ATLANTIS PRESS
Atlantis Studies in Uncertainty Modelling, volume 3
Joint Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), the 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and the 11th International Summer School on Aggregation Operators (AGOP)

Audio Surveillance of Road Traffic: An Approach Based on Anomaly Detection and Interval Type-2 Fuzzy Sets

Stefano Rovetta[a] and **Zied Mnasri**[a,b] and **Francesco Masulli**[a] and **Alberto Cabri**[a]
[a]DIBRIS, Università degli studi di Genova, Italy, {stefano.rovetta,francesco.masulli}@unige.it
[b]ENIT, University Tunis El Manar, BP 37, 1002 Tunis, Tunisia, zied.mnasri@enit.utm.tn
[a]DIBRIS, Università degli studi di Genova, Italy, alberto.cabri@dibris.unige.it

Abstract

Surveillance systems are increasingly exploiting multimodal information for improved effectiveness. This paper presents an audio event detection method for road traffic surveillance, combining generative deep autoencoders and fuzzy modelling to perform anomaly detection. Baseline deep autoencoders are used to compute the reconstruction error of each audio segment, which provides a primary estimation of outlierness. To account for the uncertainty associated to this decision-making step, an interval type-2 fuzzy membership function composed of an optimistic/upper component and a pessimistic/lower component is used. The final class attribution employs a probabilistic method for interval comparison. Evaluation results obtained after defuzzification show that, with a careful parameter setting, the proposed membership function effectively improves the performance of the baseline autoencoder, and performs better than the state-of-the-art one-class SVM in anomaly detection.

Keywords: Audio event detection, audio surveillance, anomaly detection, deep autoencoder, fuzzy membership, interval comparison.

1 Introduction

Video data have been so far much more popular than audio signals for surveillance tasks. However, in the past few years an increasing attention has been paid to audio event detection (AED). In fact, AED looks more advantageous in certain applications, such as road surveillance, for the following reasons: a) AED offers a lower installation cost, in addition to less expensive requirements in terms of bandwidth, memory and computational load; b) thanks to omnidirectional microphones and/or microphone arrays, audio surveillance has no problems with perception angles nor with luminosity/visibility conditions; c) even in presence of physical obstacles, most relevant sounds can be detected; d) audio data are more useful than video when it comes to detect certain events, like gunshots and screams, where sound bears more importance than image; e) generally, audio data are more separable than video scenes. This latter point is quite important when the task consists in detecting certain event categories.

The present paper describes the design of a machine learning-based system dealing with the problem of AED for audio surveillance of road traffic. This problem can be modelled in different ways. Two possible formalisations are: a) As a task of classification of all perceived events; b) as detection of anomalous/outlier events only. For this application, due to the strong imbalance between classes, we opt for a generative modelling approach to anomaly detection. Anomalous events are incidents, such as car accidents and other events indicating potential hazard (tire skidding, harsh braking, etc.), whereas normal events are all the rest (cars and pedestrians passing by, people talking, horn blowing, etc.). Hence the problem is how to distinguish the outliers in such a noisy environment, where: i) Practically all events are more or less masked by noise; ii) relevant events, such as car accidents, constitute an overwhelming minority in comparison to normal events. In the proposed solution, a set of autoencoders provide generative models of each class of events of interest.

Due to the complex nature of such vaguely defined classes, which is especially true of the background noise, membership to any class must be considered affected by a degree of uncertainty. This calls for an explicit treatment of such uncertainty. Type-2 fuzzy sets [9] are a natural choice for this type of problem. In this

work we opt for using interval type-2 fuzzy memberships to model classes because, apart from their inherent simplicity and their popularity, they minimise the need for arbitrary modelling decisions about the membership itself. To take the final decision, the interval-valued memberships to different classes are compared using a method of interval comparison described in [19]. In this way the decision is taken without discarding the information about uncertainty expressed by the 2-component fuzzy membership.

The rest of the paper is organized as follows: Section 2 reviews the related work, including methods and applications; Section 3 presents the utilized methods and the proposed approach; Section 4 details the experimental protocol and the obtained results. Finally the work is summarized and commented in the conclusion.

