Metric Driven Classification: A Non-Parametric Approach Based on the Henze–Penrose Test Statistic

Sally Ghanem, *Student Member, IEEE*, Hamid Krim, *Fellow, IEEE*,
Hamilton Scott Clouse, *Member, IEEE*, and Wesam Sakla, *Member, IEEE*

***Abstract*— Entropy-based divergence measures have proven their effectiveness in many areas of computer vision and pattern recognition. However, the complexity of their implementation might be prohibitive in resource-limited applications, as they require estimates of probability densities which are expensive to compute directly for high-dimensional data. In this paper, we investigate the usage of a non-parametric distribution-free metric, known as the Henze–Penrose test statistic to obtain bounds for the k-nearest neighbors (k-NN) classification accuracy. Simulation results demonstrate the effectiveness and the reliability of this metric in estimating the inter-class separability. In addition, the proposed bounds on the k-NN classification are exploited for evaluating the efficacy of different pre-processing techniques as well as selecting the least number of features that would achieve the desired classification performance.**

***Index Terms*— Dimensionality reduction, classification, divergence measures, nearest neighbor graph, pattern recognition.**

I. Introduction

COMPUTER vision and machine learning have witnessed a wealth of great research activity in image analysis and modeling for inference, all with their strengths and limitations in extracting salient features from 1/multi-dimensional signals. In imaging, a particular challenge that affects successful object discrimination is the high-dimensionality of raw sensor imagery and signature measurements. High dimensionality of sensor data is computationally demanding, and shifts limited computational resources away from other tasks, e.g. navigation, control and avionics. To further improve sensor resource management, where various sensors may be tasked to collect data, the importance and relevance of a particular sensor is tied to its discrimination capability which often needs to be quantified. The associated collected data is subsequently processed and adequately exploited for object inference applications such as multi-object classification.

Information-theoretic divergence measures have been widely used in many signal and image applications, e.g. image registration [1]–[3], image segmentation and retrieval [4], image alignment [5], speech classification [6] as well as a variety of other problems. The better known divergence measures include the Kullback-Liebler divergence [7] which is based on Shannon entropy measure and the so-called Renyi-divergence (alpha-divergence) measure [8] which is based on Renyi entropy. Other divergence measures include the Jensen-Shannon divergence [9], the Jensen-Renyi divergence [10], total variation k-dP divergence [11] and Bregman divergence [12].

A limitation common to all these measures is their typically direct use of the probability density functions whose estimation for high dimensional data is computationally prohibitive. As a result, their adoption in addressing inference problems for high dimensional data has been negatively impacted. Friedman and Rafsky [13] proposed the idea of using Minimum Spanning Trees (MST) to extend the Wald-Wolfowitz test [14], which is also known as the two-sample test, for high dimensional data. As we elaborate later, we propose to exploit this strategy by adopting a related Henze-Penrose (HP) divergence measure [15] to quantitatively estimate the number of data features required to preserve a target classification performance. This HP measure is, itself, based on the Friedman-Rafsky result. The classification of our proposed approach is evaluated by accuracy performance and its dependence on the dimensionality of the data. We further derive performance bounds for the K-Nearest Neighbors algorithm accuracy [16] in terms of the HP metric. These bounds are in turn used to gauge the performance of different feature extraction techniques.

Bounds on classification error rates have been studied and extensively reported in the machine learning literature. Chernoff α-divergence measure [17] has been used to provide an upper bound on the classification error probability. It was moreover shown that a special case of the Chernoff α divergence measure $\alpha = \frac{1}{2}$, as a Battacharya coefficient (BC) [18], could be used to upper/lower bound the Bit Error Rate (BER) [19]. The BC divergence was also shown to provide the tightest upper bound on the probability of error when the classes mildly differed from one another [20]. Other works have established connections between the Kullback-Liebler (KL) divergence and the Total Variation (TV) distance [21], [22]. Numerous other bounds on probability

Manuscript received January 17, 2018; revised July 2, 2018; accepted July 27, 2018. Date of publication August 1, 2018; date of current version September 4, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alin M. Achim. *(Corresponding author: Sally Ghanem.)*

S. Ghanem and H. Krim are with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27606 USA (e-mail: ssghanem@ncsu.edu).

H. S. Clouse is with the Air Force Research Lab, Avionics Cir, Wright-Patterson AFB, OH 45433 USA.

W. Sakla is with Data Analytics and Decision Sciences, Lawrence Livermore National Laboratory, Livermore, CA 94550 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2862352

of error have been derived in a variety of applications, and using functionals of the Probability Density Function (PDF). These for the most part, require some prior knowledge of the associated PDFs of the target classes [23], [24]. On the other hand, non-parametric approaches have been developed, and rely on the estimation of PDFs. Their various applications may be found in [25]–[27].

In this work, bounds for k-NN classifier performance were derived leveraging the HP metric to mitigate both needs for prior knowledge of the class distributions and the estimation of their PDFs. Our proposed bounds, in addition, may be used to estimate the classification error rate when the training and the test data are drawn from different distributions, as noted in [28].

The paper is organized as follows: Section 2 introduces the Friedman-Rafsky and Henze-Penrose test statistics, and provides an overview of their properties. Section 3 introduces the bounds for the k-NN ($k = 1$) classification accuracy. Section 4 describes our generalized bounds for k > 1 while Section 5 shows some other bounds for the k-NN using the inequalities derived for $k = 1$. In Section 6, we further establish relations among our derived bounds and others in the literature. Section 7 describes our dataset structure, the experimental setup and presents the simulation results of the proposed bounds, while Section 8 provides concluding remarks. The derived bounds will be exploited to quantitatively asses different image processing techniques.

II. Graph-Theoretic Data Classification

Consider two classes ω_0 and ω_1 over a space X with samples of size m and n respectively from distribution $p(x)$ and $q(x)$, both defined in $\mathbb{R}^d$, where d is a positive integer number. According to the Wald-Wolfowitz test, the null hypothesis H_o specifies that $p(x) = q(x)$ which means that both samples are drawn from the same underlying distribution. Our interest is in the case $H_1 : p(x) \neq q(x)$ where each sample belongs to a different class. The Wald-Wolfowitz test (for $d = 1$) begins by sorting the univariate observations $N = m + n$ in an ascending order with respect to their values. Each observation is then replaced by a label ω_0 or ω_1 depending upon the class to which it originally belonged. The number of runs, $R_{m,n}$, is an integer number representing the consecutive sequences of identical labels. $R_{m,n}$ provides a simple, yet effective non parametric measure of separation between the two samples of potentially distinct classes, by making use of the local characteristics of the distributions. Lower values of $R_{m,n}$ correspond to increased separation between the class distributions and vice versa.

