MULTI-SENSORY FEATURES FOR PERSONNEL DETECTION AT BORDER CROSSINGS

Po-Sen Huang[†]*, Thyagaraju Damarla*[‡]*, Mark Hasegawa-Johnson*[†]

[†]Beckman Institute, ECE Department, University of Illinois at Urbana-Champaign, U.S.A.
[‡]US Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD 20783, U.S.A.
huang146@illinois.edu, rdamarla@arl.army.mil, jhasegaw@illinois.edu

ABSTRACT

Personnel detection at border crossings has become an important issue recently. To reduce the number of false alarms, it is important to discriminate between humans and four-legged animals. This paper proposes using enhanced summary autocorrelation patterns for feature extraction from seismic sensors, a multi-stage exemplar selection framework to learn acoustic classifier, and temporal patterns from ultrasonic sensors. We compare the results using decision fusion with Gaussian Mixture Model classifiers and feature fusion with Support Vector Machines. From experimental results, we show that our proposed methods improve the robustness of the system.

Keywords: Gaussian Mixture Models, Support Vector Machines, sensor fusion, footstep detection, personnel detection

1. INTRODUCTION

Personnel detection is an important task for Intelligence, Surveillance, and Reconnaissance (ISR) [1, 2]. One might like to detect intruders in a certain area during the day and night so that the proper authorities can be alerted. For example, border crimes including human trafficking would be reduced by automatic detection of illegal aliens crossing the border. There are numerous other applications where personnel detection is important.

However, personnel detection is a challenging problem. Video sensors consume high amounts of power and require a large volume for storage. Hence, it is preferable to use non-imaging sensors, since they tend to use low amounts of power and are long-lasting. Non-imaging sensors, however, suffer from ambiguity among the footsteps of animals alone, humans alone, and of animals traveling together with humans.

Traditionally, personnel detection research concentrated on using seismic sensors. When a person walks, his/her impact on the ground causes seismic vibrations, which are captured by the seismic sensors. Previous works have relied on fundamental gait frequency estimation [3, 4]. Park et al. proposed the method of extracting temporal gait patterns to provide information on temporal distribution of gait beats [5].

At border crossings, animals such as mules, horses, or donkeys are often known to carry loads. Animal hoof sounds make them distinct from human footstep sounds. When humans and four-legged animals walk together, the sounds they make are perceptually distinguishable by human listeners. Automatic algorithms that imitate human capabilities in other acoustic event detection tasks have been constructed [6, 7, 8], e.g., using perceptual linear predictions (PLP) features coupled to tandem neural net - HMM recognizers.

Passive and active ultrasonic methods were proposed for the detection of walking personnel for ultrasound signals [9]. The passive method utilizes the footsteps' ultrasonic signals generated by friction forces, while the active method uses the human Doppler ultrasonic signature. In an outdoor scene, the passive ultrasound signals are limited in distance and are noisy. For the active ultrasound method, when a person walks, each limb is a compound pendulum and has distinct oscillatory characteristics, which in turn results in a micro Doppler effect. Similarly, the torso also oscillates at a particular frequency. The ultrasonic sensors can detect the ultrasonic signature generated by footsteps and movements of the torso. Zhang et al. reported that micro-Doppler gait signatures differ between human and four-legged animals [10]. These arise from the different physical mechanisms found in the different species. Kalgaonkar et al. analyzed spectral patterns to classify human walking (walker identification, approach vs. withdraw, male vs. female) [11].

As shown in the above literature review, existing research only uses a single sensor recorded in clean environments with a single object (a person or a four-legged animal) walking. However, in reality, when there are many objects such as people or four-legged animals walking or running in noisy environments, it is difficult to distinguish human alone vs. animals alone vs. animals and humans together using a single sensor and published approaches.

In this paper, we propose using enhanced summary autocorrelation patterns for feature extraction from seismic sensors, a multi-stage exemplar selection framework to learn acoustic classifier, and temporal patterns from ultrasonic sensors. Acoustic, seismic, and ultrasound signals are fused using decision fusion based on Gaussian Mixture Models (GMMs) and feature fusion based on Support Vector Ma-

978-0-9824438-2-8/11/$26.00 ©2011 IEEE

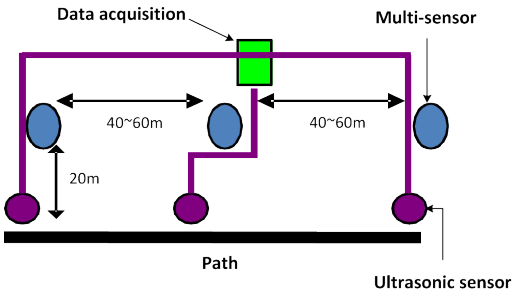

Fig. 1. Sensor layout, where a multi-sensor multi-modal system has acoustic, seismic, passive infra-red (PIR), radar, magnetic, and electric field sensors.

