CONSENSUS INFERENCE ON MOBILE PHONE SENSORS FOR ACTIVITY RECOGNITION

Huan Song[†], *Jayaraman J. Thiagarajan*[‡], *Karthikeyan Natesan Ramamurthy*[⋆], *Andreas Spanias*[†] *and Pavan Turaga*[†]
[†] SenSIP Center, ECEE, Arizona State University, Tempe, AZ
[‡] Lawrence Livermore National Labs, 7000 East Avenue, Livermore, CA
[⋆] IBM T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY

ABSTRACT

The pervasive use of wearable sensors in activity and health monitoring presents a huge potential for building novel data analysis and prediction frameworks. In particular, approaches that can harness data from a diverse set of low-cost sensors for recognition are needed. Many of the existing approaches rely heavily on elaborate feature engineering to build robust recognition systems, and their performance is often limited by the inaccuracies in the data. In this paper, we develop a novel two-stage recognition system that enables a systematic fusion of complementary information from multiple sensors in a linear graph embedding setting, while employing an ensemble classifier phase that leverages the discriminative power of different feature extraction strategies. Experimental results on a challenging dataset show that our framework greatly improves the recognition performance when compared to using any single sensor.

Index Terms— Activity recognition, Sensor fusion, Multi-layer graph, Time-delay embedding, Reference-based classification

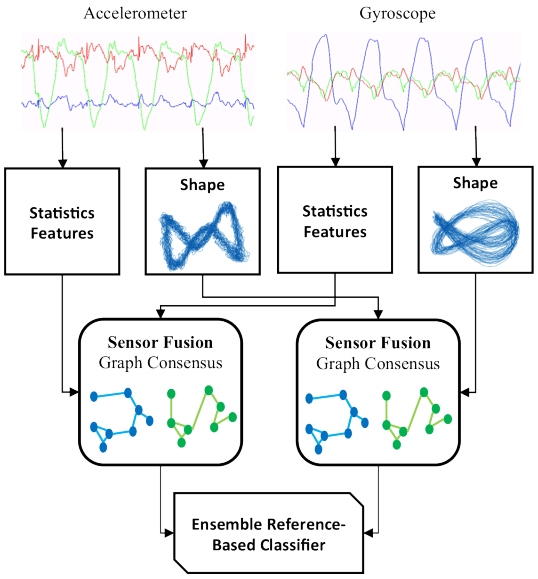


Fig. 1: Proposed two-stage architecture for activity recognition on mobile devices.

1. INTRODUCTION

The use of mobile devices and wearable sensors for activity monitoring has become an important research problem in the recent years. By enabling the design of intelligent systems, this technology has made big strides in the healthcare industry. Though the targeted applications can be very different [1, 2], the overarching goal is to analyze the data collected using the inherent sensing modalities and obtain predictive inferences using low complexity algorithms. Some of the commonly used sensors include the accelerometer, gyroscope, magnetometer and GPS (Global Positioning System), to name a few. The challenges in building effective predictive algorithms for such mobile devices are twofold. On one hand, though abundantly available, the data collected from these cheap sensors are often very noisy and unreliable. Note that, the noise and inaccuracies can be caused by the sensors themselves or due to arbitrary position/movement of the devices during the activities. On the other hand, it can be prohibitive in terms of both time and resource availability, to employ complex machine learning techniques to process this data. While there have been significant advances in activity recognition using data from high-performance, standalone sensors attached to human body [3, 4], adapting them to the case of low-cost, mobile sensors is not straightforward. However, building effective activity recognition techniques for such low-cost devices can have a significant impact in applications ranging from fitness monitoring [5] to elderly care [6, 7].

Existing approaches for activity recognition often rely on applying a variety of signal processing methods to collect a large set of statistics, and using computationally intensive feature extraction techniques to identify the most relevant features [8, 9, 10]. However, common wisdom from machine learning can show that such elaborate feature-engineering to limited training data need not generalize well to novel test data. In spite of the availability of different sensors, activity

This research was supported in part by the SenSIP center, NSF award 15400540 and Intel Corporation.

recognition is often carried out solely based on accelerometer data [9]. Using data from other sensors can potentially improve the recognition performance, and make the predictor highly robust to measurement inaccuracies. For example, in [11], Zhu *et.al.* proposed to fuse data from two inertial sensors attached to waist and foot, using a neural network and hidden Markov model classifiers.

