Expert Systems With Applications 115 (2019) 276–286

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Vehicle engine classification using normalized tone-pitch indexing and neural computing on short remote vibration sensing data[☆]

Jie Wei[a,*], Chi-Him Liu[a], Zhigang Zhu[a], Lindsay R. Cain[b], Vincent J. Velten[b]

[a] Department of Computer Science, The City College of New York, 160 Convent Ave., New York, NY 10031, United States of America
[b] Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, United States of America

ARTICLE INFO

Article history:
Received 18 August 2017
Revised 30 July 2018
Accepted 31 July 2018
Available online 1 August 2018

Keywords:
Remote sensing
Fourier transforms
Band-pass filters
Indexing
Classification algorithms
Machine learning
Neural networks

ABSTRACT

As a non-invasive and remote sensor, a Laser Doppler Vibrometer (LDV) has found a broad spectrum of applications. It is a remote, non-line-of-sight sensor to detect threats more reliably and provide increased security protection, which is of utmost importance to military and law enforcement applications. However, the use of the LDV in situation surveillance, especially in vehicle classification, lacks systematic investigations as to its phenomenological and statistical properties. In this work, we aim to identify vehicles by their engine types within a very short period of time to yield a practical expert and intelligent system to classify vehicle engines remotely using laser sensors. Based on our preliminary success on the use of tone-pitch indexes (TPI) over these data, a new normalized tone-pitch indexing (nTPI) scheme is developed to capture engine periodic vibrations by various engine types with vibration data over a much shorter period (from 1.25 to 0.2 s), which makes it possible to monitor slowly moving vehicles around 15 miles per hour. We also exploit the learning power of neural computing, including artificial neural network (ANN), Deep Belief nets (DBN), Stacked Auto-Encoder (SAE), and Convolutional Neural Networks (CNN). To apply a CNN, a two-dimensional array is formulated by stacking nTPI data in an overlapping manner, which is termed as 2DonTPI. The classification results using the proposed nTPI and 2DonTPI over a standard LDV dataset are promising: with encoding duration significantly smaller than that required by the original TPI, consistently high performance is attained for all four neural computing methods. The new vibration data representation combined with neural computing approaches gives rise to a *powerful expert and intelligent system* for vehicle engine classification, which can find a great array of applications for civil, law enforcement, and military agencies for Intelligence, Surveillance and Reconnaissance purposes that are of crucial importance to national and international security.

1. Introduction

THE use of Laser Doppler Vibrometer (LDV) as an important remote sensing instrument has been consistently growing in recent years due to the unique advantages it can provide. An LDV sensor is an active sensor based on sending and receiving laser beam signals: it works by first sending out a laser beam b to a targeted reflective surface S; if S is adequately smooth and retro-reflective, b is then reflected back by S and received by the LDV as b'. The spatial and spectral properties of *the surface's* vibrations can thus be extracted from the time difference and Doppler shift of the received b' from the out-going b. LDV sensors provide many advantages:

1. Non-invasive measurements: No mass or pressure is applied during the LDV measurement process, and the laser in our LDV is eye-safe: only some high-power types may cause damage to human eyes if viewed directly for a significant time. As an example, LDV exerts no additional pain in non-invasive medical applications such as body temperature and pulse monitoring during medical operations. By contrast, the small dose of radiation from CT or X-ray in medical applications can damage human cells; even the supposedly safer ultra-sound needs long time contact and special jellies that are inapplicable for skin burns or delicate vital sign measurements of organs during surgeries. Additionally, LDV causes no extra damage in non-intrusive civil engineering applications such as inspections of bridges, railways and buildings (Kubota, 2007; Willemann, Castellini, Revel, & Tomasini, 2004) whereas in typical ultra-

[☆] This work was supported in part by the NSF/Center for Surveillance Research under Grant No. IIP-0968922. This document was approved by Air Force Research Laboratory for public release via 88ABW-2016-0774.

[*] Corresponding author.

E-mail addresses: wei@cs.ccny.cuny.edu (J. Wei), cliu06@citymail.cuny.edu (C.-H. Liu), zhu@cs.ccny.cuny.edu (Z. Zhu), Lindsay.cain.1@us.af.mil (L.R. Cain), vincent.velten@us.af.mil (V.J. Velten).

https://doi.org/10.1016/j.eswa.2018.07.073
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

sonic tests water penetration and/or corrosion induce problematic side effects. Lastly, LDV has been used for delicate inspections of murals and antique fresco paintings in museums where LDV is the only viable means of inspection (Castellini, Paone, & Tomasini, 1996).

2. High spatial and spectral resolution: the expansive and wide range of amplitudes and frequencies offered by LDV sensors give researchers and developers valuable information in both spatial and frequency domains for intensive analysis, classification and clustering. For instance, the sampling rate for a typical LDV is up to 100 KHz whereas the measurable vibrations can be as short as less than five nanometers to several micrometers. The velocity and acceleration evaluated from these sensors are of desirable quality in space and frequency domains from microscopy analysis in biomedical, biological, and medical studies for remote situation surveillance and structural inspections. In Watson, Rhoads, and Adams (2013), the LDV was effectively employed to remotely detect bombs using actively controlled signals, e.g., chirp sounds. The surface vibrations of the suspected objects in response to the well-controlled sound are collected by the LDV situated at a safe distance, as far as 100 m, and the rich spatial and spectral contents are then analyzed to determine if the object is indeed a bomb; the initial results reported are promising. This safe method of bomb detection can replace the robots or dogs currently used by police or other security agencies, and are far more cost effective.

1.1. Use of LDV in intelligent vehicle classification

The use of LDV in many research and development subjects has become increasingly more popular due precisely to the foregoing advantages. As a non-invasive and remote sensor with the afore-mentioned benefits, LDV measured data are an ideal modality/phenomenology to detect potential threats more reliably and provide increased protection to society, which is of utmost importance to military and law enforcement institutions. Our task in this work is to identify the engine type of a vehicle from the vibration signals measured by an LDV sensor, *invariant to various nuisance conditions*: from any part of the vehicle, for any environment including weather, under any operating actions, and in as small of a period as possible. However, systematic investigations of LDV behavioral properties and effective exploitation techniques are still needed before LDV can be deployed in the field. Since summer 2014, the three authors from the City College of New York (CCNY) have been collaborating closely with the Sensor Directorate of the Air Force Research Laboratory (AFRL) at Wright Patterson Air Force Base to tackle this problem.

In Wei, Vongsy, Mendoza-Schrock, and Liu (2014), based on the 2014 collaborations between CCNY and AFRL, Wei and colleagues developed a new vibration tone-pitch index (TPI) to represent different vehicle engines, a feedforward artificial Neural Network was then trained to assign classification labels. This work was based on a data set, referred to as *summer-14* “standard” dataset, collected during summer 2014 over 12 different vehicles, which includes multiple modalities amidst vehicles of different types over five weeks. Many vehicles were collected more than once on different time/days and some vehicles share the same make and year differing only in serial numbers. Our measuring procedure proceeds as follows: The LDV sensor is maneuvered around the vehicle being measured to take 24 to 31 points, including the front bumper, above the front wheels, the front and back doors on both driver and passenger sides, and the back bumper. For each point, the driver was asked to exert five different operating states: *idle, sweep (idle to 3000 RPM), engage drive, 2000 RPM*, and *power on the AC*. For each state, the LDV recorded a 30 s time sequence. To ensure adequate data quality: (1) a retro-reflective tape was posted on the vehicles’ surface for points where the LDV collected data; (2) The distance from the LDV sensor to the collection point was set at approximately ten feet; (3) Sound noise was required to be at a minimum during the 30 s collection period. In total, there are four different types of engines involved in this data set: I4 (sedan), 11 L Diesel I6 (1-axle truck), 15.2 L diesel I6 (2-axle truck), and V6 (sedan), as illustrated in Fig. 1.

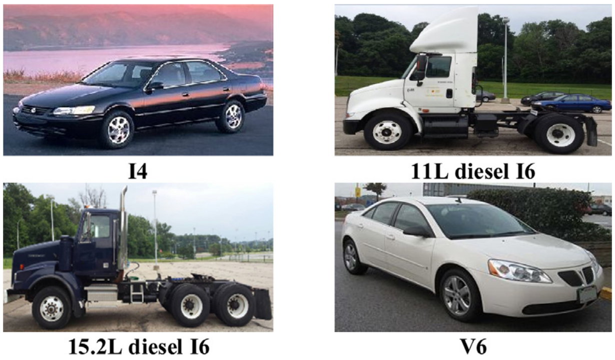

Fig. 1. The four types of vehicle engines present in the summer-14 dataset.

Based on this summer-14 dataset and the ensuing intensive analysis, the inherent features of LDV data from multiple vehicles are first studied, especially their main difference from human speech signals. The TPI scheme is then developed to capture the engine’s periodic vibrations and the associated fundamental frequencies of the vehicles’ surfaces. After extensively administering and exploring more than 10 different well-established classifiers, we discovered that a feedforward artificial neural network with 20 hidden neurons can deliver the optimal performance to classify vehicles’ engines based on the spectral tone-pitch indexes in both the cross-validation and the intensive test method. The classification results using the proposed approach over the complete LDV dataset collected by our team are exceedingly encouraging; consistently higher than 96% accuracies are attained for all four types of vehicle engines.

