KIT} = [P_1\, P_2\, P_3\, m_x\, m_y\, m_z\, m_r\, m_\theta\, v_x\, v_y\, v_z\, v_r\, v_\theta]\,. \tag{2}$$

For the IMUs, 4 sensors were attached near the wrists and knees of the subjects. Each IMU provides 30 samples per second. Again, we used the average of 3 samples to compute our features. Therefore, the IMU vector is also available at 10 samples per second. With the IMU data $I_k = [a_x\, a_y\, a_z\, a_\theta\, a_\phi]_{1\times 5}$, where $k = \{1, 2, 3, 4\}$ is the k-th IMU, we compute the following descriptors: (1) Features based on the physical parameters of the human motion [18], and (2) Statistical

Descriptors. The overall IMU descriptor is a combination of the IMU_k descriptor for each of the sensors in the network, i.e.,

$$IMUF = [IMU_1\, IMU_2\, IMU_3\, IMU_4]\,. \tag{3}$$

For the EMGs, we track the activity of 4 body muscles. We obtain the signal E_i from each muscle at a sampling frequency of 2 kHz, where i is the i-th EMG sensor. E_i is segmented by using V_j windows of 200 samples where j is the jth window. Each window V_j is concatenated to form a vector W_i and this vector is characterized by a Daubechies Wavelet transform with 35 orthogonal coefficients and 6 levels, which produces the feature vector $\mathrm{EMG}_{1\times 1300}$.

Motion Classification. We use three multi-class support vector machines with classification strategy "One-vs-All" with Gaussian kernels to separate the data. The same process is used with the Kinect, IMU, and EMG sensors.

2.2 Activity Recognition

We use our set of primitive motions described in Sect. 2.1 to classify the following activities: (1) Stand still, (2) Squat and stand up, (3) Jump, (4) Raise right hand, and (5) Jog. To classify each activity from this set, the outputs of the three SVMs are used as input to an HMM. An HMM is chosen because it has been successfully used to detect and encode sequences over time (i.e., the ones produced by the SVMs). Deep learning methods can also be explored in a future work.

Hidden Markov Model Classification (HMM). As described in Sect. 2.1, each SVM classifier generates a label that corresponds to the information provided by the different sensors. The vectors EI correspond to the network of IMUs, EK to the Kinect® device, and EE to the EMGs. The data fusion process consists of generating a EF feature vector with the labels from the SVM classifiers. EF is built by concatenating each classifier label during motion capture.

$$EF = \left[[EK_1\, EK_2 \ldots EK_{30}]\,[EI_1\, EI_2 \ldots EI_{30}]\,[EE_1\, EE_2 \ldots EE_{30}]\right]_{90\times 1} \tag{4}$$

2.3 Training and Validation Process

We train our multi-class SVM models using sequential minimal optimization (SMO). For HMM training, we used 24 states and 32 centroids for the construction of the codebook. We evaluate our models using a cross-validation strategy that partitions the database with 70% of the data for training and 30% for evaluation and generate the confusion matrix for each classifier. This process applies a Monte Carlo analysis, where the stop criterion is defined by

$$\|\mathrm{diag}\,(M_k) - \mathrm{diag}\,(M_{k-1})\|_2 < th, \tag{5}$$

where M_k is the confusion matrix at iteration k and th is the error threshold.

3 Results

We show the results to validate the performance of our method as a function of the sensors used to collect the data. Initially, we evaluate the performance of every sensor modality and the different combinations of sensors. The assessment is based on two basic steps: primitive motion analysis and activity recognition analysis. The first step carries out the performance analysis of the SVM classifiers for the proposed primitive motions. The second step validates the human activity classification using an HMM.

Table 1. Traces of the confusion matrices for the primitive motion classification analysis.

Class	Kinect	IMU		EMG	
	All joints	All	For 1, 3 and 4	All	For 1, 3 and 4
C1	86.77 ± 3.79	86.77 ± 0.73	81.33 ± 0.88	72.28 ± 5.83	62.52 ± 5.56
C2	77.57 ± 3.81	78.71 ± 2.10	75.51 ± 2.62	66.51 ± 1.13	67.81 ± 0.26
C3	71.86 ± 8.05	71.46 ± 4.33	69.29 ± 4.60	70.84 ± 4.73	73.57 ± 4.03
C4	89.95 ± 1.46	74.52 ± 0.93	83.87 ± 1.06	71.49 ± 1.33	79.34 ± 1.48
C5	90.47 ± 3.15	76.51 ± 5.25	56.85 ± 5.38	59.43 ± 5.21	45.44 ± 5.05
C6	96.78 ± 1.33	93.07 ± 2.02	0.84 ± 3.53	63.88 ± 2.35	79.03 ± 2.37
C7	75.68 ± 1.95	61.43 ± 2.91	54.68 ± 2.95	50.54 ± 1.84	51.98 ± 1.84
C8	75.63 ± 2.97	57.56 ± 2.73	55.64 ± 2.85	34.08 ± 2.50	55.71 ± 2.51
Average	84.69 ± 3.31	75.00 ± 2.63	59.76 ± 2.98	61.13 ± 3.12	64.43 ± 2.89