In this paper, we propose a novel two-stage architecture for activity recognition using multi-modal data obtained from the same mobile device. Though we describe the proposed algorithm using two sensors, this can be easily generalized to other cases. In [4], Zhang *et.al.,* developed an approach to fuse accelerometer and gyroscope sensors by simply concatenating the feature vectors. However, this naive approach is known not to exploit the inherent geometry of the different feature domains, and hence not preferred in complex recognition tasks [12, 13]. More recently, a consensus inference approach for multimodal fusion was developed in [14]. Our approach supports the use of more than one feature extraction strategy on data from each sensor. For each of the sensors, we propose to use a set of simple statistics and a shape feature that characterizes the periodic structure of the time-series data. Figure 1 illustrates the proposed algorithm using data from accelerometer and gyroscope sensors. The first stage of our architecture performs sensor fusion, for each feature independently, using a linearized variant of the multilayer graph consensus approach in [14]. In the second stage, the two sets of consensus features are used to build a reference-based ensemble classifier to make the final prediction. We tested the proposed approach with real data collected from 32 subjects performing primitive activities, and results show that the proposed two stage approach can improve the performance significantly, when compared to using a single sensor.

2. MULTI-MODAL CONSENSUS FRAMEWORK

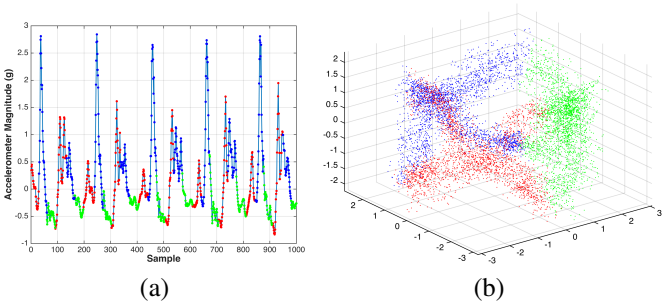
As shown in Figure 1, the proposed framework consists of three steps: (a) extract statistical and shape features from segments using windowing, (b) perform sensor fusion for each feature type across all modalities based on multilayer graphs, and (c) use ensemble reference-based classifier on the different types of fused feature for recognition.

2.1. Feature Extraction

We extract statistical features that have proved to be useful for activity recognition [4, 9, 8], and investigate the time delay embedding of the activity signals and propose to use a basic version of shape features.

2.1.1. Statistical Features

The statistical features we extracted are: mean, median, standard deviation, kurtosis, skewness, total acceleration, mean-crossing rate, autoregressive (AR) coefficients [15] and dominant frequency. Each activity signal was first windowed into 5 second non-overlapping segments. This length was chosen empirically such that there is sufficient periodic structure in each segment. The AR coefficients were extracted assuming each segment to be a stationary random signal [16]. The model order was determined to be 3 based on the Akaike information criterion [17]. The dominant frequency is defined as the frequency component having the largest FFT magnitude [8]. These statistical features were extracted separately from signals corresponding to the three axes of each sensor and then concatenated together. For both the accelerometer and gyroscope, the overall dimension of the statistical features is 31.



(a) (b)

Fig. 2: Extracting shape features - (a) Raw accelerometer data, (b) 3-D PCA representation of its delay embedding.

2.1.2. Shape Feature From Time-Delay Embedding

Given a short sequence of measurements, time-delay embedding (TDE) [18] is an approach for reconstructing the underlying system dynamics. Two important parameters for calculating the TDE are: the dimension of reconstruction space m and time delay τ. Given a time series $\mathbf{o}$, the TDE can be represented as a matrix $\mathbf{O}$ whose ith column is $[o_i, o_{i+\tau}, o_{i+2\tau}, ..., o_{i+(m-1)\tau}]$. Figure 2 visualizes the raw accelerometer signal for "fast walking" and its corresponding TDE representation. In Figure 2(b), we cluster the samples in the 3-D PCA representation of TDE and mark different clusters with specific colors. The corresponding activity samples are then marked the same color and illustrated in Figure 2(a). Notice that across the periods of the activity signal, the clusters map to very similar regions. This shows that TDE represents the periodic structure of the signal as desired and we can derive suitable features from it for the classification task.

