Multimodal Physical Activity Recognition by Fusing Temporal and Cepstral Information

Ming Li, *Student Member, IEEE*, Viktor Rozgić *Student Member, IEEE*, Gautam Thatte, *Student Member, IEEE*,
Sangwon Lee, Adar Emken, Murali Annavaram, Urbashi Mitra, *Fellow, IEEE*, Donna Spruijt-Metz,
Shrikanth Narayanan, *Fellow, IEEE*

***Abstract*—A physical activity (PA) recognition algorithm for a wearable wireless sensor network using both ambulatory electrocardiogram (ECG) and accelerometer signals is proposed. First, in the time domain, the cardiac activity mean and the motion artifact noise of the ECG signal are modeled by a Hermite polynomial expansion and principal component analysis, respectively. A set of time domain accelerometer features is also extracted. A support vector machine (SVM) is employed for supervised classification using these time domain features. Second, motivated by their potential for handling convolutional noise, cepstral features extracted from ECG and accelerometer signals based on a frame level analysis are modeled using Gaussian mixture models (GMM). Third, to reduce the dimension of the tri-axial accelerometer cepstral features which are concatenated and fused at the feature level, heteroscedastic linear discriminant analysis is performed. Finally, to improve the overall recognition performance, fusion of the multi-modal (ECG and accelerometer) and multi-domain (time domain SVM and cepstral domain GMM) subsystems at the score level is performed. The classification accuracy ranges from 79.3% to 97.3% for various testing scenarios and outperforms the state-of-the-art single accelerometer based PA recognition system by over 24% relative error reduction on our 9-category PA database.**

***Index Terms*—Physical activity recognition, Electrocardiogram, Accelerometer, Multimodal signal processing, Cepstrum.**

I. INTRODUCTION

AUTOMATIC recognition of physical activity (PA) with wearable sensors can provide feedback about an individual's lifestyle and mobility patterns. Such information can form the basis for new types of health assessment, rehabilitation, and intervention tools to help people maintain their energy balance and stay physically fit and healthy.

Recently, promising results from wearable body accelerometers in single or multiple locations for detecting PA have been presented [1]–[9]. Both [3] and [4] offer comprehensive summaries of existing accelerometer-based approaches. It has been shown in [2] that a system with five accelerometers improved the average accuracy of PA recognition by 35% compared to a system with a single accelerometer. However, placing wearable sensors in multiple body locations can be quite cumbersome when the user has to collect data on a daily basis or for longer periods of continuous monitoring. Thus, many approaches based on multiple integrated sensor modalities have been proposed, since it is much more comfortable for the user to wear a single device. Moreover, incorporating multimodal information can yield additional physiological and environmental cues, such as heart rate, light, skin resistance, temperature, audio, global positioning system (GPS) location, etc [10]–[13]. It is in this context that we examined the validity and feasibility of using multimodal wearable sensors in a laboratory setting within the KNOWME network to discriminate between various categories of PAs.

The KNOWME network [14]–[17] is developed to target technology-centric applications in health care such as pediatric obesity. The KNOWME network utilizes heterogeneous sensors simultaneously, which send their measurements to a Nokia N95 cellphone via Bluetooth, as shown in Fig. 1. Flexible sensor measurement choices can include ECG signals, accelerometer signals, heart rate, and blood oxygen levels as well as other vital signs. Furthermore, external sensor data are combined with data from the mobile phone's built-in sensors (GPS and accelerometer signal). Thus, the mobile phone can display and transmit the combined health record to a back-end server (*e.g.* Google Health Server [18]) in real time.

In this study, we use ECG and accelerometer signals in the KNOWME network to detect PA categories. This sensor choice is common and frequently used in many studies for multimodal PA recognition [12], [13], [19], [20]. ECG is a physiological signal which accompanies physical measurements and therefore has great potential to increase the accuracy of PA recognition. There already exist several commercial ECG monitors with built-in accelerometers [21]; thus, users only need to wear one single multimodal sensor of this type

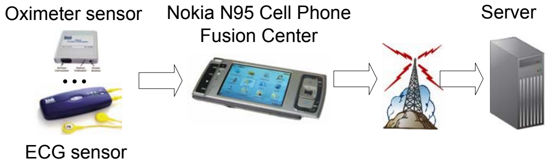


Fig. 1. KNOWME wearable body area network system

Manuscript received October 1, 2009; revised February 24, 2010; accepted May 8, 2010. This work was supported by the National Center on Minority Health and Health Disparities (supplement to P60 MD002254), Nokia and Qualcomm.

M. Li, V. Rozgić, G. Thatte, S. Lee, M. Annavaram, U. Mitra and S. Narayanan are with Viterbi School of Engineering of University of Southern California, Los Angeles, CA, 90089, USA. (email: mingli@usc.edu; rozgic@usc.edu; thatte@usc.edu; sangwonl@usc.edu; annavara@usc.edu; ubli@usc.edu; shri@sipi.usc.edu)

A. Emken and D. Spruijt-Metz are with Keck School of Medicine of University of Southern California, Los Angeles, CA, 91803, USA. (email: emken@usc.edu; dmetz@usc.edu)

and can feel more comfortable while carrying out their daily lives. Finally, the ECG is a very important diagnostic tool and is widely used in a great majority of mobile health systems. A study of the relationship between PAs and the ECG signal can be useful in health monitoring applications.

The ECG sensor measures the change in electrical potential over time. A single normal cycle of the ECG represents the successive atrial depolarization/repolarization and ventricular depolarization/repolarization. The advantage of the wearable ECG devices is that they can be used both in a hospital setting and under free living conditions. The practical challenge is that the ECG signal is often contaminated by noise and artifacts within the frequency band of interest, which can manifest with similar morphologies as the ECG itself [20]. Instant heart rate extracted from the ECG signal has been studied in distinguishing PAs in conjunction with accelerometer data [12]–[14], [22], and results showed that only modest gains were achieved [22]. Recently, it has been shown in [19], [23] that the motion artifacts in a single-lead wearable ECG signal induced by body movement of an ambulatory patient can be detected and reduced by a principal component analysis (PCA) based classification approach. Thus, in addition to heart rate details, ECG signals contain additional discriminative information about PA. In the proposed work, we extend the development in [23] by using Hermite polynomial expansion (HPE) and PCA to describe the cardiac activity mean (CAM) and motion artifact noise (MAN), respectively. Furthermore, instant heart rate variability (mean/variance) and heartbeat shape variability (noise measure within a window) are combined with HPE and PCA coefficients to generate a set of ECG temporal features and used for PA classification.

In contrast to the ECG signal, the accelerometer signal has been studied extensively for PA recognition. There exists a wide range of features and algorithms for supervised classification of PAs with accelerometer derived features. Commonly used methods in the context of activity recognition include Naive Bayes classifiers [1], [2], [6], [9], [22], C4.5 decision trees [1], [2], [6], [9], [12], [13], [22], nearest neighbor methods [1], [2], [9], boosting [6], [10], support vector machines (SVMs) [6], [8], and Hidden markov models (HMM) [5], [7]. A comparison of these methods is reported in [3], [4], [6], [9]. Moreover, a variety of features in both time and frequency domains have been adopted [1]–[4], [9]. In general, the SVM classifier based on temporal feature statistics was found to be one of the best performing systems [6], [8]. In this work, a set of conventional temporal features is extracted from accelerometer signals and used for PA classification.

The temporal features from ECG and accelerometers are modeled using a support vector machine (SVM). The generalized linear discriminative sequence (GLDS) kernel [24] was employed due to its good classification performance and low computational complexity. The GLDS kernel uses a discriminative classification metric that is simply an inner product between the averaged feature vector and model vector and thus is very computationally efficient with small model size, making it attractive for mobile device implementations.