1.2. Problems of TPI in practical use

Despite its initial success in delivering desirable classification performances, one of the major obstacles in the way of practical use is the relatively long duration of the signal window demanded by TPI: 1.25 s, that is, for this indexing and classification approach to work, the basic encoding unit must be longer than 1 s, otherwise the classification performance drops considerably from 96% to below 80%. This 1.25 s demand turns out to be a major problem for practical use of this new methodology: if the vehicle is in motion, it is impossible for an LDV to collect reliable signals from a surface for that long.

As reported in Wei, Liu, Zhu, Mendoza-Shrock, and Vongsy (2015) we made an effort to use TPI to classify vehicles in motion. However, to ensure the 1.25 s duration demanded by this approach, the vehicle must move extremely slowly: for any vehicle, the surface that can be reliably measured is less than 4 m (generally the distance covered by the front and back doors), the 1.25 s signal duration requires that the vehicle cannot move faster than 2.8 miles/h (MPH). Furthermore, to achieve reliable classification and combat noise effects, the encoding signal should be longer than 1.25 s to allow for a majority voting scheme (to be detailed in the next section). Thus, the vehicles need to move even slower, making the TPI based moving vehicle classification entirely impractical. One possible alternative is to focus the laser beam of an LDV on the front or back plate or bumper, however this strat-

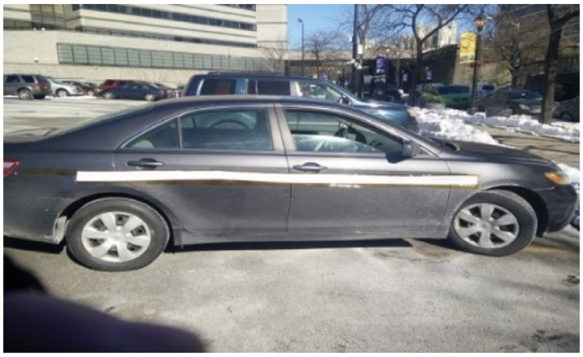

Fig. 2. Tests of moving vehicles. The long reflective tapes placed along the passenger side are needed to ensure laser beam signal quality.

egy fares no better: based on our experiments, with our LDV, if the surface displaces from the originally focused distance by more than 3 m, the laser beam emitted from the LDV must be refocused—otherwise the signal quality is too poor to be used for indexing and classification purposes. This requires the vehicle to be moving less than 3 MPH, and given that it is far more difficult to focus on the front or back bumpers in real traffic flow, this alternative is even more impractical than the lengthwise data collection wherein the laser beam focus can be easily maintained. It is thus our conclusion that absent special speed controlled zones where the vehicles are all but stationary (<2.8 MPH), the TPI based indexing and classification *per se* are inappropriate for moving vehicle cases.

As another effort to achieve vehicle classification in practical traffic flows, in Wei, Liu, Zhu, Vongsy, and Mendoza-Shrcok (2015) we attempted to utilize external objects on the road side to combat the possible counter-measures made by un-cooperative vehicles. After intensive in-lab studies from more than 20 different surfaces we found the surfaces of a steel filing cabinet yield the best signal quality that can be employed to index and classify vehicles. In Wei, Liu, Zhu, Vongsy, et al. (2015) for stationary or very slow-moving (<3 MPH) vehicles TPIs based on LDV signals collected on our cabinet surfaces planted about 10 m from the side of vehicles can produce ~90% accuracies in a consistent manner. This solves the problem of refocusing with the LDV signals directly collected from the vehicle surface as done in Wei, Liu, Zhu, Mendoza-Shrock, et al. (2015), Wei, Vongsy, et al. (2014) and Wei, Liu, and Clouse (2018) where reflective tape must be put on the vehicle surface for the LDV's laser sensor to focus on, otherwise the signals reflected from vehicle surfaces directly are almost completely saturated by noise to be of any use. As shown in Fig. 2, a long reflective tape was put on the tested car body, and special care had to be taken to ensure the focused laser beam, with diameters in several micrometers, remains inside the tape or else the LDV signals will be of no use. Use of an external object relieved this troubling burden. However, the stringent demand of extremely slow movements, i.e., <3 MPH, required by the TPI based procedure remains unsolved since measurement of the engine vibration from this external object will not be accurate if the vehicle drives away from it and may be mixed with the vibration caused by other vehicles on a crowded road. New indexes and/or classifiers need to be devised to push this line of research and development to practical use.

The major obstacle in the way of TPI's practical use in real situations is the required long data duration of 1.25 s, great effort was required to sufficiently reduce it. We discovered that by changing the original TPI to its normalized version and using the neural computing techniques, the coding duration can be reduced to as low as 0.2~0.3 s, without showing any sign of reduced classification accuracy.

There are several original contributions made in this work that play important roles to yield a *practical expert and intelligent system* to effectively classify vehicle engines remotely.

1) *Towards the first practical LDV-based engine classification approach that utilizes deep nets*. This is the first piece of work that can identify the engine type of a vehicle from the vibration signals measured by an LDV sensor, to any part of the vehicle, for any environment including weather, under more practical work conditions. The key is that the *time duration* of encoded signal window is significantly reduced from 1.25 s to 0.2 s, which makes it possible to monitor normally slowly moving vehicles in controlled regions at ~15 MPH.
2) *Novel feature extraction methods for using deep nets with small datasets.* This is particularly valuable for sensor research that utilizes specialized sensors thus collecting a large datasets is often a challenge. On this front, first a *scalable feature* vector with 120 numbers is employed to effectively represent/index the original data window which is larger by two orders of magnitude, this valuable scalability *reduces* the need of large number of training parameters thus making possible the effective development of an intelligent expert system. Second, to optimally exploit the spectral information encoded in the tone-pitch vibration index, and in the meantime facilitate deep net's data input demand, a *normalized spectral tone-pitch vibration index (nTPI)* is developed to effectively capture the rich spectral contents in the vibration signal while taking advantage of the immensely classification prowess of neural computing. The three components corresponding to the low frequency, high frequency and global variations of Fourier magnitudes, are normalized to ensure all three are adequately evaluated in the classification process, especially for deep nets, to yield promising performances. A two-dimensional overlapped nTPI (2DonTPI) is further proposed to utilize the power of CNN, where 2-D inputs are required.
3) *A systematic experimental study of the LDV data and features on various classification methods.* An extensive and rigorous *empirical study* was conducted to compare neural computing, including conventional neural network and all three variants of deep nets, with other classifiers such as k-nearest neighbors (kNN), random forest and boosting methods. It is found that our new vibration index combined with the family of neural computing approaches gives rise to a powerful *expert and intelligent system* to identify different types of engines using remote laser sensor, which can find a great array of applications for civil, law enforcement, and military agencies for Intelligence, Surveillance and Reconnaissance (ISR) purposes that are of crucial importance to national and international security.

This paper is organized as below. The LDV data used for vehicle classification and their crucial utility for engine classification are presented in the next section (Section 2). The normalized spectral tone-pitch vibration indexing (nTPI) scheme to capture the spectral and possibly fundamental frequencies on the vehicle's surfaces is detailed in the scheme section (Section 3). The deep learning based classification algorithms, namely, Deep Belief Nets (DBN), stacked auto-encoder (SAE), and CNN used in our vehicle engine categorization are described in the classification section (Section 4); 2DonTPI - 2-dimensional overlapped nTPI - is introduced to facilitate the special demands of Convolutional Neural Nets (CNN). The experimental results using the LDV data collected by our team are reported in the results section (Section 5). In Section 6, we conclude this paper with more remarks on this work and related research to be conducted in the near future.

Table 1
Summary of Existing Vehicle Engine Classification Methods.

Work	Methods	PROS	CONS
Smith et al. (2014)	Hierarchical classification using LDV	Remote hidden classification possible, LDV signals only	Low classification accuracies, very coarse engine type classification
Sigmund et al. (2012)	Auto-correlation of LDV signals to distinguish engine parameters	LDV signals only, non-invasive classification possible	Low classification accuracies to tell gas and diesel engines only
Averbuch and Hochman (2010)	Wavelet diffusion map classified using dynamic programming	High classification accuracies	Time consuming dynamic programming procedure
Ma et al. (2013)	Engine vibration using on-road accelerometer grid	High classification accuracies	Contact and invasive signal using accelerometer grid
Wang et al. (2013)	Classification using audio-visual as well as LDV features	High classification accuracies with non-contact multimodal data	LDV is combined with other data modality to yield high accuracies
Wei et al. (2014)	Engine type classification using tone-pitch vibration index	High classification accuracies with LDV data only	Long data duration, 1.25 s, hard to classify moving vehicles

2. LDV data for vehicle engine classification

The LDV data is one-dimensional data, similar to sound data. Given the great success in speaker recognition (Kinnunen & Li, 2010) and spoken language processing (Huang, Acero, & Hon, 2001), much research and development of LDV data analysis is motivated by audio signal processing. Features in the time and frequency domains are extracted using mathematical transformations, such as number of zero-crossings, short-window energy, spectral flux, linear prediction coding (LPC), short-term Fourier Transform (STFT), Mel-frequency cepstrum coefficients (MFCC) (Sigmund, Shelley, Bauer, & Heitkamp, 2012), Hidden Markov Model (HMM), dynamic time warping (DTW) (Bellman, 2003), etc.

