Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 280140, 12 pages
http://dx.doi.org/10.1155/2015/280140

Research Article

Empirical Validation of Objective Functions in Feature Selection Based on Acceleration Motion Segmentation Data

Jong Gwan Lim,[1] Mi-hye Kim,[2] and Sahngwoon Lee[3]

[1]*KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-338, Republic of Korea*
[2]*Chungbuk National University, 1 Chungdae-ro, Seowon-gu, Cheongju, Chungbuk 362-763, Republic of Korea*
[3]*Systran International, 163 Yangjaecheon-ro, Gangnam-gu, Seoul 135-855, Republic of Korea*

Correspondence should be addressed to Mi-hye Kim; mhkim@chungbuk.ac.kr

Received 5 March 2015; Accepted 14 April 2015

Academic Editor: Sanghyuk Lee

Recent change in evaluation criteria from accuracy alone to trade-off with time delay has inspired multivariate energy-based approaches in motion segmentation using acceleration. The essence of multivariate approaches lies in the construction of highly dimensional energy and requires feature subset selection in machine learning. Due to fast process, filter methods are preferred; however, their poorer estimate is of the main concerns. This paper aims at empirical validation of three objective functions for filter approaches, Fisher discriminant ratio, multiple correlation (MC), and mutual information (MI), through two subsequent experiments. With respect to 63 possible subsets out of 6 variables for acceleration motion segmentation, three functions in addition to a theoretical measure are compared with two wrappers, k-nearest neighbor and Bayes classifiers in general statistics and strongly relevant variable identification by social network analysis. Then four kinds of new proposed multivariate energy are compared with a conventional univariate approach in terms of accuracy and time delay. Finally it appears that MC and MI are acceptable enough to match the estimate of two wrappers, and multivariate approaches are justified with our analytic procedures.

1. Introduction

As one of the human computer interactions, Inertial Measurement Unit (IMU) applications have been prominently increasing in quantity [1]. Of the related technological issues, motion segmentation using accelerometers has long been a significant problem [2–6]. Motion segmentation implies the discrimination of motion-involved periods and is handled within various domains depending on the detection signal. In the IMU applications, which generally depend on accelerometers, the process can be understood as acceleration end point detection in terms of signal processing. Since linear acceleration and angular rates from IMUs are rarely used without integration, motion segmentation is inevitable because it indicates the initial and final points in the integration or the starting and ending points in the period of interest for processing [4, 7, 8].

Typical problems in motion segmentation using acceleration have been associated with how accurately both ends can be found; thereby several constraints have been reported. First, measured acceleration is corrupted with the gravitational acceleration which is intractable to separate from acceleration by body motion [2, 8, 9]. Since it is exposed to noise whose source is also body motion, such as unintentional trembles or minute motion, the estimated motion segmentation might consequently include teacher noise. Additionally, measured acceleration prevails in such low frequency bands (0–20 Hz) that spectral information is sparse. As a result, motion segmentation specialized for acceleration is temporally processed mainly [3–6, 9]. While calculating the acceleration energy in the time domain, another constraint emerges. Sample-wise linear separation between motion and nonmotion states is formidable without modifying a multivalley structure; plus, time delay produced by modifying the multivalley structure has proportional relation to accuracy [3].

The proportional tendency between accuracy and time delay in conventional approaches provokes a new requirement for rapid response time with the advent of smart devices [9, 10]. Motion segmentation obsessed with accuracy

naturally leads to requiring an appropriate trade-off between accuracy and delay. For accomplishing maximum accuracy with minimum time delay, the employment of multivariate energy appended to hyper decision boundaries has been introduced as a promising alternative [9, 11]. This approach achieves the time delay reduction by skipping energy smoothing, which is the main cause of the time delay in the previous univariate approach. Instead of an explicit energy smoothing process, a shorter time delay is produced implicitly when multivariate energy vectors are generated. The loss of accuracy resulting from the reduced time delay in this approach is compensated by motion state decision making with a nonlinear hyper decision boundary in high-dimensional space.

Consequently, accuracy is dependent on the separability between data distributions of two states represented by multivariate energy in high-dimensional space, and it is required to predict the discriminality of each data distribution represented by variables or their multidimensional combinations for building optimal multivariate energy. Because the performance of classifiers implementing a hyper decision boundary may well have a limit, it is important to find and identify variables that can have discriminant distributions between two states in multivariate space. In addition, it is so fundamental to depend on statistical regularities represented by data in pattern recognition that state separability can be used to show how well data is distributed in high-dimensional space for a given task.

2. Problem Description

The key cause of the given problem is the multivalley structure that commonly occurs in calculating temporal energy. Figure 1 shows the parsed acceleration signal $a(t)$ from a simple arm motion and its basic energy $|a(t)|$. There, the red dotted line represents the motion period, where nonzero values stand for motion state. Acceleration from arm motion has a multipeaked structure that is a representative of all human arm motion [3]. The energy calculation transforms the multipeaked structure into the multivalley structure at the bottom of Figure 1, which is commonly observed in various energy types [3–6, 9–11]. In this structure, the multiple valleys prevent a linear threshold from simply discriminating motion and nonmotion states, and this phenomenon explains why energy smoothing is required. As smoothing means to extract the desired signal by removing multiple peaks and valleys in the original signal in terms of signal processing, it represents the process to fill the valleys to make the difference between two states clear in this case. The main difference among algorithms is techniques employed to smooth these valleys: low-pass filtering including moving average, axial information integration, inactivated interval setting, extra signal addition, and so forth [2–6, 12].

The performance evaluation of motion segmentation algorithms is generally given on the basis of accuracy; however, time delay in algorithms has recently started to be taken into consideration [5, 6, 10]. A related phenomenon is explained in Figure 2, where energy is calculated by a piecewise moving variance given by Benbasat and Paradiso

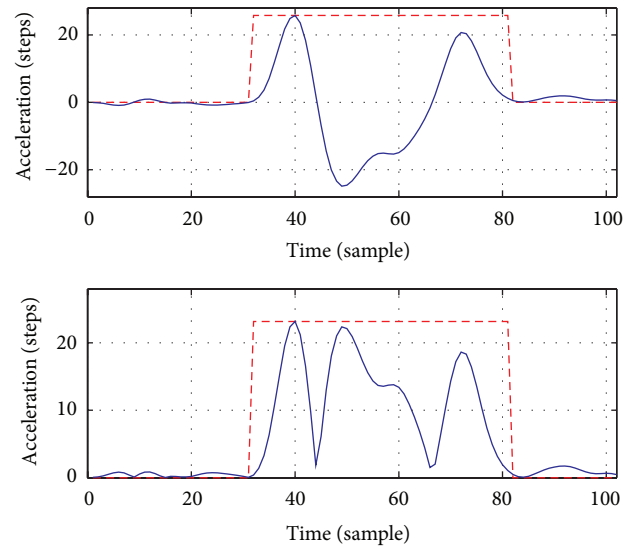


Figure 1: Acceleration and its temporally calculated energy.