3.1 Primitive Motion Analysis

We use the validation approach described in this section to obtain the confusion matrices of the Kinect®, IMU, and EMG sensors. The traces of the recognition confusion matrices of the primitive movements (using all the sensors as well as the minimum number of sensors that guarantees a reliable detection performance) are shown in Table 1. The Kinect® sensor provides the best primitive movement detection results with an average detection value of approximately 85%, which is substantially higher than those of the other sensors. The analysis of the set of IMU sensors demonstrates a comparable performance with the Kinect in the first three primitive movements. While the EMG sensors alone perform relatively poorly, they can still obtain a precision higher than 70% for classes 1, 3 and 4.

We evaluated the performance of the subsets of sensors by systematically removing the features corresponding to each sensor from our classification system. In columns 4 and 6 of Table 1, we report the results obtained from the subsets that showed the best performance. As the table indicates, while removing a single IMU sensor results in a substantial accuracy reduction for class 6 and a more modest reduction for class 5, the other activities remain mostly at the same performance level. The subset of EMG sensors, on the other hand, show comparable performance for most classes.

3.2 Activity Recognition Analysis

Table 2 shows the traces of the confusion matrices for the HMM-based activity recognition for each sensor category. The results correspond to 181 Monte Carlo iterations for each sensor. Our results demonstrate that the Kinect or the IMU sensors alone provide high classification accuracy for all the activities. The EMG sensors show high classification performance for activities 2, 3, 4 and 5. Also shown in the table are the results of 30 Monte Carlo iterations using a single IMU sensor, which demonstrate that it is possible to recognize all the activities with a single sensor.

Table 2. Traces of the confusion matrices for the activity recognition analysis.

Activity	Kinect	IMU		EMG
	All joints	All	For 1	All
1	93.50 ± 10.77	99.80 ± 1.07	100.00 ± 0.00	38.78 ± 16.33
2	90.30 ± 9.06	98.33 ± 3.49	99.49 ± 1.03	90.94 ± 11.23
3	84.96 ± 11.31	96.14 ± 7.09	92.27 ± 9.76	96.16 ± 4.63
4	97.16 ± 7.39	98.43 ± 4.43	99.60 ± 1.02	94.08 ± 3.94
5	94.86 ± 7.90	100.00 ± 0.00	87.34 ± 12.22	89.69 ± 11.07
Average	92.15 ± 9.29	98.18 ± 3.22	95.74 ± 4.81	81.93 ± 8.52

Table 3. Performance comparison for different combinations of sensors.

Sensor group	C1	C2	C3	C4	C5	Average
Kinect®+IMU+EMG	$100.0 \pm 0.00\%$	$99.60 \pm 1.53\%$	$99.62 \pm 1.03\%$	$99.09 \pm 1.59\%$	$95.76 \pm 3.36\%$	$98.81 \pm 1.81\%$
Kinect®+IMU	$100.0 \pm 0.00\%$	$100.0 \pm 0.00\%$	$97.71 \pm 2.50\%$	$97.95 \pm 2.36\%$	$98.39 \pm 2.10\%$	$98.81 \pm 1.81\%$
Kinect®+EMG	$91.00 \pm 4.78\%$	$100.0 \pm 0.00\%$	$98.8 \pm 1.82\%$	$97.68 \pm 2.51\%$	$95.51 \pm 3.45\%$	$96.81 \pm 2.93\%$
IMU+EMG	$100.0 \pm 0.00\%$	$98.14 \pm 0.02\%$	$99.46 \pm 1.23\%$	$99.67 \pm 0.96\%$	$96.65 \pm 3.00\%$	$98.78 \pm 1.78\%$

The results obtained using combined sensors are reported in Table 3, which shows the average value of the main diagonal of the confusion matrices as well as their uncertainty intervals with a confidence rate of 99%. As shown in the table, the Kinect®+IMU+EMG and the Kinect®+IMU combinations show the best overall performance, with a success rate of 100% for class 1 in both cases and comparable results for the other classes. By comparing these results

with those shown in Table 2, we can see that combining the Kinect® and EMG sensors improves the activity recognition performance by 4.66% with respect to the Kinect® sensor alone and 14.88% for the EMG sensor. The integration of the IMU and EMG sensors yields a similar performance improvement.