More recently, promising results in biometrics [25] have shown that cepstral features of stethoscope-collected heart sound signals can be used to identify different persons. This inspired us to explore the potential of cepstral domain ECG features for PA detection. Compared to time-domain fiducial points or the PCA approach, cepstral feature calculation uses short fixed length processing windows and thus does not need the pre-processing steps of heartbeat segmentation and normalization. Furthermore, for accelerometer signals, the evaluation in [4] shows that Fast Fourier Transform (FFT) features always rank among the features with the highest precision, but the FFT coefficients that attain the highest precision are different for each activity type. Therefore, combining different FFT coefficients within filter bands might provide a good compromise versus using individual spectral coefficients. Thus, in the proposed work, linear filter bank based cepstral features extracted from both accelerometer and ECG signals are used to measure the cepstral characteristics of different PAs. The cepstral features corresponding to different PA types are modeled using Gaussian Mixture Models (GMMs). We combine both temporal and cepstral information at the score level to improve the system performance. We hypothesize that cepstral features can capture the spectral envelope variations in both ECG and accelerometer signals and thus can complement conventional time domain features. Also, as described in Section II, cepstral features provide a natural way for handling convolutional noise inherent in the sensor measurements. Moreover, fusing system outputs from multiple modalities at the score level can also improve performance [26]. ECG and accelerometer cepstral features are not concatenated and fused at the feature level due to compatibility issues arising from different time shift and window length configurations and different sampling frequencies. However, the cepstral features from each axis of the accelerometer are concatenated to construct a long cepstral feature vector in each frame. Heteroscedastic linear discriminant analysis (HLDA) [27] is used to perform feature dimension reduction. As a special form of (single state) HMM, a GMM model is developed for each activity by using a sequence of feature vectors, rather than individual instances, with a view toward better capturing the temporal dynamics. As shown in Fig. 2, after the classification scores of both the temporal feature based SVM systems and the cepstral feature based GMM systems are available, the four individual system outcomes are fused at the score level to generate the final recognized activity.

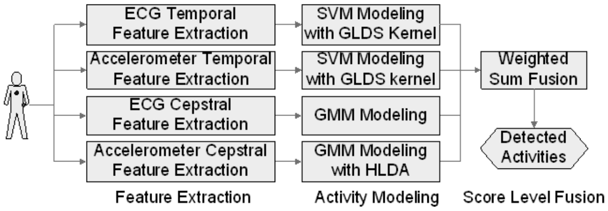

ECG Temporal Feature Extraction
Accelerometer Temporal Feature Extraction
ECG Cepstral Feature Extraction
Accelerometer Cepstral Feature Extraction
SVM Modeling with GLDS Kernel
SVM Modeling with GLDS kernel
GMM Modeling
GMM Modeling with HLDA
Weighted Sum Fusion
Detected Activities
Feature Extraction
Activity Modeling
Score Level Fusion

Fig. 2. The proposed physical activity recognition system overview

Just as individual variability can have significant impact on the interpretation of both the accelerometer and ECG data [28], [29], session variability is another important issue in PA recognition. In real life applications, many other factors can influence or even modify the desired sensor signals, such

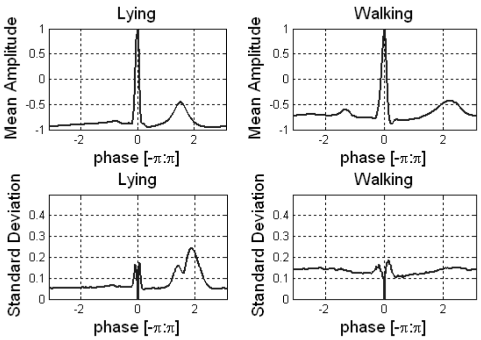


Fig. 3. The mean and standard deviation of normalized ECG signals.

as sensor placement location, user emotion, fitness, etc. Even within the same activity, an individual can perform various styles of PA, which might not appear in the training set, and thus decrease the system performance. In this study, the session variability of the ECG and accelerometer signals is studied under subject dependent modeling framework.

In summary, we address the PA recognition problem with multimodal wearable sensors (ECG and accelerometer) in this work. The contributions are as follows: (1) The cardiac activity mean (CAM) component of the ECG signal is described by Hermite polynomial expansion (HPE) in the temporal feature extraction. (2) In the SVM framework for both ECG and accelerometer temporal features, the GLDS kernel makes the classification computationally efficient with a small model size. (3) A GMM system based on cepstral features is proposed to capture the frequency domain information in a robust fashion against convolutional effects, and HLDA is used to reduce the feature dimension of tri-axial accelerometer based measurements. (4) Score level fusion of the multi-modal and multi-domain subsystems is performed to improve the overall performance (5) The effects of session variability of ECG and accelerometer measurements on PA recognition are studied.

The remainder of the paper is organized as follows. The description of the proposed multimodal PA recognition system will be provided in the following sections: feature extraction in Section II, activity modeling in Section III, and system fusion in Section IV. Section V presents the experimental setup and results followed by a discussion in Section VI. Section VII provides the paper's conclusion.

II. Feature extraction

A feature is a characteristic measurement, transform, or structural mapping extracted from the input data to represent important patterns of desired phenomena (PA in our case) with reduced dimension. For example, the standard deviation of an accelerometer reading and the mean of the instantaneous heart rate via the ECG are good candidates as PA cues or features. Furthermore, utilizing the complementary characteristics of different types of features can offer substantial improvement over single type features in the recognition accuracy depending upon the information being combined and the fusion methodology adopted [26]. In this section, we describe the proposed time domain and cepstral feature extraction process in detail.

A. Temporal feature extraction

We consider four types of temporal features. Features in the first set, which we denote as "conventional", were selected based on their efficacy as demonstrated in the literature regarding wireless body area sensor networks; for the accelerometer, the conventional features are shown in Table I, and for the ECG sensor, the mean and variance of the instantaneous heart-rate constitute the conventional features. The other three features sets are comprised of features that describe the discriminative activity information for the ECG signals. These features result from more complex processing of the ECG signal: (i) the principal component analysis (PCA) error vector, which has been previously studied in [23] for body movement activity recognition, (ii) the Hermite polynomial expansion (HPE) coefficients, and (iii) the standard deviation of multiple normalized beats which are novel to our work. These techniques model the underlying signals, and the resultant model parameters are the features. First, we describe the required pre-processing of the collected biometric ECG signal, then we describe the ECG temporal feature extraction, and finally we outline the temporal accelerometer features.

1) Pre-processing of the ECG Signal: Each type of body movement induces a particular type of motion artifact in the ECG signal. If there are M hypothesized activities, for the j^{th} heartbeat observation under the i^{th} activity, the continuous-time recorded ECG signal, $r_{ij}(t)$, is modeled as [19], [23]

$$r_{ij}(t) = \theta_i(t) + \chi_{ij}(t) + \eta_{ij}(t), \tag{1}$$

where $\theta_i(t)$ is the cardiac activity mean (CAM) which is the normal heart signal, $\chi_{ij}(t)$ is an additive motion artifact noise (MAN) due to i^{th} class of activities, and $\eta_{ij}(t)$ is the sensor noise present in the ECG signal. Since the length of each heartbeat is different due to inherent heart rate variability, the first step of pre-processing normalizes each heartbeat waveform to the same time duration (in the phase domain) and amplitude range [20], [23]. Due to the low signal to noise ratio (SNR) of the ECG signal in high intensity PA, fake peak elimination and valid beat selection [20] are performed to enhance the robustness and reduce the peak detection error.

The D-dimensional vector representation of $r_{ij}(t)$ over one heartbeat is denoted $\boldsymbol{r}_{ij}$ and the D-dimensional vector representations of the corresponding CAM, MAN, and sensor noise components are $\boldsymbol{\theta}_i$, $\boldsymbol{\chi}_{ij}$, and $\boldsymbol{\eta}_{ij}$, respectively. Fig. 3 shows the mean and standard deviation of the normalized ECG signal for different activities. One of our innovations over [23] is the recognition that both CAM and MAN carry discriminative information between different PAs.

