Exemplar Selection Methods to Distinguish Human from Animal Footsteps

Po-Sen Huang[†], Mark Hasegawa-Johnson[†], Thyagaraju Damarla[‡]
[†]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.
jhasegaw@illinois.edu, huang146@illinois.edu, rdamarla@arl.army.mil

Abstract

The "class discovery" problem is the problem of learning a classifier from a mixture of unlabeled and labeled training data, under the constraint that labeled training data exist for only N-1 of the N target classes. The task of distinguishing human from animal footsteps can be framed as a class discovery problem. When humans travel alone, every footstep sound is caused by a human foot, therefore labeled training examples for the "human" class are abundant. When humans travel with animals, their footsteps are interspersed and/or overlapped in time; without a tedious labeling effort, there are no gold-standard labels specifying which species created each of the footstep events. This paper will describe three different types of class discovery algorithm: the mixed-vs-unmixed classifier, the generative class discovery algorithm, and the class of algorithms sometimes called "self training." Experiments using the ARL/Mississippi multisensory personnel tracking database will be reported. Experimental results suggest that the mixed-vs-unmixed classifier gives the best performance in distinguishing mixed vs. unmixed test tokens (recordings containing humans alone vs. humans with animals), but that the self-training method shows promise for the task of learning to distinguish between the discrete footfall sounds of humans and animals.

Index Terms: acoustic event detection, optical flow, hidden Markov models, multi-stream HMM, coupled hidden Markov models

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.

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. Automatic algorithms that imitate human capabilities in other acoustic event detection tasks have been constructed [3, 4], e.g., using perceptual linear predictions (PLP) features coupled to tandem neural net - HMM recognizers.

Existing research considers only the case when there is a single object (a person or a four-legged animal) walking using a single sensor in clean environments. However, when multiple objects such as humans with four-legged animals travel together, the task becomes more challenging. Without a tedious labeling effort,

there are no gold-standard labels specifying which species created each of the footstep events. The task of distinguishing human from animal footsteps can be framed as a class discovery problem, which is the problem of learning a classifier from a mixture of unlabeled and labeled training data, under the constraint that labeled training data exist for only N-1 of the N target classes.

In this paper, we aim to identify the footsteps sounds of humans only and humans with animals. Especially, in the humans with animals class, there is an ambiguity among the footsteps of animals alone, of humans alone, and of animals traveling together with humans. This paper will describe three different types of class discovery algorithm: the mixed-vs-unmixed classifier using Support Vector Machines (SVM), the generative class discovery algorithm using Gaussian Mixture Models (GMM), and the exemplar selection algorithms using SVM and GMM.

The organization of this paper is as follows: Section 2 introduces the multi-sensor multi-modality data and events. Section 3 describes the acoustic feature extraction. Section 4 discusses Gaussian mixture model classifiers, Support Vector Machines, and exemplar selection methods. Section 5 describes the experiments and discussions on the multi-sensor multi-modal dataset. We conclude this paper 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 Detection and classification

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 is similar to voice activity detection in speech processing. 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 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. 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 acoustic recordings into two classes: humans only, and humans with (four-legged) animals.

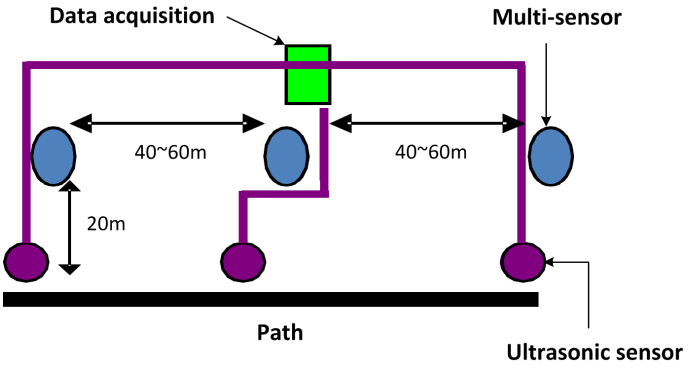

Figure 1: Sensor layout, where a multi-sensor multi-modal system has acoustic, seismic, passive infra-red (PIR), radar, magnetic, and electric field sensors

3 Features Extraction

In acoustic signals, for footsteps, 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 [5], 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 blowing in the recordings. We use spectral subtraction to reduce the effect of noise [6, 7].

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 2. 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.

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 [8, 9]. A Gaussian mixture density is a weighted sum of M component densities, as shown in the following equation,

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

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

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

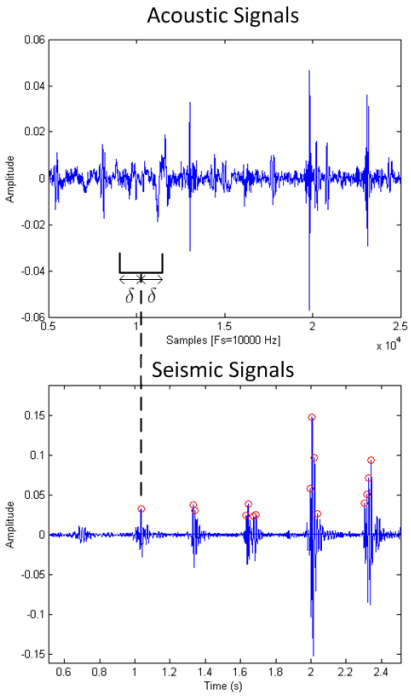

Figure 2: Using peaks of seismic signals for matching acoustic footstep sounds

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

$$\hat{N} = \underset{1 \le k \le N}{\operatorname{argmax}}\, p(X|\lambda_k) = \underset{1 \le k \le N}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k) \tag{3}$$