Friedman and Rafsky [13] generalized the Wald-Wolfowitz univariate statistic to multi-dimensional data using the number of runs computed by the MST, which they prove to be well adapted. The following is a brief description of the formalism. Let a weighted graph consist of N nodes corresponding to N pooled sample data points in $\mathbb{R}^d$. An edge weight is a measure of dissimilarity between the associated nodes, e.g. Euclidean distance. The MST of this graph is thus the subgraph of minimal total distance that provides a path between every pair of nodes. The test statistic, $R_{m,n}$, is now given by the number of connected components left after removing edges connecting nodes of different classes in the MST. $R_{m,n}$ computation for fairly separated and mixed data is depicted in Figs. (1a and 1b) respectively. Henze and Penrose [15] extended the work of Friedman and Rafsky [13] by proving that as the number of vertices $m, n \longrightarrow \infty$, a function of the statistic $R_{m,n}$ asymptotically converges to a member of the f-divergence family almost surely as shown in Eqn. (1). The Henze-Penrose divergence measure HP estimates the distributional overlap between two distributions $p(x)$ and $q(x)$, where a and $b \in$ [0,1], and $a = \frac{m}{m+n}$ and $b = 1 - a$.

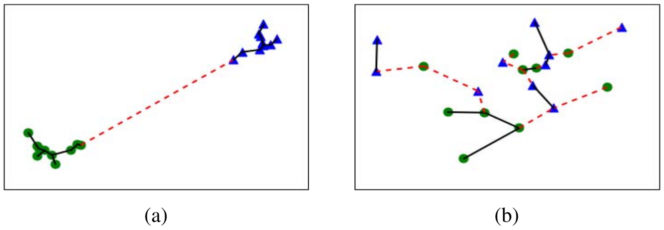


Fig. 1. Example for $R_{m,n}$ computation for fairly separated and mixed data, where the red dotted lines correspond to edges connecting nodes from different classes. (a) Separated data ($R_{m,n} = 2$). (b) Mixed data ($R_{m,n} = 12$).

$$lim_{m,n\to\infty} HP = 1 - \frac{R_{m,n}}{m+n} \longrightarrow \int \frac{a^2p^2(x) + b^2q^2(x)}{ap(x) + bq(x)} dx. \tag{1}$$

Thus, given a distance or proximity measure, HP can provide a measure of separation between two classes of objects in the original representation space. A Henze-Penrose value of 0.5 implies that the densities $p(x)$ and $q(x)$ are drawn from the same underlying distribution. As HP increases, the densities $p(x)$ and $q(x)$ are increasingly separated to the point where HP attains its maximum value at 1. Extending the capability of the Henze-Penrose metric to measure the inter-class separation for an arbitrary number of distributions is an interesting idea worthy of exploration. This has been established for other divergence measures [1], [10]. This would entail the generalization of the Henze-Penrose metric which was originally proposed for two populations, and for which we have derived bounds. While in theory, the HP measure may be thought to be readily extendable to an arbitrary number of distributions (i.e., increasing the population to 3 or more distributions) by redefining the HP metric via a modification of the integral and the establishment of its equivalence to an appropriately defined Friedman-Rafsky statistic, this is beyond the scope of the present work and would likely be part of future work.

III. Bounds on the Nearest Neighbor Classification Accuracy

The k-Nearest Neighbors algorithm (k-NN) is a method used for classification and regression [29]. In k-NN classification, an object is classified according to a majority vote of its neighbors where the object class is chosen to be the most common one among its k closest neighbors. For k = 1, the object is simply assigned to the class of the nearest neighbor and in this case, the classifier is referred to as the

Nearest Neighbor (NN) classifier. The Henze-Penrose metric can be used to provide bounds for the NN classification accuracy as follows,

Theorem 1: Let labeled classes of m and n points form a complete graph of unique distances. Given the Henze-Penrose metric HP between the two classes and the number of connected components in the nearest neighbor graph C, the nearest neighbor classification accuracy, A_{NN}, is bounded above and below by:

$$2 * HP + \frac{2}{m+n} - 1 \leq A_{NN} \leq HP + \frac{C}{m+n}. \tag{2}$$

Practically, the lower bound is more useful since it does not require calculating the number of nearest neighbor components. It proves that as the HP metric tends to 1, so does the NN classification accuracy. Both the upper and lower bounds are tight and can not be improved upon without further assumptions. See Appendix A for the full proof of Theorem 1.

A. Variations on the Bound Themes

The predicted classification accuracy on training data can be used to bound the classification accuracy on the unlabeled testing data. This is a well-known problem in the literature known by domain adaption [28]. The test data often follow a different distribution than the training data. It was shown in [28] that the empirical error on the training data can characterize the test data error of the classifier learned by this training data. The problem is formalized by viewing the training data as defining the source domain data, and interpreting the test data as the target domain data. X_S and X_T respectively represent the source and target data while y_S and y_T are their associated labels. The source and the target data are drawn from the distributions $f_S(x)$ and $f_T(x)$ respectively. The labeling function on inputs $\mathcal{X}$ is defined as $y : \mathcal{X} \to \{0, 1\}$. The probability that a hypothesis h disagrees with the label $y(x)$ is defined for the source data as:

$$\epsilon_S(h, y_S) = \boldsymbol{E_{x \sim X_S}}[|h(x) - y_S(x)|]. \tag{3}$$

where $\boldsymbol{E_{x \sim X_S}}$ is the expectation operator when x is drawn from distribution $f_S(x)$. Equation.(3) can be similarly defined for the target (or test) data. A bound on the target decision error was shown [28] as,

$$\begin{aligned}\epsilon_T(h, y_T) \leq \epsilon_S(h, y_S) &+ d_1(X_S, X_T) \\ &+ min\{\boldsymbol{E_{x \sim X_S}}[|y_T(x) - y_S(x)|], \\ &\boldsymbol{E_{x \sim X_T}}[|y_S(x) - y_T(x)|]\},\end{aligned} \tag{4}$$

where $d_1(X_S, X_T) = \int |f_S(x) - f_T(x)||h(x) - y_T(x)|dx$. Assuming that the labeling functions for the source and the target data are identical, i.e., $y_S(x) = y_T(x)$ as would follow from the covariate shift [30], yields $\boldsymbol{E_{x \sim X_T}}[|y_S(x) - y_T(x)|]$ and $\boldsymbol{E_{x \sim X_S}}[|y_T(x) - y_S(x)|]$ equal to zero.