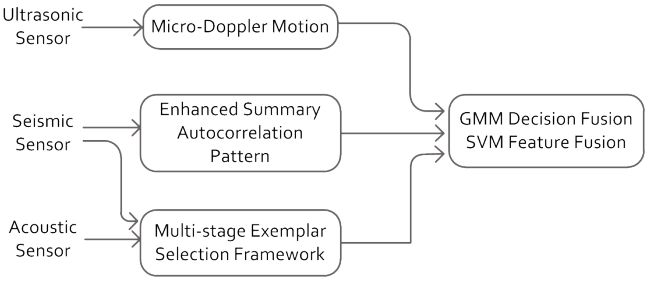

Fig. 2. The overall flow: feature extraction based on phenomenology, GMM and SVM classifiers, and decision and feature fusion.

chines in order to examine the robustness of our methods.

The organization of this paper is as follows: Section 2 introduces the multi-sensor multi-modality data and events. Section 3 discusses the feature extraction from seismic, acoustic, and ultrasonic sensors. Section 4 discusses Gaussian mixture model classifiers, decision fusion, and Support Vector Machines. Section 5 describes the experiments on the multi-sensor multi-modal dataset. We conclude this paper with future work in Section 6.

2. DATA

In this paper, we use a multi-sensor multi-modal realistic dataset collected in Arizona by the U.S. Army Research Lab and the University of Mississippi. The data are collected in a realistic environment in an open field. There are three selected vantage points in the area. These three points are known to be used by the illegal aliens crossing the border. These places where the data are collected include: (a) wash (a flash flood river bed with fine-grain sand), (b) trail (a path through the shrubs and bushes), and (c) choke point (a valley between two hills.) The data are recorded using several sensor modalities, namely, acoustic, seismic, passive infrared (PIR), magnetic, E-field, passive ultrasonic, sonar, and both infrared and visible video sensors. Each sensor suite is placed along the path with a spacing of 40 to 60 meters apart. The detailed layout of the sensors is shown in Figure 1. Test subjects walked or ran along the path and returned back along the same path.

A total of 26 scenarios with various combinations of people, animals and payload are enacted. We can categorize them as: *single person (11.6%), two people (13%), three people (21.7%), one person with one animal (14.5%), two people with two animals (15.9%), three people with three animals (17.4%), and seven people with a dog (5.9%)*, where the animals can be a mule, a donkey, a horse, or a dog, and the number in the parentheses represents the percentage of the data. The data are collected over a period of four days; each day at a different site and different environment. There is variable wind in the recording environment.

2.1. Active Sensing

The time duration for subjects passing by is short (about ten to twenty seconds at a time) compared to the whole recording time (five to six minutes recording). Without any ground truth segmentation, we would like to extract the time duration when test subjects are passing through. This problem can be formulated as an example of active sensing and learning [12, 13], which refers to sequential data selection and inference procedures that actively seek out highly informative data, rather than relying on non-adaptive data acquisition solely.

For acoustic sensors, in an outdoor scene, the signals are contaminated by wind sounds, human voices, or unexpected airplane engine sounds. Seismic and PIR sensors, on the other hand, are relatively clean. Hence, we can process seismic or PIR sensors by an energy detection to determine the time duration when test subjects pass by. If the energy in any ten-second interval exceeds a threshold, the interval is marked "**active**." Seismic and acoustic signals are pre-synchronized; therefore the acoustic active integral can be marked on the basis of seismic energy. Ultrasound is not tightly synchronized; therefore it must be independently segmented. For each recording, there are two active segments (walked or ran along the path and returned back along the same path). In this paper, we emphasize the classification of segmented multimodal recordings into two classes: **humans only**, and **humans with (four-legged) animals**.

3. FEATURES EXTRACTION

Features are extracted from seismic, acoustic, and ultrasonic sensors. The overall flow is shown in Figure 2.