2 Related work

The work in the field of audio surveillance can be reviewed either from the theoretic viewpoint, i.e. methods and techniques, or from the practical side, i.e. applications and models. As far as methods are concerned, the design of an audio surveillance system depends on the type of surveillance task, distinguishing primarily between classification or anomaly detection. In case of classification, several techniques, basically developed for speech/speaker recognition, may be useful, such as generative models (HMM and GMM) and discriminative models (SVM and neural networks). In the other case, several anomaly/outlier detection techniques have been applied to audio data, with different levels of efficiency. These methods can be classified into metric-based e.g. KL-divergence distance, reconstruction-based e.g. autoencoders, and domain-based e.g. one-class SVM.

As mentioned, we are focusing on the anomaly/outlier detection approach. Some early reviews [7, 8] use outdated categorisations. In a more recent one [13], techniques are classified into five categories: a) Probabilistic techniques, based on a density estimation of the normal class, so that a low density area in the training set may indicate a low probability of containing normal samples; b) distance-based approaches, that include nearest neighbor and clustering analysis methods, considering that normal data are closely clustered, whereas anomalies are far from their nearest neighbours; c) reconstruction-based approaches relying on training a regression model. Then the reconstruction error between the actual and the reconstructed samples indicates anomaly/novelty/outlierness; d) domain-based methods that try to characterize the training data by defining a boundary around the normal class, but without explicitly providing a distribution in high density regions; e) information theory techniques, that assume that anomaly/novelty/outlierness alters the information content in a data set; anomalies are detected by analysing the information content using information-theoretical measures, such as entropy or Kolmogorov complexity. According to this taxonomy, the present proposal is a reconstruction-based method. It is also worth noting that several feature representations have been proposed, either using hand-crafted low level descriptors (LLD), calculated in both temporal and spectral domains, or using feature embedding through auto-regressive tools, e.g. autoencoders, or from feature fusion [1].

On the application level, audio surveillance systems have taken profit from the increasing interest to anomaly detection. In particular, a variety of models have been developed for road audio surveillance applications. Among classification-based models, the CrashZam system [17] uses an in-car microphone and features/algorithm engineering, with no learning. The model designed by Foggia et al. [2] uses a two-layer representation, first low-level audio features then high-level bag-of-words; in this case a learning component, i.e. an ensemble SVM for the final event classification, is present. In [12], a universal background model (UBM) is proposed with the goal to recognize and detect a large number of audio events encountered in urban areas, with good reported results; the method uses Markov models for several classes. Previous work by the present authors has also tackled the problem of road audio surveillance from an anomaly/outlier detection perspective. In [15], an ensemble one-class SVM parallel to an MLP network is used to calculate the anomaly score for audio events. The one-class SVM yields a binary anomaly evaluation (normal if 1 and anomalous if -1), whereas the MLP output probability indicates the event class. The MLP is gated by the one-class SVM so that its task is to discriminate only between outlier (interesting) classes, disregarding the background. Also, the authors addressed the problem of data imbalance in road audio surveillance [10] by weighting, so that each event class receives a weight inversely proportional to the fraction of the samples belonging to it in the training set. An autoencoder score is used to calculate weights, where the inverse of the reconstruction error is used as a sample weight, so that the least represented classes, and thus the worst reconstructed, receive the highest weights.

The issues exhibited by these approaches may be summarised in a limited discrimination ability due to the complexity of the classes to be modelled. This point is tackled in the present proposal.

3 Methods

This work aims to detect anomalous events on roads, e.g car accident, tire skidding, harsh braking, etc. Naturally, the proportion of such events is much smaller than that of normal ones, i.e. non-hazardous events. This suggests a strategy based on anomaly detection and, rather than a discriminative classifier, on generative modelling plus adaptive weighting.

Three methods are employed: a) A state-of-the-art method, i.e. one-class SVM, used mainly for benchmarking; b) a baseline method which relies on training an autoencoder only on the normal events present in the training set, and then assessing outlierness through the comparison of the reconstruction error to a threshold; b) a proposed method that refines the baseline by introducing interval type-2 fuzzy memberships to explicitly account for the uncertainty in the generative models.

3.1 State-of-the-art method: One-class SVM

OC-SVM is a variant of SVM algorithms, which aims to estimate a function having positive values on a half-space, and negative ones on its complement. Generally speaking, OC-SVM divides the input space into normal data and outliers. However, the training is performed only on normal data. The final decision is taken using the sign function $g(x)$, calculated as in (1):

$$g(x) = \text{sgn}(w^T \phi(x) - \rho), \tag{1}$$

where ϕ is the Gaussian kernel, w is the orthogonal vector to the separating hyperplane, and ρ is a bias term. For each sample, if this function is positive, then the sample is called normal, otherwise it is an outlier. In this work, OC-SVM is utilized mainly for performance comparison with the proposed methods. A thorough description of the OC-SVM problem formulation and algorithm can be found in [18].