2) Principal Component Analysis: Principal component analysis (PCA) is used for feature extraction from the MAN component $\boldsymbol{\chi}_{ij}$. For the i^{th} activity class, we use ν_i heartbeats to estimate the CAM $\boldsymbol{\theta}_i$ (as in [23]),

$$\tilde{\boldsymbol{\theta}}_i = \frac{1}{\nu_i}\sum_{j=1}^{\nu_i} \boldsymbol{r}_{ij}. \tag{2}$$

We note that the number of heartbeats available for training, ν_i, is different for each of the activities. Subtracting the CAM

TABLE I
CONVENTIONAL TEMPORAL ACCELEROMETER FEATURES

mean absolute deviation	zero crossing rate	energy
$(20, 40, 60, 80)^{th}$ percentile	spectral entropy	kurtosis
cross correlation	mean crossing rate	median
mean of maxima	mean of minima	mean
standard deviation	root mean square	skewness

from the signal $\boldsymbol{r}_{ij}$ yields residual activity vectors $\acute{\boldsymbol{r}}_{ij}$:

$$\acute{\boldsymbol{r}}_{ij} = \boldsymbol{r}_{ij} - \tilde{\boldsymbol{\theta}}_i = \boldsymbol{\chi}_{ij} + \acute{\boldsymbol{\eta}}_{ij}, \tag{3}$$

where $\acute{\boldsymbol{\eta}}_{ij}$ includes both the sensor noise and the CAM estimation noise induced by the session variability. As noted in [23], although the signal component due to MAN has smaller amplitude than CAM, it has much greater amplitude than the sensor noise, *i.e.*, $|\boldsymbol{\eta}| \ll |\boldsymbol{\chi}_i| < |\boldsymbol{\theta}_i|$, $\forall i$ (where $|\cdot|$ is the 2-norm). Thus, the MAN has a dominant influence on the shape of the residual activity vector $\acute{\boldsymbol{r}}_i$. For each activity class i, we now compute eigenvectors and eigenvalues using the eigen-decomposition of the covariance matrix $\boldsymbol{\Sigma}_i$ of $\acute{\boldsymbol{r}}_{ij}$. Let $\mathbf{E}_i = [\mathbf{e}_{i0}, \mathbf{e}_{i1}, \cdots, \mathbf{e}_{i\kappa_i}]$ be a set of eigenvectors corresponding to the $\kappa_i < D$ largest eigenvalues, and let $\mathbf{p}_{uj}$ be a vector representation of the j^{th} normalized observation ECG heartbeat after pre-processing. We subtract the class mean $\tilde{\boldsymbol{\theta}}_i$, see (2), from $\mathbf{p}_{uj}$ to yield $\tilde{\mathbf{p}}_{ij}$. Thus, a measure of the reconstruction error in i^{th} activity's residual vector eigenspace, for the j^{th} ECG heartbeat observation, is defined as:

$$\mathrm{RE}_j^{\mathrm{PCA}}(i) = \left|\tilde{\mathbf{p}}_{ij} - (\mathbf{E}_i\mathbf{E}_i^T)\tilde{\mathbf{p}}_{ij}\right|^2, \tag{4}$$

which is summed over ν_F heartbeat observations. In the PCA approach studied in [19] and [23], the decision is assigned to the activity class label from $i = 1, \cdots, M$ for which the reconstruction error $\mathrm{RE}^{\mathrm{PCA}}(i)$ is the minimum. However, the activity class mean $\tilde{\boldsymbol{\theta}}_i$ is pre-trained and fixed in all the testing situations. This can induce session-to-session variability issues. Differences in sensor electrode placements and user emotion states can cause fluctuations of the mean vector between the training and testing data which affect the computation of the residual activity vector. Furthermore, this PCA method does not use the heart rate or other intra-beat statistical information, and focuses only on the normalized heartbeat modeling. In this work, we address this issue by adopting the PCA error vector $\mathbf{RE}^{\mathrm{PCA}} = [\mathrm{RE}^{\mathrm{PCA}}(1) \quad \mathrm{RE}^{\mathrm{PCA}}(2) \ \ldots \ \mathrm{RE}^{\mathrm{PCA}}(M)]$ as one of the temporal ECG features used for PA recognition.

3) Hermite polynomial expansion: A Hermite polynomial expansion (HPE) is used to model the CAM component $\boldsymbol{\theta}_i$ of the sampled ECG signal, and the resulting coefficients serve as another feature set for classification. Hermite polynomials are classical orthogonal polynomial sequence representations [30] and have been successfully used to describe ECG signals for arrhythmia detection [31] but do not appear to have been previously used for PA detection. In Fig. 3, the shape of the CAM component for each activity is different and thus these signals can be used to distinguish between different PA states. Rather than subtracting the activity mean to model the motion artifact noise, we average the normalized ECG signal to estimate the cardiac activity mean. Let ν_F denote the fixed number of normalized heartbeats in each running window; the CAM component of the κ^{th} window is estimated by

$$\acute{\boldsymbol{\theta}}_{i\kappa} = \frac{1}{\nu_F} \sum_{j=(\kappa)\nu_F+1}^{(\kappa+1)\nu_F} \boldsymbol{r}_{ij}. \tag{5}$$

Denote each estimated D-dimensional (D is an odd number) CAM component vector and polynomial order by $\acute{\theta}[n]$ and L, respectively. The HPE of $\acute{\theta}[n]$ can be expressed as [31]

$$\acute{\theta}[n] = \sum_{l=0}^{L-1} c_l \psi_l(n, \delta), \qquad n \in \left[-\frac{(D-1)}{2}, \frac{(D-1)}{2}\right], \tag{6}$$

where $\{c_l\}$, $l = 0, 1, \cdots, L-1$ are the HPE coefficients, and $\psi_l(n, \delta)$ are the Hermite basis functions defined as:

$$\psi_l(n, \delta) = \frac{1}{\sqrt{\delta 2^l l! \sqrt{\pi}}} e^{-n^2/2\delta^2} H_l(n/\delta). \tag{7}$$

The functions $H_l(n/\delta)$ are the Hermite polynomials [30]:

$$H_0(t) = 1, \quad H_1(t) = 2t, \tag{8}$$

$$H_l(t) = 2tH_{l-1}(t) - 2(l-1)H_{l-2}(t). \tag{9}$$

It had previously been shown in [31] that, for Hermite basis functions with different orders, the higher the order the higher is its frequency of changes within the time domain and thus resulting in a better capability for capturing morphological details of ECG signals. The HPE basis functions can be denoted by a $D \times L$ matrix $\mathbf{B} = [\boldsymbol{\psi}_0 \quad \boldsymbol{\psi}_1 \ \cdots \ \boldsymbol{\psi}_{L-1}]$; the expansion coefficients $\mathbf{c} = [c_0 \ c_1 \ \cdots \ c_{L-1}]^T$ are obtained by minimizing the sum squared error E:

$$E = \left\|\acute{\theta}[n] - \sum_{l=0}^{L-1} c_l \psi_l(n, \delta)\right\|_2^2 = \left\|\acute{\boldsymbol{\theta}} - \mathbf{B}\boldsymbol{c}\right\|_2^2 \\ \rightarrow \quad \mathbf{c} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\acute{\boldsymbol{\theta}}. \tag{10}$$

As shown in [31], HPE based reconstruction is nearly identical to the original waveform for ECG signals. The HPE coefficients $\mathbf{c}$ are also employed as ECG temporal features.

4) Standard deviation of multiple normalized beats: We had previously observed that the variance of the accelerometer measurements offered discrimination capability [14]; this feature for the ECG signal also has utility. From Fig. 3, we see that higher intensity states (walking) have a larger standard deviation than lower intensity ones (lying). If the user is lying down or sitting, then the normalized heartbeat shapes are more consistent or similar within the whole processing window, but if the user is walking or running, then the normalized ECG shape can vary dramatically and become noisy. Thus the sum of standard deviations for all the normalized bins (D bins) in the window is also employed as a feature. To our knowledge, this feature has not been previously used for PA classification.

Thus, for the temporal ECG features, not only are the PCA error vector and HPE coefficients included but also are the conventional mean and variance of instant heart rate and standard deviation of multiple normalized beats (noise measure). By using multiple measurements, this temporal ECG feature vector covers both conventionally used heart rate information and novel morphological shape information.

5) Temporal Accelerometer Features: For the tri-axial accelerometer, a set of conventional temporal features (in Table I) is extracted from the signals of each axis in every processing window. These features have been previously studied in [1]–[4], [9], employing various subsets of the features listed.

Both ECG and accelerometer temporal feature vectors are denoted as $\boldsymbol{y}$ and modeled by a support vector machine as explained in Section III-A.

B. Cepstral feature extraction

In this work, it is assumed that both ECG and accelerometer signals have quasi-periodic characteristics resulting from the convolution between an excitation (heart rate or moving pace) and a corresponding system response (ECG waveform shapes [20] or accelerometer moving patterns). Furthermore, in the acquisition of both ECG and accelerometer signals, there are many other "channel" artifacts, such as skin muscle activity, mental states variability, electrodes displacements, and so on. Cepstral analysis [32]–[34] is a homomorphic signal transform technique that transforms a convolution into an additive relationship which makes it especially conducive for mitigating convolutional effects. It has been successfully and widely used with processing many real life signals, such as speech and seismic signals [32]–[34]. Thus, in order to filter out the effects of the different paths from the source signals to the sensors, using cepstral features to model the frequency information of the native signal allows us to separate inherent convolutive effects by simple linear filtering. In the following, we explain the usage of a real cepstrum and describe the proposed linear frequency band based cepstral features in detail.

