CLASSIFICATION VIA INFORMATION-THEORETIC FUSION OF VECTOR-MAGNETIC AND ACOUSTIC SENSOR DATA

Richard J. Kozick[1] *and Brian M. Sadler*[2]

[1]Bucknell University, Lewisburg, PA 17837, kozick@bucknell.edu
[2]Army Research Laboratory, Adelphi, MD 20783, bsadler@arl.army.mil

ABSTRACT

We present a general approach for multi-modal sensor fusion based on nonparametric probability density estimation and maximization of a mutual information criterion. We apply this approach to fusion of vector-magnetic and acoustic data for classification of vehicles. Linear features are used, although the approach may be applied more generally with other sensor modalities, nonlinear features, and other classification targets. For the magnetic data, we present a parametric model with computationally efficient parameter estimation. Experimental results are provided illustrating the effectiveness of a classifier that discriminates between cars and sport utility vehicles.

Index Terms— sensor network, classification, sensor fusion, mutual information.

1. INTRODUCTION

Multimodal sensor networks are deployed in many scenarios where each node contains several sensor modalities, such as acoustic, magnetic, seismic, radar, electrostatic, infrared, optical, and others. We focus on fusing two modalities, acoustic and magnetic, for the purpose of classifying civilian vehicles such as cars, sport utility vehicles (SUVs), and trucks. In this work, the magnetic sensor is a vector magnetometer and the acoustic sensor is a single microphone, and the vehicle is moving along a road so that the range to the sensors is known approximately.

For a magnetic source moving with constant velocity, a model for the vector magnetometer output signal is available based on linear combinations of Anderson functions. We use this model to estimate the source speed and reduce the vector magnetic data to 9 parameters. A corresponding parametric model is not available for the acoustic signal from civilian vehicles, and models are not known for the joint statistical dependence between the magnetic and acoustic signals. We address this by using nonparametric probability density estimation to learn the joint statistics from training data, and then the magnetic-acoustic data is fused by extracting features for classification that maximize an information-theoretic criterion. We apply the approach with measured data from civilian vehicles, demonstrating that fusion of magnetic and acoustic data using the information-theoretic criterion improves the ability to discriminate between cars and SUVs.

2. VECTOR MAGNETIC SENSOR MODEL AND PARAMETER ESTIMATION

We begin with a review of a parametric model for the vector magnetic field observed at a sensor in a time interval around the closest point of approach (CPA) of the source, then we present a computationally-efficient algorithm for estimating the parameters.

2.1. Magnetic data model

We assume that a magnetic dipole with moment vector $\mathbf{m}$ passes CPA at time $t = 0$, as illustrated in Figure 1. The CPA range is denoted by R_{CPA}, the source velocity vector is $\mathbf{v}$, and the speed is v. Then each component of the vector magnetic field can be expressed as a linear combination of the Anderson functions [1]

$$f_n(\alpha t) = \frac{(\alpha t)^n}{\left[1+(\alpha t)^2\right]^{5/2}}, \quad n = 0,1,2, \tag{1}$$

where the time-scale α is related to the speed and R_{CPA} as

$$\alpha = v / R_{\text{CPA}}. \tag{2}$$

The magnetic field components vary with time according to [1]

$$\mathbf{B}(t) = \begin{bmatrix} B_x(t) & B_y(t) & B_z(t) \end{bmatrix} = \mathbf{F}(\alpha t)\mathbf{C} + \mathbf{e}(t), \tag{3}$$

where $\mathbf{F}(\alpha t)$ is a 1x3 vector of Anderson functions,

$$\mathbf{F}(\alpha t) = \begin{bmatrix} f_0(\alpha t) & f_1(\alpha t) & f_2(\alpha t) \end{bmatrix}, \tag{4}$$

$\mathbf{C}$ is a 3x3 matrix of coefficients, and $\mathbf{e}(t)$ is a 1x3 vector that accounts for deviations from the ideal model due to noise, sensor motion, extraneous magnetic sources, and other effects.