Many groups have explored using LDV data for effective vehicle classification. In Smith, Mendoza-Schrock, Kangas, Derking, and Shaw (2014) a hierarchical vehicle classification approached using LDV data was developed, where a wide array of aforementioned time and frequency domain features such as spectral flux, MFCC, and number of zero-crossings are tested and automatically selected to generate a decision tree for different types of vehicles. The auto-correlation function of LDV signals was employed in Sigmund et al. (2012) as the workhorse to distinguish engine type, speed, and number of cylinders. Averbuch et al. developed a diffusion map based framework to detect moving vehicles based on wavelet packets within the dynamic programming framework (Averbuch & Hochman, 2010). A prototype automatic vehicle classification system was developed in Ma et al. (2013), where a grid of accelerometers are installed on roadways to characterize road vibrations and the number of axles is classified according to ground truth with precision at about 99%. In Petrovich, Snorrason, and Stevens (2002), vehicle operating conditions were classified using 11 extracted features, including MFCC and others such as zero-crossings, dominant frequencies and Flux, which was also employed in Smith et al. (2014). Table 1 summarizes the most representative relevant pieces of work and their corresponding pros and cons.

However, the classification performance as reported in the foregoing papers remains undesirable. It has been found that LDV data has to be fused with other modality data, such as visual and range features, in order to deliver acceptable classification results, e.g., in Wang, Zhu, and Taylor (2013), Zhu and colleagues developed a method to detect and classify civilian vehicles into five classes based on multimodal audio-visual features, including visual tokens using global geometric features (aspect ratio profiles) and local structure features (HOGs), as well as various LDV features (MFCCs, short term energy, etc.). Radial-based support vector machine was utilized for classification purposes, and results of collecting data and classifying vehicles on both local roads and highways showed that using multimodal features significantly improved the classification accuracy. However, it remains inconclusive as to whether LDV data alone can serve as another data modality side by side with other modalities, mainly because of the relatively poor performance of current state-of-the-art LDV based classification methods.

The main reasons behind the outstanding efficacy of human speaker and speech recognition is the transforms such as the decibel (dB) and octave-band collecting to convert the original acoustic signals to vectors meaningful to the human auditory systems (Lerch, 2012). MFCC, by taking advantage of all the foregoing special properties of the human auditory system, is one seminal feature that has achieved great success in a wide array of speaker and speech recognition cases. However, in vehicle classification applications, the human auditory system is no longer the eventual judge for the recognition quality, these efforts to successfully “cheat” or exploit human auditory systems for optimal encoding and recognition purposes are inappropriate or even irrelevant. Instead, the vibration data collected by LDV sensors should be treated as a sequence of physical data or time series: the running vehicle engine is the periodic vibrating source propagating/dissipating its energy over the rigid surface of vehicles as vibrational waves (Kinsler, Frey, Coppens, & Sanders, 2000). The LDV sensor records the vibrations or waves propagated mechanically or acoustically to vehicle surfaces.

Great progresses have been made using modalities such as visual, infra-red, and audio data (Kinnunen & Li, 2010; Viola & Jones, 2004; Wei, 2013). However, in applications of crucial interest to military and law enforcement, e.g., detecting vehicles as potential threats, these features have major shortcomings: they are easily deceived by the enemy or criminals intent on hiding their identities—they can change or camouflage the visual, heat and audio characteristics of their vehicles to render them unidentifiable by these sensors. The data collected by LDV are reflective of the engines and surface vibrations of the vehicles and thus extremely hard, if not entirely impossible, to mask. Therefore, LDV can better penetrate the authentic identities of vehicles and thus become extremely useful for military and law enforcement for target detection and Intelligence, Surveillance and Reconnaissance (ISR) purposes that are of crucial importance to national and international security.

3. Spectral tone-pitch indexing of LDV signals and performance on summer-14 vehicle dataset

After rigorous examinations of the mathematical, statistical and mechanical engineering properties of the LDV data, in Wei, Vongsy, et al. (2014) the spectral tone-pitch indexing (TPI) scheme of LDV signals was developed, which is summarized in the following **TPI** procedure:

Procedure TPI
Input: V; **Output:** TPI_V

1) The basic encoding unit is the vibration data V with duration s s.
2) Compute the time derivatives V_t of V.
3) Apply Fourier transform and keep the magnitudes only:

$$F_V = |FFT(V_t)|, \tag{1}$$

4) Collect the first spectral *tone* index t_V with original Fourier magnitudes:

$$\boldsymbol{t_V} = F_V(h_0 : h_1), \tag{2}$$

And the second tone index s_V with Fourier magnitudes suppressed by the logarithmic transform

$$\boldsymbol{s_V} = |\log \ (F_V(h_1 + 1 : h_2))| \tag{3}$$

where $h_0 < h_1 < h_2$, the spectral bands lower than h_0 are discarded as irrelevant to the engine's vibrations, while those higher than h_2 are dropped as high frequency noise; the discounting threshold h_1 is chosen to start the logarithmic suppression transform Eq. (3) since this slightly higher band may be compromised by noise so that the magnitude suppression by logarithmic transform appears to strike a valuable compromise between signal preservation and noise mitigation.

5) Form the spectral *pitch* index $\boldsymbol{p_V}$ as defined below,

$$\boldsymbol{p_V} = |FFT(F_V)(h_0 : h_1)| \tag{4}$$

which is an actual second Fourier transform of the Fourier magnitudes.

6) The vector $TPI_V = [\boldsymbol{t_V}, \boldsymbol{s_V}, \boldsymbol{p_V}]$ is taken to be the spectral tone-pitch index (TPI) for V of duration s s.

The resultant index $\mathbf{TPI_V}$ is the actual feature reflecting the contents of LDV data V of duration s s. In our work, via intensive bootstrapping Monte Carlo studies using our available data, s in Step 1 was set at 1.25 s, h_0, h_1 and h_2 used in Steps 4 and 5 are set at 3 Hz, 43 Hz, and 82 Hz respectively. The index is hence of dimension 120.

For each data point L collected by the LDV sensor, the duration t_L of L of necessity should be longer than that demanded by Step 1 of the TPI procedure: s s, e.g., the t_L followed by our data collection in CCNY and AFRL from June 2014 to Apr. 2015 is set at 30 s. To reliably group L into different classes, the *majority rule* is employed by going through the following **Voting** procedure:

Procedure Voting
Input: *L*; Output: class label c

1) L is sliced into N overlapping frames of duration s, in our work, the overlapping length between every two adjacent frames is one-fifth of the frame size, each frame is denoted by V_i, $i = 1, \ldots, N$.
2) Compute the tone-pitch index f_i for each V_i by calling the TPI procedure.
3) Classify each f_i to a label c_i using a classification algorithm.
4) All N classification labels, i.e., $(c_1, c_2, \ldots, c_N)$, where $c_i \in [1, K]$, for K class labels, are aggregated to make a majority vote, the class label c with the maximal counts is assigned to L, i.e.,

$$c = argmax_j \ \#(j = c_i) \text{ for } \forall i \tag{5}$$

For a 30 s data slice L with basic encoding duration s = 1.25 s and 0.25 s overlapping, the number N of voting TPIs is 115. According to a classifier for K labels, each TPI_i is assigned a class label j, as dictated by Eq. (5), the j in the set of possible K labels receiving the most votes among all N units, the actual mode of all c_i's is taken to be the class label given to L. This voting scheme provides classification robust to the various conditions mentioned above.

Equipped with the foregoing two procedures and the summer-14 vehicle dataset described in Section 1. A, we conducted full circle of machine learning procedures, namely, training, cross-validation (CV) and testing, to find the optimal approach to classify the four engine types. Many (>10) cutting-edge classifiers were applied, as reported in Wei, Vongsy, et al. (2014), the CV and test accuracies delivered by the leading five classifiers and two deep nets, i.e., Deep Belief Nets (DBN) and Stacked Auto-encoder (SAE)—more details of which will be given in the next section—are tabulated in Table 2.

Table 2
Accuracy rates, in percentage, for five classifiers and two deep nets in CV and test steps using TPI and **Voting** procedures with s = 1.25 s and 0.2 s.