The sensor signal has some frequencies at which motion artifacts or sensor noise dominate. For example, the ECG baseline wanders and high frequency noises can result in drastic frame-to-frame phase changes. Furthermore, the properties of the "excitation" source of the sensor signal (*e.g.* ECG heart rate and accelerometer speed) also vary from frame to frame, which makes the phase not very meaningful. Because of this, the complex cepstrum is rarely adopted for real life signals such as speech [34]. Thus, in this activity recognition application, we use only the real cepstrum which is based on spectral magnitude information from the sensor signals. The real cepstrum of a signal $x[n]$ with spectral magnitude $|X(e^{jw})|$ is defined as [34]:

$$C[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} ln|X(e^{jw})|e^{jwn}dw. \tag{11}$$

In many applications, instead of operating directly on the signal spectrum, filter banks are employed to emphasize different portions of the spectrum separately. For example, in speech and audio processing, mel frequency cepstral coefficients are popular [33] and are derived based on nonlinear filter bank processing of the spectral energies to approximate the frequency analysis in the human ear.

Given the FFT of the input signal $x[n]$

$$X[k] = \sum_{n=0}^{\mathbb{N}-1} x[n]e^{-j2\pi nk/\mathbb{N}}, 0 \leq k < \mathbb{N}, \tag{12}$$

where $\mathbb{N}$ is the size of FFT, a filter bank with $\mathbb{M}$ filters ($m = 1, 2, \cdots, \mathbb{M}$) is adopted to map the powers of the spectrum obtained above into the mel scale using triangular overlapping windows $H_m[k]$ [34]. Thus, the log-energy at the output of each filter is computed as:

$$S[m] = ln\big[\sum_{k=0}^{\mathbb{N}-1}|X[k]|^2 H_m[k]\big], 0 \leq m < \mathbb{M}. \tag{13}$$

Finally, discrete cosine transform (DCT) of the $\mathbb{M}$ filter log-energy outputs is calculated to generate the cepstral features:

$$C[n] = \sum_{m=0}^{\mathbb{M}-1} S[m]cos(\pi n(m+1/2)/\mathbb{M}), 0 \leq n < \mathbb{M}. \tag{14}$$

The filter energies are more robust to noise and spectral estimation errors and thus have been extensively used as the golden feature set for speech and music recognition applications [34]. The perceptually motivated logarithmic mel-scale filter bands are designed for the human auditory system, which might not match the ECG and accelerometer signals. For this reason and for simplicity, in this work, we use linear frequency bands rather than the mel-scale frequency bands. Cepstral mean subtraction (CMS) and cepstral variance normalization (CVN) are adopted to mitigate convolutional filtering effects for ensuring robustness.

Specifically, due to potential inter-session variability, such as a change in electrode position or a variation in a user's emotion state, there is always a fluctuation on the "relative transfer function" as characterized by the transformation of the ground truth measurements of the PAs to the sensors' signals. Therefore, CMS is performed to mitigate this effect. The multiplication of the signal's spectrum, $X[n,k]$, and the relative transfer function's spectrum, $H[k]$, in the frequency domain is equivalent to a superposition in the cepstral domain:

$$C_y[n,k] = C_x[n,k] + C_H[k]. \tag{15}$$

And the second component $C_H[k]$ can be removed by applying long term averaging for each dimension k:

$$C_y[n,k] - \langle C_y[n,k]\rangle_{avg} = C_x[n,k] - \langle C_x[n,k]\rangle_{avg}. \tag{16}$$

Thus, cepstral features with CMS and CVN normalization are more robust against the session variability.

III. Activity Modeling

As shown in Fig. 2, the features in both temporal and cepstral domains are modeled using the SVM and GMM classifiers, respectively. The multimodal and multi-domain subsystems are fused together at the score level to improve the overall PA recognition performance.

A. SVM Classification for temporal features

An SVM is a binary classifier constructed from sums of a kernel function $K(\cdot,\cdot)$ over $\mathcal{N}$ support vectors, where $\boldsymbol{y}_i$ denotes the i^{th} support vector and t_i is the ideal output:

$$f(\boldsymbol{y}) = \sum_{i=1}^{\mathcal{N}} \alpha_i t_i K(\boldsymbol{y}, \boldsymbol{y}_i) + d. \tag{17}$$

The ideal outputs are either 1 or -1, depending upon whether the corresponding support vector belongs to class 1 or -1. By using kernel functions, an SVM can be generalized to non-linear classifiers by mapping the input features into a high dimensional feature space.

The original form of the generalized linear discriminative sequence (GLDS) kernel [24] involves a polynomial expansion, $b(\boldsymbol{y})$, with monomials (between each combination of vector components) up to a given degree p. The GLDS kernel between two sequences of vectors $\boldsymbol{Y}^1 = \{\boldsymbol{y}_t^1\}_{t=1 \cdots N_1}$ and $\boldsymbol{Y}^2 = \{\boldsymbol{y}_t^2\}_{t=1 \cdots N_2}$ is denoted as a rescaled dot product between average expansions:

$$\begin{aligned} K(\boldsymbol{Y}^1, \boldsymbol{Y}^2) &= \frac{1}{N_1} \sum_{i=1}^{N_1} b(\boldsymbol{y}_i^1)^t \cdot \boldsymbol{R}^{(-1)} \cdot \frac{1}{N_2} \sum_{j=1}^{N_2} b(\boldsymbol{y}_j^2) \\ &= (\boldsymbol{b_{y^1}}^t \boldsymbol{R}^{(-1/2)}) \cdot (\boldsymbol{R}^{(-1/2)} \boldsymbol{b_{y^2}}) \end{aligned} \tag{18}$$

where $\boldsymbol{R}$ is the second moment matrix of the polynomial expansions and its diagonal approximation is usually used for efficiency. In this work, only the first order of $b(\boldsymbol{y})$ is used for simplicity: $b(\boldsymbol{y}) = \boldsymbol{y}$. In addition, if we arbitrarily add one dummy dimension with value 1 at the head of each feature vector $\acute{b}(\boldsymbol{y}) = [1 \quad b(\boldsymbol{y})]$, $\boldsymbol{R}$ becomes $\acute{\boldsymbol{R}}$ and the scoring function of the GLDS kernel can be simplified by the following compact technique [24]:

$$f(\{\boldsymbol{Y}\}) = (\sum_{i=1}^{\mathcal{N}} \alpha_i t_i \acute{\boldsymbol{R}}^{-1} \acute{\boldsymbol{b}}_{\boldsymbol{y}^i} + \boldsymbol{d})^t \cdot \acute{\boldsymbol{b}}_{\boldsymbol{y}} = \boldsymbol{W}^t \cdot \acute{\boldsymbol{b}}_{\boldsymbol{y}}, \tag{19}$$

where $(\acute{\boldsymbol{b}}_{\boldsymbol{y}^i})^t$ are the support vectors and $\boldsymbol{d}$ is defined as $[d \quad 0 \cdots 0]^t$. Therefore, the scoring function of a target model on a sequence of observations can be calculated using the averaged observation. Furthermore, by collapsing all the support vectors down into a single model vector $\boldsymbol{W}$, each target score can be calculated by a simple inner product which makes this framework computationally efficient. In this study, the LIBSVM tool [35] and 1vsRest [24] strategy were used for the SVM model training. For each activity, a binary SVM classifier was trained against the rest $M-1$ activities using the GLDS kernel in (18). Moreover, for each binary SVM model, all the support vectors were collapsed into a single vector W by (19) to make the scoring function computationally efficient.