The model can be further elaborated to relate the elements of $\mathbf{C}$ to the magnetic dipole moment vector $\mathbf{m}$, the direction of motion $\mathbf{v}/v$, and the CPA range [1]. We do not include these detailed (and nonlinear) relationships here, but we note the following characteristics of $\mathbf{C}$ from the detailed model [1]:

1. The $\mathbf{C}$ matrix is scaled by a factor that is proportional to $\left(m/R_{\text{CPA}}^3\right)$, where m is the magnetic dipole moment magnitude. Therefore the elements of $\mathbf{C}$ have magnitude that is proportional to m and decays rapidly with range.
2. The variations from element to element in $\mathbf{C}$ depend on the orientations of the magnetic dipole vector $\mathbf{m}/m$ and the direction of motion $\mathbf{v}/v$.
3. The source speed v enters the model in (3) only through the time-scale parameter α in (2).

These observations imply that the source speed can be estimated from α if the CPA range is known. Also, $\mathbf{C}$ characterizes the magnetic dipole moment vector of the source if the sensor is placed near a road, because then the direction of motion $\mathbf{v}/v$ is fixed (except for a sign difference for left-to-right and right-to-left motion) and the CPA range is approximately known (within the width of the road). However, the $\mathbf{C}$ matrix will be different for left-to-right and right-to-left motion. Therefore we use $\mathbf{C}$ to summarize the magnetic properties of the source for classification.

2.2. Parameter estimation

The model in (3) contains 10 parameters in α and $\mathbf{C}$, where $\mathbf{B}(t)$ represents the measured vector magnetic sensor data. The parameters may be estimated by minimizing the squared-error,

$$\varepsilon^2(\alpha,\mathbf{C}) = \mathrm{Tr}\left\{\int\left[\mathbf{B}(t)-\mathbf{F}(\alpha t)\mathbf{C}\right]^T\left[\mathbf{B}(t)-\mathbf{F}(\alpha t)\mathbf{C}\right]dt\right\} \quad (5)$$

where superscript T denotes the transpose operation, Tr is the trace of the matrix, and the integration limits are from $-\infty$ to ∞. For fixed α, the least-squares estimate of $\mathbf{C}$ is

$$\hat{\mathbf{C}}\big|\alpha = \arg\min_{\mathbf{C}} \varepsilon^2(\alpha,\mathbf{C}) = \mathbf{G}(\alpha)\int\mathbf{F}(\alpha t)^T\,\mathbf{B}(t)\,dt\,, \quad (6)$$

where the 3x3 matrix $\mathbf{G}(\alpha)$ can be expressed in closed-form as

$$\mathbf{G}(\alpha) = \left[\int\mathbf{F}(\alpha t)^T\,\mathbf{F}(\alpha t)\,dt\right]^{-1} = \frac{8|\alpha|}{5\pi}\begin{bmatrix} 3 & 0 & -5 \\ 0 & 16 & 0 \\ -5 & 0 & 35 \end{bmatrix}. \quad (7)$$

Next, to find the least-squares estimate for α, we substitute (6) into (5) to eliminate $\mathbf{C}$, leading to

$$\hat{\alpha} = \arg\max_{\alpha}\ \mathrm{Tr}\left\{\int\int\mathbf{B}(s)^T\left[\mathbf{F}(\alpha s)\mathbf{G}(\alpha)\mathbf{F}(\alpha t)^T\right]\mathbf{B}(t)\,ds\,dt\right\}. \quad (8)$$

An interpretation of (8) is that α is chosen to maximize the total energy in the orthogonal projections of the components of $\mathbf{B}(t)$ onto the subspace spanned by the Anderson functions. The quantity inside the square brackets in (8) is a scalar function that can be evaluated as

$$K(s,t;\alpha) = \mathbf{F}(\alpha s)\mathbf{G}(\alpha)\mathbf{F}(\alpha t)^T = \frac{8|\alpha|}{5\pi}\,\frac{35(\alpha s)^2(\alpha t)^2 - 5(\alpha s)^2 - 5(\alpha t)^2 + 16(\alpha s)(\alpha t) + 3}{\left[1+(\alpha s)^2\right]^{5/2}\left[1+(\alpha t)^2\right]^{5/2}} \quad (9)$$

so (8) may be expressed more directly as

$$\hat{\alpha} = \arg\max_{\alpha}\ \left\{\int\int K(s,t;\alpha)\,B(s,t)\,ds\,dt\right\} \quad (10)$$

where

$$B(s,t) = B_x(s)B_x(t) + B_y(s)B_y(t) + B_z(s)B_z(t). \quad (11)$$

The operation in (10) may be viewed as a vector matched-filter on $\mathbf{B}(t)$ to estimate $\alpha = v/R_{\mathrm{CPA}}$.