In the following, we elaborate how the Henze-Penrose metric may be used to characterize the test data classification accuracy using the accuracy lower bound calculated on the training data. To do so, we use an auxiliary distance measure introduced in [31], which can be considered as a modified Henze Penrose metric demonstrated in Eqn. (1),

$$\begin{aligned}&\tilde{D_a}(p, q) \\ &= 1 - 2(\frac{R_{m,n} - 1}{m+n}) \longrightarrow 1 - 4ab \int \frac{p(x)q(x)}{ap(x) + bq(x)} dx,\end{aligned} \tag{5}$$

From Eqn.(1), it can be concluded that,

$$\tilde{D_a}(p, q) = 2HP + \frac{2}{m+n} - 1. \tag{6}$$

Proceeding as above, and given a hypothesis h, the target error $\epsilon_T(h, y_T)$ can be bounded by the source error $\epsilon_S(h, y_S)$, the difference between the labels, and a distance measure $\tilde{D}$ between the source and target distributions,

$$\begin{aligned}\epsilon_T(h, y_T) \leq \epsilon_S(h, y_S) &+ \boldsymbol{E_{x \sim X_S}}[|y_S(x) - y_T(x)|] \\ &- 2\sqrt{\tilde{D_{\frac{1}{2}}}(f_S, f_T)},\end{aligned} \tag{7}$$

where $\tilde{D_{\frac{1}{2}}}(f_S, f_T)$ assumes equiprobable data from the source and target distributions (i.e a and $b = \frac{1}{2}$). Eqn.(7) can be re-written as,

$$\begin{aligned}\epsilon_T(h, y_T) \leq \epsilon_S(h, y_S) &+ \boldsymbol{E_{x \sim X_S}}[|y_S(x) - y_T(x)|] \\ &- 2\sqrt{2HP(f_S, f_T) + \frac{2}{m+n} - 1}.\end{aligned} \tag{8}$$

If we, again, assume that the labeling functions for the source and the target are the same, the bound in Eqn. (8) reduces to:

$$\epsilon_T(h, y_T) \leq \epsilon_S(h, y_S) - 2\sqrt{2HP(f_S, f_T) + \frac{2}{m+n} - 1}. \tag{9}$$

Moreover, if the hypothesis h follows the NN error ϵ_{NN}, we can use the results from Theorem 1 to rewrite the bound in Eqn.(9) as,

$$\begin{aligned}\epsilon_T(h, y_T) \leq 2(1 &- HP(p_S, q_S) + \frac{1}{m+n} \\ &- \sqrt{2HP(f_S, f_T) + \frac{2}{m+n} - 1}),\end{aligned} \tag{10}$$

where $HP(p_S, q_S)$ is the distance between the two classes of interest in the source data domain, while $HP(f_S, f_T)$ is the distance between the source and target data. From the above equation, we may conclude that the classification error of the test (or target) data can be upper bounded using the training (or source) data even with the possibility that the training and the test data might be drawn from different distributions. In addition, using the Henze-Penrose metric would mitigate the need for prior knowledge of the test data distribution.

IV. Bounds on the k-Nearest Neighbor Classification Accuracy

The k-NN classifier assigns a point x to a particular class based on a majority vote among the classes of the k nearest training points to x. Error rates for k-NN classifiers have been extensively studied in [32]. Some of these results will be adopted in Section 5. The Henze-Penrose measure can be further exploited to bound the k-NN accuracy for the cases where k > 1 by following a different approach than the one employed in Section 3. Denote by w the total number of wrong votes for all nodes in the graph, where a wrong vote on a node among the k closest neighbors, implies its connection to a node that belongs to a wrong class. In addition, we assume no relation between the MST and the k-NN graph as previously done for $k = 1$. Moreover, we consider D as the edit distance [33] between the MST and k-NN graph, i.e. the total number of edges in the k-NN graph and absent in the MST together with the number of edges in the MST but absent in the k-NN graph. If Ed_{kNN} represents the set of edges in k-NN and Ed_{MST} is the set of edges in MST, the edit distance D can be described as follows,

$$D = |(Ed_{kNN} \cup Ed_{MST}) - (Ed_{kNN} \cap Ed_{MST})|.$$

Theorem 2: Let labeled classes of m and n points form a complete graph. Given the Henze-Penrose metric HP between the two classes and the edit distance D between the MST and the k-NN graph, the k-nearest neighbor classification accuracy, A_{kNN}, is bounded above and below by:

$$1 - \frac{2}{m+n}\left\lceil \frac{(1-HP)(m+n)+D-1}{\lfloor \frac{k}{2} \rfloor} \right\rceil \leq A_{kNN} \leq 1 - \frac{1}{m+n}\left\lceil \frac{(1-HP)(m+n)-D-\lfloor \frac{k}{2} \rfloor (m+n)}{\lceil \frac{k}{2} \rceil} \right\rceil. \tag{11}$$

See Appendix B for the proof of Theorem 2.

V. Other Bounds on the k-NN Accuracy

In the following, we will be showing some other bounds on the k-NN classification accuracy. Using the bounds discussed in [32], we can generalize the lower bound derived for k = 1 in Section III to obtain tighter accuracy lower bounds for the k>1 case. For all odd k and all distributions, it was proved in [32] that:

$$A_{kNN} \geq A_{NN} - \frac{1}{\sqrt{ke}}, \tag{12}$$

where $e = exp(1)$. From eqn. (2), we have:

$$A_{NN} \geq 2HP + \frac{2}{m+n} - 1, \tag{13}$$

yielding,

$$A_{kNN} \geq 2HP + \frac{2}{m+n} - \frac{1}{\sqrt{ke}} - 1. \tag{14}$$

In addition, we can use the bound that ties the Bayes classification accuracy A_B with the k-NN classification accuracy. It has been shown in [32] that for all distributions and all odd k,

$$A_{kNN} \geq A_B - \sqrt{(\frac{2(1-A_{NN})}{k})}, \tag{15}$$

Since $A_{NN} \leq A_B$, we invoke the following,

$$A_{kNN} \geq A_{NN} - \sqrt{(\frac{2(1-A_{NN})}{k})}, \tag{16}$$

which results in,

$$A_{kNN} \geq 2(HP + \frac{1}{m+n} - \sqrt{(\frac{1-HP-\frac{1}{m+n}}{k})} - \frac{1}{2}). \tag{17}$$

Moreover, we have the following for $k > 3$ [34],

$$A_{kNN} \geq 1 - (1 + \sqrt{\frac{1}{k}})(1 - A_B), \tag{18}$$

which results in,

$$A_{kNN} \geq 1 - (1 + \sqrt{\frac{1}{k}})(1 - A_{NN}), \tag{19}$$

$$A_{kNN} \geq 1 - (1 + \sqrt{\frac{1}{k}})(2 - 2HP - \frac{2}{m+n}). \tag{20}$$

From all the above, we may conclude that the three different lower bounds for the k-NN classification accuracy in Eqns. (14), (17) and (20), are all dependent on the HP metric, the number of nearest neighbor k, and the number of elements in each class.