4 Conclusions

We developed an automatic method for human activity recognition based on multi-modal data fusion from a network of IMU and EMG sensors and a Kinect® sensor. Our approach uses multi-class support vector machines for primitive movement detection and subsequently classifies the activity according to the sequences provided by the SVM outputs over a time interval using an HMM. This work studies the contribution of each sensor to the recognition task by evaluating the performance of different sensor configurations. To perform robust activity recognition, it is necessary to use all the sensors due to the potential failures that these devices might show during the process. These failures include partial occlusions or self-occlusions from the Kinect® or connection losses in the wireless communication systems, which are commonly used to acquire data from the IMU or EMG sensors. Multi-modal information from every sensor might mitigate mistakes caused by such failures. The proposed approach was tested in an annotated dataset that was created specifically for this work, because there was no publicly available database with synchronized recording of these three sensor modalities. We made the dataset publicly available to facilitate comparisons and accelerate the research in this area. In the future, the database must be expanded to validate our approach on a wider set of activities.

Acknowledgments. The authors want to thank the support from the Master's program in Electrical Engineering at Universidad Tecnologica de Pereira.

References

1. Bayat, A., Pomplun, M., Tran, D.A.: A study on human activity recognition using accelerometer data from smartphones. In: The 11th International Conference on Mobile Systems and Pervasive Computing (2014)
2. de Castro, D.M.: Aplicación Android para el reconocimiento automático de actividades físicas en tiempo real. Master's thesis, Universidad Carlos III de Madrid Departamento de Informática (2012)
3. Leightley, D., Darby, J., Li, B., McPhee, J.S., Yap, M.H.: Human activity recognition for physical rehabilitation. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 261–266, October 2013
4. Destelle, F., et al.: Low-cost accurate skeleton tracking based on fusion of kinect and wearable inertial sensors. In: 2014 Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), pp. 371–375, September 2014
5. Feng, S., Murray-Smith, R.: Fusing Kinect sensor and inertial sensors with multi-rate Kalman filter. In: Data Fusion & Target Tracking 2014: Algorithms and Applications (DF&TT 2014), pp. 1–8 (2014)

6. Ferguson, S., Dunlop, R.G.: Grasp recognition from myoelectric signals. In: Australasian Conference on Robotics and Automation, November 2002
7. Gaglio, S., Re, G.L., Morana, M.: Human activity recognition process using 3-D posture data. IEEE Trans. Hum. Mach. Syst. **45**(5), 586–597 (2015)
8. Ghasemzadeh, H., Barnes, J., Guenterberg, E., Jafari, R.: A phonological expression for physical movement monitoring in body sensor networks. In: 5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2008, pp. 58–68, September 2008
9. Helten, T., Muller, M., Seidel, H.P., Theobalt, C.: Real-time body tracking with one depth camera and inertial sensors. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 1105–1112, December 2013
10. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, 1st edn. Prentice Hall PTR, Upper Saddle River (2001)
11. Kang, W.J., Shiu, J.R., Cheng, C.K., Lai, J.S., Tsao, H.W., Kuo, T.S.: The application of cepstral coefficients and maximum likelihood method in emg pattern recognition [movements classification]. IEEE Trans. Biomed. Eng. **42**(8), 777–785 (1995)
12. Frank, K., Nadales, M.J.V., Robertson, P., Angerman, M.: Reliable real-time recognition of motion related human activities using MEMS inertial sensors. In: 23rd International Technical Meeting of the Satellite Division of the Institute of Navigation (2010)
13. Bocksch, M., Seitz, J., Jahn, J.: Pedestrian activity classification to improve human tracking and localization. In: Fourth International Conference on Indoor Positioning and Indoor Navigation (2013)
14. Pancholi, S., Agarwal, R.: Development of low cost EMG data acquisition system for Arm Activities Recognition. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2465–2469, September 2016
15. Tao, G., Archambault, P., Levin, M.: Evaluation of Kinect skeletal tracking in a virtual reality rehabilitation system for upper limb hemiparesis. In: 2013 International Conference on Virtual Rehabilitation (ICVR), pp. 164–165, August 2013
16. Wu, H., Pan, W., Xiong, X., Xu, S.: Human activity recognition based on the combined svm amp; hmm. In: 2014 IEEE International Conference on Information and Automation (ICIA), pp. 219–224, July 2014
17. Zhang, M., Sawchuk, A.A.: Motion primitive-based human activity recognition using a bag-of-features approach. In: ACM SIGHIT International Health Informatics Symposium, IHI 2012, pp. 631–640. ACM, New York (2012)
18. Zhang, M., Sawchuk, A.A.: A feature selection-based framework for human activity recognition using wearable multimodal sensors. In: International Conference on Body Area Networks (BodyNets), Beijing, China (2011)
19. Zhang, Z., Liu, Y., Li, A., Wang, M.: A novel method for user-defined human posture recognition using Kinect. In: 7th International Congress on Image and Signal Processing (CISP), pp. 736–740, October 2014