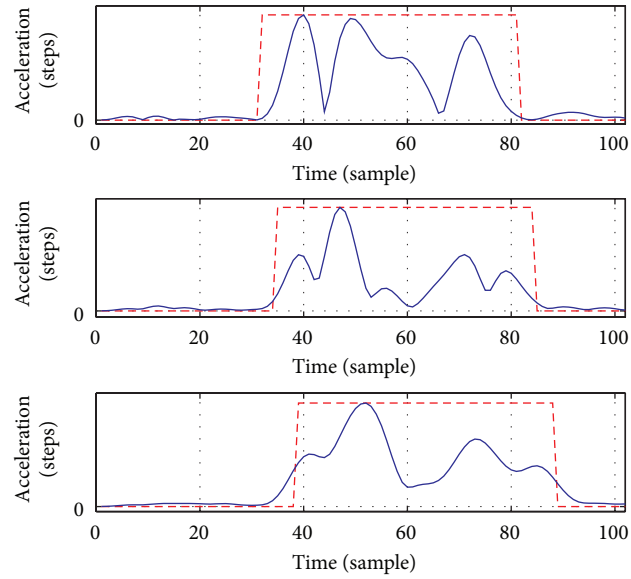


Figure 2: Multivalley structure change over time delay increase in smoothing.

[3]. In this approach, the length of a sliding window is directly proportional to the size of the time delay. The graph (without time delay) at the bottom of Figure 1 is again shown at the top of Figure 2 for comparison, and each smoothed energy variation with time delays of 70 ms and 150 ms, respectively, follows by turn. It is clearly shown that the discrimination between two states gets easier by a simple threshold, as the time delay increases.

Theoretically, in this situation, accuracy equates to indicating the exact motion starting and ending points; practically, however, the whole detected motion period is compared with the one given by the target label (red dotted line), which measures their overlap with the number of successfully detected samples with respect to full samples or similarity

measurements between two time series [5, 6]. If accuracy is 100%, the annotated and estimated motion periods must be coincident. When fluctuation in acceleration is occasionally extreme and energy smoothing is disabled to flatten the valleys fully, the motion discontinuity happens in the estimated motion period, and such a phenomenon needs to be considered a detection failure regardless of accuracy. To avoid this, energy smoothing is reinforced, thereby increasing time delays.

Time delay in this paper results solely from algorithms excluding computation and communication. It is determined by the past data length stored in short-term memory and the group delay for digital filtering, regardless of hardware enhancements. It is basic in statistical inference to make a decision based on previous data. The capacity to store the previous data for processing current data is called the short-term memory [13]. In signal processing, sliding windows implement this by generating a time delay proportional to a window length for the derivative of the signal with respect to time, moving average/variance, and digital filtering often found in algorithms. Group delay is an integrated measurement of the time delay by frequency band when the signal goes through filters. Filtering produces group delay $\partial\phi/\partial\omega$, where ϕ and ω represent phase shift and radian frequency, respectively [14]. Moving average is a special case of low-pass filtering to generate group delay. Moving variance can be interpreted as a case of moving average with additional operations since the moving average is used in its calculation. Given that motion segmentation is generally a part of full interaction, time delay by motion segmentation should be much less than the optimal delay of 150–200 ms reported by event-related brain potential measurements for a computer response to a user action [15, 16].

The minimized time delay requirement turns efficient energy smoothing in previous approaches into an estimate of the probability at two states in high-dimensional space by expanding univariate energy to multivariate. Borza [11] and Lim et al. [9, 17] introduce motion segmentation based on this idea, but multivariate energy and state decision methods in their approaches differ. While Borza's approach emphasizes axial integration and the difference between only two time sequences given in (2) for generating variables, Lim et al. are interested in various variables and their combinations, including the time series of a certain length without axial integration as shown in Table 1. Consider the following:

$$\{\widehat{a}(t-1), \widehat{a}(t), \Delta\widehat{a}(t)\}, \tag{1}$$

where

$$\widehat{a}(t) = \sqrt{a_x^2(t) + a_y^2(t) + a_z^2(t)},$$

$$\Delta\widehat{a}(t) = \sqrt{\left(a_x(t) - a_x(t-1)\right)^2 + \cdots + \left(a_z(t) - a_z(t-1)\right)^2}. \tag{2}$$

The interest in various candidates of Lim et al. naturally induces the question of how to choose the best combination, and feature subset selection in machine learning is consequently employed to build multivariate energy in motion segmentation [13, 18–20]. They adopt a naïve sequential feature selection to estimate the predictability between each candidate and target values with correlation coefficients as an objective function. It will very rarely work since the estimation of multivariate feature subsets is discarded in this strategy by not accounting for variable dependence.

Table 1

	Candidates
Variables	$\{a_i(t)\}$, $\{\lvert a_i(t)\rvert\}$, $\{\Delta a_i(t)\}$, $\{\lvert\Delta a_i(t)\rvert\}$, $\{\Delta^2 a_i(t)\}$, $\{\lvert\Delta^2 a_i(t)\rvert\}$, where $\Delta a_i(t) = a_i(t) - a_i(t-1)$, $\Delta^2 a_i(t) = \Delta a_i(t) - \Delta a_i(t-1)$
Combinations	$\{\lvert a_i(t)\rvert, \lvert\Delta a_i(t)\rvert\}$, $\{\lvert a_i(t)\rvert, \lvert\Delta^2 a_i(t)\rvert\}$, $\{\lvert\Delta a_i(t)\rvert, \lvert\Delta^2 a_i(t)\rvert\}$, $\{\lvert a_i(t)\rvert, \lvert\Delta a_i(t)\rvert, \lvert\Delta^2 a_i(t)\rvert\}$, $\{\lvert\Delta a_i(t-n)\rvert, \lvert\Delta a_i(t-(n-1))\rvert, \ldots, \lvert\Delta a_i(t)\rvert\}$

Feature subset selection is the process of identifying and eliminating as much irrelevant and redundant information as possible [13, 20, 21]. Diminishing the dimensionality of the data may allow learning algorithms to operate faster and more effectively, and, in most cases, final classification accuracy can be improved and data can be easily interpreted as a representation of the target concept. Filter and wrapper methods, which vary in how to estimate feature subset candidates, are generally accepted. Filter methods are the earliest approaches to feature selection within machine learning. They use additional objective functions based on general characteristics of the data to evaluate the merit of feature subsets, whereas wrapper strategies use a learning algorithm to estimate such merit. As a result, filter methods are generally much faster than wrapper methods and, as such, are more practical for use on high-dimensional data. The rationale for wrapper approaches is that the task-dependent induction algorithms should provide a better estimate of accuracy than a separate measure with inductive bias. Despite the better estimate of wrappers tuned to the specific interaction between an induction algorithm and its training data, they tend to be much slower than filter strategies because feature selection must be accompanied by a model selection process for the induction algorithm used.

In this study, filter strategy is scrutinized with several causes. Our problem is the investigation as to how to choose relevant variables for multivariate energy construction to reduce time delays with superior or equivalent accuracy guaranteed. For the given task, an evaluation of the estimate by a few objective functions is required. Another underlying goal that can be accomplished during this investigation is the justification of a multivariate approach compared with the previous univariate approach. To achieve this, we put more emphasis on the understanding of general characteristics of acceleration data than on a learning algorithm. The comprehension of data distribution is followed by designing a hyper decision boundary that should be so independent that more various applications can be expected; however since

Table 2: Univariate energy and multivariate energy.