VI. Bounds on the Bayes Error and Bhattacharyya Distance Using the Henze Penrose

Classification Bayes error is a desirable entity as it represents a good benchmark performance. Bounds on Bayes error rate based on non-parametric divergence measures were investigated in [31]. $\tilde{D_a}(p,q)$, given in Eqn. (5), was shown useful in approximating the upper and lower bounds on the Bayes error rate as follows,

$$0.5 - 0.5\sqrt{\tilde{D_a}(p,q)} \leq \epsilon_{Bayes} \leq 0.5 - 0.5\tilde{D_a}(p,q), \tag{21}$$

which may, equivalently, be expressed as,

$$0.5 - 0.5\sqrt{2HP + \frac{2}{m+n} - 1} \leq \epsilon_{Bayes} \leq 1 - HP - \frac{1}{m+n}. \tag{22}$$

This result is of interest in our paper, as by merely comparing the upper bound on the k-NN error ϵ_{NN} for k = 1 (or equivalently the lower bound on the NN classification accuracy A_{NN}) given by Eqn. (2),

$$\epsilon_{NN} \leq 2 - 2HP - \frac{2}{m+n},$$

with the upper bound for the Bayes error in Eqn. (22),

$$\epsilon_{Bayes} \leq 1 - HP - \frac{1}{m+n}.$$

From the above equations, we may then conclude that the bounds derived in this paper are consistent with the results in [31], namely,

$$\epsilon_{NN} \leq 2\epsilon_{Bayes}, \tag{23}$$

where $\epsilon_{Bayes} \rightarrow 0$, $\epsilon_{NN} \approx 2\epsilon_{Bayes}$.

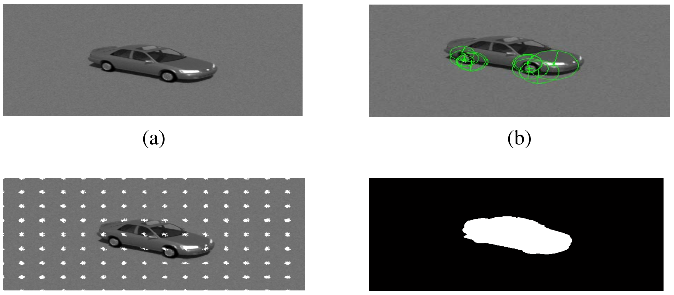
VII. Experimental Validation

We demonstrated the derived bounds on a synthetic dataset. The synthetic imagery data-set we use, consists of a variety of vehicles, some with high variability and others with high similarity. It is comprised of 7056 images for fourteen different vehicles; ten of them are civilian vehicles and the other four are military vehicles. Six of the civilian vehicles are sedans,

TABLE I

THE DATASET DESCRIPTION

	Number of vehicles	Number of images
Class 1	4 sedans	2016
Class 2	2 SUVs and 2 trucks	2016
Class 3	4 military vehicles	2016



(a) (b) (c) (d)

Fig. 2. The different adopted features. (a) Originl imaage. (b) SURF features. (c) HoG features. (d) Mask.

two are sport utility vehicles (SUVs), one is a minivan and the other one is a pickup truck. Two of the military vehicles are treaded tanks, and the other two are armored carriers with wheels. We divided our dataset into three classes as discussed in Table I.

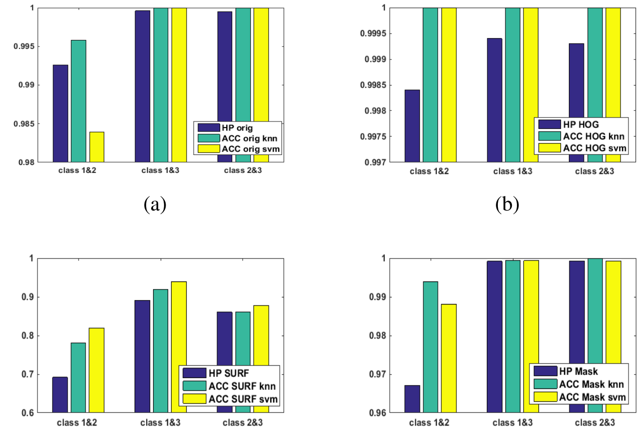
The images for each vehicle were collected at seven different elevations and seventy two different azimuth values such that each elevation level has 72 different viewpoints to represent the same vehicle. This resulted in 504 images for each vehicle. All the images were converted to gray-scale values and cropped to 76×76 pixels to increase the computational efficiency. The images were vectorized, $x_i = \text{vectorize}(Image_i)$, and the Euclidean distance, $d(x_i, x_j) = \|x_i - x_j\|_2$, was used as a proximity measure for all the experiments.

We subsequently applied some common feature extraction techniques like Speeded Up Robust Features (SURF) [35] and Histogram of Gradients (HoG) [36], to assess their effectiveness in preserving separation between each pair of classes in our dataset. We also computed the gradient mask (or the silhouette) for the vehicles through contrasting the vehicle from the background. Changes in contrast can be detected by operators that calculate the gradient of the image. Furthermore, a threshold was applied to create a binary mask containing the segmented vehicle after filling the interior gaps inside the vehicle as shown in Fig. (2d). The adopted features are shown in Figs. (2a-2d). In Fig. (2b), the SURF features are displayed as green circles centered on the key feature and the diameter of the circle indicates the extent of the feature histogram at the key scale while in Fig. (2c), HoG features are visualized using a grid of uniformly spaced rose plots. The cell size and the size of the image determines the grid dimensions. Each rose plot shows the distribution of gradient orientations within a HoG cell. The length of each petal of the rose plot is scaled to indicate the contribution each orientation makes within the cell histogram. As previously stated in Section II, the Henze-Penrose metric estimates the distributional overlap between two distributions, implicitly accounting for the

TABLE II

THE HENZE-PENROSE METRIC VALUES FOR THE DIFFERENT FEATURE SPACES

	Orig. data	SURF	HoG	Mask
HP_{12}	0.9926	0.6921	0.9984	0.9671
HP_{13}	0.9996	0.8915	0.9994	0.9992
HP_{23}	0.9995	0.8606	0.9993	0.9993



(a) (b) (c) (d)

Fig. 3. The Henze-Penrose metric and the classification accuracy for SVM and kNN classifiers computed between each pair of classes using original data, HoG, SURF and Mask respectively. (a) Original dataset. (b) HoG. (c) SURF. (d) Mask.

marginal distributions when given a distance measure between the data points from the two classes. This hence obviates the explicit computation of the distributions, instead, allowing us the direct use of the observed data. Levels of separation between each pair of classes, represented by the HP metric, for different feature spaces are depicted in Table II.

