IEICE TRANS. INF. & SYST., VOL.E97–D, NO.8 AUGUST 2014

LETTER

Activity Recognition Based on an Accelerometer in a Smartphone Using an FFT-Based New Feature and Fusion Methods

Yang XUE[†a)], **Yaoquan HU**[†], ***Nonmembers***, ***and*** **Lianwen JIN**[†], ***Member***

SUMMARY With the development of personal electronic equipment, the use of a smartphone with a tri-axial accelerometer to detect human physical activity is becoming popular. In this paper, we propose a new feature based on FFT for activity recognition from tri-axial acceleration signals. To improve the classification performance, two fusion methods, minimal distance optimization (MDO) and variance contribution ranking (VCR), are proposed. The new proposed feature achieves a recognition rate of 92.41%, which outperforms six traditional time- or frequency-domain features. Furthermore, the proposed fusion methods effectively improve the recognition rates. In particular, the average accuracy based on class fusion VCR (CFVCR) is 97.01%, which results in an improvement in accuracy of 4.14% compared with the results without any fusion. Experiments confirm the effectiveness of the new proposed feature and fusion methods.

key words: *acceleration data, activity recognition, feature extraction, fusion method, tri-axial accelerometer*

1. Introduction

Recently, the development of personal electronic equipment has allowed for the popular use of personal companion devices such as smartphones with embedded sensing and computing power to detect physical activities. When smartphones are carried by people in pockets or bags, they are moving at the pace of the human body; thus, they appear to be the ideal platforms for detecting physical activities such as sitting, walking, and running [1]. However, the study of activity recognition using an accelerometer-embedded smartphone is still very limited, and there are still many difficulties that have greatly prevented it from mass adoption thus far. Therefore, it is very important to pay more attention to the research in activity recognition based on accelerometer-embedded smartphones.

It is known that robust features and fusion methods play a very crucial role in determining the accuracy of activity recognition and the flexibility and practicality of an application. Some researchers have focused on feature extraction approaches and fusion methods and have applied them into the large-scale applications of activity recognition. Yan et al. [2] proposed a classification algorithm called A3R based on time- and frequency-domain features. They saved approximately 50% of the energy of a continuous sensing engine running at full power, and fused features on the basis of specific activities and the adapted-needed sampling rate. They achieved a recognition rate of about 75%–92% for some activities such as normal walking and descending stairs. Chung et al. [3] applied a hierarchical classification method which could be regarded as a classifier fusion method, to recognize the three basic activities of standing, walking, and running with an overall accuracy of 82.8%. Ward et al. [4] proposed three fusion methods using class ranking after classification, which were called highest rank, Borda count, and logistic regression. The greatest advantage in [4] was considering the number of calculations when fusing.

The fusion idea has been studied in some of recent papers. In general, the widely used fusion method is to simply concatenate the different type of features together [2], [3]. Some researchers also concentrated on the classifier based fusion method for activity recognition and have proposed different fusion strategies [4]. However, we proposed two fusion methods based on time- and frequency- domain features, and we also discussed the performance of two fusion applications in activity recognition. The contributions of this paper include the following: 1) a new feature based on FFT. The feature make use of the mean of the FFT coefficients and the difference between the maximum and minimum of acceleration signals using logarithmic and averaging operations, and 2) two new fusion methods, minimal distance optimization (MDO) and variance contribution ranking (VCR), for improving the classification performance of five activities. The MDO fusion searches the fusion rankings of each class for a specific classifier. The VCR fuses features through the extraction of base vectors and the clustering of center vectors. The fusion weights obtained through the modified variance contribution strengthen the discrimination of different features.

2. Recognition Approach

2.1 Data Collection

Two smartphones manufactured by HTC and Samsung with Android OS were used as a platform for data collection. The subjects were 87 volunteers (44 male and 43 female) with a mean age of 22 from a local university. Data were collected in a less-noisy, broad, and flat place outside of the laboratory, which means that the environment was less-controlled and user-annotated. The subjects placed the smartphone on their body, alternating between the waist belt, shirt pocket, and trousers pocket respectively, for data collection. Each of the subjects recorded their own tri-axial acceleration signal

Manuscript received January 10, 2014.
Manuscript revised April 9, 2014.

[†]The authors are with South China University of Technology, China.
a) E-mail: yxue@scut.edu.cn

DOI: 10.1587/transinf.E97.D.2182

Table 1 Accuracy for different features.