3.2 Baseline method: Normal event-based autoencoders

The autoencoder is a neural network whose objective approximates the identity function. It is commonly used as an unsupervised learning technique, that aims to extract features from unlabeled data. To achieve this goal, the autoencoder optimizes the weights to minimize the mean square difference between the given input and the obtained output; then, the value of a hidden layer is used as an encoded representation of the input. A simple autoencoder has only one hidden layer (cf. Figure 1). It is therefore parametrised by weights ($w \in \mathbb{R}^{mxn}$, $\tilde{w} \in \mathbb{R}^{nxm}$) and biases ($b, \tilde{b} \in \mathbb{R}^m$), as follows:

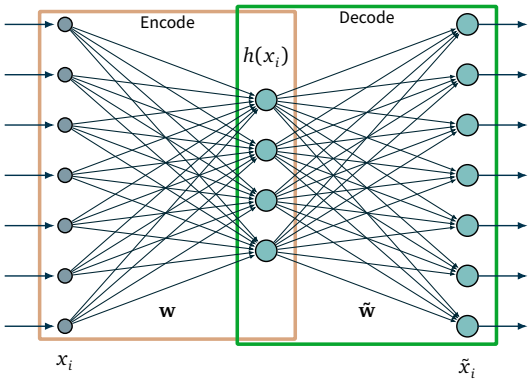

Figure 1: Architecture of an autoencoder

$$\begin{cases} h = f(wx + b), \\ \tilde{x} = \tilde{f}(h\tilde{w} + \tilde{b}), \end{cases} \tag{2}$$

where $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m) \in \mathbb{R}^m$ and $h = (h_1, h_2, \ldots, h_n) \in \mathbb{R}^n$ are respectively the inputs, the outputs and the hidden layer code, and f, $\tilde{f}$ are non linear activation functions, such as the sigmoid function, $f(z) = \frac{1}{1+e^{-z}}$ [11].

It can be shown that the encoding obtained from a simple *linear* autoencoder, i.e. with $f(z) = \tilde{f}(z) = z$, spans the n principal components of the data space, recovering therefore the same embedding as PCA of order n. In this sense we may state that an autoencoder is a nonlinear generalization of PCA. Deep autoencoders with several hidden layers are also possible, although this may imply an excessive overparameterization with increased risk of overfitting, or, correspondingly, the need for exponentially more data.

In this work, the autoencoder is used as a baseline anomaly detection technique, as it can approximate the identity function, so that it generates an image of the input. Thus, the reconstruction error indicates whether the input pattern is a normal or an outlier. Outliers are expected to have a different behaviour than normal samples, and thus their reconstruction error should be higher.

In a similar approach to one-class SVM, the baseline autoencoder is trained on normal data, i.e. non-hazardous events, only. Then the output RMSE error is calculated for the aforementioned terms as in (3)::

$$\varepsilon = \sqrt{\frac{\sum_{k=1}^{m} (x_k - \tilde{x}_k)^2}{m}}, \tag{3}$$

The comparison of the output error ε_i to a preset threshold τ_0 indicates whether the input sample i is normal or anomalous as in (4):

$$Event(i) = \begin{cases} \text{'Anomalous'} & \text{if} \quad \varepsilon_i > \tau_0, \\ \text{'Normal'} & \text{if} \quad \varepsilon_i \leq \tau_0. \end{cases} \quad (4)$$

Regarding the architecture, we opted to train deep feedforward autoencoders. Deep autoencoders were trained on feature vectors, comprising Mel-frequency cepstral coefficients (MFCC) and log-Energy, with their first and second derivatives (Δ and Δ-Δ). The choice of such features is motivated by their outstanding results in the state-of-the-art methods of speech recognition [14], audio event detection [12] and in particular road traffic surveillance [16].