A. Experimental Results

We evaluate the inter-class separability versus the classification accuracy using two classification methods: the k-NN classifier ($k = 1$) [37] and the Support Vector Machine (SVM) classifier [38]. A representative one-third of the dataset was used to train each classifier, and the rest of the data was used to carry out the testing. The accuracy of classification for each pair of classes was computed using the previously mentioned classifiers along with the inter-class HP metric. The results are shown in Figs. (3a-3d). As expected; the accuracy for both classifiers increases as the HP value increases. This is intuitive since as the separation between different classes increases, it becomes easier for the classifier to efficiently perform the discrimination task. Furthermore, the HP values and the NN classification accuracy nicely track, to coincide with the bounds derived in Section III.

Looking at Fig. (3), we observe that the HoG descriptor, which gives special attention to geometric features, outperformed the other feature extraction techniques. The mask (or the silhouette), which only preserves the most general geometric features performed slightly worse than the original dataset. The SURF descriptor, which focuses more on regional features such as texture, was the worst performing feature extraction technique. This is reasonable because in our dataset

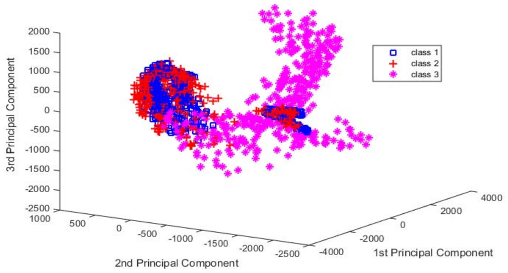
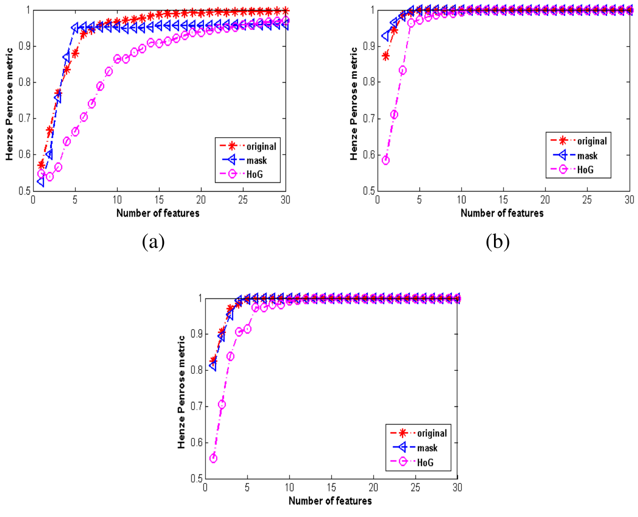


Fig. 4. The three-dimesional embedding for the original dataset using PCA.



(a)

(b)

(c)

Fig. 5. The Henze-Penrose metric for each pair of classes as a function of the dimensionality for the different feature spaces. (a) Class 1 and 2. (b) Class 1 and 3. (c) Class 2 and 3.

geometric information is better suited than texture information for vehicle classification.

We next consider applying Principal Component Analysis (PCA) [39], a well-known linear dimension reduction technique, to the different feature spaces. The three dimensional embedding for the original dataset is shown in Fig. (4). We propose using the bounds derived in Section 3 to predict the number of features (or principle components) that would achieve a desired classification performance. Figs. (5a-5c) display the HP score for each pair of classes as a function of the number of principal components for the original dataset, HoG and mask. For a NN accuracy of at least 0.95, we require a HP value greater than 0.94975 between each pair of classes. We selected the minimum number of features that would achieve our desired HP metric values. After choosing the number of principal components that would represent each image, we applied the NN classifier on the reduced-dimension dataset to compute the actual classification accuracy. The results are shown in Table III. We may conclude that using the dataset in original sensed space is more efficient than the extracted features since it involves less number of PCA components along with higher actual accuracy. Moreover, we considered fixing the number of features obtained for the different pre-processing techniques we used to evaluate the efficacy of each one of them fairly with respect to the dimensionality of the feature space. The results are outlined in Table IV. From the results, we can conclude that using the original dataset with no prior preprocessing, again, is better than extracting the HoG or mask descriptors. In Figs. (6a-6c), we show the tightness of the predicted lower bound in Eqn. (2) versus the actual NN classification accuracy as a function of the number of features for our dataset in the original sensed space. It is obvious from the figures that the tightness of the lower bound increases as we increase the number of the PCA components. In addition, they almost overlap after certain number of components which demonstrates the effectiveness

TABLE III

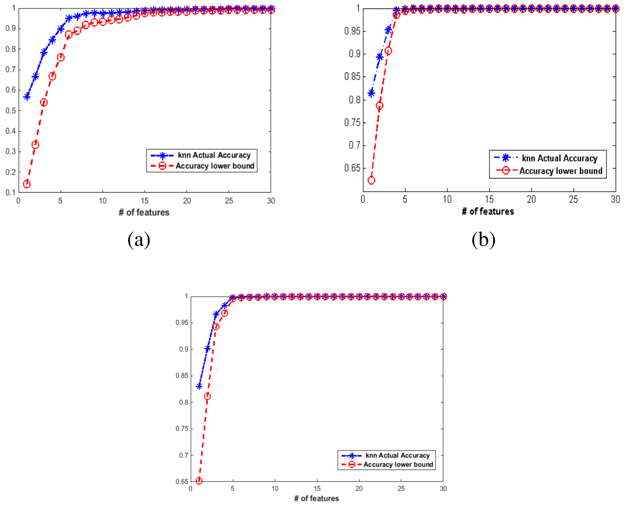
RESULTS FOR NN CLASSIFICATION ACCURACY VERSUS THE PREDICTED LOWER BOUND FOR ACCEPTABLE CLASSIFIER PERFORMANCE OF AT LEAST 95%

	No. of features	HP	Acc. low. bound	Actual accuracy.
Orig_{12}	10	0.9752	95.09%	98.66%
HoG_{12}	31	0.9762	95.29%	99.06%
Mask_{12}	132	0.9750	95.04%	99.13%
Orig_{13}	3	0.9864	97.32%	98.83%
HoG_{13}	5	0.9792	95.88%	98.46%
Mask_{13}	3	0.9824	96.53%	98.39%
Orig_{23}	4	0.9836	96.78%	98.36%
HoG_{23}	7	0.9750	95.04%	98.14%
Mask_{23}	4	0.9931	98.66%	99.50%

TABLE IV

RESULTS FOR NN CLASSIFICATION ACCURACY VERSUS THE PREDICTED LOWER BOUND FOR A FIXED NUMBER OF FEATURES

	No. of features	HP	Acc. low. bound	Actual accuracy.
Orig_{12}	10	0.9752	95.09%	98.66%
HoG_{12}	10	0.8864	77.33%	92.68%
Mask_{12}	10	0.9559	91.22%	97.50%
Orig_{13}	10	0.9995	99.95%	100%
HoG_{13}	10	0.9938	98.81%	99.63%
Mask_{13}	10	0.9993	99.90%	99.98%
Orig_{23}	10	0.9995	99.95%	100%
HoG_{23}	10	0.9926	98.56%	99.50%
Mask_{23}	10	0.9993	99.90%	99.95%