Abbreviation	Energy	Dim.
BENBASAT.n	$\left\{\sum_{t-n}^{t} a_i^2(k) - \left(\sum_{t-n}^{t} a_i(k)\right)^2\right\}$	1
LIM1,. . .,LIM6	$\{a_i(t)\}$, $\{\lvert a_i(t)\rvert\}$, $\{\Delta a_i(t)\}$, $\{\lvert\Delta a_i(t)\rvert\}$, $\{\Delta^2 a_i(t)\}$, $\{\lvert\Delta^2 a_i(t)\rvert\}$	1
LIM4.n	$\{\lvert\Delta a_i(t-n)\rvert, \lvert\Delta a_i(t-(n-1))\rvert, \ldots, \lvert\Delta a_i(t)\rvert\}$	$n+1$
LIM7,. . .,LIM63	{LIM1, LIM2}, {LIM1, LIM3}, . . . , {LIM1, LIM2, LIM3, LIM4, LIM5, LIM6}	2, . . .,6
LIM7(LIM2.n)	{LIM1, LIM2.n}	$n+2$

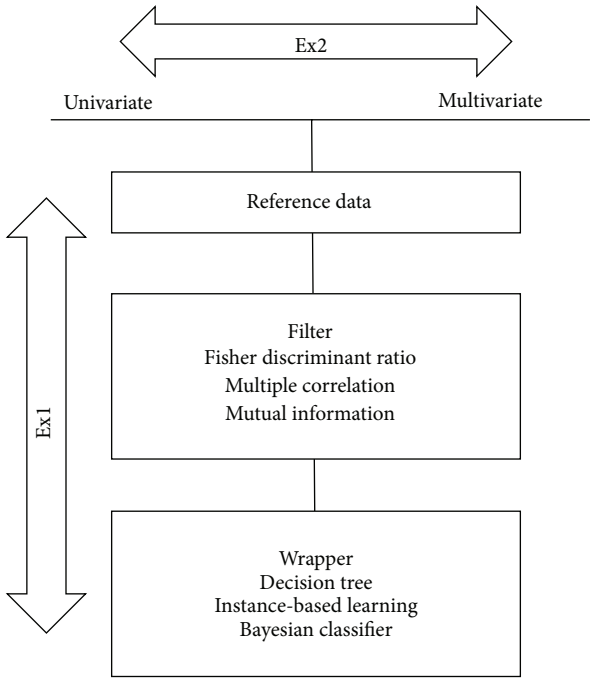


Figure 3: Overall experiment process.

wrapper methods are generally accepted to provide better estimates of feature subsets, the reliability and limitations in discriminality of filter strategies need to be compared with those of wrapper strategies.

3. Experiment

With respect to handwriting acceleration, univariate energy proposed by Benbasat and Paradiso [3] and multivariate energy by Lim et al. [9, 17] are created, and each separability estimate is measured by filter and wrapper processes. For the rigorous comparison, theoretical errors are calculated as reference data based on the conditional probability density function of both motion and nonmotion states. A detailed explanation of experimental conditions will be provided. Figure 3 shows overall experiments. Throughout the experiments, the following questions are pursued:

(i) Can filter approaches estimate accurately enough to predict discriminality between motion and nonmotion states?

(ii) Can it be justified that multivariate energy guarantees superior time delay and accuracy to univariate energy?

(iii) Can the analysis of the above results offer the understanding of the underlying structure of data distributions?

3.1. Data. A total of 294 handwriting measurements are collected with a 3D pen embedded with three-axis accelerometer MMA 7260Q (Freescale) from 7 subjects (male 4, female 3) thrice when drawing the numbers from 0 to 9 and four kinds of symbols. In data acquisition by microcontroller Atmega8 (Atmel), two least significant bits are discarded to cancel the noise effect for 10-bit quantization and 100 Hz sampling. Samplewise motion state annotations paired with acceleration profiles, that is, target values, are measured by subjects pushing a button to mark when drawing [22]. Collected data has been finally grouped into training (98 pieces, 17189 samples), validation (98 pieces, 16728 samples), and test set (98 pieces, 17489 samples). Since acceleration and its paired target label are considered at a single axis, acceleration profiles at three axes integrate to their samplewise mean.

3.2. Energy Generation by Univariate and Multivariate Approaches. The univariate energy used by Benbasat and Paradiso [3] and multivariate energy by Lim et al. [9, 17] have been chosen for the investigation. In the approach by Benbasat and Paradiso, the energy is calculated by piecewise moving variance, which combines energy calculation and smoothing. It is an upgraded version of earlier energy calculation of absolute conversion or squared acceleration and is widely accepted as one of the baseline methods considering that several variations have been created.

For multivariate approaches the multivariate energy in Lim et al. [9, 17] are mainly used for the experiment. Note that every type of energy is abbreviated as in Table 2 for clarity and simplicity hereafter. For feature subset selection and strongly relevant variable identification, the subsets of LIM1~LIM63 are employed in the experiment 1, and, after selecting the best subset, it is compared with BENBASAT.n in experiment 2 (Figure 3). While LIM1~LIM6 are subsets including each basic variable, LIM7~LIM63 are subsets composed of the combinations of basic variables.

3.3. Theoretical Measure. To compare filter and wrapper estimates, we need a theoretical reference. We define the likelihood of each state as a conditional probability density function with two assumptions: each is Gaussian distributed and variables v_i of subset $\mathbf{V} = \{v_1, v_2, \ldots, v_n\}$ are independent

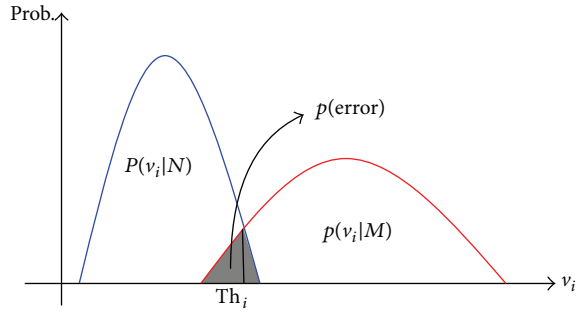


Figure 4: Likelihood of each state.

and identically distributed for $i \neq j$. The density distribution of each state is given as follows and in Figure 4:

$$N = \{v_i \mid v_i \in \text{non-motion state}\} \sim N\left(\mu_1, \sigma_1^2\right),$$
$$M = \{v_i \mid v_i \in \text{motion state}\} \sim N\left(\mu_2, \sigma_2^2\right), \tag{3}$$

where

$$\mu_1 \leq \mu_2,$$
$$\sigma_1 \leq \sigma_2,$$
$$p\left(v_i \mid N\right) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-1/2((v_i-\mu_1)/\sigma_1)^2}, \tag{4}$$
$$p\left(v_i \mid M\right) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-1/2((v_i-\mu_2)/\sigma_2)^2}.$$

In this condition, error results from the overlap between the two states are given by the following equation and are depicted by the dark region in Figure 4:

$$p\left(\text{error}\right) = \int_{-\infty}^{\text{Th}_i} p\left(v_i \mid M\right) dv_i + \int_{\text{Th}_i}^{\infty} p\left(v_i \mid N\right) dv_i, \tag{5}$$

where a threshold Th_i is found by satisfying $p(\text{Th}_i \mid N) = p(\text{Th}_i \mid M)$. Consider

$$\frac{1}{\sqrt{2\pi}\sigma_1} e^{-1/2((\text{Th}_i-\mu_1)/\sigma_1)^2} = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-1/2((\text{Th}_i-\mu_2)/\sigma_2)^2} \tag{6}$$

$$\left(\sigma_1^2 - \sigma_2^2\right)\text{Th}_i{}^2 - 2\left(\mu_2\sigma_1^2 - \mu_1\sigma_2^2\right)\text{Th}_i + \left(\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2\right) + 2\sigma_1^2\sigma_2^2 \ln\frac{\sigma_1}{\sigma_2} = 0. \tag{7}$$