3.3 Proposed method: Anomaly detection based on deep autoencoders, fuzzy membership and interval comparison

In order to improve the performance of the baseline method applied to audio surveillance of road traffic, a method based on combining reconstruction-based learning through autoencoders, and anomaly detection via a fuzzy membership function is proposed in this work. Thus, the method proceeds as follows:

- For each subset containing only one type of events, e.g. background noise, car accidents, tire skidding, etc., an autoencoder is trained.
- In the test phase, for each signal i and each event class $j = 1 \ldots N_{\text{classes}}$, the RMSE error $\varepsilon_{i,j}$ is calculated between the input, i.e. the feature vector representing the signal frame, and the outputs, i.e. its reconstructed image by the corresponding autoencoders.
- For each input signal i, the output error of each autoencoder, i.e. $\varepsilon i, j$ is evaluated using a dedicated fuzzy membership function, which value indicates how close the signal is to the event of interest, i.e. the event on which the autoencoder model had been trained.
- For each type of events, the autoencoder's output error is associated to a membership function composed of a low/pessimistic component and an upper/optimistic component. The values of both components form the membership function interval (cf. Figure 2).
- Finally, a probabilistic method for interval comparison [23] is applied to detect the corresponding event, and hence to detect outlierness.

3.3.1 Fuzzy membership function

The membership of type-2 fuzzy sets can be expressed as a 2-variable membership function $\mu_A(x,u)$ where

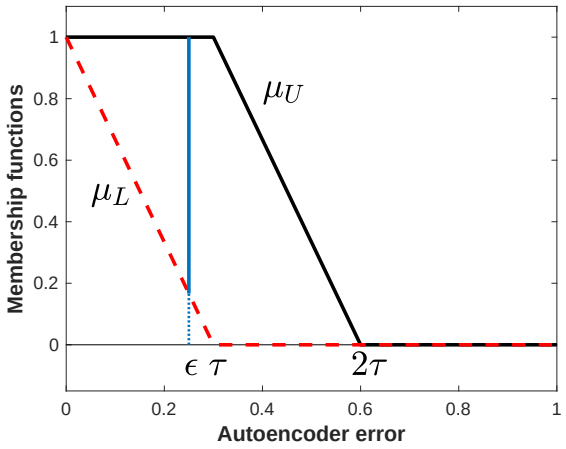

Figure 2: Example of the proposed reconstruction-error-based membership function. Continuous line (μ_U): optimistic membership. Dashed line (μ_L): pessimistic membership. Vertical line at ε: interval values of membership corresponding to error ε.

$\forall\, x \in X, \forall\, u \in J_x \subseteq [0,1], \tilde{A} = \{(x,u), \mu_A(x,u) | \forall x \in X, \forall u \in J_x \subseteq [0,1]\}$ and $0 \leq \mu_{\tilde{A}}(x,\mu) \leq 1$. $\mu_{\tilde{A}}(x,u)$ is called the second grade. When all second grades are equal to 1, then $\tilde{A}$ is called an interval type-2 fuzzy set [9].

Membership to each event is modeled through a fuzzy membership function based on the corresponding autoencoder output error. For each input signal i and for each event-related autoencoder j, the fuzzy membership is composed of two membership functions: a) Pessimistic/Lower membership $\mu_{L,j}$ that is minimum if the sample is an outlier, i.e. its autoencoder error is above the defined threshold (cf. (5)), and b) Optimistic/Upper membership $\mu_{U,j}$, which is maximum when the sample is considered as normal, i.e. its autoencoder error is below the threshold (cf. (6)).

$$\mu_{L,j}(\varepsilon_{i,j}) = \begin{cases} \alpha_j - \frac{\varepsilon_{i,j}}{\tau_j} & \text{if} \quad \varepsilon_j \leq \tau_j, \\ 0 & \text{if} \quad \varepsilon_j > \tau_j, \end{cases} \quad (5)$$

$$\mu_{U,j}(\varepsilon_{i,j}) = \begin{cases} \beta_j - 1 & \text{if} \quad \varepsilon_j \leq \tau_j, \\ \beta_j - \frac{\varepsilon_{i,j}}{\tau_j} & \text{if} \quad \varepsilon_j > \tau_j, \end{cases} \quad (6)$$

where α_j and $(\beta_j - 1)$ are the the upper bounds of $\mu_{L,j}$ and $\mu_{U,j}$, respectively; τ_j and $\varepsilon_{i,j}$ are the threshold and the error of input signal i corresponding to the autoencoder trained on the samples of event class j only, respectively. See Figure 2.