B. GMM modeling for cepstral features

A Gaussian Mixture model (GMM) is used to model the cepstral features of the ECG and accelerometer signals. A Gaussian mixture density is a weighted sum of N component densities and is given by

$$p(\boldsymbol{C}|\lambda) = \sum_{j=1}^{N} \omega_j p_j(\boldsymbol{C}) \tag{20}$$

where $\boldsymbol{C}$ is a $\mathbb{D}$-dimensional random vector, $p_j(\boldsymbol{C}), j = 1, \cdots, N$ are the component densities and $\omega_j, j = 1, \cdots, N$ are the mixture weights. Each component density is a $\mathbb{D}$-variate Gaussian function of the following form:

$$p_j(\boldsymbol{C}) = \frac{1}{(2\pi)^{\mathbb{D}/2} |\boldsymbol{\Sigma}_j|^{1/2}} exp\{-\frac{1}{2}(\boldsymbol{C} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{C} - \boldsymbol{\mu}_j)\} \tag{21}$$

with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. The mixture weights satisfy the constraint that $\sum_{j=1}^{N} \omega_j = 1$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices, and mixture weights from all component densities. These parameters are collectively represented by the notation λ_i for activity $i, i = 1, \cdots, M$, and are explicitly written as

$$\lambda_i = \{p_j, \vec{\mu}_j, \Sigma_j\}, \qquad j = 1, \cdots, N. \tag{22}$$

For subject-dependent PA identification using the cepstral features of sensor signals, each activity performed by every subject is represented by a GMM and is referred to by its model λ_i. In the proposed work, since the training data for each activity of each subject is too limited to train a good GMM, a Universal Background Model (UBM) in conjunction with a Maximum A Posteriori (MAP) model adaptation approach [33] is used to model different PAs in a supervised manner. The UBM model is trained using all the training data including all the activities and all the subjects; then the subject-dependent activity model is derived using MAP adaptation from the UBM model with subject specific activity training data. The expectation maximization (EM) algorithm is adopted for the UBM training.

Under the framework of GMM, during testing, each signal segment with T frames is scored on all the activities' models from the same subject. By using logarithms and the independence between observations, the GMM system outputs the recognized activity by maximizing log likelihood criterion:

$$\hat{S} = \arg \max_{1 \le i \le M} \sum_{t=1}^{T} log\{p(\boldsymbol{C}_t|\lambda_i)\}. \tag{23}$$

IV. System Fusion

In a multimodal activity recognition system, fusion can be accomplished by utilizing the complementary information available in each of the modalities. In the proposed work, both feature level fusion and score level fusion are studied.

A. Feature level fusion

Feature level fusion requires the feature sets of multiple modalities to be compatible [26]. Let $\mathbf{q} = \{q_1, q_2, \cdots, q_m\}$ and $\mathbf{s} = \{s_1, s_2, \cdots, s_n\}$ denote two feature vectors ($\boldsymbol{q} \in R^m$ and $\boldsymbol{s} \in R^n$) representing the information extracted from two different modalities. The goal of the feature level fusion is to fuse these two feature sets in order to yield a new feature vector $\boldsymbol{z}$ with better capability to represent the PA. The l-dimensional vector $\boldsymbol{z}$, $l \le (m+n)$, can be generated by first augmenting vectors $\boldsymbol{q}$ and $\boldsymbol{s}$ and then performing feature selection or feature transformation on the resultant feature vector in order to reduce the feature dimensionality.

In the proposed work, we only studied the feature level fusion with different axis features of accelerometer in the cepstral domain. It is because the cepstral feature may not be compatible with temporal features and the window length for the temporal feature calculation is significantly larger than for the cepstral features. Furthermore, ECG and accelerometer cepstral features are not concatenated and fused at the feature level due to the compatibility issues arising from different time shift and window length configurations and different sampling frequencies. However, the cepstral features from each axis of the accelerometer are concatenated to construct a long cepstral feature vector in each frame. Heteroscedastic linear discriminant analysis (HLDA) [27] is used to perform feature dimension reduction.

B. Score level fusion

Multimodal information can also be fused at the score level rather than the feature level. The match score is a measure of similarity between the input sensor signals and the hypothesized activity. When these match scores generated by subsystems based on different modalities are consolidated in order to generate a final recognition decision, fusion is done at the score level. Since some multimodal feature sets are not compatible and it is relatively easy to access and combine the scores generated by different subsystems, information fusion at the score level is the most commonly used approach in multimodal recognition systems [26].

Let there be $\mathbb{K}$ input PA recognition subsystems (as shown in Fig. 2, $\mathbb{K} = 4$ in this work), each acting on a specific sensor modality and feature set, where the k^{th} subsystem outputs its own normalized log-likelihood vector $l_k(\boldsymbol{x}_t)$ for every trial. Then the fused log-likelihood vector is given by:

$$\acute{l}(\boldsymbol{x}_t) = \sum_{k=1}^{\mathbb{K}} \beta_k l_k(\boldsymbol{x}_t) \tag{24}$$

The weight, β_k, is determined by logistic regression based on the training data [26].

V. Experimental setup and results

A. Data acquisition and evaluation

Data collection was conducted using an ALIVE heart rate monitor [21] and a Nokia N95 cell phone. The single lead ECG signal is collected by the heart rate monitor with electrodes on the chest, and at the same time the heart rate monitor with built in accelerometer is placed on the left hip to record the accelerometer signal. The placement of electrodes and accelerometer is shown in Fig. 4. Both signals are synchronized and packaged together to transmit to the cell phone through a Bluetooth wireless connection [14], [15], [21]. The sampling frequencies of the ECG and the accelerometer are 300 Hz and 75 Hz, respectively. In this work, only one tri-axial (heart rate monitor built-in) accelerometer signal and one single lead ECG signal are used for analysis. For each session, the subject was required to wear the sensors and perform 9 categories of PA following a predetermined protocol [17], [36] of lying, sitting, sitting fidgeting, standing, standing fidgeting, playing Nintendo Wii tennis, slow walking, brisk walking, and running. The last 3 activities were performed on a treadmill with subjects' own choices of speed (around 1.5 mph for slow walking and around 3 mph for brisk walking). The activities selected here are based on a version of the System for Observing Fitness Instruction Time (SOFIT), considered a gold standard for physical activity measurement [36]. These basic activities are believed to make up or represent a majority of real life physical activities. Furthermore, since measurements are based on a laboratory protocol, the modeling and recognition of these categories can be considered as a foundational baseline. Subjects wore the sensors for 7 minutes in each of the 9 PAs with inter-activity rest as needed. Data from 5 subjects (2 male, 3 female, ages ranging from 13 to 30) who participated in the experiment are reported in this paper. Each subject performed 4 sessions on different days and at different times. Thus the data reflect variability of electrodes positions and a variety of environmental and physiological factors. In the following, the proposed approach is evaluated in both closed set and open set classification tasks.

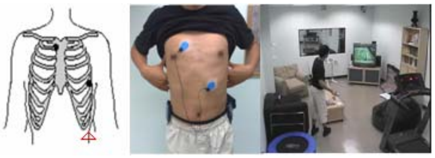


Fig. 4. Placement of electrodes (Black filled circles) and accelerometer (Red open triangle) and data collection environment.

First, the proposed PA recognition is formulated as a subject-dependent closed set activity identification problem, so the performance is measured by classification accuracy. For each subject, there are data from 4 sessions. Thus we established 3 different settings to evaluate our methods: **Setting 1**: For each subject and session, training was based on data from the first half and testing from the second half. **Setting 2**: For each subject, training was on one session's data and testing was on another session. **Setting 3**: For each subject, training was on 3 sessions' data and testing was on the remaining session. In the following, evaluations of our feature extraction and supervised modeling as shown in Table II,III, and V are performed by using setting 3 in which training and testing data are from different days/times and training/testing data are rotated 4 times (for cross validation). The performance reported is based on the average of all the subjects and all the rotation tests. In addition, score level fusion and session variability regarding all 3 settings are studied and demonstrated in Table IV.

Second, in real life free living conditions, there might be situations that do not quite fit in our 9-category PA protocol. Thus, 3 different open set task experiments were conducted to evaluate the generalizability of the results to everyday, ambulatory monitoring by testing the ability to correctly reject activities that do not fall within the set categories. All 3 open set tasks are based on subject dependent modeling of the previously described Setting 3. First, task 1 is formulated as an activity verification task (*e.g.* walking or not) by testing each in-set hypothesis activity's likelihood against a global threshold. **Task 1**: For each time, 8 activities are considered as in-set target activities while the remaining one activity is assigned as out-of-set activity for rejection purpose. This out of set activity is excluded from any training process. The setup was rotated 9 times to calculate the average performance. Equal Error Rate (EER) is used to evaluate the performance. Second, rather than identifying/rejecting activities based on thresholds, **Tasks 2**

TABLE II
PERFORMANCE (% CORRECT) OF SVM SYSTEM BASED ON TEMPORAL ECG FEATURES (HR:HEART RATE, NM:NOISE MEASUREMENT)

ECG	1 PCA	2 HPE	3 HR+NM	2+3	1+2+3
10 beats	46.9	51.7	43.3	56.7	60.8
20 seconds	49.4	54.4	44.0	60.0	64.2

and 3 employed closed set classification with the usage of an "others" activity model to classify all other activities that do not belong in the desired closed set. As shown in Table V, task 2 is focused on distinguishing sedentary activities while rejecting unknown vigorous activities by using the "others" model trained using data from the "standing fidgeting" activity, and vice versa for Task 3.