Let the coefficient of each term in (7) be A, B, and C, respectively,

$$\text{Th}_i = \max\left\{\frac{-B \pm \sqrt{B^2 - 4AC}}{2A}\right\}. \tag{8}$$

Since $\mathbf{V}$ is multivariate, for example, $\mathbf{V} = \{v_1, v_2\}$, (5) is rewritten as

$$\begin{aligned}\text{TM} &= \int_{-\infty}^{\text{Th}_\mathbf{V}} p\left(v_1, v_2 \mid M\right) d\mathbf{V} + \int_{\text{Th}_\mathbf{V}}^{\infty} p\left(v_1, v_2 \mid N\right) d\mathbf{V} \\ &= \int_{-\infty}^{\text{Th}_2}\int_{-\infty}^{\text{Th}_1} p\left(v_1, v_2 \mid M\right) dv_1 dv_2 + \int_{\text{Th}_2}^{\infty}\int_{\text{Th}_1}^{\infty} p\left(v_1, v_n \mid N\right) dv_1 dv_2 \\ &\cong \frac{n\left(\{\mathbf{V} = \langle v_1, v_2\rangle \mid v_1 \leq \text{Th}_1, v_2 \leq \text{Th}_2, \mathbf{V} \in M\}\right) + n\left(\{\mathbf{V} = \langle v_1, v_2\rangle \mid v_1 > \text{Th}_1, v_2 > \text{Th}_2, \mathbf{V} \in N\}\right)}{n\left(M \cup N\right)}.\end{aligned} \tag{9}$$

Therefore, the approximate error, which is estimated by the likelihood of both states, can be counted to the summation of the number of samples that belong to motion state depicted by the bright grey area and samples that belong to nonmotion state depicted by the dark grey area in Figure 5. The actual boundary is located in the line orthogonal to the connecting line between mean vectors at both states, because the state membership of each sample is determined by Mahalanobis distances from each mean vector of two Gaussian distributions. Accordingly error should be also estimated by this linear boundary, but we simplify it in the way of (9) due to computation convenience, which indicates minimum error with highly probable occurrence only.

3.4. Feature Subset Selection: Filter Approaches. Traditional feature subset selection process includes two steps of subset generation and subset evaluation. Since $\sum_{k=1}^{n} {}_nC_k$ subset candidates can be produced with respect to n variable candidates, various greedy search strategies are generally used to reduce computation. In our study, the subset candidates are fixed so that search strategy is not considered. We concentrate only on how to evaluate each subset.

Typical objective functions in filter approaches are based on distance measures, dependence measures, and information measures [18, 20]. Fisher Discriminant Ratio (FDR) or Fisher criterion is exemplary in distance measures and is defined by the ratio of the between-class scatter $\mathbf{S}_B$ to the within-class scatter $\mathbf{S}_W$, where there is a total of N instances

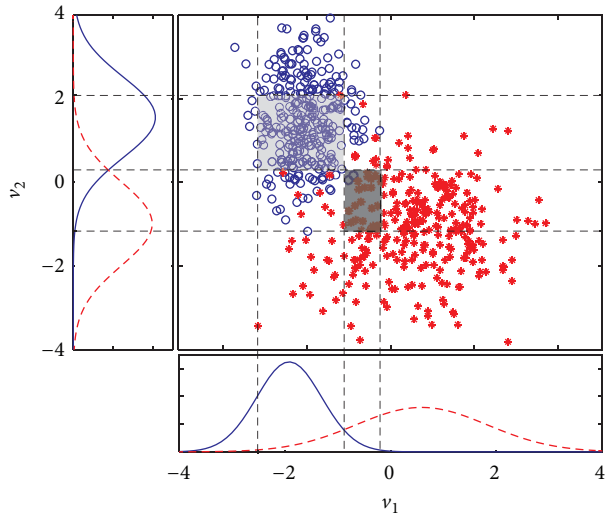


FIGURE 5: Error approximation of multivariate data distribution.

and n_i $(i = 1, 2, \ldots, C)$ in C classes, as shown in the following equations:

$$\text{FDR} = \frac{\text{tr}\left(\mathbf{S}_B\right)}{\text{tr}\left(\mathbf{S}_W\right)}, \tag{10}$$

where

$$\mathbf{S}_W = \sum_{i=1}^{C} \frac{n_i}{N} \sum_{j=1}^{n_i} \frac{1}{n_i} \left(\mathbf{x}_j - \mathbf{m}_i\right)\left(\mathbf{x}_j - \mathbf{m}_i\right)^T, \tag{11}$$

$$\mathbf{S}_B = \sum_{i=1}^{C} \frac{n_i}{N} \left(\mathbf{m}_i - \mathbf{M}\right)\left(\mathbf{m}_i - \mathbf{M}\right)^T, \tag{12}$$

where

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_j, \tag{13}$$

$$\mathbf{M} = \sum_{i=1}^{C} \frac{n_i}{N} \mathbf{m}_i. \tag{14}$$

A multiple correlation coefficient is a multivariate extension of a traditional correlation coefficient, which is a typical statistical technique to measure linear dependence between variables [18–20]. In statistics, the multiple correlation coefficient measures how well a given variable can be predicted by a set of other variables using the ratio of the correlation between variable vectors $\mathbf{x}_i$ and target values $\mathbf{y}_i$ and the correlation between each variable in (15)–(18). As a result, the process of multiple correlation is equivalent to the rationale that a good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other:

$$\text{MCC} = \mathbf{C}^T \mathbf{R}_{xx}^{-1} \mathbf{C}, \tag{15}$$

where

$$\mathbf{C} = \frac{(1/N)\sum_{i=1}^{N}\left(\mathbf{x}_i - \mathbf{m}_\mathbf{x}\right)\left(\mathbf{y}_i - \mathbf{m}_\mathbf{y}\right)}{\sigma_\mathbf{x}\sigma_\mathbf{y}}, \tag{16}$$

$$\mathbf{R}_{\mathbf{xx}} = \frac{(1/N)\sum_{i=1}^{N}\left(\mathbf{x}_i - \mathbf{m}_\mathbf{x}\right)\left(\mathbf{x}_i - \mathbf{m}_\mathbf{x}\right)^T}{\sigma_\mathbf{x}\sigma_\mathbf{x}^T}, \tag{17}$$

where

$$\sigma_\mathbf{x} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}_i - \mathbf{m}_\mathbf{x}\right)^2}. \tag{18}$$

Correlation is capable of measuring linear dependence only. A more powerful method, which measures nonlinear dependence, is the mutual information $I(V_k; \omega_C)$ in (19) under the condition that k subset candidates $\mathbf{V}_k$ and C classes ω_C are given [18–20]:

$$\begin{aligned} \text{MI} &= I\left(\mathbf{V}_k; \omega_C\right) = H\left(\omega_C\right) - H\left(\omega_C \mid \mathbf{V}_k\right) \\ &= \sum_{i=1}^{C} \int_{\mathbf{V}_k} p\left(\mathbf{V}_k, \omega_i\right) \log \frac{p\left(\mathbf{V}_k, \omega_i\right)}{p\left(\mathbf{V}_k\right) p\left(\omega_i\right)} d\mathbf{V}_k, \end{aligned} \tag{19}$$

where $H()$ is the entropy function. Intuitively, the mutual information method measures the information that $\mathbf{V}_k$ and ω_C share: it measures the amount by which the uncertainty in the class $H(\omega_C)$, prior uncertainty, is decreased by knowledge of the subset $H(\omega_C \mid \mathbf{V}_k)$, expected posterior uncertainty. For the high order density estimation from limited data, we apply a mixture of three Gaussian distributions.