3.1. Seismic

Seismic sensors capture the vibrations in the ground caused by the motion of the targets or ground coupling of acoustic waves. The gait patterns of humans and four-legged animals differ. Previous approaches do not consider the case for multiple human and/or four-legged animals [3, 5]. When there

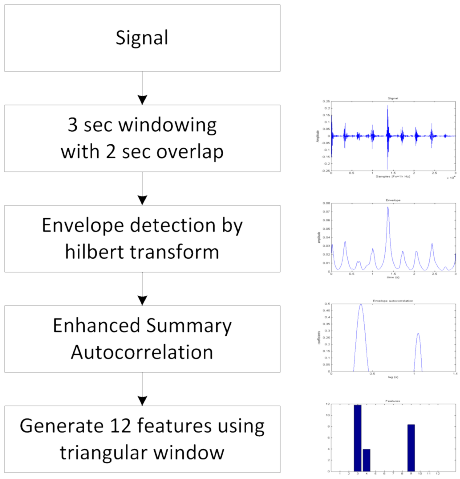

Fig. 3. Seismic feature extraction algorithm.

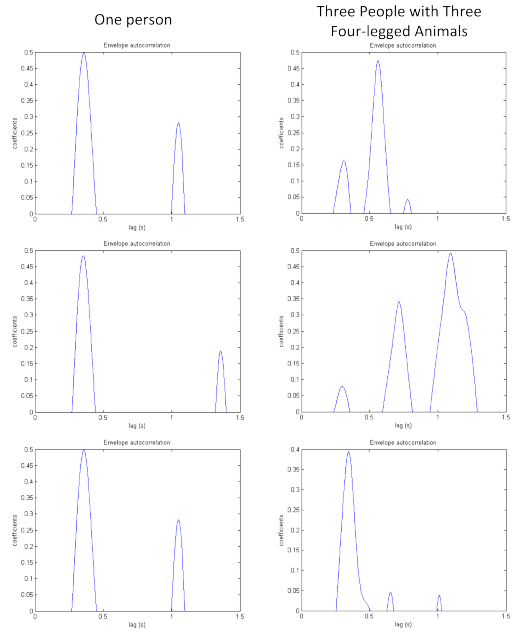

Fig. 4. Examples of enhanced summary autocorrelation of seismic signals. The left column shows examples of the feature vector for one person, and the right column is by three people with three four-legged animals at three different time frames.

are multiple human and/or four-legged animals, it is not reliable to estimate the gait period based on the single pitch (fundamental frequency) detection method [14, 15]. Inspired by Park's temporal gait pattern approach [5] and the progress in multipitch analysis [16], we propose a gait pattern feature extraction method based on **enhanced summary autocorrelation** [16], as shown in Figure 3. A typical example of enhanced summary autocorrelation function is shown in Figure 4, where the same subjects generate similar enhanced summary autocorrelation patterns. We form analytic signals by Hilbert transform and then use full wave rectification followed by low-pass filtering and down-sampling for envelope detection. Finally, we use enhanced summary autocorrelation to estimate the gait pattern and generate a 12-dimensional feature vector using 12 triangular windows.

The idea of enhanced summary autocorrelation is to prune the periodicity of the autocorrelation function. The procedure is the following: First, from the envelope signals, the autocorrelation function is computed within each channel (2 channels in the model of Tolonen and Karjalainen [16]). Second, the autocorrelation functions are summed up across the channels to form a summary autocorrelation function. Third, the summary autocorrelation function is clipped to positive values, then time-scaled by a factor of two, and subtracted from the original clipped function. Then, the same procedure is repeated with other integer factors so that repetitive peaks at integer multiples can be removed. The resulting function is called the enhanced summary autocorrelation.

3.2. Acoustic

In acoustic signals, the hoof sounds of animals such as horses, donkeys, or mules are perceptually distinct from human footstep sounds. In order to imitate the perceptual discrimination abilities of human listeners, we begin by using Perceptual Linear Predictive (PLP) features [17], which are common features in speech recognition. As mentioned in Section 2, the data are recorded in an open field. There are noisy wind sounds in the recordings. We use spectral subtraction to reduce the effect of noise [18, 19].