The testing duration of all the evaluation experiments is fixed at 20 seconds. HPE order, PCA eigenvector dimension, and the normalized heartbeat sample length D were empirically chosen to be 60, 40 and 201, respectively.

B. Results

1) SVM system based on temporal features: Table II shows the results of the ECG temporal feature based SVM system. Compared to the conventional PCA method [23], the proposed HPE coefficients together with heart rate (HR) and noise measurement (NM) features achieved nearly 10% improvement in accuracy. Furthermore, fusing PCA, HPE, HR, and NM features together achieves an additional 4% improvement.

2) GMM system based on cepstral features: In Table III, the results of the GMM system based on different configurations of cepstral features are shown. Before feature extraction, the DC baseline is removed by a high pass filter. ECG IDs $(1, 2, 8)$ show that smaller shifts and window sizes have better performances while ECG IDs $(2, 3, 4)$ show that the number of cepstral coefficients used for recognition does not have to be the number of spectral bands because DCT calculation in cepstral feature extraction can be seen as a hidden dimension reduction method. Moreover, ECG IDs $(3, 5, 6)$ demonstrate that 50% overlap and first order delta in cepstral extraction is necessary. Finally, ECG IDs $(7, 8, 9)$ illustrate the performance against different numbers of Gaussian components. In this case, GMM with 64 components together with a 120 milliseconds window, 24 cepstral coefficients, 48 frequency bands, 50% overlap, and first order delta derivatives give us the best performance of 63.45%.

Evaluation of the accelerometer (ACC) cepstral features in Table III yields similar results: smaller window sizes yield higher accuracy. Since the sampling frequency of the accelerometer is only 75 Hz, we set the minimum window length to be 480 milliseconds which is exactly $1/4^{th}$ of the ECG feature window size. However, in ACC IDs (1,6), the best setup for the number of cepstral coefficients is 20 rather than 7. So the final feature dimension is 120 because of the addition of a first order delta and tri-axial feature vector combination. ACC IDs $(1, 7, 8, 9)$ show the results of the HLDA dimension reduction method in the accelerometer cepstral domain. Results show that the system is not sensitive to the final reduced dimension, and the accuracy is improved from 74.76% to 77.56% when the dimension is reduced to 72.

3) Score level fusion: Performance of the score level fusion at different settings is shown in Table IV. In setting 3, firstly, fusion of ECG temporal and cepstral systems improves the accuracy from 64.17% to 68.49% while fusion of accelerometer temporal and cepstral systems achieves accuracy improvement from 84.85% to 90.00%. Secondly, using the same kind of features, fusing both ECG and accelerometer information together can also improve the results. We can see that, in the temporal domain, fusion of the ECG SVM system and the accelerometer SVM system increases the accuracy only by 1% while, in the cepstral domain, fusion of both modalities improves the accuracy from 77.56% to 82.30%. Finally, we fuse all 4 individual systems together to further improve the PA recognition performance which results in 91.40% accuracy for setting 3. It is shown that our fusion method has 6.55% absolute improvement (from 84.85% to 91.40%) compared to the conventional accelerometer temporal-features based SVM system. Similar results are also shown in settings 1 and 2.

4) Session variability study: In Table IV, the performances in setting 2 are noticeably lower than in setting 1 because of the mismatch between training and testing data due to the session variability. The ECG systems can drop their performance by up to 30% while the accelerometer systems are relatively more robust with only a 15% decrease. This might be because the ECG signal varies due to a range of factors, such as electrode placement, mental stress, emotion, and so on, while the accelerometer only measures the physical movement and thus only varies by different movement types or patterns. However, by adding more training data from different sessions, this variability can be mitigated and the system can be made more robust. This is demonstrated by observing the 10%-21% improvement from setting 3 to setting 2. The accuracy standard deviations of different subjects are also shown in Table IV. The individual standard deviation is also improved along with the average accuracy in score level fusion. Furthermore, in terms of accuracy for fusion system (ID 9), the p-values [37] of null hypothesis that setting 1 is equal to setting 2 and setting 3 is equal to setting 2 are 0.00003 and 0.0009, respectively. Thus, with the influence of individual variability, session variability is verified with 0.0009 significance level.

5) Open set tasks study: Table V clearly shows that, in the open set tasks, score level fusion of the multi-modal and multi-domain subsystems significantly improves performance. Based on the similar accuracy results between closed set classification and open set tasks 2 and 3, it can be observed that with the usage of the "others" activity model, the proposed approach can effectively identify the activities of interest as well as reject out of set activities.

VI. DISCUSSION

This work addresses the PA recognition problem with multimodal wearable sensors (ECG and accelerometer). The contributions are as follows:

(1) The cardiac activity mean (CAM) component of the ECG signal is described by HPE in the temporal feature

TABLE III
EVALUATION OF GMM SYSTEMS BASED ON DIFFERENT CONFIGURATIONS OF CEPSTRAL FEATURE EXTRACTION. (ACC=ACCELEROMETER)

ECG ID	number of cepstra	number of spectral bands	window length in second	window shift in second	first order delta	HLDA	cepstral feature dimension	GMM Gaussian components	accuracy $P_c(\%)$
1	36	64	0.5	0.25	yes	no	64	32	55.49
2	36	64	0.25	0.125	yes	no	64	32	61.83
3	24	64	0.25	0.125	yes	no	48	32	61.45
4	64	64	0.25	0.125	yes	no	128	32	57.64
5	24	64	0.25	0.25	yes	no	48	32	59.23
6	24	64	0.25	0.125	no	no	24	32	59.17
7	24	48	0.12	0.06	yes	no	48	16	61.26
8	24	48	0.12	0.06	yes	no	48	32	63.20
9	**24**	**48**	**0.12**	**0.06**	**yes**	**no**	**48**	**64**	**63.45**

ACC ID	number of cepstra	number of spectral bands	window length in second	window shift in second	first order delta	HLDA	cepstral feature dimension	GMM Gaussian components	accuracy $P_c(\%)$
1	**20**	**20**	**0.48**	**0.24**	**yes**	**no**	**120**	**32**	**74.76**
2	20	20	0.96	0.48	yes	no	120	32	67.48
3	20	20	0.24	0.24	yes	no	120	32	70.73
4	20	20	0.48	0.24	yes	no	120	16	75.01
5	20	20	0.48	0.24	yes	no	120	64	72.45
6	7	20	0.48	0.24	yes	no	42	32	72.00
7	20	20	0.48	0.24	yes	yes	96	32	77.52
8	20	20	0.48	0.24	yes	yes	84	32	77.34
9	**20**	**20**	**0.48**	**0.24**	**yes**	**yes**	**72**	**32**	**77.56**

extraction. It can be observed in Table II that HPE features perform better than conventional PCA features and adding PCA, HPE, HR, and NM features together achieves significant improvement. This is because the pre-trained activity mean in the PCA approach might be different from the testing condition due to session variability which can decrease system performance. Moreover, PCA and HPE model the MAN and CAM part of the normalized ECG waveform, respectively, while HR and NM measure the heart rate and inter-beats noise level, and this information is complementary.

(2) In the SVM framework for both ECG and accelerometer temporal features, the GLDS kernel makes the classification computationally efficient with a small model size. We can see that the single lead ECG signal has more activity discrimination information than provided by just the heart rate, but as shown in Table IV, the performance is still relatively low compared with accelerometer based methods. Therefore, fusing the information from both modalities is necessary.

(3) A GMM system based on cepstral features is proposed to capture the frequency domain information, and HLDA is used to reduce the feature dimension of tri-axial accelerometer based measurements. In Table IV, compared to the ECG temporal feature based SVM system, the GMM approach with ECG cepstral features achieved almost the same performance in setting 3 and, in fact, 10% better in setting 2 because cepstral features together with CMS are more robust to session variability. Furthermore, because there is no need for preprocessing steps, such as peak detection and segmentation which are inherently noisy and computationally expansive, cepstral feature calculation is faster and more efficient than temporal feature extraction. Compared to the result of the accelerometer temporal feature based SVM system (84.85%), this GMM-cepstral approach achieved a lower performance (77.56%). This is due to the characteristics of the cepstral feature and CMS normalization, in which the mean of the accelerometer signal is removed. The mean of the tri-axial accelerometer signal corresponds to the gravity along different directions; thus different static positions of activity might have different mean values because of the sensor rotation. By analysis of this mean value, the performance of the accelerometer temporal SVM system is enhanced. However, comparing the results from both setting 1 and setting 2 in Table IV, it is clear that the cepstral features based system is less sensitive to session variability than the temporal features based system.