In summary, the global solution to the least-squares minimization in (5) is obtained by first solving (10) for $\hat{\alpha}$, and then using $\hat{\alpha}$ in (6) to find $\hat{\mathbf{C}}$. The continuous-time formulation facilitates the evaluation of the closed-form expressions in (7) and (9). In practice, sampled data is used, so $\mathbf{B}(t)$ is replaced by an *N*x3 matrix, $\mathbf{B}_N = [\mathbf{B}_x, \mathbf{B}_y, \mathbf{B}_z]$, containing *N* samples of each vector magnetic sensor data component with spacing T_s sec between samples. Then the model in (3) becomes

$$\mathbf{B}_N = \mathbf{F}_N(\alpha)\mathbf{C} + \mathbf{e}_N\,, \quad (12)$$

where the *N*x3 matrix $\mathbf{F}_N(\alpha)$ contains samples of the Anderson functions in (1). We define the 3x3 matrix

$$\mathbf{G}_N(\alpha) = \left[T_s\,\mathbf{F}_N(\alpha)^T\,\mathbf{F}_N(\alpha)\right]^{-1} \approx \mathbf{G}(\alpha) \quad (13)$$

and the *N*x*N* matrix

$$\mathbf{K}_N(\alpha) = \mathbf{F}_N(\alpha)\,\mathbf{G}_N(\alpha)\,\mathbf{F}_N(\alpha)^T \approx \mathbf{K}(\alpha) \quad (14)$$

where $\mathbf{K}(\alpha)$ is an *N*x*N* matrix obtained by sampling the function in (9). The approximations in (13) and (14) become exact as the sample spacing $T_s \to 0$ and the processing time interval $\to \infty$ The approximations reduce computations by eliminating the matrix products and inverse in (13) and (14). The least-squares estimates for $\hat{\alpha}$ and $\hat{\mathbf{C}}$ with discrete-time data are then

$$\begin{aligned}\hat{\alpha} &= \arg\max_{\alpha}\ \left\{\mathbf{B}_x^T\,\mathbf{K}_N(\alpha)\mathbf{B}_x + \mathbf{B}_y^T\,\mathbf{K}_N(\alpha)\mathbf{B}_y + \mathbf{B}_z^T\,\mathbf{K}_N(\alpha)\mathbf{B}_z\right\} \\ &\approx \arg\max_{\alpha}\ \left\{\mathbf{B}_x^T\,\mathbf{K}(\alpha)\mathbf{B}_x + \mathbf{B}_y^T\,\mathbf{K}(\alpha)\mathbf{B}_y + \mathbf{B}_z^T\,\mathbf{K}(\alpha)\mathbf{B}_z\right\}\end{aligned} \quad (15)$$

and

$$\hat{\mathbf{C}} = T_s\,\mathbf{G}_N(\hat{\alpha})\mathbf{F}_N(\hat{\alpha})^T\mathbf{B}_N \approx T_s\,\mathbf{G}(\hat{\alpha})\mathbf{F}_N(\hat{\alpha})^T\mathbf{B}_N \quad (16)$$

where the approximations in (13) and (14) are used to reduce computation. The model-based estimate is then

$$\hat{\mathbf{B}}_N = \mathbf{F}_N(\hat{\alpha})\hat{\mathbf{C}} \quad (17)$$

and $\hat{\mathbf{B}}_N$ can be compared with the data $\mathbf{B}_N$ to assess the fit of the model to the data. Figure 2 shows a good fit between the model and measured data for a car traveling at 15 mph with CPA range 19 ft. The estimated speed is 15.2 mph, which agrees closely with the ground truth. Similar fits to the source speed and model were obtained for 25 different vehicles in the measured data set.