From the active segments we extracted in Section 2.1, we further extract acoustic features from short-time footstep sounds by incorporating seismic signals. Since there are no labels for the exact time of footstep sounds, we have to use the seismic sensor information, assuming that the peaks in the seismic signals correspond to footsteps. Suppose there are n groups of peaks (if some peaks are close to each other, we count them as one group) in the seismic signal, whose times are t_i, for $i = 1, \ldots, n$. We choose a small time δ around the peaks and extract PLP features within the time duration $(t_i - \delta, t_i + \delta)$, for $i = 1, \ldots, n$, as shown in Figure 5. In each time period, we extract 13 PLP features using 186ms Hamming windows with 75% overlap, where 186ms is approximately equal to the time duration of a single footstep (from heel strike to toe slap). Delta and delta-delta coefficients are appended to create a 39-dimensional feature vector.

Our goal is to classify *humans only* vs. *humans with animals*. In the *humans with animals* class, there are instances of human footstep sounds. Therefore, there are some overlap

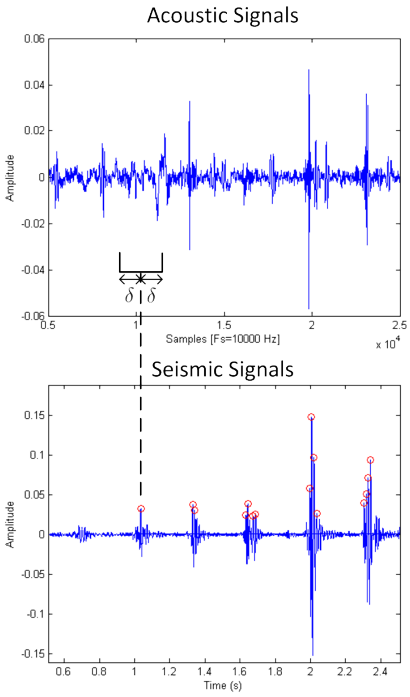

Fig. 5. Using peaks of seismic signals for matching acoustic footstep sounds.

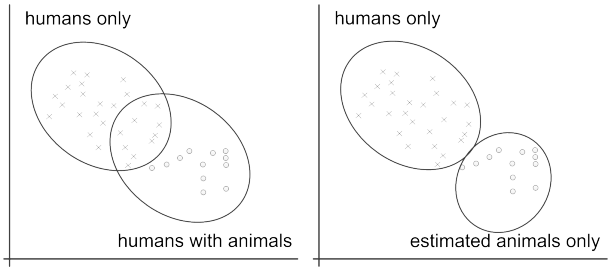

Fig. 6. Left: an example of feature space of *humans only* and *humans with animals* class. Right: an example of feature space of *humans only* and *estimated animals only* class, after exemplar selection.

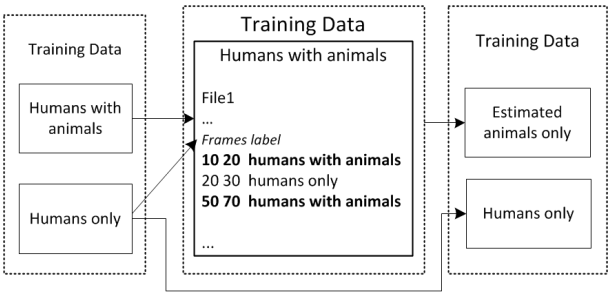

Fig. 7. Multi-stage framework for acoustic exemplar selection.

between the two classes in the feature space, as shown on the left hand side of Figure 6. Regularized discriminative methods such as support vector machines (SVM) explicitly trade off the degree of class overlap vs. the complexity of the decision boundary in order to minimize an estimate of expected risk. Generative models, on the other hand, model overlap only to the extent permitted by the specified generative model.

In order to improve the classifiers' ability to compensate for class overlap, therefore, we propose a multi-stage algorithm for exemplar selection, as shown in Figure 7; this framework is similar to the "self-training" methods used in semi-supervised learning. The idea of the framework is to select the exemplar frames in the *humans with animals* class which are dissimilar to the features in the *humans only* class. With the exemplar selection method, classifiers are easier to learn the distinctive features between classes as shown on the right hand side of Figure 6. The algorithm is as follows:

1. Train an exemplar selection classifier (SVM or GMM) for *humans only* and *humans with animals* using training data as shown in the left block of Figure 7.

2. Label the training data of the *humans with animals* class using the trained models as shown in the middle block of Figure 7. Each frame in the training data is labeled as either the *humans only* class or the *humans with animals* class.

3. Keep the frames which were labeled as *humans with animals*; in other words, discard the frames which were labeled as *humans only*.

4. Train a new classifier (SVM or GMM) between the *estimated animals only* class and the *humans only* class as shown in the right block of Figure 7.