3.3.2 Interval comparison

The key idea is to compare the interval between the upper and the lower membership functions $[\mu_{L,j}(\varepsilon_{i,j}), \mu_{U,j}(\varepsilon_{i,j})]$ for each event-related autoencoder $j = 1,\ldots,N$. Such an interval could be interpreted as a confidence measure. Hence, the smaller is the interval, the higher is the confidence, and thus the tighter is the membership function. Fuzzy number comparison in general, and interval comparison in particular have been broadly investigated since several years [4, 6], using different approaches, either probabilistic [3, 19], possiblistic [5] or based on fuzzy set theory [24]. A comprehensive review of interval and fuzzy number comparison is presented in [21, 22].

The goal of interval comparison is to rank real-number interval, or fuzzy numbers, based on the values of their boundaries. A heuristic approach developed in [23] has an advantage in that it does not rely on midpoints for interval comparison, which makes sense only in the case of fuzzy numbers or confidence intervals.

The comparison between two intervals $A = [a_1, a_2]$ and $B = [b_1, b_2]$ is expressed by the degree of preference of A over B, denoted $P(A > B)$ defined by [23], as reported in [19], using (7):

$$P(A > B) = \frac{\max(0, a_2 - b_1) - \max(0, a_1 - b_2)}{(a_2 - a_1) + (b_2 - b_1)}. \quad (7)$$

Reciprocally, the degree of preference of B over A is defined by (8):

$$P(B > A) = \frac{\max(0, b_2 - a_1) - \max(0, b_1 - a_2)}{(a_2 - a_1) + (b_2 - b_1)}. \quad (8)$$

Hence we obviously have

$$P(A > B) + P(B > A) = 1, \quad (9)$$

and

$$\begin{cases} \text{if} \quad A \equiv B \quad \text{then} \quad P(A > B) = P(B > A) = 0.5, \\ \text{if} \quad a_2 < b_1 \quad \text{then} \quad P(B > A) = 1. \end{cases} \quad (10)$$

Using equations (7) to (10), we measure the membership function as a degree of preference of intervals. Thus, for each event-related autoencoder $j = 1,\ldots,N$, the pessimistic/lower and the optimistic/upper membership functions calculated for the autoencoder's error of the sample i, i.e. $\varepsilon_{i,j}$, form an interval $[\mu_{L,j}(\varepsilon_{i,j}), \mu_{U,j}(\varepsilon_{i,j})]$, to be compared to all other intervals formed by membership functions related to the rest of autoencoders. Finally, the event class corresponds to the interval selected as the least preferred one, as given by (11):

$$Event(i) = \arg \min_{j=1,\ldots,N} \{P(A_j > A_{k \neq j})\}, \quad (11)$$

where $A_j = [\mu_{L,j}(\varepsilon_{i,j}), \mu_{U,j}(\varepsilon_{i,j})]$, $A_{k \neq j} = [\mu_{L,k}(\varepsilon_{i,k}), \mu_{U,k}(\varepsilon_{i,k})] \; \forall \, k \neq j$.

Figure 2 illustrates the principle of the proposed method. Let's assume we have only two categories of events, i.e. 'Normal' and 'Anomalous'. The data belonging to each class is trained to yield two autoencoder models. For each autoencoder, a threshold $\tau_{j=1,2}$ is set. In the test phase, the error of each input sample i generated by each autoencoder j is used through (5) and (6) to calculate the interval $A_{i,j} = [\mu_{L,j}(\varepsilon_{i,j}), \mu_{U,j}(\varepsilon_{i,j})]$. Then $A_{i,j}$ is compared to the other intervals obtained for the same sample i by other autoencoders. Hence, in this case, 2 intervals are obtained, i.e. $\{A_{i,1}, A_{i,2}\}$ for each sample i. Interval comparison is then performed using (7) for all the obtained interval for sample i, yielding in this case $\{P(A_{i,1} > A_{i,2}), P(A_{i,2} > A_{i,1})\}$. Finally, the predicted event is obtained by applying (11).

4 Experiments and results

4.1 Audio materials

MIVIA dataset [2] has been designed for an audio-based road surveillance system. Recordings were realized in a real road environment at 23 locations in the province of Salerno, Italy, covering city center, highways and country roads. The recorded sounds were labeled manually, indicating the audio event and its onset and offset times. Two audio events are considered anomalous, i.e. car crash and tire skidding, whereas all other events are considered as normal, such as cars and pedestrians passing by, people talking, and background street noise. The total duration of the database is approximately one hour, segmented into 57 audio clips.