We extract a simple shape function based on the geometric distance property, and use it to derive our feature. The shape function we consider measures the pair-wise distance between samples in the TDE space, calculated as $\mathbf{S}_{ij} = \|\mathbf{o}_i - \mathbf{o}_j\|_2$ [19]. A histogram is constructed on these distances with specified bin size to obtain the feature.

2.2. Sensor Fusion Using Multilayer Graph

The goal of the sensor fusion is to obtain a unified feature for each activity segment by fusing similar features from the two modalities (accelerometer and gyroscope). We adapt the multilayer graph consensus approach in [14], where each layer represents a single modality containing an intra- and inter-class graph corresponding to the class relationships of the activities. We estimate linear local discriminant embeddings (LDE) instead of kernel embeddings on the graphs to keep the process computationally simple. Figure 3 shows the overview of this process for a given feature type. Note that we will obtain separate consensus projections for the two feature types, namely statistical and shape.

Denote the T modalities in one feature type as $\{\mathbf{X}_t\}_{t=1}^T$, where the columns of $\mathbf{X}_t \in \mathbb{R}^{M_t \times N}$ correspond to the features extracted from each activity segment. The label for an activity segment i is denoted by l_i. We construct the intra- and inter-class graphs for modality t and represent the adjacency matrices as $\mathbf{W}_t$ and $\mathbf{W}'_t$, whose elements are defined using the Gaussian RBF with parameter γ,

$$w_{t,ij} = \begin{cases} e^{-\gamma \|\mathbf{x}_{t,i} - \mathbf{x}_{t,j}\|^2}, & \text{if } l_i = l_j, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

$$w'_{t,ij} = \begin{cases} e^{-\gamma \|\mathbf{x}_{t,i} - \mathbf{x}_{t,j}\|^2}, & \text{if } l_i \neq l_j, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

The idea of linear LDE [20] is to construct the low-dimensional embedding $\mathbf{V}_t = \mathbf{U}_t^T \mathbf{X}_t$, $\mathbf{U}_t \in \mathbb{R}^{M_t \times D}$ (D is the dimension of projection) being the projection matrix, such that the neighboring points of same class in the ambient space are still close, whereas the neighboring points from different classes are distant. Defining the Laplacian matrices $\mathbf{L}_t = \mathbf{D}_t - \mathbf{W}_t$, and $\mathbf{L}'_t = \mathbf{D}'_t - \mathbf{W}'_t$, where $\mathbf{D}_t$ and $\mathbf{D}'_t$ are the respective diagonal degree matrices, the individual discriminant projections can be computed as the trace-ratio maximization [21],

$$\mathbf{U}_t = \underset{\mathbf{U}_t : \mathbf{U}_t^T \mathbf{U}_t = \mathbf{I}}{\arg\max} \frac{\mathrm{Tr}(\mathbf{U}_t^T \mathbf{X}_t \mathbf{L}'_t \mathbf{X}_t^T \mathbf{U}_t)}{\mathrm{Tr}(\mathbf{U}_t^T \mathbf{X}_t \mathbf{L}_t \mathbf{X}_t^T \mathbf{U}_t)}. \tag{3}$$

Since the individual projections belong to the Grassmann manifold, the consensus projection $\mathbf{U}$ can be obtained as geometric mean of the individual projections with respect to the chordal distance [22],

$$d^2_{proj} = D - \sum_{t=1}^{T} \mathrm{Tr}(\mathbf{U}\mathbf{U}^T \mathbf{U}_t \mathbf{U}_t^T). \tag{4}$$

We also require the consensus projection to be discriminative across all the modalities. Combining this with (4), the final optimization is