Note that the acoustic features capture short-time footstep sounds as features, while seismic and ultrasonic features utilize temporal pattern information. Therefore, the multi-stage exemplar selection framework applies for acoustic features only.

3.3. Ultrasound

Ultrasonic sensors, also known as acoustic Doppler sensors [9], emit acoustic waves toward objects and receive reflected responses from objects. Benefits of using ultrasonic sensors include low cost ($5 USD in 2011) and low power. The limitation is that, because of the rapid attenuation of high-frequency acoustic waves, ultrasonic sensors have a limited range on the order of ten meters.

By measuring the frequency shift of a wave scattered or radiated by a moving object, the velocity of the object relative to an observer can be calculated; this is known as the Doppler

effect. If the object contains moving parts, each moving part will result in a modulation of the base Doppler frequency shift, which is known as the micro-Doppler effect. Given an acoustic wave transmitted by an observer, the frequency of the received wave by a single point scatterer is

$$f = f_0 \left(1 + \frac{2v}{c}\right) \tag{1}$$

where f_0 is the frequency of the transmitted acoustic wave, v is the velocity of the scattered wave relative to the observer and c is the speed of sound. The Doppler frequency shift, $\Delta f = \frac{2v}{c}$, is proportional to the scattered wave velocity relative to the observer.

A human body is an articulated object, comprising a number of rigid bones connected by joints. When a continuous tone is incident on an animal or a walking person, the reflected signal contains a spectrum of frequencies by the Doppler shifts of the carrier tone because of the velocities of various moving body parts.

As reported in Zhang et al. [10], based on different physical walking mechanisms, the micro-Doppler gait signatures between a person and a four-legged animal are different. We use this concept to extract features in order to distinguish between humans and four-legged animals.

For ultrasound signal processing, given the data with two channels, 25 kHz and 40 kHz, we first use a band-pass filter with stopband at 20 kHz and 30 kHz and passband at 22.5 kHz and 27.5 kHz for 25 kHz channel, and a band-pass filter stopband at 30 kHz and 45 kHz and passband at 37.5 kHz and 42.5 kHz for 40 kHz channel. Then, we use Hilbert transform demodulating the captured Doppler signals to emphasize the contributions of various velocities. Finally, we use cepstral coefficients for representing the patterns in the spectrogram [11]. We use 62ms Hamming window with 75% overlap. The 80-dimensional feature vector includes as cepstral coefficients and their deltas.

4. METHODS

4.1. Gaussian Mixture Model Classifiers

The motivation for using Gaussian mixture densities is that a sufficiently large linear combination of Gaussian basis functions is capable of representing any differentiable sample distribution [20, 21].

A Gaussian mixture density is a weighted sum of M component densities, as shown in the following equation,

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{2}$$

where $\vec{x}$ is a D-dimension random vector, $b_i(\vec{x})$, $i = 1, \ldots, M$, are the component densities and p_i, $i = 1, \ldots, M$, are the mixture weights. Each component density is a D-variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i)\} \tag{3}$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i. The mixture weights are constrained by $\sum_{i=1}^{M} p_i = 1$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices (we use diagonal covariance matrices here) and mixture weights from all component densities. These parameters are collectively represented by the notation $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}, i = 1, \ldots, M$. For classification, each class is represented by a GMM parameterized by λ.

Given training data from each class, the goal of model training is to estimate the parameters of the GMM. Maximum likelihood model parameters are estimated using the Expectation-Maximization (EM) algorithm. Generally, ten iterations are sufficient for parameter convergence.

The objective is to find the class model that has the maximum *a posteriori* probability for a given observation sequence X. Assuming equal likelihood for all classes (i.e., $p(\lambda_k) = 1/N$) , the classification rule simplifies to

$$\hat{N} = \underset{1 \le k \le N}{\operatorname{argmax}}\, p(X|\lambda_k) = \underset{1 \le k \le N}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k) \tag{4}$$

where the second equation uses logarithms and the independence between observations. T is the number of observations.

4.2. Decision Fusion

GMMs are trained for each modality and their log probabilities are combined as

$$s_\lambda(\vec{x}) = \sum_{m \in M} w_m \log P(\vec{x}_m|\lambda) \tag{5}$$

where $M = \{a, s, u\}$, a, s, u represents acoustic, seismic, and ultrasound modalities, respectively. If all likelihood functions were correctly trained, and if the vectors $\vec{x}_a$, $\vec{x}_s$, and $\vec{x}_u$ were conditionally independent given class label, then the Bayes-optimal mode weights would be $w_m = 1$. In practive the likelihood functions tend to be overconfident; therefore, we scale them using $0 \le w_m \le 1$, $\sum_{m \in M} w_m = 1$.