3. JOINT MAGNETIC-ACOUSTIC FEATURES THAT MAXIMIZE MUTUAL INFORMATION

In this section, we consider classification of vehicles by jointly processing vector magnetic field data measured with a magnetometer and single-channel acoustic data measured with a microphone. The steps in our approach for linear feature extraction are described first, followed by an algorithm for finding features that maximize a mutual information criterion.

3.1. Procedure for linear feature extraction

(1) Estimate the CPA time using the peak of the total field, $\hat{t}_{CPA} = \arg\max_t B(t) = \arg\max_t \left[B_x(t)^2 + B_y(t)^2 + B_z(t)^2\right]^{1/2}$.

Take a window of vector magnetic field samples $\mathbf{B}_N$ and acoustic

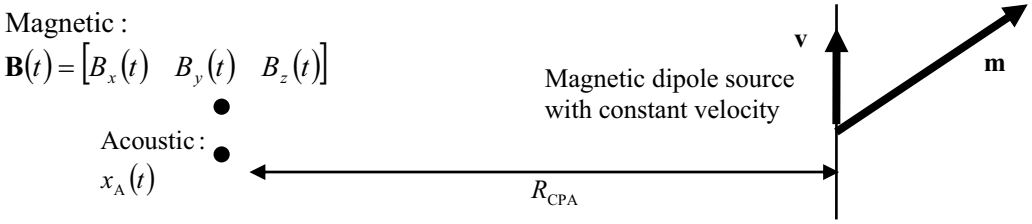

Figure 1: Illustration of acoustic-magnetic source moving with constant velocity near a vector magnetometer and microphone.

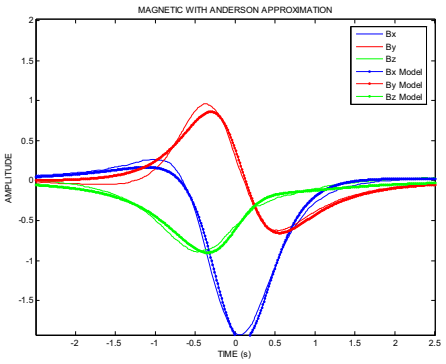

Figure 2: Comparison of the fit between measured vector-magnetic data and Anderson function model (15)-(17) for a car.

samples $\mathbf{X}_A$ with CPA at the center of the window, and redefine the time axis for the samples so that $t = 0$ at CPA.

(2) Process the window of vector magnetic field data $\mathbf{B}_N$ as in (15) and (16) to estimate the model parameters $\hat{\alpha}$ and $\hat{\mathbf{C}}$. The columns of $\hat{\mathbf{C}}$ are stacked into a 9x1 vector $\mathbf{X}_{\text{Mag}}$ that represents the magnetic data. The source speed may be estimated from $\hat{\alpha}$ using (2) if the CPA range is known.

(3) The window of acoustic samples is placed into a vector $\mathbf{X}_A$ with $N_A \times 1$ samples. Parametric models are not available for the acoustic signals corresponding to civilian vehicles at CPA, so it is not obvious how the acoustic data may be reduced to a few parameters that are analogous to $\mathbf{X}_{\text{Mag}}$ for the magnetic field data.

(4) The $\mathbf{X}_{\text{Mag}}$ and $\mathbf{X}_A$ vectors are stacked into a joint magnetic-acoustic vector, $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{\text{Mag}} \\ \mathbf{X}_A \end{bmatrix}$. We focus on magnetic and acoustic data in this paper, but other sensor modalities may be included.

(5) The magnetic-acoustic data in $\mathbf{X}$ is processed by a linear transformation to extract a low-dimensional feature vector $\mathbf{Y}$ with dimension $N_Y \times 1$,

$$\mathbf{Y} = \mathbf{A}^T \mathbf{X} \tag{18}$$

where $\mathbf{A}$ is a matrix with dimension $(9 + N_A) \times N_Y$. Information-theoretic criteria for choosing $\mathbf{A}$ to maximize the classification information in $\mathbf{Y}$ are described in Section 3.2.

3.2. Maximum mutual information (MMI) features

We begin with a review of several desirable properties of features that maximize a mutual information (MMI) criterion. Then we review a particular algorithm [2] for extracting MMI features that uses nonparametric probability density function (pdf) estimation to learn the joint statistical dependence between the magnetic and acoustic measurements from the training data.