	Mean	Var.	Diff.	PSD	FFT	RCEP	SEF
SF	62.8%	49.6%	54.3%	92.2%	91.5%	74.4%	91.5%
SFR	91.5%	90.7%	91.5%	89.9%	89.9%	90.7%	91.5%

data with an approximate time duration of approximately 90s for each of the following activities: jumping, walking, running, ascending stairs, and descending stairs.

2.2 Feature Extraction

The recognition rate of activities is different with different features. Therefore, we select some efficient features that have been demonstrated to be successful in previous works [5]–[11]. The features include the mean [6]–[9], the variance (Var.) [6]–[9], the difference between the maximum and the minimum (Diff.) [10], the power spectral density (PSD) [3], and the FFT coefficients [2], [5]–[7]. In addition, the real cepstrum (RCEP) of a signal is also included. The RCEP is defined as the inverse Fourier transform of the real logarithm of the magnitude of the Fourier transform of a signal, which is very useful for speech signal processing [11]. These six features are referred to as the set of efficient features (SEF).

We perform experiments to compare the performance of six features and the SEF in our dataset. In our experiments, we recognize the five activities using a Bayesian network (BayesNet) [12] classifier for a comparison with the experiments for the fusion methods in Sect. 3. Table 1 lists the average classification accuracy of the five activities for different features. The "SF" row specifies the recognition rate based on the specific feature solely, while the "SFR" row specifies the recognition rate with any specific feature removed.

From Table 1, we find that the recognition rate based solely on the FFT or PSD feature is high. The decrease of recognition rate is more noticeable with FFT or PSD feature removed. This means that FFT feature and PSD feature are more useful to classify five activities. The FFT feature describes a variation of the signal strength, while the PSD feature describes the variation of the signal magnitude versus frequency [3]. Although the classification accuracies using the Mean, Var., and Diff. features are not high enough, these three features still remain as equilibrium features in the SEF. From Table 1, we also find that the recognition rate decreases 0.7% when the PSD feature combines with the other features. In view of the higher discriminative ability of the FFT feature, a new feature based on an FFT (*NewF*) is proposed and defined as

$$NewF = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\log^2_{Diff}(mean(FFTcoefficients_n(1:k)))} \quad (1)$$

where N is the number of sliding windows, and $N = [\frac{K-(L-1)}{L*overlap}+1]$, $overlap = 50\%$, K is the length of raw acceleration signal, and $L = 512$ is window length. The partial expression $mean(FFTcoefficients_n(1:k))$ of the *NewF* means the average of the first k FFT coefficients of the acceleration signal of the nth sliding window per axis, while *Diff* is the difference between the maximum and the minimum values of the acceleration signal within each sliding window. In statistics, *Diff* suggest how diversely spread out the data values. By computing *Diff*, we can get an estimate of the spread of the signal. The $mean(FFTcoefficients_n(1:k))$ in Eq. (1) is designed to capture features in frequency-domain, and the *Diff* is time-domain feature. Further, for better combination of two different domain features, logarithmic and averaging operations are used. All of the features are extracted from the raw tri-axial acceleration signal using a sliding window size of 512 samples with 50% overlap.

Table 2 Accuracy for *NewF* and SEF+*NewF*.

	NewF	SEF+*NewF*
SF	92.41%	92.87%
SFR	91.5%	92.87%

Table 3 Confusion matrix using the *NewF* feature.

Class	Walk	Run	Jump	Upstairs	Downstairs
Walk	80	0	0	4	3
Run	1	80	2	3	1
Jump	0	1	82	3	1
Upstairs	1	3	2	81	0
Downstairs	2	0	3	3	79

To validate the effectiveness of the proposed *NewF* feature, experiments are conducted to compare the performance of *NewF* and SEF+*NewF*. A BayesNet is used as a classifier. Table 2 lists the average accuracy for the five activities.

The average accuracy based solely on the *NewF* is 92.41%, which is better than using the PSD feature, the FFT feature, and the SEF (listed in Table 1). From Table 2, the recognition rate based on SEF+*NewF* is 92.87%, increasing by 1.37% compared with the results based on the SEF, while there is a noticeable decrease in the recognition rate with *NewF* feature removed. These results demonstrate that the *NewF* feature has better classification performance. To investigate this further, the aggregate confusion matrix is summarized in Table 3, which also shows that the proposed feature has better performance.