3.5. Feature Subset Selection: Wrapper Approaches. Of the numerous classification algorithms available, Bayes Classifiers (BCs) and k-nearest neighbor classifiers (KNNs) have been chosen because both of them, proposed relatively early, have small numbers of parameters for performance optimization, and their competence has been generally accepted enough to regard them as one of filters based on recognition rates [13]. In addition, since these statistical classification algorithms have their own statistical models quite different from one another, each of their results helps us to comprehend the characteristics of high-dimensional data distribution. Note that their parameters are tuned with the validation set after training, and the error rates are finally counted in the test set.

A BC is a statistically parametric classifier based on applying Bayes' theorem, such as the naïve Bayes classification given in the reference data section. Due to the assumption of strong independence between feature variables, its performance can be improved by removing redundant features. The identical conditions and data distributions given in Section 3.3 are applied except for the consideration of data dimension using mean vector $\mathbf{m}_i$ given in (13) and covariance matrix $\mathbf{\Sigma}_i$ by the inside summation term in (11). Given i classes with n dimensional data, each Gaussian multivariate density $f_i(\mathbf{x})$ is given in (20), and its second-order discrimination

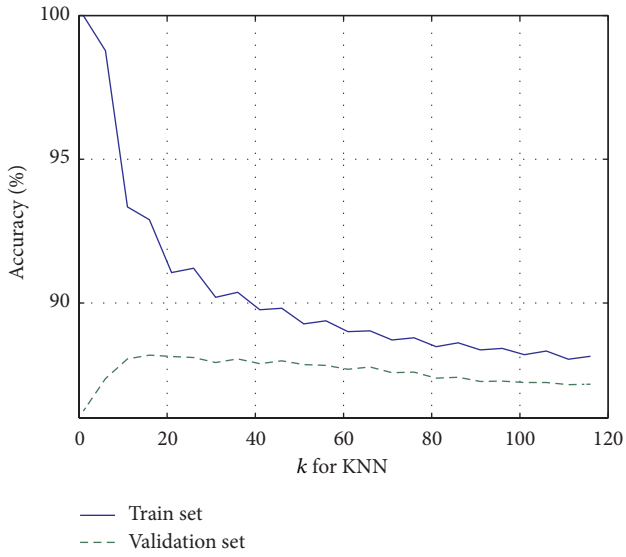


Figure 6: Likelihood of each state.

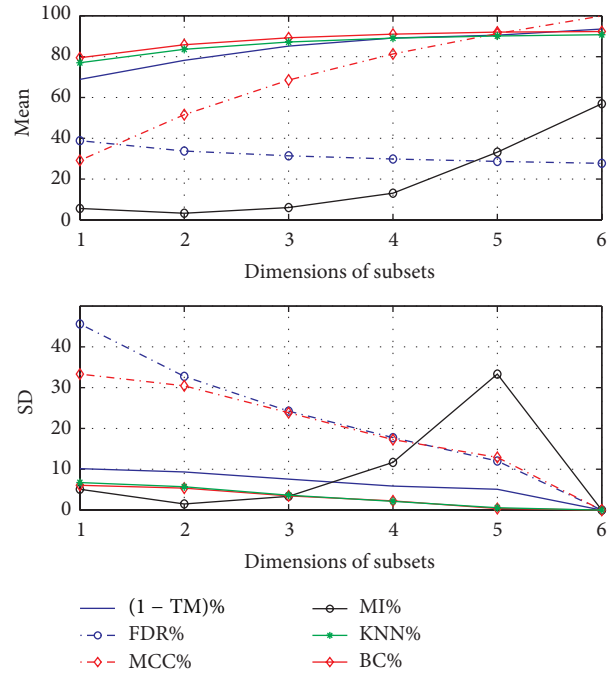


Figure 7: Estimated measure averages and standard deviations.

function $g_i(\mathbf{x})$ is given by taking the natural logarithm of each side of (20) and simplifying it for classification in (21):

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{m}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}-\mathbf{m}_i\right)\right], \quad (20)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\left(\mathbf{x}-\mathbf{m}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}-\mathbf{m}_i\right) - \frac{1}{2}\ln\left(\left|\boldsymbol{\Sigma}_i\right|\right). \quad (21)$$

A KNN is a representative of nonparametric methods and is a type of instance-based learning used in classification and regression [13]. In both cases, the input instance is classified by a majority vote of its k closest training samples, neighbors, in the feature space, with the instance being assigned to the most common class among its k nearest neighbors. If $k = 1$, the instance is simply assigned to the class of the single nearest neighbor. The density function is locally approximated, and all computation is deferred until classification. A KNN is likely to be overfit so that k is chosen with an extra validation set (Figure 6).

4. Result and Analysis

4.1. Experiment 1

4.1.1. Descriptive Statistics. With respect to 63 possible subsets out of LIM1, LIM2, LIM3, LIM4, LIM5, and LIM6, we apply three filters of FDR, MCC, and MI and two wrappers of KNN and BC in addition to the theoretical measure, TM. Figure 7 presents the descriptive statistics of the object function scores estimated by each method by showing the average (top) and standard deviation (bottom) of 6 groups which are categorized according to the dimension of subsets. Except for KNN and BC, since all measures are linearly adjusted to bring all of them into proportion with one another for the normalization, note that it is meaningless to compare the scores estimated by different strategies in Figure 7.

It appears that every filter shows similar estimates to two wrappers with respect to 63 individual subsets by the statistical analysis. Of the filters, MCC is evidently correlated with two wrappers of KNN and BC with the general trend that the closer to six the subset dimension gets, the greater the average estimates are and the smaller their variances get excluding the mean of FDR and the standard deviation of MI. Despite the dissimilar tendency of FDR and MI in Figure 7, the correlations between the methods give another insight with respect to 63 individual subsets in Table 3. The significant correlations between MI and two wrappers imply that MI has just ill-fitting scales but tends to bring about analogous scores. Likewise FDR might be explained to record similar scores to two wrappers with respect to the individual subsets considering the correlation with MCC, which is significantly correlated with two wrappers and TM.

However it turns out that the poorer scores FDR tends to underestimate the higher dimensions the subsets have because the distributions of the variables from LIM1 to LIM6 has much narrower mean differences compared to variances in (11). As the subset dimension consequently increases, this asymmetric proportion gets worse by summing the diagonal components of the covariance matrix. Even though MCC uses a similar covariance matrix to that of FDR in (17), the covariance matrix for MCC is normalized by the product of standard deviations and all of elements are included to calculate the influence of each variable.

4.1.2. Network Analysis. General feature selection employs various heuristic greedy search strategies to find an optimal subset, but we conduct an exhaustive full search to

Table 3: Correlations between each method.