For simplicity, we choose weights by a grid-search of global weights on validation sets [22]. Note that Equation (5) corresponds to a linear combination in the log-likelihood domain; however, it does not represent a probability distribution in general, and will be referred to as a score.

4.3. Support Vector Machines

A Support Vector Machine (SVM) estimates decision surfaces, $g(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$, directly [23], rather than modeling a probability distribution from the training data. Given

training feature vectors $\mathbf{x_i} \in R^n$, $i = 1, \ldots, k$ in two classes with label $y_i \in \{1, -1\}$, $i = 1, \ldots, k$, a SVM solves the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{k}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, \ldots, k \end{aligned}$$

where $\phi(\mathbf{x_i})$ maps $\mathbf{x_i}$ onto a higher dimensional space, $C \geq 0$ is the regularization parameter, and ξ_i is a slack variable, which measures the degree of misclassification of the datum $\mathbf{x_i}$.

The solution can be written as $\mathbf{w}$ satisfies $\mathbf{w} = \sum_{i=1}^{k} y_i\alpha_i\phi(\mathbf{x_i})$, where $0 \leq \alpha_i \leq C$, $i = 1, \ldots, k$, and the decision function is

$$h(x) = \text{sgn}\left(\sum_{i=1}^{k} y_i\alpha_i K(\mathbf{x_i}, \mathbf{x}) + b\right) \tag{6}$$

where $K(\mathbf{x_i}, \mathbf{x}) = \phi(\mathbf{x_i})^T\phi(\mathbf{x})$ is the kernel function. In this paper, we use LIBSVM with Radial Basis Function (RBF) kernels, that is, $K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma||\mathbf{x_i} - \mathbf{x_j}||^2)$ [24].

5. EXPERIMENTS

In this section, we describe three experiments in order to compare our proposed methods with previous approaches in classifying *humans only* vs. *humans with four-legged animals*. There are 69 recordings in the dataset. We divide the recordings into four groups and choose two for training and two for testing at a time, resulting in a six-fold cross-validation. In each fold, we randomly select a part of recordings from training and testing sets as a validation set. We choose the best mixture count for the GMM classifier and parameters γ and C for the SVM, according to the validation set. The experimental results are represented by mean $\pm$ standard error.

5.1. Seismic features

As describe in Section 3.1, we compare our gait pattern features based on enhanced summary autocorrelation with the temporal gait pattern [5] under the same experimental setup. The experimental results are shown in Table 1.

Feature	Accuracy (%)	
	GMM	SVM
Temporal gait pattern [5]	71.883±4.607	79.010±4.648
Enhanced summary autocorrelation pattern	81.707±2.564	84.446±2.868

Table 1. Classification accuracy using seismic features.

From the experimental results of Table 1, our proposed method using enhanced summary autocorrelation pattern outperforms the previous method [5] in both GMM and SVM classifiers, because the previous method did not consider the case of multiple objects. Compared with GMM classifiers [5], the experimental results show that SVM has a better discrimination between the two classes for seismic features.

5.2. Acoustic features

As described in Section 3.2, we want to examine the effect of using (1) spectral subtraction, (2) seismic peaks with different δ's, and (3) our proposed multi-stage exemplar selection framework using GMM and SVM classifiers as the first step of the algorithm. The experimental results are shown in Table 2.

The first row *PLP features without (1)(2)(3)(4)* in Table 2 represents using the active audio segments, without using the duration estimated by the peaks of seismic signals, and without using spectral subtraction. Spectral subtraction (row 2) improves the performance for both classifiers.

It is helpful to further extract audio features from the time durations marked by peaks of seismic signals. This method utilizes both the characteristics of acoustic and seismic sensor in the sensor suites. Without using this method, there are many silence or noise segments in the audio signals, and the silence or noise signals make both classifiers ill-trained.

Moreover, different values of δ capture different amounts of acoustic information. The results show that δ=0.3s has the best performance compared with δ=0.1s and δ=0.5s. The seismic sensor and acoustic sensor are not at exactly the same place and the rates of propagation are different. Therefore, there are asynchronies between acoustic and seismic signals. Specifically, with δ=0.1s, the acoustic segment does not contain the entire footstep sound. On the other hand, with δ=0.5s, the acoustic signals include too much unrelated noise. These reasons may explain the performance variation of both classifiers.