Method	CV 1.25	Test	CV 0.20	Test
K Nearest Neighbor (kNN)	**89**	55	77	40
Random Forest (RF)	**91**	67	92	37
AdaBoost (AB)	**92**	70	90	42
LogitBoost (LB)	**91**	72	88	45
Artificial Neural Network (ANN)	**94**	**96**	**86**	**84**
Deep belief nets (DBN)	0	0	0	0
Stacked auto-encoder (SAE)	0	0	0	0
Convolutional Neural Net (CNN)	86	88	**87**	**85**

As can be observed in Table 2, in the CV step—the training data consist of four randomly chosen data measures, while the remainder 20∼27 data measures serve as the CV data—most classifiers, except two deep learning methods: DBN and SAE, achieved desired performance: of the five tested classifiers, all of them (in bold fonts) delivered accuracies higher than 89%. This should not be too surprising: the training and CV data are from the same vehicle measured at almost the same time, they are thus highly correlated. The test data are far more challenging: they are taken from different vehicles of similar engine types or same vehicles measured on different days with different weather or noise levels, although efforts were made to minimize background noises that were beyond experimental control. In the test phase, among[1] all eight approaches based on the **TPI** and **Voting** procedures only one - the conventional (feedforward) artificial neural network (ANN) with 20 nodes in the hidden layer - can deliver performance comparable to the ones in the CV step namely 96% vs. 94%. None of the other four conventional classifiers and two deep nets (DBN & SAE) can attain test accuracies higher than 72%: from 55% by *k Nearest Neighbors (kNN),* where k = 5 in our tests, to 67% by random forest, 70% and 72% for the other two boosting methods: AdaBoost (AB) and LogitBoost (LB) .[2] Besides the performance in the summer-14 dataset, we also tested on datasets collected by our own LDV sensor on the CCNY campus from Oct. 2014-Apr. 2015. For this CCNY data the ANN and these classifiers delivered performance similar to those reported in Table 2 (Wei, Vongsy, et al., 2014). CNN performance using TPI is second only to ANN with CV and Test accuracies of 86% and 88%.

When the encoding duration is s = 0.20 s, the CV results are still strong, which is likely due to over-fitting. Thus, CV performance is not quite useful as a measure of maturity. The test results of the four conventional classifiers are in the 30 and 40 percentage, not much better than random guesses. The ANN and CNN both delivered far better accuracies well above 80%, but still less than 90% and hence inappropriate for military applications/deployment.

The performance attained by the two deep nets, DBN and SAE, are far from ideal. From Table 2, the test accuracies, as defined

[1] CNN was not tested in our previous papers since TPI is 1-D while CNN demands 2-D. The results reported here are based on the 2-D version of TPI using the same method to generate 2DonTPI described in Eq. (7), Section 4.2.

[2] kNN, random forest, adaboost and logitboost are available in Matlab's statistics toolbox; neural network is available in the neural network toolbox; DBN, SAE and CNN are available as Deep Learning Toolbox provided by R.B. Palm in Mathwork's Matlab Central.

in Eq. (9) in Section 4, for both nets are 0 since one engine type *11* L *diesel V6* was entirely classified as the inline-4. In Wei, Vongsy, Mendoza-Schrock, and Liu (2014) we argued that this subpar performance by deep nets may be caused by the lack of data volume since our summer-14 data only have about 2200 data measurements, while the outstanding results achieved by deep nets in other contexts (LeCun, Bengio, & Hinton, 2015) are mostly done for far larger data volumes. However, to collect this "small," clean, and lab-quality (with consistent noise suppression efforts and long measuring time, 30 s) summer-14 dataset, it took a team of six researchers and students from CCNY and AFRL, with the additional support of 10 colleagues for test support, and more than one month to complete the data collection. This includes planning, scheduling, and weather delays (cannot collect on rainy/stormy days). Therefore, it is not trivial to collect a data volume comparable to other big data projects with millions or even more data points, as summarized in LeCun et al. (2015). The only viable means is to refine and revise our LDV representation TPI, as described in the **TPI** procedure, so that even with a reasonably small dataset, the deep nets can still show their great discriminative and cognitive power. As to be seen in the sequel, the performance delivered by deep nets can be significantly improved by a more refined TPI (below) to encode the LDV data.

4. Deep learning via TPI and Voting procedures

As discussed in Section 3, the main obstacle to the practical use of the **TPI** coding procedure is the choice of encoding unit s in Step 1: after extensive Monte Carlo studies using the summer-14 dataset, the classification performance is peaked at s = 1.25 s, shorter durations reduce performance as reported in Table 2 across the board: the accuracies attained by the two leading approaches, CNN and ANN, are consistently well below 90% if the unit s is set at < 1 s. We already determined that to ensure a single 1.25 s measurement, vehicles must move at < 2.8 MPH to achieve reliable classification labeling, and according to the **Voting** procedure described in Section 3, the number N TPIs for each measured data vector *L* should be relatively large to be resilient to noise effects. In the summer-14 dataset and ensuing CCNY data collections, with $t_L = 10{\sim}30$ s and N > 10, the **Voting** procedure can yield desirable classification results. It is unrealistic to always demand a very large N as the smallest possible N in principle should be 5 with the corresponding t_L at 2.25 s in order that a majority vote can be determined. Therefore, in practical situations, the actual moving speed of vehicles must be further halved to < 1.4 MPH, which is obviously impossible for any realistic traffic situation. If we want to apply our work in real-world military and law enforcement applications, means must be found to significantly reduce the encoding unit s; otherwise the **TPI** and **Voting** procedures developed in Wei, Liu, Zhu, Mendoza-Shrock, et al. (2015) and Wei, Vongsy, et al. (2014) can only stay inside a laboratory as a methodology for classifying stationary vehicles, which is not our research objective.

4.1. Deep nets: a powerful machine learning framework

Of all the classifiers tried in our explorations, three of them reported in Table 2 belong to the methodology of *Deep learning* or *deep nets* including Deep Belief Net (DBN) (Hinton & Salakhutdinov, 2006), Stacked Auto-Encoder (SAE) (Bengio, 2009), and Convolutional Neural Net (CNN) (LeCun et al., 2015), which has won most competitions in recent years since Hinton's insights of a decade ago. The gist of all deep learning methods is the use of more layers than the conventional ANN, which is generally composed by three layers: the input, the hidden, and the output layer. In our choice of feedforward ANN (Wei, Vongsy, et al., 2014), we only need to run a Monte Carlo study to determine the optimal number of nodes in the hidden layer since the input and output layers are pre-determined by the dimensionality of the TPI vectors, of 120 dimension, and the number, 4, of the engine classes to be grouped. In our work, the patternnet function available in the Neural Network toolbox of Matlab was employed. In theory there is no limit of the number of layers in constructing an ANN: indeed, in the **patternnet** function, users can define as many hidden layers as desired. We tentatively tried it out in our Monte Carlo studies; we found that addition of hidden layers to render the ANN "deeper" did not produce better performance. Instead, the best CV and test accuracies attained by a 4-layer ANN, in the same manner as those reported in Table 2 are 93% and 91% respectively which are worse than the single hidden layer ANN. Further increasing the number of hidden layers of ANN only reduces the accuracy. Thus, simply increasing the number of hidden layers in the ANN framework is not a viable choice. Our foregoing observations are merely verifications of the claim made by Hinton and Salakhutdinov (2006): the ANN constructed by the back-propagation algorithm, an actual gradient descent algorithm, will likely be trapped in local minimum with more hidden layers and the variations in the distant output layers cannot be effectively propagated to earlier layers since most gradients vanish. The high likelihood of trapping in local minimum and vanishing gradients make it easy for a "deeper" ANN to get stuck in over-fitting trouble: fitting the training data too well while lacking generalization power. Hence the backpropagation enabled ANN is unfit for a network with deep hidden layers. In general, ANN and the three deep nets still belong to the *neural computing* framework since they are all inspired by the human cognitive process.

It is well established by neural scientists (Utgoff & Stracuzzi, 2002) that the neurons in human brains achieve learning and cognition by putting neurons in deep nets in order to disentangle the innate regularities and granularities of concepts and tasks; shallow ones like the neural nets or support vector machines are ill equipped to represent these complexities. Hence to achieve better learning capacity and address cognitive tasks, a new algorithm with drastically different perspectives is thus needed. Deep learning is the actual framework in response to this call: the specific details of the three main deep nets, DBN, SAE, and CNN, may be entirely different, they took their inspirations from totally different subjects, e.g., DBN from stochastic modeling, SAE based on innately sparse datasets, whereas CNN is targeted to ensure invariant features using local receptive fields, shared weights and spatial/temporal sub-sampling (LeCun, Bottou, Bengio, & Haffner, 1998).