		N	Mean	SD	Correlations					
					1	2	3	4	5	6
1	TM	63	83.5513	9.9085	1					
2	FDR	63	31.9509	26.0949	0.2034	1				
3	MCC	63	66.4364	29.5854	$0.7352^{\dagger}$	$0.5946^{\dagger}$	1			
4	MI	63	9.3350	14.1683	0.2756^{*}	−0.0711	0.2443	1		
5	KNN	63	88.2365	5.2062	$0.6654^{\dagger}$	−0.0633	$0.3898^{\dagger}$	0.2984^{*}	1	
6	BC	63	86.1483	5.4823	$0.6985^{\dagger}$	−0.0293	$0.4362^{\dagger}$	0.2967^{*}	$0.9926^{\dagger}$	1

df = 61, $^{*}p < 0.05$, $^{\dagger}p < 0.01$.

Table 4: Variable evaluation by objective functions.

	Strongly relevant	Weakly relevant	Irrelevant	Subset
KNN	LIM1, LIM3, LIM4	LIM2, LIM5		LIM48 = {LIM1, LIM3, LIM4, LIM5}
MI	LIM1, LIM2, LIM4	LIM3, LIM5, LIM6		LIM61 = {LIM1, LIM3, LIM4, LIM5, LIM6}
FDR	LIM2, LIM4	LIM6	LIM5	LIM36 = {LIM2, LIM4, LIM6}
BC	LIM1, LIM3, LIM4	LIM2, LIM5		LIM57 = {LIM1, LIM2, LIM3, LIM4, LIM5}
MCC	LIM2, LIM4	LIM1, LIM3, LIM5, LIM6		LIM42 = {LIM1, LIM2, LIM3, LIM4}
TM	LIM2, LIM4, LIM6	LIM3		LIM46 = {LIM1, LIM2, LIM4, LIM6}

understand the attributes of each objective function. Instead of omitting this procedure, we analyze the interrelation between six variables of LIM1~LIM6 with a social network analysis technique based on the same data used for statistical analysis. As a result, this analysis reveals the underlying attributes of each measure.

We regard the variables and the subset as keywords and a link, respectively, for network visualization. To begin with, we identify the affirmative influences of each variable on the discriminality estimation. After the subsets are ranked in order of scores, we choose 10% of total subsets with highest scores and split variables from the subsets. Its influence is then counted by a vote because the stronger influence the variable has, the more frequently it appears in the above selection. After the negative influences are identically identified in another selection of 10% of total subsets with lowest scores, we subtract two votes in each variable and normalize their scales into the range between −1 and 1. With two selections, the network influences are analyzed by counting links between variables again and the links with votes below average are finally removed for clarity. Figure 8 shows the network analysis visualization of KNN, MI, FDR, BC, MCC, and TM.

First similarity among them is that every measure does not specify irrelevant variables because all of variables record nonnegative scores except for FDR. It is interpreted that each variable is strongly or weakly relevant given that the scores of two wrappers tend to be proportional to the subset dimension in the above statistical analysis. Another analogy comes from the network connections between variables which the linearity of each measure causes. It is observed that KNN has a resemblance to MI, and BC does to MCC with few differences in the network topology and this similarity is prominent by classifying respective variables into strongly relevant, weakly relevant, and irrelevant variables based on the network analysis visualization (Table 4).

Such a topological analogy also explains why MI records higher correlation with KNN than with BC and MCC vice versa in Table 3. In addition, although FDR is poor at estimating subsets with different dimensions, it appears that the significant correlation with MCC is achieved by the fact that FDR and MCC share the common strongly relevant variables of LIM2 and LIM4. Note that the subsets in Table 4 are just transformed from each network topology with links and nodes into the description of subsets and variables.

4.2. Experiment 2. Based on the previous analysis, we validate the possibility that multivariate approaches can be a better alternative to a conventional univariate in experiment 2. The energy of BENBASAT.n based on the piecewise variance tends to produce improved accuracy, as the size of sliding window increases. For comparison, we propose four multivariate energy candidates combining LIM48 = {LIM1, LIM3, LIM4, LIM5}, which are identified as the best subset in Table 4 when using a KNN, to the idea of the time series with $n + 1$ lengths of the previous data. After using LIM3.n as a multivariate energy basis because LIM3 are identified as the most strongly relevant variable as a result of a KNN evaluation, it is bound with LIM1, LIM4, and LIM5 one after another in the order of the variable influences in Figure 8(a), only to create the multivariate energy candidates of LIM3.n, LIM9 (LIM3.n) = {LIM1, LIM3.n}, LIM26 (LIM3.n) = {LIM1, LIM3.n, LIM4}, and LIM48 (LIM3.n) = {LIM1, LIM3.n, LIM4, LIM5}. Finally as n increases, the changes in accuracy are compared with BENBASAT.n, along with time delay. In this way, we examine the potential of the energy with high dimension and reconfirm the fidelity of network analysis in the experiment 1 simultaneously.

Figure 9 shows the comparison between BENBASAT.n and four multivariate approaches in accuracy and time delay. In the result of a KNN, BENBASAT.14 records the best accuracy of 94.95% at $n = 14$ while LIM48 (LIM3.5) shows its