4.2 Feature set

In [20], a standard set of features was proposed in the IEEE challenge for detection and classification of acoustic scenes and events (DCASE 2013 challenge), including temporal (energy, zero-crossing rate), spectral (spectral roll-off, flux, entropy, variance, aperiodicity bands energy, etc.) and cepstral features (Mel-frequency cepstral coefficients (MFCC)), in addition to time-frequency features, extracted from the wavelet analysis, such as Perceptual Wavelet Packet (PWP). However, it has been demonstrated in [15] that some of these features are not quite discriminatory. In particular, for real-word data, where target events are intrinsically mixed with background noise, some of these

features contribute in worsening the classification performance instead of improving it.

Therefore, after a fine analysis of the discriminatory power of each type of the aforementioned features, we opted to keep only the MFCC coefficients and the Log-Energy. MFCC coefficients have been used for speech and speaker recognition since a long time for their considerable efficiency and their ability to capture the gross spectral characteristics of an audio event [14]. Usually, 13 MFCC coefficients are extracted from the Mel-log spectrum, in addition to log-energy, along with their first and second derivatives (Δ and Δ-Δ).

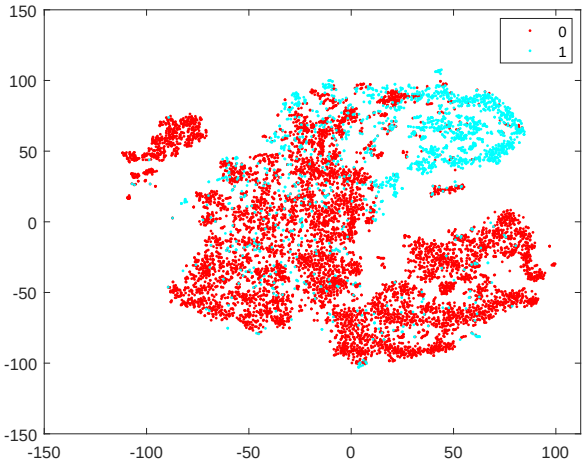

Figure 3: t-SNE distribution of MFCC and Log-Energy features for normal events (0) vs. anomalous events (1)

4.3 Parameter setting

Part	Parameter	Value
All	Event weight w_j	$1/p_j$
Baseline autoencoder	Error threshold τ_0	$\tau_0 \in]0,1[$
Event-based autoencoders	Error threshold τ_j	$\tau_0 \times w_j$
Fuzzy membership	Upper bound for $\mu_{L,j}$ (α_j)	1
	Upper bound for $\mu_{U,j}$ (β_j)	2

Table 1: Parameter setting for the autoencoder error and the fuzzy membership function (p_j is the proportion of Class j samples in the training set)

Since different parameters are involved either in the autoencoder architecture or in the fuzzy membership function (cf. (5),(6)), a special care has been addressed to setting such parameters before presenting the final results. Naturally, different values were tried out in the draft version of the algorithm, however only the parameter values that give the best results are presented (cf. Table 1).

Regarding the baseline autoencoder, i.e. without a fuzzy membership function, the main parameter to be tuned is the threshold τ_0 above which the sample is considered as an outlier. This parameter has been set inside the interval $]0,1[$ since the baseline autoencoder error was normalized using min-max rescaling. In our case, the best results were obtained for $\tau_0 = 1/2$. However, the choice may depend on the distribution of the autoencoder's error.

For the event-based autoencoders, as the dataset is highly imbalanced towards the class of background noise samples, the error thresholds were pondered using the inverse of the proportion of each class as a weighting coefficient. Therefore, we opted to set the threshold τ_j for each class $j = 1,\ldots,N$ as the baseline threshold autoencoder's threshold τ_0 pondered by the corresponding class weight $w_j = 1/p_j$, where p_j is the proportion of samples of Class j.

For the fuzzy membership function, two other parameters have to be set: the upper and the lower bounds, i.e. α_j and β_j, for each event $j = 1,\ldots,N$. For the same reason advanced for the choice of τ_0, i.e. using normalised output error of the autoencoder, we opted to set $\alpha_j = 1$ and $\beta_j = 2$, so that $0 \leq \mu_{L,j} \leq 1$ and $0 \leq \mu_{U,j} \leq 1 \forall j = 1,\ldots,N$ (cf. Table 1).