Unlike the single back-propagation training algorithm for ANN, deep nets resorted to different procedures to train the weights and bias for the multiple hidden layers. Some are creative reuse of the back-propagation procedure, i.e., SAE and CNN. For instance, SAE is a revolutionary unsupervised use of the original ANN: using the input signals as *both* input and output signals, train the ANN with one hidden layer H_1 whose number of nodes is generally less than that of the input/output signals. Afterwards use the newly trained H_1 as the input and output of a new ANN and train the second hidden layer H_2. This procedure can iterate until the desired number of layers, the local minimum trapping and vanishing gradients troubles are avoided since the back-propagation procedure was performed for one hidden layer per iteration so that the long chain rule used in the back-propagation procedure is elegantly evaded. This layer-wise optimization idea can be found in the training of DBN and CNN as well. The number of hidden layers is hence not a bottleneck any longer within this deep net framework. In the past five years, deep net based methods have excelled in many international image and speech understanding competitions (LeCun et al., 2015).

4.2. Normalized TPI as a better data index

Given that our **TPI** and **Voting** procedure based vehicle engine classification attained impressive performance by using ANN, it is natural to try the supposedly better neural nets, the deep nets, to render a practical vehicle classification. That is, to achieve performance comparable to the ANN as reported in Table 2, with both CV and Test accuracies being comfortably above 90%, while considerably reducing the encoding unit s used by **TPI** procedure. As formerly discussed, the number N in the **Voting** procedure should be at least 5. A normally slowly moving vehicle, e.g., moving at 15 MPH, travels 6.7 m/s. Then t_L for the 4 m (front to back door) is 0.6 s. Assuming the overlapping duration between succeeding L remains one fifth of L, the encoding unit s should thus be at most 0.3 s. It is thus clear that we should endeavor to develop a new TPI with s < 0.3 s in Step 1 of **TPI** procedure using deep nets that is able to deliver the CV and Test accuracies like those shown in the last row of Table 2.

Although apparently better nets, using **TPI** and **Voting** procedure based on DBN, SAE, or CNN cannot blindly deliver better results than the "shallow" ANN, as initially reported in Wei, Vongsy, et al. (2014), the CV and Test accuracy attained by DBN and SAE are in the range 72~91% which is significantly worse than the ANN's 96~97% accuracy. Since CNN needs 2-D matrices as inputs, the LDV data cannot be plugged in for classification until a data transform is conducted to yield a 2-D representation for a 1-D data vector L. When the encoding unit s is reduced from the default 1.25 s the classification accuracies by SAE and DBN further degenerated to as low as 40%; not significantly better than random guesses. In consequence the procedures described in Section 3 (TPI in particular) need to be revised to attain the promise of deep nets that have proved themselves in many different applications (LeCun et al., 2015).

As can be observed in the **TPI** procedure in Section 3, the TPI representation, a 120-D numerical vector, essentially consists of three 40-element components: the first 40 elements, $\boldsymbol{t_V}$, of TPI are the magnitudes of the lower frequency (No. 3 to 42) for the Fourier transform of $\mathrm{V_t}$, according to Eqs. (1, 2); while the next 40 elements, $\boldsymbol{s_V}$, are the magnitudes of those higher frequencies (No. 43 to 82) that went through a logarithmic transform, cf. Eqs. (1, 3); the third component $\boldsymbol{p_V}$ encodes the changes in the magnitudes of the Fourier transform, cf. Eqs. (1, 4). Combined in TPI these components encode the rich content of the data signal V. Despite the initial success of this indexing scheme achieved by ANN, as shown in Table 2, careful investigation is still needed if we want to take better advantage of this framework. One major problem of this TPI representation is its heterogeneous nature: the three components reflect data in different domains that may each admit a different phenomenology: $\boldsymbol{t_v}$ is the original Fourier magnitude for low frequencies, $\boldsymbol{s_v}$ is the logarithmic value of the Fourier magnitude for slightly higher frequencies; whereas $\boldsymbol{p_v}$ is the magnitudes of the Fourier transform of the Fourier magnitude. These three components encode different aspects of the data vector V so that no one should be given any preference. However, the TPI representation simply stacks these three 40-element components together as one long 120-element index for V, in most cases a single component, e.g., $\mathrm{t_V}$, encompasses considerably larger values, thence the other components are all but useless in the training and classification process: the training is essentially an over fitting to one component only with information from others discarded altogether. In machine learning and multimedia database applications (Wei, 2002) this unbalanced contribution from only a subset of an object index should be corrected. One popular approach is via *normalization by range* to ensure no component is overly emphasized or ignored thus yielding a more balanced similarity measure or distance measure (Wei, 2004), the new normalized TPI for data vector V, referred to as $\mathrm{nTPI_V}$, following the same notation as those in the **TPI** procedure:

$$nTPI_V = \left(\frac{\boldsymbol{t_V}}{\max(\boldsymbol{t_V})}, \frac{\boldsymbol{s_V}}{\max(\boldsymbol{s_V})}, \frac{\boldsymbol{p_V}}{\max(\boldsymbol{p_V})}, \right) \tag{6}$$

Indeed, this nTPI representation can be evaluated simply as a post-processing of the **TPI** procedure presented in Section 3. As defined in Eq. (6), all elements in $\mathrm{nTPI_V}$ are in the range of 0 to 1 and since each of the three components are normalized individually by its own range the resultant representation is a more balanced reflection of the three different types of phenomenology of low frequency, slightly high frequency, and variations of Fourier magnitudes. As is reported in the next section, this balanced version of TPI has considerably more discriminative power than the original non-normalized TPI.

The new $\mathrm{nTPI_V}$ is a 120-element column vector that can be processed directly by any classifiers including DBN and SAE. However, the CNN—originally developed by LeCun and colleagues for visual object classification and recognition (LeCun et al., 1998) and having been found to be exceedingly successful in recent speech and visual object recognition applications (LeCun et al., 2015)—can only handle 2-D inputs. In view of the special needs of CNN, we next developed a 2-D representation based on the 1-D column vector $\mathrm{nTPI_V}$. Given that the CNN achieves its efficacy by conducting local convolutions, weight sharing, and pooling to produce scale and space invariant features, the 2-D signatures from 1-D nTPI should suffice to make CNN applicable. In consequence, for the nTPI vector with 120 elements, a 36 by 32 2-D overlapping "icon" matrix M is generated according to the following equation:

$$\begin{aligned} \mathrm{M}(12{*}\mathrm{i}+\mathrm{j},\ 1:32) = \ & \mathrm{nTPI}(\mathrm{i}{*}40+\mathrm{j}+1:\ \mathrm{i}{*}40+\mathrm{j}+32), \\ & \mathrm{i}=0,1,2,\ \mathrm{j}=1:12, \end{aligned} \tag{7}$$

According to the above conversion rule, the three components normalized respectively by their own range, namely, $\mathrm{t_V}$, $\mathrm{s_V}$, and $\mathrm{p_V}$, are independently placed in their own 12 × 32 2-D blocks, controlled by index i. To induce spatial redundancies thus making the convolving operator used in CNN more effective, the 40-element vector for each component is assembled in an overlapping manner to generate its corresponding 12 × 32 matrix. This 2-D overlapping matrix was derived by trial and error in our Monte Carlo studies to generate classification results comparable to other neural computing methods. It is denoted by 2DonTPI since it is the 2-D overlapping nTPI specifically designed for CNN to classify or recognize. As shown in the next section CNN based on this representation can produce performance competitive with other deep nets. The overall flowchart summarizing all steps proposed in this work is diagrammatically shown in Fig. 3 for train and test data with a data coding duration of 0.2-0.3 s.

5. Experimental results

To obtain the performance of the nTPI, and it 2-D variant 2DonTPI as developed in Section 4, with short encoding duration s which should be less than 0.3 s to be of practical utility in vehicle surveillance of normal traffic, we applied the step-by-step procedures as shown in Fig. 3 based on the summer-14 dataset following the same data partitioning as reported in Table 2 (Wei, Vongsy, et al., 2014).

To determine multi-class classification performance, we employ a measure based on the confusion matrix C, a K × K matrix with each row normalized to one:

$$accuracy = \frac{1}{K} \sum_{i=1}^{K} C_{ii}, \tag{8}$$

which is the mean value of the diagonal entries in the confusion matrix. However, in many applications this mean operator is