For our proposed multi-stage exemplar selection framework, using GMM for exemplar selection improves the accuracy around 1$\sim$2% for GMM classifiers; on the contrary, using GMM for exemplar selection degrades the accuracy for SVM classifiers. A possible reason is that SVM implicitly chooses support vectors for the hyperplane in the feature space. Using GMM selected features, the SVM has less information, and hence has worse performance. On the other hand, using SVM for exemplar selection degrades performance in all cases. A possible explanation is that the SVM cannot select proper exemplar in the case of overlapping feature space in the first stage.

Feature	Accuracy (%)	
	GMM	SVM
PLP features without (1)(2)(3)(4)	73.768±2.230	65.337±1.896
PLP features with (1)	76.105±4.098	71.698±4.572
PLP features with (1)(2), δ=0.1s	74,975±5.079	78.093±1.699
PLP features with (1)(2)(3), δ=0.1s	75.737±2.936	76.604±2.179
PLP features with (1)(2)(4), δ=0.1s	72.735±4.585	75.090±2.577
PLP features with (1)(2), δ=0.3s	77.555±4.268	80.578±3.113
PLP features with (1)(2)(3), δ=0.3s	79.015±3.799	72.638±2.727
PLP features with (1)(2)(4), δ=0.3s	75.325±3.739	77.196±1.706
PLP features with (1)(2), δ=0.5s	75.392±3.376	76.214±4.396
PLP features with (1)(2)(3), δ=0.5s	77.688±3.149	74.507±3.634
PLP features with (1)(2)(4), δ=0.5s	74.800±4.523	71.313±3.456

Table 2. Classification accuracy using acoustic features, where (1) represents spectral subtraction, (2) represents the use of seismic peaks with different δ second (s), and (3) represents the use of our proposed multi-stage exemplar selection framework using a GMM classifier as the first step of the algorithm. (4) represents the use of our proposed multi-stage exemplar selection framework using a SVM classifier as the first step of the algorithm.

5.3. Decision fusion and feature fusion with seismic, acoustic, and ultrasonic features

We perform multimodal fusion in a classifier-dependent fusion: decision fusion with GMMs, feature fusion (vector concatenation) with SVM. Note that, for ultrasonic data, within 186ms, there are eight moving windows resulting in a 640-dimensional feature vector. We use principal component analysis (PCA) keeping 99% of the energy, and reduce features to 7 dimensions.

We compare our proposed methods using GMM and SVM classifiers, as shown in Table 3. Row 1 of Table 3 represents the use of ultrasonic features, enhanced summary autocorrelation pattern, PLP features with spectral subtraction, seismic peaks with δ=0.3s, and the multi-stage exemplar selection framework using GMM classifiers; Row 2 of Table 3 represents the use of the same seismic, ultrasonic features as Row 1, and acoustic features without the multi-stage exemplar selection. Row 3 of Table 3 represents the use of temporal gait pattern [5], PLP features without spectral subtraction, using the whole active segments, and without the multi-stage exemplar selection. Row 4 of Table 3 represents the use of ultrasonic features.

In Table 3, our proposed method, using seismic and acoustic features along with ultrasonic features, greatly improves the robustness compared with previous approaches. With the exemplar selection framework, GMM classifiers achieve the best fusion accuracy. The SVM, however, performs worse with exemplar selection, as mentioned above. The classification task, using only ultrasonic features (last row), is roughly 7% better with SVM classifiers compared with GMM classifiers.

We analyze the errors in the *(1)(3)(5)* in the GMM decision fusion case. Among the six-fold cross-validations, the recordings of the event, *seven people with a dog*, are all incorrectly classified as *human only*. This accounts for 52.6% of all errors. A possible explanation is that, dogs have padded feet (instead of hoofs) and are relatively small. It is difficult to tell dogs from humans because the classifier has learned to recognize hoof sounds. The limited amount of data for this event means that the classifier is unable to learn its distinctive pattern.