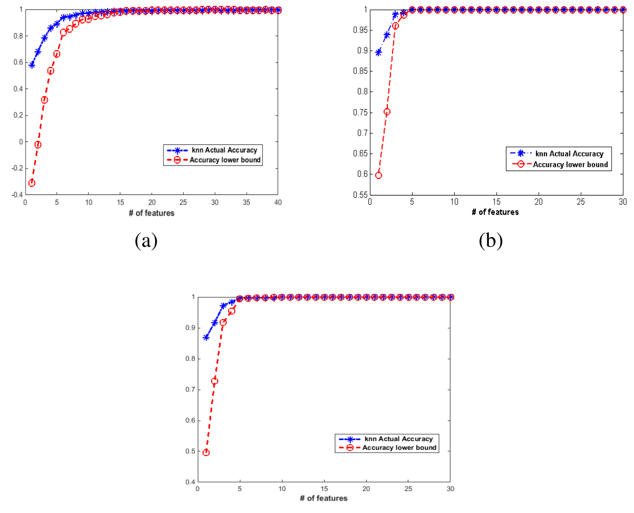


(a)

(b)

(c)

Fig. 6. Results showing the tightness of the lower bound for the case when k = 1, where the x-axis represents the number of features and the y-axis shows the kNN actual accuracy versus the lower bound. (a) Class 1 and 2. (b) Class 1 and 3. (c) Class 2 and 3.



(a) (b)

(c)

Fig. 7. The achieved accuracy versus the predicted lower bound for k = 5, where the x-axis represents the number of features needed to achieve the corresponding accuracy. (a) Class 1 and 2. (b) Class 1 and 3. (c) Class 2 and 3.

TABLE V

RESULTS FOR k-NN CLASSIFICATION ACCURACY VERSUS THE PREDICTED LOWER BOUND FOR ACCEPTABLE CLASSIFIER PERFORMANCE OF AT LEAST 95% AND FOR k = 5

	No. of features	HP	Acc. low. bound	Actual accuracy.
Orig_{12}	12	0.9831	95.19%	98.19%
HoG_{12}	38	0.9829	95.12%	98.51%
Mask_{12}	inf	–	–	–
Orig_{13}	3	0.9864	96.12%	98.78%
HoG_{13}	7	0.9834	95.26%	98.69%
Mask_{13}	4	0.9970	99.21%	99.75%
Orig_{23}	4	0.9836	95.33%	98.41%
HoG_{23}	9	0.9831	95.19%	98.29%
Mask_{23}	4	0.9931	98.06%	99.38%

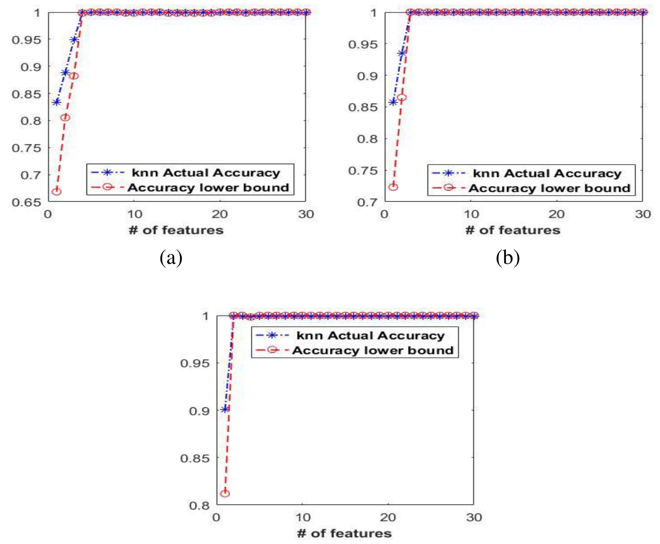
TABLE VI

RESULTS FOR k-NN ACTUAL ACCURACY VERSUS THE PREDICTED LOWER BOUND FOR A FIXED NUMBER OF FEATURES AND FOR k = 5

	No. of features	HP	Acc. low. bound	Actual accuracy.
Orig_{12}	10	0.9752	92.89%	97.17%
HoG_{12}	10	0.8864	67.19%	91.82%
Mask_{12}	10	0.9559	87.29%	96.90%
Orig_{13}	10	0.9995	99.93%	100%
HoG_{13}	10	0.9938	98.28%	99.38%
Mask_{13}	10	0.9993	99.86%	99.98%
Orig_{23}	10	0.9995	99.93%	100%
HoG_{23}	10	0.9926	97.92%	99.26%
Mask_{23}	10	0.9993	99.86%	99.95%

and the reliability of the lower bound in predicting the actual classification accuracy. Besides, we performed the same experiment for $k = 5$ using the bound in Eqn. (20) and the results are shown in Figs. (7a-7c).

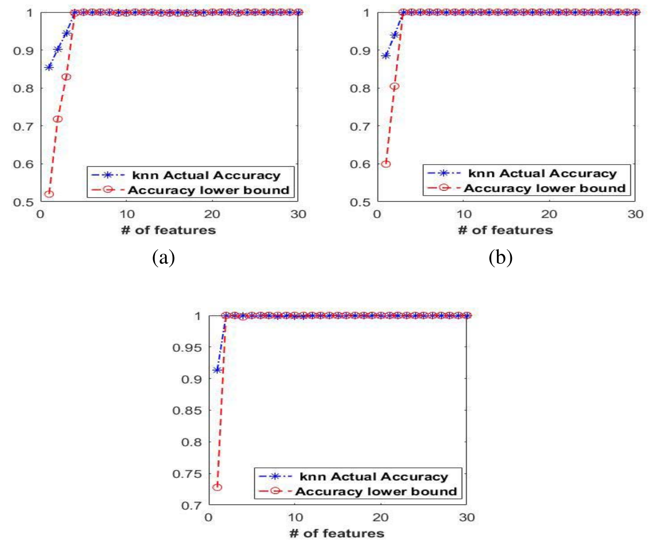
Furthermore, we use the bound in Eqn. (20) to evaluate the number of features that would retain a favorable classification performance of at least 95%, much like the experiment that was carried out for the case of k = 1. The numerical results are depicted in Table (V). We next fix the number of features obtained for the different pre-processing techniques. The results are outlined in Table VI. From the results, again, we can conclude that using the original dataset without any prior preprocessing is better than extracting the HoG or mask descriptors for our dataset.