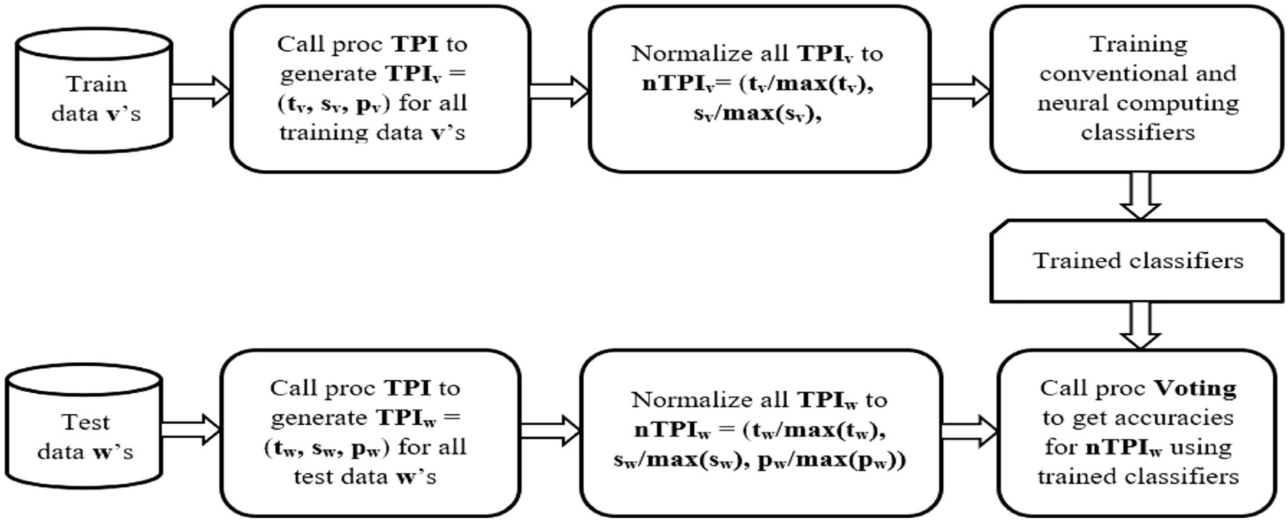
Train data v's
Call proc TPI to generate TPIv = (tv, sv, pv) for all training data v's
Normalize all TPIv to nTPIv= (tv/max(tv), sv/max(sv),
Training conventional and neural computing classifiers
Trained classifiers
Test data w's
Call proc TPI to generate TPIw = (tw, sw, pw) for all test data w's
Normalize all TPIw to nTPIw = (tw/max(tw), sw/max(sw), pw/max(pw))
Call proc Voting to get accuracies for nTPIw using trained classifiers

Fig. 3. The overall flowchart of the proposed nTPI based methods. The input train and test data are data with a short coding duration, e.g., 0.2–0.3 s.

Table 3
Accuracy rates, in percentage, for five classifiers and three deep nets in CV and test steps using nTPI and **Voting** procedures with s = 1.25 s, same as Table 2. accuracy differences from those using TPI as reported in Table 2 are shown for comparison.

Method	CV	CV Diff	Test	Test Diff
kNN	96	+7	74	+19
RF	91	0	79	+12
AB	96	+4	85	+15
LB	95	+4	82	+10
ANN	98	+4	**97**	+1
DBN	96	+96	**97**	+97
SAE	95	+95	**94**	+94
CNN	93	+7	**91**	+3

inappropriate since a weak result for one class could be masked by strong result for other classes which could cause undesirable consequences. A more suitable and conservative measure is:

$$accuracy = \min_{i\varepsilon[1,K]} \{C_{ii}\}, \tag{9}$$

This accuracy is the minimal entry in the confusion matrix C. Accuracy defined in this manner provides more confidence in the performance of a classifier since it guarantees that instances for all classes can be classified with probability limited below by this value. All classification accuracies reported in this work are defined this way.

With the new nTPI (and 2DonTPI for CNN only), the topology of ANN, and especially DBN, SAE, and CNN, should first be determined. After Monte Carlo studies, the one yielding the optimal performance is used. They are enumerated below:

1. For ANN, we found that performance has a plateau when the number of hidden nodes lies within 10 to 24. We still use 20 to stay consistent with the TPI work reported in Wei, Liu, Zhu, et al. (2014).
2. For all three deep nets we used two hidden layers—thus far our Monte Carlo studies have not yielded better performance by using more than two hidden layers—and we used an exhaustive grid search to locate the optimal number of nodes in these two layers. The resultant numbers of nodes in the 1st and 2nd hidden layers for both DBN and SAE plateau around 40 and 18, respectively, little performance difference was observed near these values. The numbers for the two hidden layers for DBN and SAE are thus set at [40, 17] and [40, 19], respectively, in evaluating their classification accuracy.

Since CNN deals with 2-D data inputs, each layer consists of a convolution and sub-sampling layer. From our Monte Carlo studies, we readjust the number of training epochs to find the optimal results. Based on our observations, the classification accuracies go up as the number of epoch increase to around 50, above which the accuracies start to fluctuate. We thus set the epoch number at 50.

We further examine nTPI in comparison to the original TPI developed and applied in Wei, Liu, Zhu, Mendoza-Shrock, et al. (2015), Wei, Liu, Zhu, Vongsy, et al. (2015) and Wei, Vongsy, et al. (2014) to justify the theoretical benefits of the normalization in the new indexing scheme as discussed in Section 4: the balanced contributions of all three components should yield better classification performance. In Table 3, the accuracies using nTPI delivered by the same eight approaches as given in Table 2, with the same encoding unit s = 1.25 s, are reported.

By comparing Tables 2 and 3 with nTPI the CV accuracies by the five conventional classifiers are better but only slightly so, with gains ranging from 0 to + 7 percentage points, over TPI except for the two deep nets: unlike TPI where both DBN and SAE totally failed with 0 (misclassified one whole class) now their accuracies, 96 and 95 for DBN and SAE respectively, are consistent with the other classifier accuracies. In CV the nTPI achieved desirable results since it not only preserved the great performance of the five conventional classifiers, but also significantly improved the performance of DBN and SAE to elevate them to be on par with other classifiers. More exciting advances are made in the more important Test performance: All eight approaches delivered significantly better performance using nTPI except ANN and CNN with only + 1% and + 3% increase, respectively (which is unsurprising given their already high baseline values). The other six approaches gained at least 10 percentage points. The two deep nets now join ANN and CNN with valuable classification performance. Therefore, at the encoding unit duration s = 1.25 s, the nTPI based classification performance is empirically better than TPI across the board, as our theoretical reasoning in Section 4 suggested.

In Table 4, of all eight different methods tested with the encoding unit *s* set at 0.2 s, which is lower than the desired maximal duration 0.3 s and thus of crucial practical interest for military and law enforcement application such as ISR and critical situation surveillance. The four approaches within the neural comput-

Table 4
Accuracy rates, IN PERCENTAGE, for FIVE CONVENTIONAL classifiers and three deep nets in CV and test steps using nTPI and 2DonTPI with encoding duration s = 0.20 s.

Method	CV	CV Diff	Test	Test Diff
kNN	82	+5	78	+38
RF	66	−26	64	+27
AB	91	+1	73	+31
LB	91	+1	75	+30
ANN	89	+3	**90**	**+6**
DBN	95	+95	**93**	**+93**
SAE	91	+91	**90**	**+90**
CNN	86	−1	**89**	**+4**

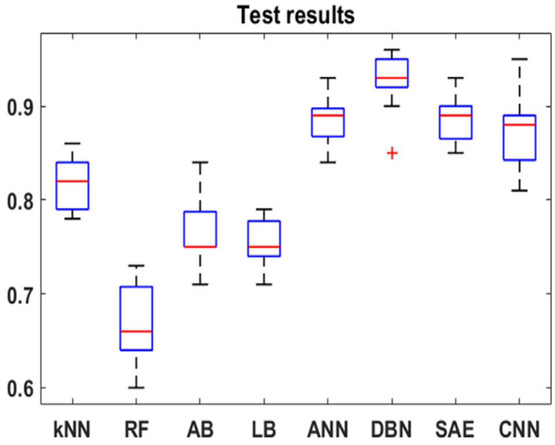

Fig. 4. Boxplot of 11 test results for s = 0.20 to 0.30 s by step 0.01 s, from the eight methods as reported in Tables 1–3.

ing family, namely, the “shallow” ANN and three deep nets DBN, SAE and CNN, all achieved around 90% accuracies for the challenging test column. It is worth pointing out that with nTPI the accuracy achieved by ANN is still quite resilient: although it is arguably worse than DBN in both CV and Test accuracies with 89% vs. 95% and 90% vs. 93%, respectively. However, it is almost indistinguishable with SAE: with 89% vs 91% and 90% vs 90% for CV and Test accuracies. Compared with Table 2 where the TPI was used, the accuracy of the test difference by use of nTPI are promising: except ANN and CNN whose increases are modest +6% and +4% respectively, the other six methods improved by great margins.

To further inspect the performance of the eight different methods, in Fig. 4, the box plots of 11 test accuracies generated by the eight methods tabulated in Tables 2–4 with 11 choices of encoding unit s increasing from 0.20 to 0.30 s with step 0.01 s, are reported. Since CV results are not indicative of anticipated field performance, here only the test accuracies are used to assess performance. From Fig. 4 it can be concluded that the performance is similar to that reported in Table 4 when s = 0.20 s. That is, DBN is still the best among all eight methods; the performance of the other three back-propagation based neural nets: ANN, SAE, and CNN are statistically indistinguishable. Interestingly CNN, the only deep learning method that can achieve acceptable performance using the old TPI, cf. Table 2, has the largest spread under 2DonTPI: this is likely caused by our simplistic 1-D to 2-D conversion. More work is still needed to improve the 2DonTPI transform thus taking full advantage of the power of CNN, which has been dominating many machine learning contests at the international level (LeCun et al., 2015). The other four conventional classifiers are considerably worse than the four methods in the neural computing family, especially DBN, with the simple kNN taking the lead while the Random forest lags far behind. The feedforward ANN, although not within the deep learning framework, not only delivered the best performance using TPI with s = 1.25 s, but also attained competent accuracies no worse than the deep nets SAE and CNN.