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

4.2 Support Vector Machines

A Support Vector Machine (SVM) estimates decision surfaces directly [10], 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

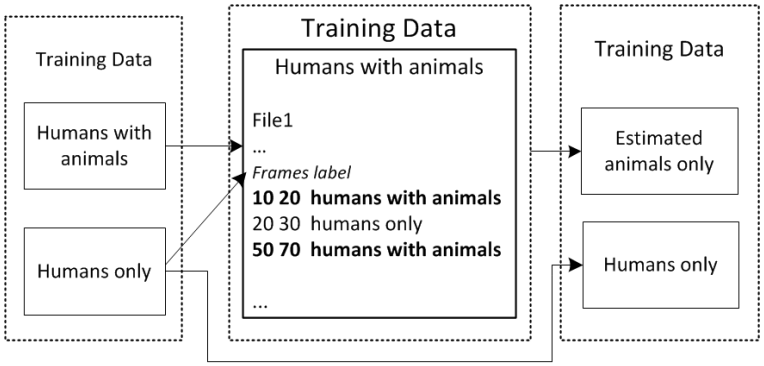

Figure 3: Multi-stage framework for acoustic exemplar selection

label $\mathbf{y} \in R^k$, where $y_i \in \{1, -1\}$, 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 $\mathbf{w} = \sum_{i=1}^{k} y_i\alpha_i\phi(\mathbf{x_i})$, and the decision function is

$$h(x) = \ sgn\left(\sum_{i=1}^{k} y_i\alpha_i K(\mathbf{x_i}, \mathbf{x}) + b\right) \tag{4}$$

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, $K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma||\mathbf{x_i} - \mathbf{x_j}||^2)$ [11].

4.3 Exemplar Selection Methods

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 will be some overlap between the two classes in the feature space, as shown on the left hand side of Figure 4. 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 3; 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 4.
The algorithm is as follows:

1. Train a classifier (SVM or GMM) for *humans only* and *humans with animals* using training data as shown in the left block of Figure 3.

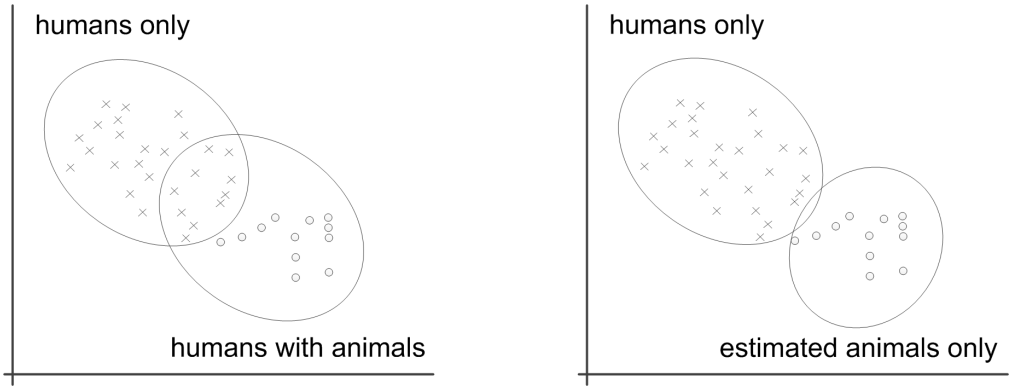

Figure 4: Left: feature space of humans only and humans with animals class. Right: feature space of humans only and estimated animals only class

2. Label the training data of the *humans with animals* class using trained models as shown in the middle block of Figure 3. 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 the *humans with animals* class; in other words, discard the frames which were labeled as the *humans only* class.

4. Train a new classifier (SVM or GMM) between *estimated animals only* and the *humans only* class as shown in the right block of Figure 3.

5 Experiments

In this section, we describe the experiments in order to compare our proposed methods with previous approach in classifying *humans only* vs. *humans with 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.

As described in Section 4.3, we want to examine the effect of using (1) spectral subtraction, (2) seismic peaks with different δ's, (3) and (4) our proposed multi-stage exemplar selection framework using GMM and SVM classifiers for exemplar selection as the first step of the algorithm in Section 4.3, respectively. The experimental results are shown in Table 1.

The first row *PLP features without (1)(2)(3)(4)* in Table 1 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

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 1: 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 GMM classifier. (4) represents the use of our proposed multi-stage exemplar selection framework using SVM classifier.

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 accuracies around 1$\sim$2% for GMM classifiers; on the contrary, using GMM for exemplar selection degrades the accuracies for SVM classifiers. A possible reason is that the 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 classifiers for exemplar selection degrades performance for all cases. It might because the SVM classifiers cannot select proper exemplar in the overlapping feature space case in the first stage.

6 Conclusion

In this paper, we use a challenging realistic multi-sensor multi-modal dataset for personnel detection focusing on classification between humans only and humans with animals. To reduce the ambiguity between the two classes, this paper explores the method of multi-stage exemplar selection method. Experimental results suggest that the SVM classifier gives the best performance in distinguishing mixed vs. unmixed test tokens, but that the self-training method shows promise for the task of learning to distinguish between the discrete footfall sounds of humans and animals.

7 Acknowledgments

This research is supported by ARO MURI 2009-31.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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

[11] 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.