Digital Signal Processing 31 (2014) 69–78

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Universal background modeling for acoustic surveillance of urban traffic

Stavros Ntalampiras

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, 20133, Italy

ARTICLE INFO

Article history:
Available online 9 May 2014

Keywords:
Acoustic signal processing
Generalized sound recognition
Universal background modeling
Hidden Markov model
Wavelet packets
Intelligent transportation systems

ABSTRACT

Traffic congestion in modern cities is an increasing problem having significant consequences in our daily lives. This work proposes a non-intrusive, passive monitoring framework based on the acoustic modality which can be used either autonomously or as a part of a multimodal system and provide valuable information to an intelligent transportation system. We consider a large number of audio classes which are typically encountered in urban areas. We introduce a combination of a powerful audio representation mechanism based on time, frequency and wavelet domain features with universal background modeling which leads to higher recognition accuracies and detection rates (in terms of false alarm and miss probability rates) with respect to commonly employed methodologies. The basic advantage of a class-specific model derived using the universal background modeling logic is its tolerance to data which belong to other sound classes. Another important feature of the proposed system is its ability to detect crash incidents, which apart from their catastrophic impact on human life and property, have negative consequences on the traffic flow. Our experiments are based on the concurrent usage of professional sound effect collections which include audio recordings of high quality. We thoroughly examine the performance of the proposed system on isolated sound events as well as continuous audio streams using confusion matrices and detection error trade-off curves.

1. Introduction

Modern societies in general and especially major cities are currently facing the constantly increasing problem of traffic congestion. A characteristic statistic is that the number of vehicles which traveled within the U.S. borders between 1980 and 1998 increased 72% while at the same time interval the number of lane lines increased only 1% [33]. Since the addition of more lanes is becoming less and less feasible, contemporaneous approaches are concentrated on using the existing infrastructure more efficiently based on reliable traffic information. In order to limit the specific problem several Intelligent Transportation System (ITS) programs have been initiated, e.g. the Federal ITS program by the U.S. Government in 1991. The overall aim is the development of technologies which maximize the traffic capacity of a certain area and concurrently minimize the delay of the transportation. However, the current technological solutions cannot meet the traffic demands with respect to most major urban areas.

In general, ITS systems are employed for: a) information provision to travelers using the variable message sign (VMS) technology, b) free-way and arterial management, c) crisis situation management, d) parking management and d) increase traffic safety. It is evident that the particular systems heavily rely on the traffic conditions of the network. Typically an ITS employs electronic, computer and communication technologies into vehicles and roadways for increasing traffic safety and reducing congestion under the prism of improving the quality of life of the general public. In this context, information about the traffic conditions is extremely useful, thus there is an immediate need for automated monitoring of urban traffic in order to facilitate ITS methodologies. For improved ITS performance, monitoring should operate in real-time, be diverse (different kind of information should be available) and provide reliable scene analysis under a wide range of environmental conditions. We suggest that traffic surveillance may be based on the acoustic emissions of each vehicle or urban environmental sound event, a process which is non-invasive and involves signal processing and pattern recognition algorithms. Even though the current traffic surveillance systems are mostly based on optical information (e.g. Smart-Cam [3,20,11,10]), the acoustic modality could be used in parallel for improved performance and better quality of service, following the logic of [26].

The motivation behind an urban traffic monitoring system based solely on the acoustic modality is that it may assist the measuring of various traffic parameters such as flow and density of

E-mail addresses: sntalampiras@upatras.gr, dalaouzos@gmail.com.

http://dx.doi.org/10.1016/j.dsp.2014.05.003
1051-2004/

vehicles inside the area of interest. Subsequently these parameters can be used by experts through a decision support interface for reducing congestion and/or pollution by applying a more appropriate traffic management plan. Alternative routes may be suggested to the drivers and potentially save valuable time and energy. Overall, such monitoring systems allow for a more efficient usage of existing infrastructures, which nowadays is the main target as road building is no longer seen as an acceptable solution [16].

This type of monitoring may quicken traffic incident response times and expedite the recovery process. It could also be proven useful to manage the urban development by taking into account the sonic information and constructing sound maps for categorizing and indexing sources of urban noise pollution [31,2]. Other applications include informing travelers about the possibility of heavy traffic or more efficient routing (direct traffic to particular routes or areas or give priority to specific categories of vehicles), pavement maintenance, adaptive road signal management based on real-time traffic flow, etc.