Based on experimental results reported here in Tables 2–4 and Fig. 4, together with those using original TPI approach (Wei et al., 2014), among conventional classifiers, ANN is consistently the best in classifying vehicle engines with different choices of signal durations. For other conventional classifiers, their performances fluctuated wildly: while kNN is among the worst for original TPI, cf. (Wei et al., 2014); as shown in Fig. 4, it beats the other three sophisticated conventional classifiers using the new nTPI index. On the other hand, random forest was the best based on TPI, whereas using nTPI it is the worst, by a large margin, among the four conventional classifiers. Deep nets needs more data to be effective, that is why when the 1.25 s signal duration is used in TPI, the training data size is inadequate thus the performance failed gravely, see Wei et al. (2014) and Table 2. However, with enough training data (with a requirement of a much shorter duration of time of each sample), their prowess can be released, hence the consistently outstanding results as evidenced in Table 4 and Fig. 4. Despite a simple variant of deep net, ANN nonetheless remains a competent classifier in the neural computing family, which should not be dismissed in the expert and intelligent systems when a decision is to be made from data.

6. Conclusion

As a non-invasive remote sensor with data measurements that are hard to deceive, the Laser Doppler Vibrometer (LDV) has fostered increasing interest for military and law enforcement applications due to its immense precision in both spatial and spectral domains combined with its non-contact remote sensing ability. To tap into the prowess of this new sensor modality, in our earlier work (Wei, Vongsy, et al., 2014) via the collaboration between CCNY and AFRL, we developed a tone-pitch indexing (TPI) scheme to represent the LDV data for vehicle engine classification. As reported in Wei, Vongsy, et al. (2014), the artificial neural network (ANN) with 20 hidden nodes can achieve an impressive 96% accuracy over our summer-14 vehicle dataset. However, one main obstacle of this method is the demand of a long coding duration of 1.25 s, which is impractical to monitor vehicles in real-world traffic since for reliable classification purposed as required by our scheme, vehicles can only move at no more than 1.4 miles per hour. In this work, in view of the different natures of the three components in TPI, to achieve a balanced representation without over-stressing or ignoring any of the three components, we found that by normalizing each component to the range [0, 1] yields a better index. Based on this new normalized TPI and the neural computing methods, we found that high classification performance can be delivered by reducing the coding duration from 1.25 s to merely 0.2 s, whence a vehicle moving at about 15 miles per hour can be effectively monitored. Promising performance has been observed in our tests by all four neural computing algorithms: DBN, SAE, CNN, and ANN, with DBN slightly better than the other three. Further ANN, the non-deep neural net, still stands firmly against the other more popular deep nets. This might be because the current classification task only has four classes and the engine vibration itself has no deep structure so that a shallow neural net still works well.

This paper reports our endeavor to classify vehicle engines by use of scalable feature vectors combined with neural computing. As intensive studied in Peteiro-Barral, Bolon-Canedo, Alonso-Betanzos, Guijarro-Berdinas, and Sanches-Marono (2013), the scalability of feature vectors within the neural computation framework is of utmost importance in the development of intelligent expert systems. In this work, instead of using the original data window of size ~4000 data items, a merely 120-dimension vector is produced to represent the spectral property of a vibration signal of only 0.2 s. This relatively short time duration gives rise to a practical intelligent system that can classify vehicle engines using remotely and hidden laser sensor for slowly moving vehicles. Neural computing, including ANN and deep nets, has been among the most effective classifiers in many existing intelligent expert systems, such as recent work on audio clip classification (Murthy & Koolagudi, 2018), and land cover classification (Gumus, 2018), where ANN was found to be superior to other classifiers by both groups for different application domains. In this work, after extensive empirical studies, we have found that besides the ANN, the deep nets, the DBN in particular, can all produce better performances than other conventional classifiers, such as kNN, random forest and boosting approaches. Therefore, we believe our study add new insights in using the neural computing methodology as a valuable choice in intelligent expert systems.

Despite the promising classification performances achieved by this new representation and neural computing, there are some weakness and limitations to the current work: (1) *Sensitivity to deceiving efforts*: The 0.2 s encoding window collected from the vehicle surface can effectively handle slowly moving vehicles in controlled regions. However, if the vehicle owners camouflage or conduct some special treatments to the vehicle surfaces, the signals collected by LDV are of little discriminating power thus seriously compromising the classification performance. More robust means should be examined to render this system more intelligent and robust. (2) *Civilian vehicle only*: The current methodology is applied to civilian vehicles, for military ones, esp. armored ones, which are of utmost interest to military applications, the vibration signatures could be considerably different. To make our expert and intelligent system applicable to more scenarios, more studies on the data phenomenology and learning methods should be further explored. (3) *Incomplete use of deep learning methodology:* In this work only CNN, DBN and SAE are examined, some recent progress has been made within the deep learning framework, such as *generative adversarial networks (GAN)* (Goodfellow, 2014), *variational autoencoder (VAE)* (Kingma, 2014), and long short-term memory (LSTM) or Gated Recurrent Units (GRU) based Recurrent Neural Net (RNN) (Goodfellow, Bengio, & Courville, 2017), which has delivered state-of-the-art performance for a broad spectrum of applications. (4) *Issues with small data*: The data currently available to us is barely sufficient for the neural computing methods, if deeper nets are needed of interest to military applications, available training and validation data with size larger by several orders of magnitudes is necessitated. However, it is exceedingly hard to make such large military dataset available, systematic means that is not covered by this current work are thus needed to combat the size problem in deep learning.

To address the foregoing limitations of the current work, in subsequent work we would like to further establish the feasibility of our new LDV data representation. In the near future several lines of investigation will be conducted to further expand and generalize the approach developed here. They are detailed below:

1) We will extend our empirical studies to include military vehicles to further determine if the nTPI, deep nets, DBN, LSTM or GRU based RNN, and **Voting** procedure based classification framework can be deployed in real world conditions such as normal or high-traffic flow roads for ISR and situation surveillance purposes. We would also like to explore the potential of deep nets when the numbers of classes and running conditions increased to better reflect real-world scenarios of practical interest to military and law enforcement applications. Deep nets are expected to exhibit their knowledge representation prowess in handling these challenging data complexities, e.g., the temporal dependencies existing in the adjacent signals could be effectively captured by use of LSTM/GRU RNN; whereas shallow ANN may have trouble dealing with these more sophisticated situations.
2) To address the small data problem in military applications, two different routes will be taken: a) *transfer learning* (Goodfellow et al., 2017) using previously well trained deep nets, the available small data is employed to train several layers while freezing most layers; b) using GAN (Olmschenk, Tang, & Zhu, 2018) and VAE to simulate more data to significantly increase the available data size. For transfer learning, in a more recent work, Li, Grandvalet, and Davoince (2018) discovered that instead of freezing some individual parameters, one may freeze some groups of parameters corresponding to channels of convolutional kernels, and apply an L2 penalty using the pretrained model as a reference, i.e., the starting point, thus yielding a L2-SP target function. With empirical evidence and theoretical justifications, Li and colleagues demonstrated better performances than other standard transfer learning schemes that used weight decay on a partially frozen network. This promising transfer learning scheme could be of crucial interest to our ongoing investigations on transfer learning.
3) Direct measurement of LDV on the moving vehicle body without reflective tape admits signal weakness thereby simplifying adversary evasion and masking efforts that compromise classification performance. The most reliable way is to use external objects controlled and owned by the measuring party. Following the line of attack set in Wei, Liu, Zhu, Vongsy, et al. (2015), we will further explore the utilities of our nTPI and deep learning framework on external objects under full control of the measuring agencies so that there is much less opportunity for adversaries to mask the LDV sensor return.
4) We will endeavor to address some more fundamental studies of different scales or geometric transformations, such as similarities and affine transforms, that may arise in both training and testing data sets. One way is to employ multiple neural nets (CNN) for different scales, as done by van Noord and Postma (2016), to achieve scale and geometric transformation invariance. The other is to exploit scale and transform *invariant* representations, e.g., Laplacian pyramid or wavelets, or features such as SIFT (scale-invariant feature transform) and HOG (Histogram of Oriented Gradients) (Szeliski, 2011; Wang et al., 2013; Wei, 2007) as input data fed to neural nets. More careful setup of data collections and methodology investigations are called for on this line of work, which will significantly reduce the complexity of neural nets, be it shallow or deep ones, and thus gaining deeper insights into the data.

Author contributions

Jie Wei: conceptualization; data curation; formal analysis, funding acquisition; methodology; original draft; review & editing.

Chi-Him Liu: data curation; investigation; software; validation; visualization.

Zhigang Zhu: conceptualization; funding acquisition; original draft; review & editing.

Lindsay Cain: funding acquisition; project administration; resources; supervision; original draft.