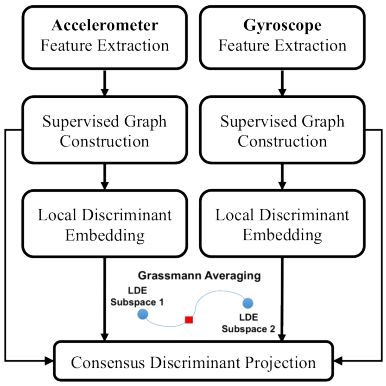


Fig. 3: Multilayer graph consensus algorithm for fusing features from two different sensors.

$$\min_{\mathbf{U}} \quad \sum_{t=1}^{T} \mathrm{Tr}(\mathbf{U}^T \mathbf{X}_t \mathbf{L}_t \mathbf{X}_t^T \mathbf{U}) - \alpha \sum_{t=1}^{T} \mathrm{Tr}(\mathbf{U}\mathbf{U}^T \mathbf{U}_t \mathbf{U}_t^T)$$

$$\text{s.t.} \quad \sum_{t=1}^{T} \mathrm{Tr}(\mathbf{U}^T \mathbf{X}_t \mathbf{L}'_t \mathbf{X}_t^T \mathbf{U}) = c, \mathbf{U}^T \mathbf{U} = \mathbf{I}$$

where α is the trade-off parameter. This can be posed as the trace-ratio maximization,

$$\max_{\mathbf{U}:\mathbf{U}^T\mathbf{U}=\mathbf{I}} \frac{\mathrm{Tr}\left(\mathbf{U}^T \left(\sum_{t=1}^{T} \mathbf{X}_t \mathbf{L}'_t \mathbf{X}_t^T\right) \mathbf{U}\right)}{\mathrm{Tr}\left(\mathbf{U}^T \left(\sum_{t=1}^{T} \mathbf{X}_t \left(\mathbf{L}_t - \alpha \mathbf{U}_t \mathbf{U}_t^T\right) \mathbf{X}_t^T\right) \mathbf{U}\right)}$$

and solved using the decomposed Newton's or the iterative trace ratio method [21]. The out-of-sample projection for the test data $\{\mathbf{Y}_t\}_{t=1}^T$ is obtained as $\mathbf{Z} = \sum_{t=1}^{T} \mathbf{U}^T \mathbf{Y}_t$.

2.3. Ensemble Reference-Based Classification

Given different types of consensus features, it is important that the classification mechanism can effectively combine them. We extend the reference-based classification in [23] using an ensemble classification approach. Different from [23], we use the whole training data as the reference set and we perform inference directly based on the combined similarity matrix between a probe sample and the reference set. This simplifies the classification and also takes into consideration characteristics of both features. The detailed steps are:

1. Denote a probe sample as $\mathbf{z}$ and the F consensus features as $\{\mathbf{V}_f\}_{f=1}^F$. We construct the similarity vector $\mathbf{s}_f$, where each element s is the similarity between the probe sample and one sample $\mathbf{v}$ in the training feature $\mathbf{V}_f$[23], $s = 1 - \frac{\gamma(\frac{k}{2}, \frac{d^{\mathbf{v}}_{\mathbf{z}}}{2})}{\Gamma(\frac{k}{2})}$. Here $d^{\mathbf{v}}_{\mathbf{z}}$ is the Euclidean distance between the probe sample $\mathbf{z}$ and the reference sample $\mathbf{v}$. Γ is the Gamma function and γ is the lower incomplete Gamma function with parameter k.

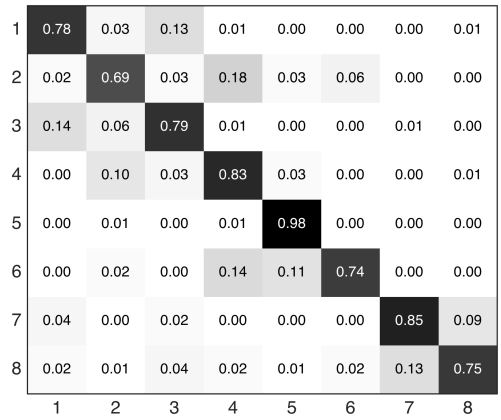


Fig. 4: Average confusion matrix of the proposed recongition algorithm.