TABLE IV
SCORE LEVEL FUSION: THE MEAN ± STANDARD DEVIATION OF ACCURACIES $P_c(\%)$ FOR DIFFERENT SUBJECTS

System ID and name		Setting 1	Setting 2	Setting 3
1	ECG-Temporal-SVM	88.05±5.0	43.39±8.9	64.17±6.3
2	ACC-Temporal-SVM	95.13±4.1	72.76±7.9	84.85±7.8
3	ECG-Cepstral-GMM	85.43±9.1	53.81±15.1	63.45±9.8
4	ACC-Cepstral-GMM	78.93±7.4	63.69±5.7	77.56±5.4
5	Fusion (1+3)	92.52±3.0	54.05±13.9	68.49±7.4
6	Fusion (2+4)	96.17±3.9	79.04±5.4	90.00±4.1
7	Fusion (1+2)	97.02±2.8	71.81±7.6	85.49±5.8
8	Fusion (3+4)	90.78±8.8	66.12±5.8	82.30±6.2
9	**Fusion (1+2+3+4)**	**97.29±2.4**	**79.30±4.8**	**91.40±3.4**

TABLE V
THE CONFIGURATION AND PERFORMANCE OF OPENSET TASKS.

Experiment Setup			Performance	EER(%)	Accuracy(%)	
Activities	T2	T3	System ID	T1	T2	T3
Lying	▲	□	1: ECG-Tem	14.7	66.8	68.8
Sitting	▲	□	2: ACC-Tem	6.6	84.6	80.6
Sit Fidgeting	▲	□	3: ECG-Cep	23.2	70.0	49.1
Standing	▲	□	4: ACC-Cep	12.7	64.3	76.4
Stand Fidgeting	△	□	5: Fuse1,3	14.5	72.5	68.8
Playing Wii	□	△	6: Fuse2,4	5.3	88.2	83.4
Slow Walking	□	▲	7: Fuse1,2	6.5	87.0	83.5
Brisk Walking	□	▲	8: Fuse3,4	9.8	75.8	82.3
Running	□	▲	**9: Fuse1,2,3,4**	**5.0**	**91.4**	**86.5**

▲ is in set target activity, △ is "others" model activity, □ is out of set activity. T1,T2 and T3 denote Task 1,2 and 3, respectively.

(4) Score level fusion of the multi-modal and multi-domain subsystems is performed to improve the overall recognition performance. We demonstrated in Section V-B3 that fusing both temporal and cepstral information in each single modality can improve the overall system performance. This

result substantiates our assumption that temporal information and cepstral information are complementary. Additionally, fusing both ECG and accelerometer information together can also increase the accuracy. Therefore, fusing both modalities is also useful. Compared to the conventional accelerometer temporal feature based approach (System ID 2), the proposed multimodal temporal and cepstral information fusion method (System ID 9) achieved 44%, 24%, and 43% relative error reduction for setting 1,2, and 3, respectively.

(5) The effects of session variability of ECG and accelerometer measurements on PA recognition were studied. Session variability compensation in the PA recognition application might become an important and challenging research question where many algorithms need to be designed and applied to increase the system robustness. For example, the nuisance attribute projection (NAP) [38] method in the SVM modeling has already been successfully and widely used in speaker recognition to reduce the influence of different channels. In this study with hypotheses testing, we just showed that results in setting 1 (within session recognition) can not reflect the performance in real PA recognition applications such as in the across session condition of setting 2. But adding more training data from multiple sessions can mitigate this variability and improve the real system performance. This also underscores the need for dynamic adaptation to changing data conditions.

VII. Conclusion

In this work, a multimodal physical activity recognition system was developed by fusing both ECG and accelerometer information together. Each modality is modeled in both temporal and cepstral domains. The main novelty is that by fusing both modalities together, and fusing both temporal and cepstral domain information within each modality, the overall system performance is shown to improve significantly in both accuracy and robustness. We also show that the ECG signals are more sensitive to session-to-session variability than the accelerometer signals, and by adding more multi-session training data, the session variability can be mitigated and the system can become more robust in real life usage conditions. Future work includes validating the results with data collected under free living conditions.

References

[1] U. Maurer, A. Rowe, A. Smailagic, and D. Siewiorek, "Location and Activity Recognition Using eWatch: A Wearable Sensor Platform," *Lecture Notes in Computer Science*, vol. 3864, p. 86, 2006.

[2] L. Bao and S. Intille, "Activity recognition from user-annotated acceleration data," *Lecture Notes in Computer Science*, vol. 3001, pp. 1–17, 2004.

[3] A. Godfrey, R. Conway, D. Meagher, and G. ÓLaighin, "Direct measurement of human movement by accelerometry," *Medical Engineering and Physics*, vol. 30, no. 10, pp. 1364–1386, 2008.

[4] D. Huynh, "Human Activity Recognition with Wearable Sensors," *Ph.D. Thesis*, 2008.

[5] J. He, H. Li, and J. Tan, "Real-time daily activity classification with wireless sensor networks using Hidden Markov Model," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2007, pp. 3192–3195.

[6] N. Ravi, N. Dandekar, P. Mysore, and M. Littman, "Activity recognition from accelerometer data," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 20, no. 3, 2005, p. 1541.

[7] A. Krause, D. Siewiorek, A. Smailagic, and J. Farringdon, "Unsupervised, dynamic identification of physiological and activity context in wearable computing," in *IEEE International Symposium on Wearable Computers*, 2005, pp. 88–97.

[8] T. Huynh, U. Blanke, and B. Schiele, "Scalable recognition of daily activities with wearable sensors," *Lecture Notes in Computer Science*, vol. 4718, p. 50, 2007.

[9] L. Jatoba, U. Grossmann, C. Kunze, J. Ottenbacher, and W. Stork, "Context-aware mobile health monitoring: Evaluation of different pattern recognition methods for classification of physical activity," in *30th Annual International Conference of the IEEEE engineering in Medicine and Biology Society, EMBS.*, 2008, pp. 5250–5253.

[10] J. Lester, T. Choudhury, and G. Borriello, "A practical approach to recognizing physical activities," *Lecture Notes in Computer Science*, vol. 3968, pp. 1–16, 2006.

[11] P. Lukowicz, H. Junker, M. Stager, T. Von Buren, and G. Troster, "WearNET: A distributed multi-sensor system for context aware wearables," *Lecture notes in computer science*, pp. 361–370, 2002.

[12] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, I. Korhonen, V. Technol, and F. Tampere, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[13] M. Ermes, J. Parkka, J. Mantyjarvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 1, pp. 20–26, 2008.

[14] M. Annavaram, N. Medvidovic, U. Mitra, S. Narayanan, G. Sukhatme, Z. Meng, S. Qiu, R. Kumar, G. Thatte, and D. Spruijt-Metz, "Multimodal sensing for pediatric obesity applications," *International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems, UrbanSense*, 2008.

[15] S. Lee, M. Annavaram, G. Thatte, R. V., M. Li, U. Mitra, S. Narayanan, and D. Spruijt-Metz, "Sensing for Obesity: KNOWME Implementation and Lessons for an Architect," in *Workshop on Biomedicine in Computing: Systems, Architectures, and Circuits*, 2009.

[16] G. Thatte, M. Li, A. Emken, U. Mitra, S. Narayanan, M. Annavaram, and D. Spruijt-Metz, "Energy-Efficient Multihypothesis Activity-Detection for Health-Monitoring Applications," in *International Conference of the IEEE engineering in Medicine and Biology Society, EMBS*, 2009.

[17] G. Thatte, V. Rozgic, M. Li, S. Ghosh, U. Mitra, S. Narayanan, M. Annavaram, and D. Spruijt-Metz, "Optimal Allocation of Time-Resources for Multihypothesis Activity-Level Detection," in *IEEE International Conference on Distributed Computing in Sensor Systems, DCOSS*, 2009.

[18] "Google Health Service," *http://www.google.com/health/*.

[19] T. Pawar, N. Anantakrishnan, S. Chaudhuri, and S. Duttagupta, "Impact of ambulation in wearable-ECG," *Annals of Biomedical Engineering*, vol. 36, no. 9, pp. 1547–1557, 2008.