2.3 Fusion Methods

Fusion is a vital step for eliminating false clustering due to the excessive sensitivity of the acceleration signals and for further improvement of the state-of-the-art recognition rates. We propose two fusion methods as follows:

1) Minimal Distance Optimization (MDO): We first assign the rankings a linear order, with "1" being the highest and the lowest equaling the number of classes [4]. Second, we

choose a classifier, which may result in great classification performance according to previous works [5], [13], to train the feature set fused with the rankings. Through a simple traversal of the different rankings, we obtain the optimized fusion ranking suitable for the specific classifier. Further, the rankings of each class assigned by the different classifiers are near the center ranking. Thus, the problem of determining the center ranking of each class is described as a minimal distance optimization problem:

$$\arg\min \sum_{i=1}^{N} |r - r_i| \tag{2}$$

where r_i is the ranking of each class assigned by the ith classifier, and r is the center ranking of each class. We take the integer part of r as the final ranking for fusion.

2) Variance Contribution Ranking (VCR): The variable contribution rate refers to the contribution of one factor in proportion to the total contribution, and the variance contribution describes its fluctuations. Based on VCR, we propose two fusion methods: feature-fusion-based VCR (FFVCR) and class-fusion-based VCR (CFVCR).

FFVCR: We choose one basic feature set $F = [F_x, F_y, F_z]$ from seven features, which include the SEF and the new proposed feature *NewF*. Then, this basic feature set is expressed as the combination of vectors $X = [X_1, X_2, \ldots, X_p]$,which are a set of base vectors extracted from the feature space matrix $[F_x; F_y; F_z]$:

$$\begin{cases} F_x = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p + \varepsilon_1 \\ F_y = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p + \varepsilon_2 \\ F_z = a_{31}X_1 + a_{32}X_2 + \cdots + a_{3p}X_p + \varepsilon_3 \end{cases} \tag{3}$$

where ε_i is for error balance, and a_{ij} is a coefficient with $i = 1, 2, 3$, $j = 1, 2, \ldots, p$.

Therefore, the problem of solving the base vectors X and the coefficients matrix A is equivalent to minimize the cost function $J(A, X)$.

$$J(A, X) = \|AX - F\|_2^2 + \lambda\|X\|_1 + \gamma\|A\|_2^2 \tag{4}$$

where λ and γ are control parameters.

Thus, the variance contribution is defined as

$$VC_i^2 = \sum_{j=1}^{p} (a_{ij} - \mu_i)(a_{ij} - \mu_i)^T, \quad i = 1,\ 2,\ 3 \tag{5}$$

where $\mu_i = \dfrac{1}{p}\sum\limits_{j=1}^{p} a_{ij}$.

A large variance contribution indicates that the features of different activities may be superposed. This means that classification is more difficult. Thus, the variance contribution is modified by

$$NewVC_i^2 = \frac{VC_{\max}^2 - VC_{\min}^2}{VC_i^2 - VC_{\min}^2} \tag{6}$$

Through amplitude compression for the tri-axial feature fusion weights, the efficient fusion features (*EFF*) of one basic feature set are consequently computed as follows:

$$EFF = [F_x, F_y, F_z]\frac{[NewVC_1^2, NewVC_2^2, NewVC_3^2]^T}{\sum\limits_{j=1}^{3} NewVC_i^2} \tag{7}$$

CFVCR: We divide the entire sample set into T testing sets $\{S_1, S_2, \cdots S_T\}$. Thus, each testing set has m clustering centers $\{CC_1, CC_2, \cdots, CC_m\}$, where CC_i $(i = 1, 2, \ldots m)$ denotes the feature vector of the ith activity after clustering, and m is the number of activities. We regard $\{CC_1, CC_2, \cdots, CC_m\}$ as a set of basis vectors. Thus, the feature of each person in each testing set is expressed as the combination of this set of basis vectors:

$$\begin{cases} F_1 = a_{11}CC_1 + a_{12}CC_2 + \cdots a_{1m}CC_m + \varepsilon_1 \\ F_2 = a_{21}CC_1 + a_{22}CC_2 + \cdots a_{2m}CC_m + \varepsilon_2 \\ \vdots \\ F_q = a_{q1}CC_1 + a_{q2}CC_2 + \cdots a_{qm}CC_m + \varepsilon_q \end{cases} \tag{8}$$

where F_i, $i = 1, 2, \ldots, q$ denotes the activity feature set of each subject in the feature subset, and q denotes the number of sampling subjects in the feature subset.

Taking into account the effects of the clustering center on the recognition rate, the variance contribution is defined as

$$\hat{VC}_j^2 = \sum_{i=1}^{q}(a_{ij} - \frac{1}{q}\sum_{l=1}^{q}\mu_l)(a_{ij} - \frac{1}{q}\sum_{l=1}^{q}\mu_l)^T, \tag{9}$$
$$j = 1, 2, \ldots, m$$

where $\mu_l = \dfrac{1}{m}\sum\limits_{j=1}^{m} a_{ij}$.