(a) (b)

(c)

Fig. 8. The achieved accuracy versus the predicted lower bound for k = 1. (a) Class 1 and 2. (b) Class 1 and 3. (c) Class 2 and 3.



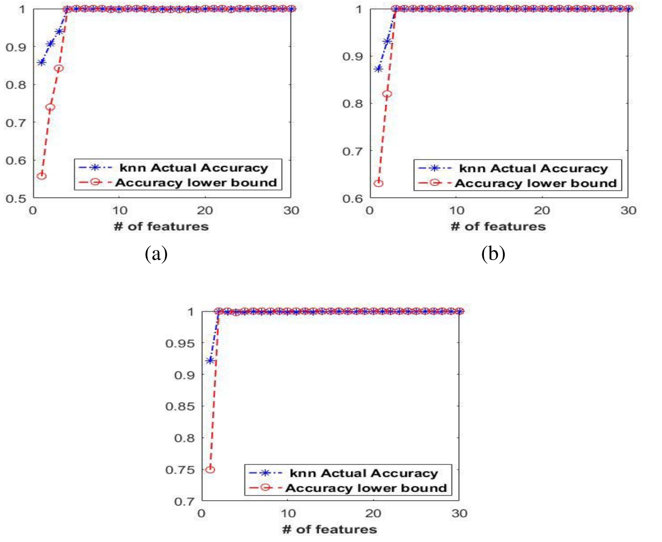
(a) (b)

(c)

Fig. 9. The achieved accuracy versus the predicted lower bound for k = 5. (a) Class 1 and 2. (b) Class 1 and 3. (c) Class 2 and 3.

B. Using Extended Yale Face Database

To further substantiate our proposed approach, we apply our bounds to the data from the Extended Yale Face Database B [40]. The Extended Yale Face Database B contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions. We used three sets of images that belong to three different people. Images belonging to one person were considered as one class. Each class consists of 585 different images representing the face in different illumination conditions with some changes in the background. The images were downsampled to 240 × 320 pixels and then vectorized. We next applied PCA to the data in the original



(a) (b) (c)

Fig. 10. The achieved accuracy versus the predicted lower bound for k = 9. (a) Class 1 and 2. (b) Class 1 and 3. (c) Class 2 and 3.

sensed space. We used the lower bounds in Eqs. (2) and (20) to compare them against the actual k-NN classification accuracy as a function of the number of principal components for the dataset in the original sensed space. The results are shown in Figs. (8a-8c), (9a-9c) and (10a-10c) for k = 1, 5 and 9 respectively. Similar to the previous experiments, we can see that the tightness of the lower bound increases as the number of the PCA components increases. Moreover, they overlap past a number of a few number of components, hence demonstrating the fidelity of estimating the actual classification accuracy using our derived bounds for k $\geq$ 1.

VIII. Conclusion

In this paper, we derived new bounds on the k-NN classifier performance using a non parametric distribution free metric, known as the Henze-Penrose metric. The metric was exploited to quantify the inherent separation between classes from a labeled set of high-dimensional synthetic vehicle imagery data. The derived bounds for k $\geq$ 1 allow to incorporate prior knowledge regarding the performance of the k-NN classifier assuming no prior information about the class distributions and the estimation of their PDFs. The predicted bounds on the labeled training data have been used to bound the classification accuracy on the unlabeled test data. In short, distribution-free metrics such as Henze-Penrose and Friedman-Rafsky provide an avenue to quantify the efficiency of a particular dataset with regards to its discrimination capability, which also affords the minimum number of reduced dimensions required to maintain a desired discrimination that is comparable to the original dataset.

Appendix A
Proof of Theorem 1

Proof: Suppose two classes of m and n points form a complete graph of unique distances. Therefore, the minimum spanning tree is unique and the NN graph will be a subset of the MST. We proceed by first relating the number of misclassified points, E, the number of edges which are present in the MST but not in the NNG, $C-1$, and the number of edges connecting different classes in the MST, $R_{m,n}-1$. Each edge connecting different classes in the MST causes at most two classification errors, and each edge connecting different classes in the MST either does not exist in the NNG or causes at least one classification error,

$$\frac{E}{2} \leq R_{m,n} - 1 \leq C - 1 + E.$$

We next prove the lower bound by writing:

$$\begin{aligned} A_{NN} &= 1 - \frac{E}{m+n}, \\ &\geq 1 - \frac{2(R_{m,n}-1)}{m+n}, \\ &\geq 1 - \frac{2((1-HP)(m+n)-1)}{m+n}, \\ &\geq 2HP + \frac{2}{m+n} - 1. \end{aligned}$$

Lastly, we prove the upper bound by proceeding as:

$$\begin{aligned} A_{NN} &= 1 - \frac{E}{m+n}, \\ &\leq 1 - \frac{R_{m,n}-C}{m+n}, \\ &\leq 1 - \frac{(1-HP)(m+n)-C}{m+n}, \\ &\leq HP + \frac{C}{m+n}. \end{aligned}$$

□

Appendix B
Proof of Theorem 2

Proof: We first establish a relationship between the number of wrong votes, w, the edit distance, D, and the number of edges connecting different classes in the MST, $R_{m,n}-1$. Each edge connecting different classes in the MST causes at most two wrong votes and can either be included among the D edges or causes at least one wrong vote,

$$\frac{w}{2} - D \leq R_{m,n} - 1 \leq w + D - 1.$$

In addition, the number of misclassification errors will be bounded above and below by:

$$\left\lceil \frac{max((w - \lfloor \frac{k}{2} \rfloor (m+n)), 0)}{\lceil \frac{k}{2} \rceil} \right\rceil \leq E \leq \left\lceil \frac{w}{\lceil \frac{k}{2} \rceil} \right\rceil.$$

We next prove the lower bound by writing,

$$\begin{aligned} A_{NN} &= 1 - \frac{E}{m+n}, \\ &\geq 1 - \frac{1}{m+n} \left\lceil \frac{w}{\lceil \frac{k}{2} \rceil} \right\rceil, \\ &\geq 1 - \frac{2}{m+n} \left\lceil \frac{(1-HP)(m+n)+D-1}{\lceil \frac{k}{2} \rceil} \right\rceil. \end{aligned}$$

Lastly, we prove the upper bound by proceeding as:

$$
\begin{aligned}
&A_{NN}\\
&= 1 - \frac{E}{m+n},\\
&\leq 1 - \frac{1}{m+n}\left\lceil \frac{max((w - \lfloor \frac{k}{2} \rfloor (m+n)), 0)}{\lceil \frac{k}{2} \rceil} \right\rceil,\\
&\leq 1 - \frac{1}{m+n}\left\lceil \frac{max((m+n)((1-HP) - \lfloor \frac{k}{2} \rfloor) - D, 0)}{\lceil \frac{k}{2} \rceil} \right\rceil.
\end{aligned}
$$

□

Acknowledgment

The first author would like to thank Dr. Erik Skau for his helpful conversations and comments about this work.