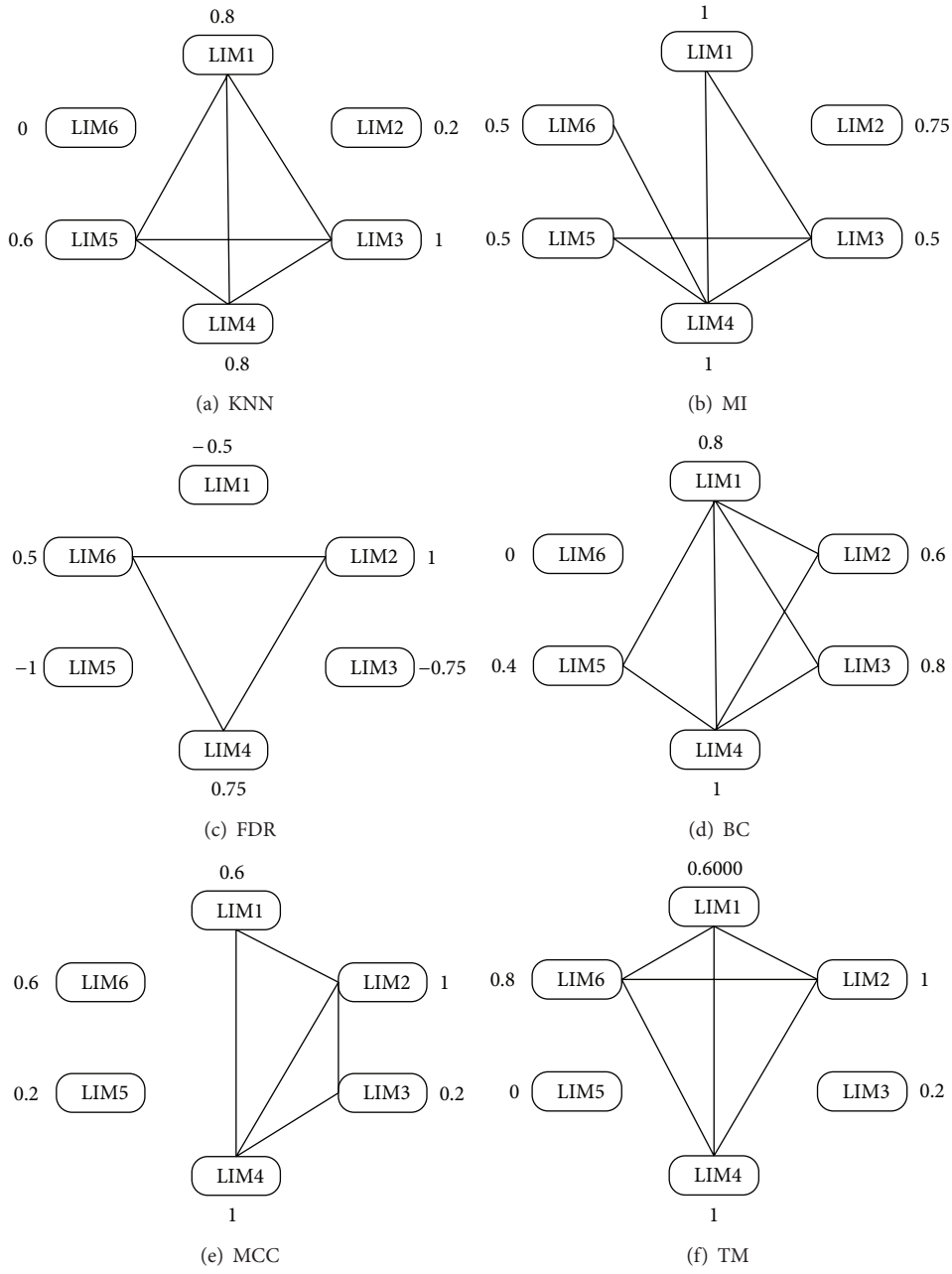
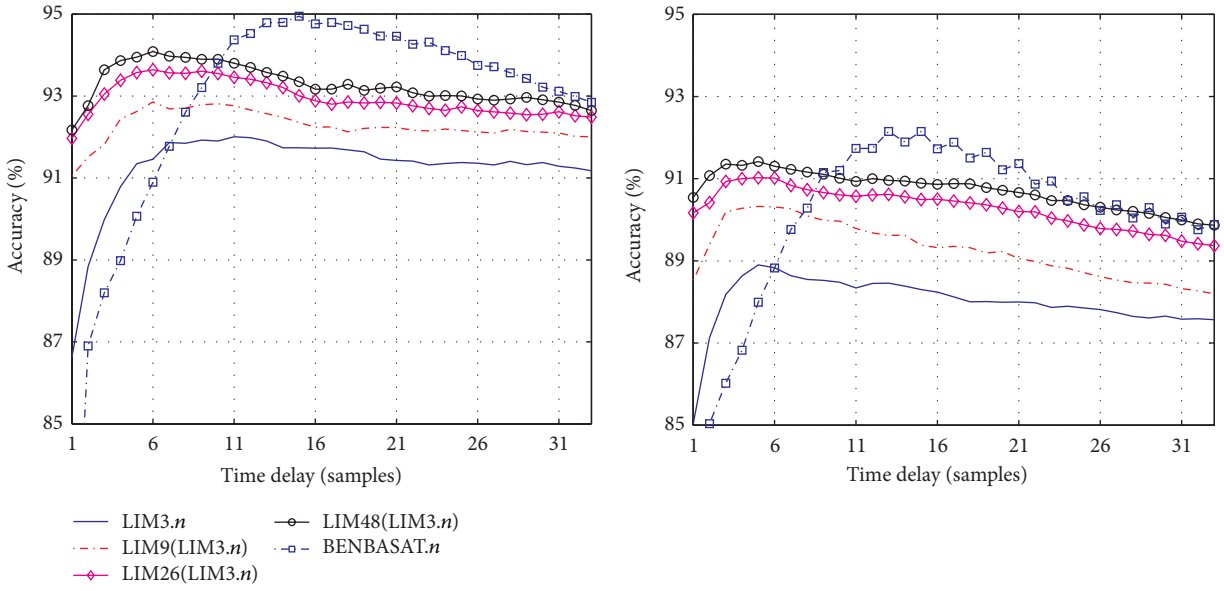


Figure 8: Network analysis visualization.

best performance of 94.09% at $n = 5$. If the comparison is only evaluated in the aspect of accuracy, no doubt BENBASAT.14 is the best choice and the proposed multivariate energy is purposeless; however the identical consequence can have the contrasting significance when time delay is taken into account. When $n = 14$, time delay caused by the size of short-term memory and group delay is counted to 150 ms. Noted that we deal with the delay caused by a pure algorithm alone excluding all delays which result from computation and communication. In spite of satisfying the optimal delay condition of 150–200 ms for a computer response to a user action, it is easily concluded that the performance of 94.95% in accuracy and 150 ms in delay is not accepted to be excellent considering motion segmentation is usually used as one component in the entire interaction system.

On the contrary, the changes in the evaluation criteria are led to reevaluate LIM48 (LIM3.5) with 94.09% in accuracy and 60 ms in delay. If one tries to reduce the time delay of BENBASAT.14 as less as that of LIM48 (LIM3.5), the risk of accuracy reduction by 4% needs to be taken, and this is one of benefits which LIM48 (LIM3.n) possesses because its changes in accuracy is not rapid with respect to the changes in time delays. Even considering the minimum time delay of 10 ms in LIM48 (LIM3.0) at $n = 0$, its accuracy of 92.17% is remarkably excellent compared to 78.37% in BENBASAT.0. The identical tendency appears in the result of a BC in Figure 9 except for



(a) KNN (b) BC

Figure 9: Accuracy and time delay comparison between multivariate and univariate approaches.

accuracy differences. In addition, the fact that a nonlinear KNN shows better estimates than a linear BC in Figure 9 implies that the estimate is so dependent on the choice of classifiers that a new nonlinear classifier might record better accuracy than a KNN, given that we simply employ it for the comparison with filters only because of its simplicity in modeling before its excellence in accuracy. For the further investigation on the enhancement in accuracy, the cutting-edge nonlinear classifiers will be more likely to be used, and the details are again discussed in the conclusion. Note that the ultimate goal of our study in this paper is not the improvement of motion segmentation performance but the validation of a few objective functions in filter strategies to replace wrapper approaches with larger computations.

Before finalizing the comparison, it is worthy of mentioning that the increment in accuracy as a variable is added to the basis subset of LIM3.*n* one after another. This result does not merely mean the dimension increments are led to the improvement in accuracy but implies the improvement in accuracy critically has to do with the selection of proper variables in that Figure 9 shows the dimension increment up to 33 in LIM3.*n* has the limitation of performance improvement. Therefore this result clearly justifies our analysis of experiment 1.

5. Conclusion

The goal of our study is to validate the reliability of a few objective functions to be used in finding optimal multivariate energy for motion segmentation in accelerometer applications. To achieve this goal, Fisher discriminent ratio, multiple correlation, and mutual information are tested by comparing them with a theoretical measure and two wrappers of KNNs and BCs in two experiments. Its analysis finally enables us to answer to three questions which have arisen during the investigation and this study is concluded giving summarized explanation to those questions instead of the formal conclusion.

(1) Can filter approaches estimate accurately enough to predict discriminality between motion and nonmotion states?