The dimensionality-reducing feature extraction processing in (18) is not strictly necessary for classification, since the classifier can operate directly on the higher-dimensional data vector, $\mathbf{X}$. However, with small training sets, classifiers often generalize better when they are trained with low-dimensional features that retain the information for classification. In addition, features derived from information-theoretic criteria have recently been found [4] to achieve lower classification error on several data sets than systems that *jointly* derive the features and the classifier, such as [5]. An advantage of designing the features independently from the classifier is that the features may be applied subsequently to any of a large number of classifiers [3].

A theoretical basis for using mutual information for feature extraction is provided by bounds on the probability of classification error, P_e. Suppose that $\mathbf{S}$ is a discrete random variable with alphabet {1, 2, …, M} representing the labels of M classes. The mutual information (MI) between the feature vector $\mathbf{Y}$ and the class label $\mathbf{S}$ is denoted by $I(\mathbf{S}, \mathbf{Y})$. It has been shown that P_e is bounded above [6] and below [7] by functions of the MI, where the bounds are *decreasing* functions of the MI. Therefore maximizing the MI in the features minimizes the bounds on P_e. The upper and lower bounds in [6,7] are stated in terms of Shannon mutual information, but the bounds have recently been extended to Renyi mutual information in [8]. The MMI algorithm that we use from [2] maximizes a form of Renyi mutual information. Further justification for mutual information as a metric for feature extraction has recently been presented in [9], where several commonly used linear feature extraction methods are formulated in a unified information-theoretic framework.

The desirable properties of MMI features have been known for some time, but computational difficulties have prevented widespread use until recently. As above, let $\mathbf{S}$ be a discrete random variable with alphabet {1, 2, …, M} that represents the class label, $\mathbf{Y}$ is the feature vector, $f_{\mathbf{S}}(i)$ is the *a priori* probability that $\mathbf{S} = i$, $f_{\mathbf{Y}|\mathbf{S}}(\mathbf{y} \mid i)$ is the probability distribution for the features in class i, and $f_{\mathbf{Y}}(\mathbf{y}) = \sum_{i=1}^{M} f_{\mathbf{Y}|\mathbf{S}}(\mathbf{y} \mid i) f_{\mathbf{S}}(i)$ is the distribution of all the classes. The definitions of Renyi entropy with order α and Shannon entropy for a random vector $\mathbf{X}$ with probability distribution $f_{\mathbf{X}}(\mathbf{x})$ are [11]

$$\text{Renyi}: H_\alpha(\mathbf{X}) = \frac{1}{1-\alpha} \log \mathsf{E}_{\mathbf{X}} \left\{ f_{\mathbf{X}}(\mathbf{X})^{\alpha - 1} \right\} \tag{19}$$

$$\text{Shannon}: H(\mathbf{X}) = -\mathsf{E}_{\mathbf{X}} \{ \log f_{\mathbf{X}}(\mathbf{X}) \}, \tag{20}$$

where $\alpha > 0, \alpha \neq 1$, and $\lim_{\alpha \to 1} H_\alpha(\mathbf{X}) = H(\mathbf{X})$. We will use Renyi's quadratic entropy with $\alpha = 2$.

The Shannon and Renyi entropies can be used to define the classical Shannon MI and Renyi's quadratic MI, and both of these MI forms have been used for feature extraction. In [4], Renyi's quadratic MI is used and combined with the stochastic information gradient from [12] to reduce complexity. In [13], Shannon MI is used to extract *nonlinear* MMI features via the "kernel induced feature space" (KIFS). Other recent works [14, 15] have connected information-theoretic learning with kernel methods.