After providing a thorough review of the related work, we experiment on the classification of nine categories (*car*, *motorcycle*, *aircraft*, *crowd*, *thunder*, *wind*, *train*, *horn* and *crash*) based on the experience gained from [29]. The novel aspects of the proposed work are the following:

1. The utilization of a multidomain feature set (time, frequency and wavelet) in the context of urban audio signal processing.
2. The classification stage relying on generative models which are adapted versions of a universal one. The universal background model (UBM) is an effective framework and has received little attention from the generalized audio recognition community.
3. The proposition of a reference dataset which is appropriate for the specific task and will permit the reliable comparison of approaches with similar goals.
4. Last but not least, the incident detection component (to the best of our knowledge a work which addresses the particular problem has not been previously addressed). The particular component caries significant importance as it can notify the authorized manager of a potentially hazardous situation and may assist limiting its consequences including the activation of an alternative traffic management plan.

The rest of this article is organized as follows: Section 2 provides an overview of the related literature while the emphasis is being placed on approaches which include audio signals belonging to vehicles often encountered in urban areas. Section 3 analyses the modules which comprise the proposed surveillance framework with special attention to the universal background modeling. The next two sections (5 and 6) examine the detection capabilities of the proposed approach in a thorough and concise way. Finally Section 7 offers our conclusions as well as ideas for future works.

2. Related literature

Several studies have been conducted which fall into the general area of processing of urban audio signals. However there are no many studies which address the problem through the traffic management point of view nor they consider a large number of audio classes as the present work does. An interesting approach is described in [22] where the continuous usage of features coming from the time, spectral and cepstral domains is applied for the distinction of cars, vans and trucks. Their classification system is based on the Support Vector Machine algorithm. Another approach is explained in [40] where the "eigenfaces method" (borrowed from the field of face recognition) is employed. The main characteristic is that the frequency spectrum of about 200 ms is treated as a vector in a high-dimensional frequency feature space. When a new vector is processed, its spectrum is compared to the already processed spectrums and the difference vector is projected onto the principal component axes to compute the residual. Their experiments are focused on discriminating cars from other vehicles. A Classification and Regression Tree (CART) classifier is proposed in [1] which is based on the minimal distance. The input to the CART is an acoustic signature representative of the distribution of the energies among blocks which consist of its wavelet packet coefficients. The categories which are considered are cars, trucks and vans. Another approach which tries to create sound maps of the monitored area based on passive detection of sounds emitted by road vehicles is presented in [7]. The particular method does not perform any kind of vehicle identification (and thus is only indirectly connected to the present article) but rather tries to approximate parameters reflecting the road conditions, e.g. vehicle count, speed, etc. In the end, the authors suggest that further research is required to compute these parameters in a secure and reliable manner from the sound field maps. An approach which is based on Gaussian mixture models (GMM) is detailed in [28]. The focus is placed upon the classification of two classes, light vs. heavy vehicles where light vehicles include cars, SUVs, minivans and light trucks while the heavy ones include heavy diesel trucks and buses. Their feature set is formed by the output energy levels of a generalized parametric non-linear filterbank. The GMM method is compared to a linearly weighted discriminator and it is concluded that they may be used in a collaborative manner.

A narrower problem which falls into the category of urban audio signal processing is addressed in [15]. The authors used the Time Encoded Signal Processing and Recognition (TESPAR) method combined with the archetypes technique. A variety of Butterworth low pass filters was explored while their recordings included six different models of cars. In [13], the authors explain the usage of time-varying autoregressive (TVAR) modeling to analyze acoustic signatures of six different classes of moving vehicles. They employ an artificial neural network for the classification fed with the TVAR parameters expanded by a low-order discrete cosine transform. Another interesting work presented in [17,18] uses different classifiers (MLPs) trained on individual noise types in the context of computational auditory scene analysis.

It is worth mentioning that there is also a substantial amount of work done in the area of classification of acoustic emissions coming from *military* vehicles (e.g. [39,27]). However the associated signals exhibit special properties and therefore they are treated differently in terms of features and classifiers. Another line of thought is followed in [22] where the problem of separating large trucks, small trucks and cars is dealt by the concurrent usage of audio and visual information. The merits of fusing data on feature level are explained since they allow to decrease the number of learning samples in order to obtain the same classification accuracy with mono-modal data. In this work we are based solely on the acoustic modality which in many cases may provide information that is difficult or even impossible to capture by any other means while, in general, algorithms of lower computational complexity are involved. Furthermore the present article considers signals which may help the vehicle identification task, unlike [37] where the focus is on cues present in the cumulative acoustic signal acquired from a roadside installed single microphone.

3. System blocks

The specific section describes the two main modules of the proposed surveillance system. It is divided into two parts following the fundamental logic behind generalized sound recognition, which states that each sound source distributes its energy across different frequencies in a unique way. Initially we try to capture this way by extracting acoustic features and then model them using statistical