2. Select the top $K = 30$ closest samples from every class and form the new similarity vectors $\{\mathbf{s}_f^{'}\}_{f=1}^{F}$.

3. Denote the elements containing similarities to class c as $(\mathbf{s}_f^{'})^c$. We perform the ensemble for measuring the closeness of the probe sample to class c as, $S^c = \sum_f \sum_n (\mathbf{s}_f^{'})^c$. The inference can then be directly carried out by assigning the label of the class having largest S^c value.

3. EXPERIMENTS AND RESULTS

3.1. Description of the Data

Human movement in daily activities are complex in nature. Even for the same activity, the styles can be largely different among people. Hence, we collected data from a set of 32 subjects with diversity in gender, age, weight, and height. The statistics of these demographic factors are shown in Table 1. Each subject performed 5 different activities, - slow walking, fast walking, running, slow biking, and fast biking - using a treadmill or biking machine. The first three activities were performed twice with the subjects carrying the mobile phones first in their front pockets and then in their back pockets, whereas biking activities were performed with mobile phone only in front pockets. As a result, the dataset contains data from 8 classes in total. The labeling also follows this order. The duration of each activity was 75 seconds and the speeds were fixed. The Nexus 4 Android phone that we used had one 3-axis accelerometer, and one 3-axis gyroscope to measure the amount of rotation. We set the sampling rates of the sensors at 200Hz through the Android APK interface.

3.2. Results

We performed 5-fold cross-validation on this dataset by a random split of data according to subject label. In other words, in each validation the training and testing data do not come from the same subject. This setting increases the challenge of the task but better simulates real-world applications.

Table 1: Demographic statistics of the subjects that participated in our data collection experiment.

Statistics	Mean	STD	Range
Age	30.5	7.8	20-52
Height (cm)	174.8	9.5	155-191
Weight (kg)	73.9	14.0	42-108

Table 2 compares the recognition rates in each step of our framework, i.e., using the two sensors independently with each of the features, consensus of the two sensors for each of the features, and finally our two stage architecture. We observed that, with both the feature extraction strategies, sensor fusion performs better than using any sensor alone. However, using the ensemble classifier improves the performance significantly, providing an improvement of around 10% over the best results obtained with a single sensor. Figure 4 plots the confusion matrix for the 8 classes, obtained using the proposed algorithm.

Table 2: Activity recognition performance obtained using different combinations of sensors and features, in comparison to the proposed two stage architecture.

Sensor	Feature	Recognition Rate
Accelerometer	Shape (LDE)	57.8
Gyroscope	Shape (LDE)	51.66
Shape (Consensus)		68.73
Accelerometer	Stats (LDE)	70.05
Gyroscope	Stats (LDE)	69.95
Stats (Consensus)		73.2
Two Stage Approach		**80.14**

4. CONCLUSIONS

In this paper, we developed a novel approach for activity recognition by fusing multiple distinct features from multiple sensors. In particular, we presented a linearized variant of the multilayer graph consensus technique and effectively combined the discriminative capabilities of multiple sensors. Also we adopted a simple, reference-based classifier and fused the decisions from two distinct feature sets. We observed from our results that the framework can produce high quality recognition performances. Though we demonstrated our setup with this particular choice of sensors and features, the proposed two stage architecture is generally enough to be adapted to other applications as well.

5. REFERENCES

[1] A. Spanias, T. Painter, and V. Atti, *Audio signal processing and coding*, John Wiley & Sons, 2006.

[2] J.J. Thiagarajan, K.N. Ramamurthy, P. Turaga, and A. Spanias, "Image understanding using sparse representations," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 7, no. 1, pp. 1–118, 2014.

[3] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *ARCS, 2010 23rd international conference on*. VDE, 2010, pp. 1–10.