Feature	Accuracy (%)	
	GMM	SVM
(1)(3)(5)	86.092±2.313	84.446±2.868
(1)(2)(5)	84.928±2.790	85.307±3.405
(4)(5)	81.903±3.144	81.041±1.754
(5)	75.528±3.564	82.188±3.466

Table 3. Classification accuracy using decision fusion (GMM classifier) and feature fusion (SVM classifier), where (1) represents the enhanced summary autocorrelation pattern, (2) represents PLP features with spectral subtraction and seismic peaks with δ=0.3s, (3) represents (2) with the multi-stage exemplar selection framework using a GMM classifier as the first step of the algorithm, (4) represents the use of temporal gait pattern [5], PLP features without spectral subtraction, using the whole active segments, and without the multi-stage exemplar selection, and (5) represents ultrasonic features.

6. CONCLUSION

In this paper, we use a challenging realistic multi-sensor multi-modal dataset for personnel detection. Based on phenomenology of the differences (gait pattern, footstep sound, and micro-Doppler motion) between humans and four-legged animals, we propose using a new seismic feature extraction method based on enhanced summary autocorrelation, a multi-stage acoustic exemplar selection framework, and temporal patterns from ultrasonic sensors. Experimental results show that the combination of multi-modal sensors improves the robustness of the system over previous approaches. Since it is inexpensive to deploy unattended ground sensors such as acoustic, seismic, and ultrasonic sensors in target areas; it is possible to further extend the current fusion system to create a tracking system based on sensor network fusion.

7. ACKNOWLEDGMENTS

The authors thank J.-T. Huang and Dr. X. Zhuang for helpful discussions. This research is supported by ARO MURI 2009-31.

8. REFERENCES

[1] T. Damarla, "Sensor fusion for ISR assets," M. A. Kolodny, Ed., vol. 7694. SPIE, 2010.

[2] T. Damarla, L. Kaplan, and A. Chan, "Human infrastructure & human activity detection," in *Information Fusion, 2007 10th International Conference on*, 9-12 2007, pp. 1 –8.

[3] J. M. Sabatier and A. E. Ekimov, "Range limitation for seismic footstep detection," E. M. Carapezza, Ed., vol. 6963. SPIE, 2008.

[4] K. M. Houston and D. P. McGaffigan, "Spectrum analysis techniques for personnel detection using seismic sensors," E. M. Carapezza, Ed., vol. 5090. SPIE, 2003, pp. 162–173.

[5] H. O. Park, A. A. Dibazar, and T. W. Berger, "Cadence analysis of temporal gait patterns for seismic discrimination between human and quadruped footsteps," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pp. 1749–1752, 2009.

[6] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using gaussian mixture models and gmm supervectors," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pp. 69–72, 2009.

[7] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543 – 1551, 2010.

[8] P.-S. Huang, X. Zhuang, and M. A. Hasegawa-Johnson, "Improving acoustic event detection using generalizable visual features and multi-modality modeling," in *Acoustics, Speech and Signal Processing. ICASSP 2011. IEEE International Conference on*, 2011.

[9] A. Ekimov and J. M. Sabatier, "Human detection range by active doppler and passive ultrasonic methods," E. M. Carapezza, Ed., vol. 6943. SPIE, 2008.

[10] Z. Zhang, P. Pouliquen, A. Waxman, and A. Andreou, "Acoustic micro-doppler gait signatures of humans and animals," in *Information Sciences and Systems, 2007. CISS '07. 41st Annual Conference on*, 14-16 2007, pp. 627 –630.

[11] K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recognition," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, 5-7 2007, pp. 27 –32.

[12] D. J. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 590–604.

[13] R. Castro, C. Kalish, R. Nowak, R. Qian, T. Rogers, and X. Zhu, "Human active learning," *NIPS*, 2008.

[14] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 25, no. 1, pp. 24 – 33, feb 1977.

[15] P. D. L. Cuadra and A. Master, "Efficient pitch detection techniques for interactive music," in *In Proceedings of the 2001 International Computer Music Conference, La Habana*, 2001.

[16] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 6, pp. 708 –716, Nov. 2000.

[17] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[18] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, vol. 4, Apr. 1979, pp. 208 – 211.

[19] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504 –512, Jul. 2001.

[20] L. R. Rabiner, B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using hidden markov models with continuous mixture densities." *AT Technical Journal*, vol. 64, no. 6 pt 1, pp. 1211–1234, 1985.

[21] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257 –286, Feb. 1989.

[22] Guillaume, G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. Int. Conf. Acous. Speech Sig. Process*, 2002, pp. 853–856.

[23] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, ser. COLT '92. New York, NY, USA: ACM, 1992, pp. 144–152. [Online]. Available: http://doi.acm.org/10.1145/130385.130401

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