For each testing set, the final feature vector (*FFV*) is calculated by

$$FFV_j = CC_j \frac{VC_{\max}^2 - VC_{\min}^2}{VC_j^2 - VC_{\min}^2}$$
$$FFV = [FFV_1; FFV_2; \ldots FFV_m] \tag{10}$$

3. Experiments and Analysis

3.1 Experimental Setup

First, we preprocessed the 3D acceleration signals. Window filtering and normalization were used to remove noise and weaken the influence of physical factors among the different subjects, respectively. Second, the feature matrix was extracted and input into the classifier according to the new proposed feature and fusion methods in Sect. 2. Finally, the proposed methods were tested using five-fold-cross-validation method. For one round of cross-validation, we chose 70 subjects for each activity class, resulting in 350 samples for training the classifiers, and the remaining 17 subjects (85 samples) were used for testing each class. Five

Table 4 Classification accuracy for the different classifiers.

Class	J48	BayesNet	MLP
walking	91.95%	91.95%	89.65%
running	89.65%	91.95%	97.70%
jumping	94.25%	96.55%	91.95%
going upstairs	90.80%	94.25%	89.65%
going downstairs	94.25%	89.65%	89.65%
average	92.18%	**92.87%**	91.72%

Table 5 Accuracy for the different smartphone locations.

Locations	No fusion	MDO	FFVCR	CFVCR
Waist	92.87%	94.48%	95.17%	97.01%
Shirt pocket	92.18%	93.56%	95.86%	95.86%
Trousers pocket	92.41%	94.02%	96.55%	95.17%

rounds of cross-validation were performed using different partitions, and the validation results were averaged over the five rounds.

3.2 Experimental Results

To validate the classification performance of the two proposed fusion methods, we performed experiments to select a better classifier. We compared the performance of a J48 decision tree, multilayer perceptron (MLP), and BayesNet classifier using the mixed set of features, which contained the *NewF* and the SEF. Table 4 summarizes a performance comparison of the different classifiers for each of the five activities.

We find that the BayesNet classifier has the best recognition rate (92.87%) among the three classifiers. Thus, BayesNet is used as a classifier for evaluating the performance of the different fusion methods for the different smartphone locations, which are listed in Table 5.

It can be seen that our fusion method can significantly improve the recognition rates. The accuracy based on the two fusion methods, MDO and VCR, outperforms that without fusion. Furthermore, the accuracy based on the VCR fusion method outperforms that using MDO. From Table 5, it is noticeable that the recognition rate is relatively high for the smartphone located at the waist, while the recognition rate fluctuates for the smartphone placed in the trousers pocket. Intuitively, the smartphone placed in the trousers pocket should be the most powerful, since the majority of activities involve heavy use of the legs. However, as the smartphone is not fixed to the body, it may move randomly in the pocket (e.g. rotate), thereby producing more variations during the data collection process.

Table 6 shows the recognition results based on different fusion methods for the smartphone located at the waist.

It can be seen that, for each of the five activities, the accuracy using CFVCR fusion method is much higher than using MDO and FFVCR. Furthermore, the CFVCR fusion method produces the highest recognition rate of 97.01%, an increase of 4.14% compared with that without fusion. From

Table 6 Accuracy based on different fusion methods.

Class	MDO	FFVCR	CFVCR
walking	94.25%	91.95%	97.70%
running	91.95%	97.70%	100%
jumping	97.70%	94.25%	95.40%
going upstairs	96.55%	94.25%	97.70%
going downstairs	91.95%	97.70%	94.25%
average	94.48%	**95.17%**	**97.01%**

Tables 5 and 6, the CFVCR fusion method exhibits better performance because it strongly attracts the clustering center to samples within the class.

3.3 Analysis and Discussion

A BayesNet consists of a directed acylic graph and conditional probability table [12] and can train a best matching network to classify root nodes. The *NewF* feature and fusion feature vectors are both nodes which are strong dependent on each other. This strong dependence relies on the quality and character of the selected or fused feature with high recognition accuracy. Furthermore, the enhanced probability distribution also strengthens the causal relationship among learning variables. A BayesNet learns by causal edges. According to our analysis, the BayesNet performs the best.