[20] G. Clifford, F. Azuaje, and P. McSharry, *Advanced methods and tools for ECG data analysis*. Artech House, 2006.

[21] "Alive Heart Monitor," *http://www.alivetec.com/products.htm*.

[22] E. Tapia, S. Intille, W. Haskell, K. Larson, J. Wright, A. King, and R. Friedman, "Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart monitor," in *IEEE International Symposium on Wearable Computers, ISWC*, 2007.

[23] T. Pawar, S. Chaudhuri, and S. Duttagupta, "Body movement activity recognition for ambulatory cardiac monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 874–882, 2007.

[24] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, pp. 210–229, 2006.

[25] K. Phua, J. Chen, T. Dat, and L. Shue, "Heart sound as a biometric," *Pattern Recognition*, vol. 41, no. 3, pp. 906–919, 2008.

[26] A. Ross, K. Nandakumar, and A. Jain, *Handbook of multibiometrics*. Springer, 2006.

[27] N. Kumar and A. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech communication*, vol. 26, no. 4, pp. 283–298, 1998.

[28] S. Israel, J. Irvine, A. Cheng, M. Wiederhold, and B. Wiederhold, "ECG to identify individuals," *Pattern Recognition*, vol. 38, pp. 133–142, 2005.

[29] D. Gafurov, K. Helkala, and T. Søndrol, "Biometric gait authentication using accelerometer sensor," *Journal of computers*, vol. 1, 2006.

[30] G. Arfken, H. Weber, and H. Weber, *Mathematical methods for physicists*. Academic press New York, 1985.

[31] T. Linh, S. Osowski, and M. Stodolski, "On-line heart beat recognition using Hermite polynomials and neuro-fuzzy network," *IEEE Transactions on Instrumentation and Measurement*, vol. 52, no. 4, pp. 1224–1231, 2003.

[32] D. Childers, D. Skinner, and R. Kemerait, "The cepstrum: a guide to processing," *Proceedings of the IEEE*, vol. 65, pp. 1428–1443, 1977.

[33] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[34] X. Huang, A. Acero, and H. Hon, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR Upper Saddle River, NJ, USA, 2001.

[35] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

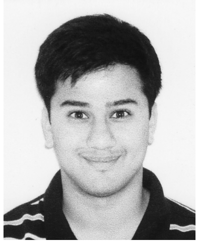
[36] T. McKenzie, J. Sallis, and P. Nader, "SOFIT: System for observing fitness instruction time," *J Teach Phys Educ*, vol. 11, pp. 195–205, 1991.

[37] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences (5th Edition)*. Prentice Hall, 2006.

[38] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, vol. 1, 2006.

Ming Li received the B.S. degree in communication engineering from Nanjing University, China, in 2005 and the M.S. degree in signal processing from Institute of Acoustics, Chinese Academy of Sciences, in 2008. Currently, he is working toward the Ph.D. degree in electrical engineering at USC. His research interests are in the areas of multimodal signal processing, audio-visual joint biometrics, speaker verification, language identification, audio watermarking and speech separation.

Viktor Rozgic received Dipl.Eng. degree in electrical engineering from the University of Belgrade, Serbia, in 2001, and the M.S. degree from the University of Southern California, Los Angeles, in 2007. He is currently a Ph.D. candidate in the Signal Anlayis and Interpretation Laboratory at the University of Southern California, Los Angeles. His research interests include audio-visual signal processing, multi-modal fusion, multimedia content analysis, sequential and Markov chain Monte Carlo filtering methods, and multi-target tracking algorithms.



Gautam Thatte received the B.S. degree (distinction) in engineering from Harvey Mudd College (HMC), Claremont, CA, in 2003 and the M.S. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2004. Currently, he is working toward the Ph.D. degree in electrical engineering at USC. His current research interests are in the areas of estimation and detection in sensor networks and computer networks.

Sangwon Lee is a PhD student in the department of computer science at the University of Southern California (USC), Los Angeles. In 2008, he was working in LG Electronics as a senior research engineer. He received his MS degree in computer science from USC in 2008 and his B.A degree in Computer Science from Seoul National University of Technology, South Korea, in 2000. Before his studies at USC, he worked as a system architecture and a DBA for 6 years. He established his own company, Interrush Korea Inc. in 2002. His general interest is in mobile applications and wireless sensor networks.

B. Adar Emken received the B.S. degree in psychobiology (magna cum laude, with honors and with distinction) from Ohio State University in 2001 and the Ph.D. degree in neuroscience from the University of California, Irvine, in 2008. In 2004, she received a NSF Graduate Research Fellowship. She is currently a post-doctoral researcher at the University of Southern California and her research interests include objective measurement of physical activity and the effects of physical activity on cognitive function.

Dr. Donna Spruijt-Metz received her PhD in Adolescent Medicine from the Vrije Universitiet Amsterdam in 1996. She is Associate Professor at the Keck School of Medicine's Department of Preventive Medicine. Her research focuses on pediatric obesity. Current studies include a longitudinal study of the impact of puberty on insulin dynamics, mood and physical activity in African American and Latina girls (funded by NCI) , a study examining the impact of simple carbohydrate versus complex carbohydrate meals on behavior, insulin dynamics, select gut peptides, and psychosocial measures in overweight minority youth (funded by NCHMD), and the KNOWME Networks project, studying WBANs developed specifically for minority youth for non-intrusive monitoring of metabolic health, vital signs such as heart rate, and physical activity and other obesity-related behaviors (funded by NCHMD).

Urbashi Mitra received the B.S. and the M.S. degrees from the University of California at Berkeley in 1987 and 1989 respectively, both in Electrical Engineering and Computer Science. From 1989 until 1990 she worked as a Member of Technical Staff at Bellcore in Red Bank, NJ. In 1994, she received her Ph.D. from Princeton University in Electrical Engineering. From 1994 to 2000, Dr. Mitra was a member of the faculty of the Department of Electrical Engineering at The Ohio State University, Columbus, Ohio. In 2001, she joined the Department of Electrical Engineering at the University of Southern California, Los Angeles, where she is currently a Professor. Dr. Mitra is currently an Associate Editor for the IEEE Transactions on Information Theory and the Journal of Oceanic Engineering. She was an Associate Editor for the IEEE Transactions on Communications from 1996 to 2001. Dr. Mitra served two terms as a member of the IEEE Information Theory Society's Board of Governors (2002-2007). She is the recipient of: Best Applications Paper Award C 2009 International Conference on Distributed Computing in Sensor Systems, IEEE Fellow (2007), Texas Instruments Visiting Professor (Fall 2002, Rice University), 2001 Okawa Foundation Award, 2000 Lumley Award for Research (OSU College of Engineering), 1997 MacQuigg Award for Teaching (OSU College of Engineering), and a 1996 National Science Foundation (NSF) CAREER Award. She has co-chaired the IEEE Communication Theory Symposium at ICC 2003 in Anchorage, AK and the ACM Workshop on Underwater Networks at Mobicom 2006, Los Angeles, CA. Dr. Mitra has held visiting appointments at: the Technical University of Delft, Stanford University, Rice Unviersity, and the Eurecom Institute. She served as co-Director of the Communication Sciences Institute at the University of Southern California from 2004-2007.

Murali Annavaram 's research focuses on energy efficiency and reliability of computing platforms. On the mobile platform end, his research focuses on energy efficient sensor management for body area sensor networks for continuous and real time health monitoring. He also has an active research group focused on computer systems architecture exploring reliability challenges in the future CMOS technologies. Prior to his teaching career, Annavaram worked in industrial research labs for 6 years; first at the Intel Microprocessor Research Labs as a senior researcher and then at the Nokia Research Center as a visiting research faculty.

Shrikanth (Shri) Narayanan is the Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics and Psychology. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC he directs the Signal Analysis and Interpretation Laboratory. His research focuses on human-centered information processing and communication technologies. Shri Narayanan is a Fellow of IEEE, the Acoustical Society of America, and the American Association for the Advancement of Science (AAAS). Shri Narayanan is also an Editor for the Computer Speech and Language Journal and an Associate Editor for the IEEE Transactions on Multimedia, IEEE Transactions on Affective Computing and the Journal of the Acoustical Society of America. He was also previously an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000-04) and the IEEE Signal Processing Magazine (2005-2008). He is a recipient of a number of honors including Best Paper awards from the IEEE Signal Processing society in 2005 (with Alex Potamianos) and in 2009 (with Chul Min Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010-11.