4.4 Experimental protocol

A series of experiments was conducted to detect audio events on roads, based on the classification into normal vs. anomalous events of the provided samples in MIVIA database [2]. However, as expected, the proportion of the normal (non-hazardous) event samples, is much bigger than that of anomalous (hazardous) events. Therefore, data augmentation was achieved by segmenting the audio signals into short frames, with a duration of 250 ms, with a high overlap rate, i.e. 75%. Hence, the 57 audio clips of approx. 1 min each yielded 57090 frames, in which 45081 belong to Class 1 (normal) and 12009 frames belonging to Class 2 (anomalous), more precisely 4440 frames account for tire skidding and 7569 frames for car accidents. However, it should be emphasized that all training segments, whether normal or anomalous, contain nearly the same level of background street noise.

An architecture of a deep autoencoder was utilized, relying on a feedforward neural network. The results of the autoencoder were analyzed with and without the fuzzy membership function mentioned in (5) and (6). For both cases, the input is the vector of 13 × MFCC

Method	w_{norm}	w_{anom}	*Accuracy*	P_1	P_2	R_1	R_2	$F1_1$	$F1_2$
One-Class SVM			**0.84**	**0.94**	**0.59**	**0.86**	**0.77**	**0.90**	**0.67**
Deep autoencoder	$1-p_{norm}$		**0.85**	**0.86**	**0.79**	**0.97**	**0.38**	**0.91**	**0.51**
only (Baseline)	1/2		0.81	0.81	0.85	1.00	0.08	0.89	0.14
	p_{norm}		0.79	0.79	0.52	1.00	0.01	0.88	0.01
Deep autoencoder	$1-p_{norm}$	$1-p_{anom}$	**0.87**	**0.89**	**0.77**	**0.95**	**0.56**	**0.92**	**0.65**
with fuzzy	1/2	1/2	0.75	0.96	0.45	0.70	0.90	0.81	0.6
membership	p_{norm}	p_{anom}	0.27	0.99	0.22	0.08	1.00	0.14	0.36

Table 2: Results of anomalous event detection using autoencoders and fuzzy membership function for 'Normal' vs. 'Anomalous' event classification ($p_{norm} = 0.79$ and $p_{anom} = 0.21$ are the proportions of normal and anomalous samples in the training set); For OC-SVM, the parameters $\nu =$ and γ are set to 0.14 and 2.5e-5, respectively, for their high performance.

and Log-Energy features, with their Δ and Δ-Δ derivatives. Training and validation of the autoencoders were processed on 80% of the available data, whereas test was conducted on the remaining 20%.

4.5 Analysis of results

Table 2 lists the results obtained for different experimental settings, including the methods used: state-of-the-art OC-SVM (used for benchmarking), baseline autoencoder and event-based autoencoder with fuzzy membership, and varying the values of the the event weights $\{w_j\}_{j=1,\ldots,N}$. The yielding results are expressed in terms of overall accuracy (*Acc*), precision (P), recall (R) and $F1$ scores, defined as in (12):

$$P_j = \frac{c_j}{e_j}, R_j = \frac{c_j}{r_j}, F1_j = \frac{2P_jR_j}{P_j+R_j}, \qquad (12)$$

where r_j, e_j and c_j are the number of ground-truth, estimated and correctly detected events for each class $j = 1,\ldots,N$, respectively.

In Table 2, the results of both proposed methods, i.e. baseline autoencoder and event-based autoencoder with fuzzy membership, show the contribution of the fuzzy membership function to improve anomaly detection. The effects of using fuzzy membership can be listed as follows:

- Both proposed methods perform better than the state-of-the-art OC-SVM, in terms of overall accuracy and balance between class-based metrics.
- Overall accuracy rates are enhanced, reaching 88% for 'Normal' vs. 'Anomalous' event detection (cf. Table 2).
- Precision, recall and F1 score are more balanced between 'Normal' and 'Anomalous' classes when the fuzzy membership is used (cf. Table 2).
- The effect of event wights is more evidenced, with higher accuracy for $w_j = 1 - p_j$. Hence, the prediction of least abundant class, i.e. 'Anomalous', is the most enhanced.
- For such a balanced weighting e.g. $w_j = 1 - p_j$, precision, recall and F1 scores are the highest.