Of three objective functions and one theoretic measure we suggest, it turns out that multiple correlation, mutual information, and theoretic measure are competent enough to replace two wrappers. With respect to 63 subsets found in literatures, all of them excluding Fisher Discriminant Ratio clearly show that they are significantly correlated with the estimates produced by two wrappers. Furthermore the network analysis for the identification of strongly relevant variables clarifies that each function offers similar interpretation with respect to all possible 63 subsets from six variables. Since each distribution of motion and nonmotion states built by six basic variables from acceleration has too narrow mean differences and wide variance, Fisher Discriminant ratio tends to underestimate their separability, as the dimension of subsets increases. In addition, mutual information turns out to show reliable estimates enough to replace the wrappers, but it is so unstable that it varies dramatically from time to time due to the intractability of density estimation, as data dimension increments. This phenomenon comes from the computation complexity of high order density estimation using Gaussian mixture models, and we suggest calculating stable multivariate density estimation in the way to use variable box size over the corresponding variable space like [23, 24] instead of expectation-maximization algorithm.

(2) Can it be justified that multivariate energy guarantees superior time delay and accuracy to univariate energy?

In the comparison between one conventional univariate and our multivariate approaches, we justified the superiority of multivariate approach. In our experiment the univariate approach just showed better accuracy than ours by about 0.9%, but the rapid processing in our multivariate approach outperformed the univariate one by 100% more. It is also observed that the risk of the serious loss in accuracy is required to be taken for the reduction in time delay for the univariate approach while the performance of our multivariate approaches lies in stable ranges.

(3) Can the analysis of the above results offer the understanding of the underlying structure of data distributions?

Using four linear and two nonlinear measures to estimate the separability between motion and nonmotion states with acceleration data, we have concluded that data is distributed linearly and separably considering that multiple correlation works successfully in estimating the discriminality. Despite the linearity, since two distributions are located too closely, the messy condition in the excessively overlapped spaces hinder linear BCs from outperforming nonlinear KNNs. The distribution of two states varies from variables. Since acceleration data without absolute conversion consists of two distributions with nearly identical means but different variances while absolutely converted acceleration data is distributed relatively far distant each other, as a result, linear measures tend to identify variables with absolute conversion as strongly relevant ones and nonlinear estimators vice versa. Overall it seems that motion segmentation using acceleration needs to be achieved by classifiers with a nonlinear hyper boundary such as multilayer perceptrons or support vector machines prior to classifiers depending on Mahalanobis distance kernel such as radial basis functions or BCs, and it is because statistically Gaussian modeling is inefficient when data lie on or near a nonlinear manifold in the data space. Modeling data that lie very close to the surface of a sphere only requires a few parameters using an appropriate model, but it requires a very large number of diagonal Gaussians or a fairly large number of full-covariance Gaussians.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] A. D. King, "Inertial navigation—forty years of evolution," *GEC Review*, vol. 13, no. 3, pp. 140–149, 1998.

[2] R. Baron and R. Plamondon, "Acceleration measurement with an instrumented pen for signature verification and handwriting analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 38, no. 6, pp. 1132–1138, 1989.

[3] A. Y. Benbasat and J. A. Paradiso, "An inertial measurement framework for gesture recognition and applications," in *Gesture and Sign Language in Human-Computer Interaction*, vol. 2298 of *Lecture Notes in Computer Science*, pp. 9–20, Springer, Berlin, Germany, 2002.

[4] W.-C. Bang, W. Chang, K.-H. Kang, E.-S. Choi, A. Potanin, and D.-Y. Kim, "Self-contained spatial input device for wearable computers," in *Proceedings of the 16th International Symposium on Wearable Computers*, pp. 26–34, IEEE Computer Society, 2003.

[5] J. G. Lim, Y.-I. Sohn, and D.-S. Kwon, "Real-time accelerometer signal processing of end point detection and feature extraction for motion detection," in *Proceedings of the International Federation of Automatic Control-Human Machine System*, pp. 4–6, 2007.

[6] Y. Zhou, Z. Cheng, and L. Jing, "Threshold selection and adjustment for online segmentation of one-stroke finger gestures using single tri-axial accelerometer," *Multimedia Tools and Applications*, pp. 1–20, 2014.

[7] Y. K. Thong, M. S. Woolfson, J. A. Crowe, B. R. Hayes-Gill, and D. A. Jones, "Numerical double integration of acceleration measurements in noise," *Measurement*, vol. 36, no. 1, pp. 73–92, 2004.

[8] J. Huddle, "Trends in inertial systems technology for high accuracy AUV navigation," in *Proceedings of the Workshop on Autonomous Underwater Vehicles (AUV '98)*, pp. 63–73, IEEE, Cambridge, Mass, USA, August 1998.

[9] J. G. Lim, S.-Y. Kim, and D.-S. Kwon, "Pattern recognition-based real-time end point detection specialized for accelerometer signal," in *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM '09)*, pp. 203–208, IEEE, July 2009.

[10] S. Kim, G. Park, S. Yim, S. Choi, S. Jeon, and G. Han, "Gesture-recognizing hand-held interface with vibrotactile feedback for 3D interaction," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1169–1177, 2009.

[11] P.-V. Borza, *Motion-based gesture recognition with an accelerometer [Ph.D. thesis]*, Babes-Bolyai University, Faculty of Mathematics and Computer Science, 2008.

[12] E.-S. Choi, W.-C. Bang, S.-J. Cho, J. Yang, D.-Y. Kim, and S.-R. Kim, "Beatbox music phone: gesture-based interactive mobile phone using a tri-axis accelerometer," in *Proceedings of the IEEE International Conference on Industrial Technology (ICIT '05)*, pp. 97–102, IEEE, Hong Kong, December 2005.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2012.

[14] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, vol. 2, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[15] S. L. Smith and J. N. Mosier, *Guidelines for Designing User Interface Software*, Mitre Corporation, Bedford, Mass, USA, 1986.

[16] H. Nittono, "Event-related brain potentials corroborate subjectively optimal delay in computer response to a user's action," in *Engineering Psychology and Cognitive Ergonomics*, pp. 575–581, Springer, Berlin, Germany, 2007.

[17] J. G. Lim, S.-Y. Kim, and D.-S. Kwon, "Real-time end point detection specialized for acceleration signal," in *Proceedings of*

the International Joint Conference (ICCAS-SICE '09), pp. 5331–5335, IEEE, August 2009.

[18] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[19] M. A. Hall, *Correlation-based feature selection for machine learning [Ph.D. thesis]*, University of Waikato, Hamilton, New Zealand, 1999.

[20] S. Theodoridis and K. Koutroumbas, "Feature selection," in *Pattern Recognition*, pp. 139–179, Academic Press, San Diego, Calif, USA, 1998.

[21] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proceedings of the 11th International Conference on Machine Learning (ICML '94)*, New Brunswick, NJ, USA, 1994.

[22] J. G. Lim, F. Sharifi, and D.-S. Kwon, "Fast and reliable camera-tracked laser pointer system designed for audience," in *Proceedings of the 5th International Conference on Ubiquitous Robots and Ambient Intelligence*, pp. 529–534, 2008.

[23] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, vol. 33, no. 2, pp. 1134–1140, 1986.

[24] A. Al-Ani and M. Deriche, "Feature selection using a mutual information based measure," in *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 4, pp. 82–85, IEEE, 2002.