A different form of quadratic MI is defined in [2] for the purpose of extracting MMI features. The quadratic mutual information in [2] is motivated from a quadratic divergence measure between $f_{\mathbf{Y},\mathbf{S}}(\mathbf{y}, i) = f_{\mathbf{Y}|\mathbf{S}}(\mathbf{y} \mid i) f_{\mathbf{S}}(i)$ and $f_{\mathbf{Y}}(\mathbf{y}) f_{\mathbf{S}}(i)$,

$$\begin{aligned} I_T(\mathbf{Y};\mathbf{S}) &= \sum_{i=1}^{M} \int \left[f_{\mathbf{Y},\mathbf{S}}(\mathbf{y}, i) - f_{\mathbf{Y}}(\mathbf{y}) f_{\mathbf{S}}(i) \right]^2 d\mathbf{y} \\ &= \sum_{i=1}^{M} f_{\mathbf{S}}(i)^2 \left[\int f_{\mathbf{Y}|\mathbf{S}}(\mathbf{y}|i)^2 \, d\mathbf{y} + \int f_{\mathbf{Y}}(\mathbf{y})^2 \, d\mathbf{y} - 2 \int f_{\mathbf{Y}|\mathbf{S}}(\mathbf{y}|i) f_{\mathbf{Y}}(\mathbf{y}) \, d\mathbf{y} \right] \end{aligned} \tag{21}$$

This same divergence measure is proposed and studied in [14, 15].

Nonparametric Parzen windows may be used to estimate the probability distributions in (21). The Parzen window method places a "kernel function" around each training sample and adds the kernels to yield a continuous function. We will use a Gaussian kernel function with dimension N_Y and diagonal covariance $\sigma^2\mathbf{I}$,

$$G_{\sigma^2}(\mathbf{y}) = (2\pi\sigma^2)^{-N_Y/2}\exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y}\right). \qquad (22)$$

As in [2], the following notation is used for the features $\mathbf{Y} = \mathbf{A}^T\mathbf{X}$ corresponding to the training data. Let T_i denote the number of training samples for class i, for $i = 1,\ldots,M$, and let $T = T_1 + \cdots + T_M$ be the total number of training samples. The features corresponding to the training data for class i are denoted by $\mathbf{y}_t^i$ for $t = 1,\ldots,T_i$ for $i = 1,\ldots,M$. The superscript is the class label and the subscript is the index of the vector within the class. When the class label is not important, the training samples are indexed with a single subscript, $\mathbf{y}_t$, for $t = 1,\ldots,T$.

Using (22) to estimate the pdfs in (21) yields the MMI criterion in [2],

$$\hat{I}_T(\mathbf{Y};\mathbf{S}) = \frac{1}{T^2}\sum_{i=1}^{M}\sum_{t=1}^{T_i}\sum_{u=1}^{T_i}G_{2\sigma^2}(\mathbf{y}_t^i - \mathbf{y}_u^i) - \frac{2}{T^2}\sum_{i=1}^{M}\frac{T_i}{T}\sum_{t=1}^{T}\sum_{u=1}^{T_i}G_{2\sigma^2}(\mathbf{y}_t - \mathbf{y}_u^i) + \frac{1}{T^2}\left[\sum_{i=1}^{M}\left(\frac{T_i}{T}\right)^2\right]\sum_{t=1}^{T}\sum_{u=1}^{T}G_{2\sigma^2}(\mathbf{y}_t - \mathbf{y}_u) \qquad (23)$$

We maximize (23) with respect to the matrix **A** to obtain the MMI features, with the constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. The maximization of quadratic MI can also be applied with *nonlinear* feature mappings $\mathbf{Y} = \mathbf{g}(\mathbf{X};\boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a parameter vector.

4. RESULTS USING MEASURED DATA

We have applied the information-theoretic fusion of magnetic and acoustic data with measured data to classify vehicles. The experiments consisted of cars, SUVs, and trucks traveling along a road in both directions, left-to-right (L2R) and right-to-left (R2L), with CPA ranges from 19 to 28 ft, and at speeds of 15 mph and 25 mph. The data set was too small to estimate the probability of classification error, so we examined scatter plots of three-dimensional feature vectors to evaluate discrimination between cars and SUVs. We observed the following results from processing the measured data. **(1)** The MMI features significantly improve discrimination compared with simple Fisher's LDA [3] features. **(2)** Fusion of magnetic and acoustic data allows discrimination, but using magnetic data alone does not allow discrimination. **(3)** Incorporation of simple information about the vehicle's track (speed & direction) improves feature extraction for classification. The magnetic signatures vary with the vehicle's direction and the acoustic signatures vary with the speed, so features that are matched to the direction & speed perform better.