In our experiments, the BayesNet classifier has the best recognition performance listed in the Table 4. But we cannot ensure that the best results are still obtained from a BayesNet when the the number of classes is larger or the feature changes. Because the scoring function in a BayesNet is the minimal description length (MDL) [13]. A larger class will result in a larger error for the MDL. Further, the MDL may score and automatically remove vital attribute variables for classification. Thus, Our goal is to simply validate the effectiveness of the new feature and fusion methods.

The MDO fusion searches the fusion rankings of each class for a specific classifier, which is a horizontal improvement among the classifiers with high accuracy. FFVCR and CFVCR fuse features through the extraction of base vectors and the clustering of center vectors, respectively, and considerably reduce the impact of a small amount of low-quality data through the modified variance contribution. Coupled with the BayesNet classifier, the fused feature training set indicates a stronger and more correct probability distribution network. On the other hand, the fusion weights obtained through the modified variance contribution strengthen the discrimination of different features, which inevitably results in perfect performance for the five activities recognition.

4. Conclusion

In this paper, a new feature based on an FFT (*NewF*) for activity recognition from tri-axial acceleration signals has been proposed. The average accuracy of recognizing five

activities using *NewF* was 92.41%, which was better than six traditional time- or frequency-domain features. When combined with the set of traditional efficient features, the *NewF* feature further improved the recognition accuracy. To improve the classification performance, two fusion methods (MDO and VCR) were proposed. The average accuracy based on the CFVCR fusion method was 95.17%, an increase of 0.69% compared to the MDO fusion method. The recognition results demonstrated that FFVCR achieved the best recognition performance. The average accuracy was 97.01%. The experimental results have confirmed the effectiveness of the proposed feature and fusion methods.

Acknowledgments

This work is supported in part by the NSFC (grant nos. 61201348, 61075021), the National Science and Technology Support Plan (2013BAH65F01-2013BAH65F04), the GDSTP (no. 2012A010701001), and Guangdong Natural Science Foundation (grant no. S2012040008016). We sincerely thank the helpful comments and suggestions given by the anonymous reviewers.

References

[1] L. Sun, D. Zhang, B. Li, and S. Li, Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations, pp.548–562, Springer-Verlag Berlin Heidelberg, 2010.

[2] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, and K. Aberer, "Energy-efficient continuous activity recognition on mobile phones: an activity-adaptive approach," IEEE Symp. Wear. Computers, pp.1550–4816, 2012.

[3] W.Y. Chung, A. Purwar, and A. Sharma, "Frequency domain approach for activity classification using accelerometer," 30th Ann. Int. Conf. of IEEE EMBS, pp.1120–1123, Aug. 2008.

[4] J.A. Ward, P. Lukowicz, G. Troster, and T.E. Starner, "Activity recognition of assembly tasks using body-worn microphones and accelerometers," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, pp.1553–1567, 2006.

[5] Y. Xue and L. Jin, "A naturalistic 3D acceleration-based activity dataset & benchmark evaluations," IEEE Int. Conf. System Man and Cybernetics, pp.4081–4085, 2010.

[6] S. Liu, R. Gao, D. John, J. Standenmayer, and P. Freedson, "Multisensor data fusion for physical activity assessment," IEEE Trans. Biomed. Eng., vol.59, no.3, pp.687–696, 2012.

[7] Y. Xue and L. Jin, "Discrimination between upstairs and downstairs based on accelerometer," IEICE Trans. Inf. & Syst., vol.E94-D, no.6, pp.1173–1177, June 2011.

[8] C. Zhu and W. Sheng, "Realtime recognition of complex human daily activities using human motion and location data," IEEE Trans. Biomed. Eng., vol.59, pp.2422–2430, 2012.

[9] M.A. Ayu, T. Mantoro, A.F.A. Matin, and S.S.O. Basamh, "Recognizing user activity based on accelerometer data from a mobile phone," IEEE Symposium on Computers & Informatics, pp.617–621, 2011.

[10] B. Das, B.L. Thomas, A.M. Seelye, D.J. Cook, L.B. Holder, and M.S. Edgecombe, "Context-aware prompting from your smart phone," IEEE Consumer Communication and Networking Conference, pp.56–57, 2012.

[11] A.V. Oppenheim and R.W. Schafer, Discrete-Time Signal Processing, Third ed., Publishing House of Electronics Industry, 2011.

[12] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifier," Mach. Learn., vol.29, pp.131–163, 1997.

[13] T. Inomata, F. Naya, N. Kuwahara, F. Hattori, and K. Kogure, "Activity recognition from interactions with objects using dynamic bayesian network," CASEMANS 2009, pp.39–42, 2009.