[4] M. Zhang and A. A Sawchuk, "Human daily activity recognition with sparse representation using wearable sensors," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 3, pp. 553–560, 2013.

[5] I. Anderson, J. Maitland, S. Sherwood, L. Barkhuus, M. Chalmers, M. Hall, B. Brown, and H. Muller, "Shakra: tracking and sharing daily activity levels with unaugmented mobile phones," *Mobile Networks and Applications*, vol. 12, no. 2-3, pp. 185–199, 2007.

[6] C. Orwat, A. Graefe, and T. Faulwasser, "Towards pervasive computing in health care–a literature review," *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, pp. 26, 2008.

[7] B. Das, C. Chen, A.M. Seelye, and D.J. Cook, "An automated prompting system for smart environments," in *Toward Useful Services for Elderly and People with Disabilities*, pp. 9–16. Springer, 2011.

[8] M. Zhang and A. A Sawchuk, "A feature selection-based framework for human activity recognition using wearable multimodal sensors," in *Proceedings of the 6th International Conference on Body Area Networks*. ICST, 2011, pp. 92–98.

[9] J. R Kwapisz, G. M Weiss, and S. A Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[10] J. Frank, S. Mannor, and D. Precup, "Activity and gait recognition with time-delay embeddings.," in *AAAI*. Citeseer, 2010.

[11] C. Zhu and W.H. Sheng, "Human daily activity recognition in robot-assisted living using multi-sensor fusion," in *Robotics and Automation, 2009. ICRA. IEEE International Conference on*. IEEE, 2009, pp. 2154–2159.

[12] J.J. Thiagarajan, K.N. Ramamurthy, and A. Spanias, "Multiple kernel sparse representations for supervised and unsupervised learning," *Image Processing, IEEE Transactions on*, vol. 23, no. 7, pp. 2905–2915, July 2014.

[13] N. Kovvali, M. Banavar, and A. Spanias, "An introduction to kalman filtering with matlab examples," *Synthesis Lectures on Signal Processing*, vol. 6, no. 2, pp. 1–81, 2013.

[14] K.N. Ramamurthy, J.J. Thiagarajan, R. Sridhar, P. Kothandaraman, and R. Nachiappan, "Consensus inference with multilayer graphs for multi-modal data," in *Signals, Systems and Computers, 2014 48th Asilomar Conference on*, Nov 2014, pp. 1341–1345.

[15] A. Spanias, "Digital signal processing: An interactive approach," Tech. Rep., ISBN 978-1-4675-9892-7, Morrisville, NC: Lulu Press On-demand Publishers, 2014.

[16] Z.-Yu He and L.-Wen Jin, "Activity recognition from acceleration data using ar model representation and svm," in *Machine Learning and Cybernetics, 2008 International Conference on*. IEEE, 2008, vol. 4, pp. 2245–2250.

[17] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[18] T. Sauer, J. A Yorke, and M. Casdagli, "Embedology," *Journal of statistical Physics*, vol. 65, no. 3-4, pp. 579–616, 1991.

[19] V. Venkataraman, P. Turaga, Nicole Lehrer, Michael B., Thanassis R., and S.L. Wolf, "Attractor-shape for dynamical analysis of human movement: Applications in stroke rehabilitation and action recognition," in *CVPRW, 2013 IEEE Conference on*. IEEE, 2013, pp. 514–520.

[20] H.-Tzong Chen, H.-Wei Chang, and T.-Luh Liu, "Local discriminant embedding and its variants," in *CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 2, pp. 846–853.

[21] Y.Q. Jia, F.P. Nie, and C.S. Zhang, "Trace ratio problem revisited," *Neural Networks, IEEE Transactions on*, vol. 20, no. 4, pp. 729–735, 2009.

[22] K. Ye and L.H. Lim, "Distance between subspaces of different dimensions," *arXiv preprint arXiv:1407.0900*, 2014.

[23] Q. Li, H.G. Zhang, J. Guo, B. Bhanu, and L. An, "Reference-based scheme combined with k-svd for scene image categorization," *Signal Processing Letters, IEEE*, vol. 20, no. 1, pp. 67–70, 2013.