Figure 3 contains a representative result, where CAR-L and CAR-R are training data for (different) cars moving L2R and R2L, respectively, and SUV-L and SUV-R are corresponding training data for SUVs. The features are derived separately for vehicles moving L2R and R2L, giving rise to the top and bottom panels. The LSUV points in Figure 3 correspond to a new "light SUV" that is different than the SUVs in the training data, where LSUV-L and LSUV-R are traveling L2R and R2L, respectively. Note that the LSUV-L is closely clustered with the SUV-L training data and LSUV-R is closely clustered with SUV-R, indicating discrimination of the SUV from the cars.

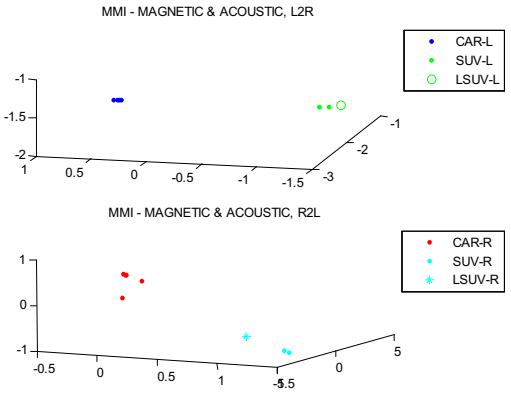

Figure 3: MMI features for new (non-training) magnetic-acoustic data from a light SUV (LSUV) moving L2R and R2L at 15 mph.

REFERENCES

[1] W.M. Wynn, "Detection, Localization, and Characterization of Static Magnetic-Dipole Sources," Chapt. in *Detect. and Ident. of Visually Obscured Targets*, C.E. Baum (Ed.), Taylor & Francis, 1999.

[2] K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization," *Jrnl. of Machine Learning Research*, vol. 3, pp. 1415-1438, March 2003.

[3] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification* (2nd Edition), Wiley, 2000.

[4] K.E. Hild, D. Erdogmus, K. Torkkola, and J.C. Principe, "Feature Extraction Using Information-Theoretic Learning," *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 28, iss. 9, Sept. 2006.

[5] A. Biem, S. Katagiri, and B.-H. Juang, "Pattern Recognition Using Discriminative Feature Extraction," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 500-504, Feb. 1997.

[6] M.E. Hellman and J. Raviv, "Probability of error, equivocation and the Chernoff bound," *IEEE Trans. Info. Theory*, vol. 16, 1970.

[7] R.M. Fano. *Transmission of Information: A Statistical theory of Communication,* Wiley, New York, 1961.

[8] D. Erdogmus and J.C. Principe, "Lower and Upper Bounds for Misclassification Probability Based on Renyi's Information," *The Journal of VLSI Sig. Proc.*, Vol. 37, Nos. 2-3, June 2004.

[9] S. Petridis and S.J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern Recognition*, Vol. 34, No. 5, May 2004.

[10] J.C. Principe, D. Xu, and J.W. Fisher III, "Information Theoretic Learning," Chapter 7 in *Unsupervised Adaptive Filtering*, Simon Haykin (Editor), pp. 265-319, Wiley, 1999.

[11] J.C. Principe, D. Xu, Q. Zhao, and J.W. Fisher III, "Learning from examples with information theoretic criteria.," *The Journal of VLSI Sig. Proc.*, Vol. 26, Nos. 1-2, pp. 61-77, Aug. 2000.

[12] D. Erdogmus, K.E. Hild II, and J.C. Principe, "On-Line Entropy Manipulation: Stochastic Information Gradient," *IEEE Signal Processing Letters*, vol. 10, no. 8, pp. 242-245, Aug. 2003.

[13] U. Ozertem, D. Erdogmus, R. Jenssen, "Spectral Feature Projections That Maximize Shannon Mutual Information with Class Labels," *Pattern Recognition*, vol. 39, pp. 1241-1252, 2006.

[14] R. Jenssen, D. Erdogmus, J. C. Principe and T. Eltoft, "Towards A Unification of Information Theoretic Learning and Kernel Methods," *IEEE Workshop on Mach. Learning for Sig. Proc.*, Sept. 2004.

[15] R. Jenssen, J.C. Principe, D. Erdogmus, and T. Eltoft, "The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels," *Jrnl. Franklin Inst.,* to appear 2006.