Vincent Velten: funding acquisition; methodology; project administration; resources; original draft; review & editing.

References

Averbuch, A., & Hochman, K. (2010). A diffusion framework for detection of moving vehicles. *Digital Signal Process, 20*(1), 111–122.

Bellman, R. E. (2003). *Dynamic programming*. Princeton: Univ. Press.

Bengio, Y. (2009). Learning deep architecture for AI. *Foundations and trends in machine learning*: 2.

Castellini, P., Paone, N., & Tomasini, E. P. (1996). The laser Doppler vibrometer as an instrument for non-intrusive diagnostic of works of art: Application to fresco paintings. *Optics & Lasers in Engineering, 25*, 227–246.

Goodfellow, I. (2014). Generative adversarial nets. *NIPS'14*.

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. Cambridge, MA: MIT Press.

Gumus, E. K. (2018). Selection of spectral features for land cover type classification. *Expert System with Applications, 102*, 9.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507. doi:10.1126/science.1127647.

Huang, X. D., Acero, A., & Hon, H. W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice-Hall.

Kingma, D. W. ,M. (2014). Auto-encoding variational bayes. *International conference on learning representations (ICLR)*.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 12–40.

Kinsler, L. E., Frey, A. R., Coppens, A. B., & Sanders, J. V. (2000). *Fundamentals of acoustics*. Wiley.

Kubota, K. (2007). Development of a remote non-contact measurement system combining laser doppler vibrometer and total station for monitoring of structures. *Paper presented at the the 3rd International; conference on structural health monitoring of intelligent infrastructure*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 9.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In Proceeding of the IEEE *(November)* (p. 46).

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. Berlin, Germany: Wiley.

Li, X., Grandvalet, Y., & Davoince, F. (2018). Explicit inductive bias for transfer learning with convolutional networks. *arXiv:1802.01483*, 12 pages.

Ma, W., Xing, D., McKee, A., Bajwa, R., Flores, C., & Varaiya, P. (2013). A wireless accelerometer-based automatic vehicle classification prototype. *IEEE Transactions on Intelligent Transportation Systems, 15*(1), 104–111.

Murthy, Y. K., & Koolagudi, S. G. (2018). Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (GAFS). *Expert System with Applications, 106*, 15.

Olmschenk, G., Tang, H., & Zhu, Z. (2018). Crowd counting with minimal data using generative adversarial networks for multiple target regression. *Paper presented at the IEEE winter conference on application of computer vision (WACV)*.

Peteiro-Barral, D. B.-C., Bolon-Canedo, V., Alonso-Betanzos, A., Guijarro-Berdinas, B. &, & Sanches-Marono, N. (2013). Toward the scalability of neural networks through feature selection. *Expert System with Applications, 40*, 10.

Petrovich, D., Snorrason, M., & Stevens, M. (2002). Identifying vehicles using vibrometry signatures. *Paper presented at the proceedings of the 16th international conference on pattern recognition*: Vol. 3.

Sigmund, K. J., Shelley, S. J., Bauer, M., & Heitkamp, F. (2012). Analysis of vehicle vibration sources for automatic differentiation between gas and diesel piston engines. *Paper presented at the SPIE 8391, automatic target recognition XXII*.

Smith, A., Mendoza-Schrock, O., Kangas, S., Derking, M., & Shaw, A. (2014). Vechicle classification using laser-vibrometry. *Paper presented at the SPIE DSS 111 ground/air multisensor interoperability, integration, and networking for persistent ISR V*.

Szeliski, R. (2011). *Computer vision: Algorithms and applications*. Springer.

Utgoff, P. E., & Stracuzzi, D. J. (2002). Many-layered learning. *Neural Computing, 14*(10), 2497–2529. doi:10.1162/08997660260293319.

van Noord, N., & Postma, E. (2016). Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognition, 61*, 10.

Viola, P., & Jones, M. (2004). Robust real-time object object detection. *International Journal of Computer Vision, 57*(2), 137–154.

Wang, T., Zhu, Z., & Taylor, C. N. (2013). A multimodal temporal panorama approach for moving vehicle detection, reconstruction and classification. *Computer Vision and Image Understanding, 117*, 1724–1735.

Watson, C., Rhoads, J., & Adams, D. E. (2013). Structural dynamic imaging through interfaces using piezoelectric actuation and laser vibrometry for diagnosing the mechanical properties of composite materials. *Paper presented at the ASME dynamic systems & control conference*.

Wei, J. (2002). Color object indexing and recognition in digital libraries. *IEEE Transactions on Image Processing, 11*(8), 912–922.

Wei, J. (2004). Markov edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(3), 311–321.

Wei, J. (2007). Shape indexing and recognition based on regional analysis. *IEEE Transactions on Multimedia, 9*(5), 1049–1061.

Wei, J. (2013). Small moving object detection from video sequences. *International Journal of Image and Graphics, 14*(3).

Wei, J., Liu, C.-H., Zhu, Z., Mendoza-Shrock, O., & Vongsy, K. (2015a). Classification of uncooperative vehicles with sparse laser doppler vibrometry measurements. *Paper presented at the SPIE-DSS, No. 9464-34*.

Wei, J., Liu, C., & Clouse, H. (2018). Spectral eigen index: Military vehicle fingerprinting using eigen analysis in spectral domain. *Pattern Recognition Letters, 112*, 6.

Wei, J., Liu, C. H., Zhu, Z., Vongsy, K., & Mendoza-Shrcok, O. (2015b). Engine classification using vibrations measured by laser Doppler vibrometer on different surfaces. *Paper presented at the SPIE-DSS, No. 9474-48*.

Wei, J., Vongsy, K., Mendoza-Schrock, O., & Liu, C. (2014). Vehicle engine classification using spectral tone-pitch vibration indexing and neural network. *International Journal of Surveillance & Monitoring Research Technology, 2*(3), 29 *Special Issue on Machine Learning and Sensor Fusion Techniques*.

Willemann, D. P., Castellini, P., Revel, G. M., & Tomasini, E. P. (2004). Structural damage assessment in composite material using laser Doppler vibrometry. *Paper presented at the international conference on vibration measurements by laser techniques*.

Jie Wei received the B.S. degree in computer science from the University of Science and Technology of China, Hefei, the M.S. degree in computer science from the Institute of Software, Chinese Academy of Sciences, Beijing, and the Ph.D. degree in computing science from Simon Fraser University, Burnaby, Canada, in 1989, 1992, and 1999, respectively. He has been on the faculty of the Department of Computer Science, City College and graduate center, City University of New York, New York, since 1999, where he is currently a full professor. His current research interests include image processing, computer vision, machine learning and multimodal computing. His research has been supported by NSF, NIH, Air Force Research Laboratory, Air Force Office of Scientific Research, Office of Naval Research and City Seeds.

Chi-Him Liu received his B.S. degree in applied math, minoring in computer science, and M.S. degree in computer science, from the City College of New York (CCNY) in 2013 and 2016, respectively. He is currently pursuing his Ph.D. study in computer science at the Graduate Center, City University of New York since Fall 2016. His research interests include artificial intelligence, image processing, machine learning, and robotics.

Zhigang Zhu received his B.E., M.E. and Ph.D. degrees, all in computer science from Tsinghua University, Beijing, China, in 1988, 1991 and 1997 respectively. Dr. Zhu is the Herbert G. Kayser Chair Professor of Computer Science, at the CUNY City College and the Graduate Center. He directs the City College Visual Computing Laboratory, and co-directs Master's Program in Data Science and Engineering at CCNY. His research interests include 3D computer vision, multimodal sensing, virtual/augmented reality, video representation, and various applications in assistive technology, environment, robotics, surveillance and transportation. He has published over 170 technical papers in the related fields. He is an Associate Editor of the Machine Vision Applications Journal, Springer and IFAC Mechatronics Journal, Elsevier, and was a Technical Editor of the ASME/IEEE Transactions on Mechatronics.

Lindsay R. Cain received her B.S. in electrical engineering from the Northern Arizona University in Flagstaff, AZ. She was commissioned as an officer in the United States Air Force through ROTC on 19 May 2012. After graduation she was selected to attend the Air Force Institute of Technology (AFIT) to pursue a M.S. in electrical engineering with an emphasis on signal processing in the area of automatic target recognition, machine learning, and artificial intelligence. Following her graduation, she was assigned to the Air Force Research Laboratories Sensors Directorate with the EO Exploitation Technology Branch (RYAT). She was a member of the technical staff in AFRL/RYAT from March 2014–April 2016.

Vincent J. Velten received the B.S. degree in mathematics, B.S. degree in electrical engineering, M.S. degree in mathematics, and the Ph.D. degree in electrical engineering from the University of Dayton in 1982, 1982, 1987, and 2002 respectively. He has been with the Air Force Research Laboratory Sensors Directorate since 1982 where he is currently the technical advisor for the Electro-Optical (EO) Exploitation Technology Branch. His current research interests include computer vision, machine learning, compressive sensing, and EO and Radar-Frequency synthetic image simulation.