References

[1] A. B. Hamza and H. Krim, "Image registration and segmentation by maximizing the Jensen–Rényi divergence," in *Proc. Int. Workshop Energy Minimization Methods Comput. Vis. Pattern Recognit.* Berlin, Germany: Springer, 2003, pp. 147–163.

[2] B. Ma, A. Hero, J. Gorman, and O. Michel, "Image registration with minimum spanning tree algorithm," in *Proc. Int. Conf. Image Process.*, Sep. 2000, pp. 481–484.

[3] H. Neemuchwala, A. Hero, S. Zabuawala, and P. Carson, "Image registration methods in high-dimensional space," *Int. J. Imag. Syst. Technol.*, vol. 16, no. 5, pp. 130–145, 2006.

[4] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 267–272.

[5] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

[6] A. Wisler, V. Berisha, J. Liss, and A. Spanias, "Domain invariant speech features using a new divergence measure," in *Proc. Spoken Lang. Technol. Workshop (SLT)*, Dec. 2014, pp. 77–82.

[7] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[8] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, in Contributions to the Theory of Statistics, vol. 1. Oakland, CA, USA: Regents Univ. California, 1961, pp. 547–561.

[9] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[10] Y. He, A. B. Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1211–1220, May 2003.

[11] D. Stowell and M. D. Plumbley, "Fast multidimensional entropy estimation by k-d partitioning," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 537–540, Jun. 2009.

[12] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. Math. Phys.*, vol. 7, no. 3, pp. 200–217, 1967.

[13] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests," *Ann. Statist.*, vol. 7, no. 4, pp. 697–717, 1979.

[14] A. Wald and J. Wolfowitz, "On a test whether two samples are from the same population," *Ann. Math. Statist.*, vol. 11, no. 2, pp. 147–162, 1940.

[15] N. Henze and M. D. Penrose, "On the multivariate runs test," *Ann. Statist.*, vol. 27, no. 1, pp. 290–298, 1999.

[16] D. Eppstein, M. S. Paterson, and F. F. Yao, "On nearest-neighbor graphs," *Discrete Comput. Geometry*, vol. 17, no. 3, pp. 263–282, 1997.

[17] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, no. 4, pp. 493–507, 1952.

[18] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhyā, Indian J. Statist.*, vol. 7, no. 4, pp. 401–406, 1946.

[19] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, no. 1, pp. 52–60, Feb. 1967.

[20] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Alpha-divergence for classification, indexing and retrieval (revised)," Commun. Signal Process. Lab., Univ. Michigan, Ann Arbor, MA, USA, Tech. Rep. CSPL-328, 2001.

[21] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungarica*, vol. 2, pp. 299–318, Jan. 1967.

[22] S. Kullback, "A lower bound for discrimination information in terms of variation (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 126–127, Jan. 1967.

[23] W. A. Hashlamoun, P. K. Varshney, and V. N. S. Samarasooriya, "A tight upper bound on the Bayesian probability of error," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 2, pp. 220–224, Feb. 1994.

[24] H. Avi-Itzhak and T. Diep, "Arbitrarily tight upper and lower bounds on the Bayesian probability of error," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 1, pp. 89–91, Jan. 1996.

[25] K. Sricharan, R. Raich, and A. O. Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4135–4159, Jul. 2012.

[26] X. Nguyen, M. J. Wainwright, and M. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5847–5861, Nov. 2010.

[27] V. Berisha and A. O. Hero, "Empirical non-parametric estimation of the Fisher information," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 988–992, Jul. 2015.

[28] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.

[29] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.

[30] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Statist. Planning Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[31] V. Berisha, A. Wisler, A. O. Hero, and A. Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 580–591, Feb. 2016.

[32] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31. Springer, 2013.

[33] S. Baloch and H. Krim, "Object recognition through topo-geometric shape models using error-tolerant subgraph isomorphisms," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1191–1200, May 2010.

[34] L. Devroye, "On the asymptotic probability of error in nonparametric discrimination," *Ann. Statist.*, vol. 9, no. 6, pp. 1320–1327, 1981.

[35] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[37] D. Coomans and D. L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. K-nearest neighbour classification by using alternative voting rules," *Anal. Chim. Acta*, vol. 136, pp. 15–27, Jan. 1982.

[38] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[39] F. R. S. K. Pearson, "On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.

[40] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

Sally Ghanem was born in Raleigh, NC, USA. She received the B.Sc. degree in electrical engineering from Alexandria University in 2013 and the M.Sc. degree in electrical and computer engineering from North Carolina State University in 2016, where she is currently pursuing the Ph.D. degree. Her research interests include the areas of computer vision, digital signal processing, video, and image processing.

Hamid Krim received the B.Sc. degree from University of Southern California and University of Washington in 1979, the M.Sc. degree from University of Washington in 1980, and the Ph.D. degree from Northeastern University in 1991, all in electrical engineering. He was a Technical Staff Member with AT&T Bell Labs, where he conducted research and development in the areas of telephony and digital communication systems/subsystems. He was an NSF Post-Doctoral Fellow with the Foreign Centers of Excellence, LSS/University of Orsay, Paris, France. He was a Research Scientist with the Laboratory for Information and Decision Systems, MIT, Cambridge, MA, USA, where he was performing and supervising research. He is currently a Professor of electrical engineering with the Electrical and Computer Engineering Department, North Carolina State University, Raleigh, where he leads the Vision, Information and Statistical Signal Theories and Applications Group. His research interests are in statistical signal and image analysis and mathematical modeling with a keen emphasis on applied problems in classification and recognition using geometric and topological tools. He served for the SP Society Editorial Board and TCs. He was an SP Distinguished Lecturer from 2015 to 2016.

Hamilton Scott Clouse has worked with partners in industry, academia, and both public and private research, for over a decade, to provide innovative and state-of-the-art automation and data science solutions for a host of complex challenges. He currently leads the Autonomy Technical Team at AFRL and he is a faculty member with Wright State University. Fervor for research in artificial intelligence and data science, bolstered by formal training in machine learning and data analysis, drives him to stay at the cutting edge of these fields through frequent presentation, publication, and participation in the global research community as a member of AAAI, CVF, IEEE, SIAM, and SPIE.

Wesam Sakla received the Ph.D. degree in electrical engineering from Texas A&M University. He is a Computer Vision Scientist at Lawrence Livermore National Laboratory. He is currently working for the Global Security Directorate at LLNL, where he specializes in applying the state of-the-art deep machine learning algorithms to multimodal imagery acquired from aircraft and satellite imagery. His current research interests include the use of deep convolutional neural networks for automated detection and localization, recognition, and classification applications.