5 Discussion and conclusion

In this paper, a novel method of anomaly detection has been proposed and applied to road traffic surveillance, in the aim to allow detecting hazardous events such as car accidents using audio signals. The proposed method relies on combining two anomaly detection tools, i.e. autoencoders and interval type-2 fuzzy sets. In the training phase, an autoencoder model is learned for each class of events, to be used to generate the reconstruction error in the test phase. The baseline model uses the reconstruction error calculated on the 'Normal' class only, to compare it to a preset threshold. Hence, the autoencoder model that provides the highest reconstruction error corresponds to the class of 'Anomalous' events. Then, this baseline has been improved by adding a membership function stage, where the reconstruction errors computed on each class are leveraged to provide a membership score. To do so, the membership function has been calculated using a couple of optimistic/upper and a pessimistic/lower membership components. Both components are used to define intervals of confidence, which are compared using a probabilistic method. Hence, the least preferred/smallest interval corresponds to the 'Anomalous' class. For evaluation purposes, metrics such as accuracy, precision, recall and F1-score have been calculated. Results show that with i) a good selection of the input features, ii) an adequate choice of the membership function parameters, and iii) a fine tuning of the event-related class weights, the anomalous events can be correctly detected with a comparable performance of state-of-the-art anomaly detection methods,

such as OC-SVM. However, the proposed model could be further improved by becoming less supervised, e.g. without relying on class weights, or fully unsupervised by getting around the class-based autoencoder model.

Acknowledgement

This work was carried out in the framework of the project *Xpert* funded by the University of Genova.

References

[1] S. Chandrakala, S. Jayalakshmi, Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies, ACM Computing Surveys (CSUR) 52 (3) (2019) 1–34.

[2] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, Audio surveillance of roads: A system for detecting anomalous sounds, IEEE transactions on intelligent transportation systems 17 (1) (2015) 279–288.

[3] V.-N. Huynh, Y. Nakamori, J. Lawry, A probability-based approach to comparison of fuzzy numbers and applications to target-oriented decision making, IEEE Transactions on Fuzzy Systems 16 (2) (2008) 371–387.

[4] M. Jiménez, Ranking fuzzy numbers through the comparison of its expected intervals, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 4 (04) (1996) 379–388.

[5] A. Kasperski, A possibilistic approach to sequencing problems with fuzzy parameters, Fuzzy Sets and Systems 150 (1) (2005) 77–86.

[6] E. Lee, R.-J. Li, Comparison of fuzzy numbers based on the probability measure of fuzzy events, Computers & Mathematics with Applications 15 (10) (1988) 887–896.

[7] M. Markou, S. Singh, Novelty detection: a review—part 1: statistical approaches, Signal processing 83 (12) (2003) 2481–2497.

[8] M. Markou, S. Singh, Novelty detection: a review—part 2:: neural network based approaches, Signal processing 83 (12) (2003) 2499–2521.

[9] J. M. Mendel, R. I. B. John, Type-2 fuzzy sets made simple, IEEE Transactions on Fuzzy Systems 10 (2) (2002) 117–127.

[10] Z. Mnasri, S. Rovetta, F. Masulli, Audio surveillance of roads using deep learning and autoencoder-based sample weight initialization, in: 2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON), IEEE, 2020, pp. 99–103.

[11] Ng, Andrew, Sparse autoencoder, `https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf`, online; accessed 29 March 2020 (2011).

[12] S. Ntalampiras, Universal background modeling for acoustic surveillance of urban traffic, Digital Signal Processing 31 (2014) 69–78.

[13] M. A. Pimentel, D. A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Processing 99 (2014) 215–249.

[14] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[15] S. Rovetta, Z. Mnasri, F. Masulli, Detection of hazardous road events from audio streams: An ensemble outlier detection approach, in: 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), IEEE, 2020, pp. 1–6.

[16] A. Saggese, N. Strisciuglio, M. Vento, N. Petkov, Time-frequency analysis for audio event detection in real scenarios, in: 2016 13th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, 2016, pp. 438–443.

[17] M. Sammarco, M. Detyniecki, Crashzam: Sound-based car crash detection., in: Proceedings of Vehicle Technology and Intelligent Transport Systems (VEHITS), 2018, pp. 27–35.

[18] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, Advances in neural information processing systems 12 (1999) 582–588.

[19] P. Sevastianov, Numerical methods for interval and fuzzy number comparison based on the probabilistic approach and dempster–shafer theory, Information Sciences 177 (21) (2007) 4645–4661.

[20] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M. D. Plumbley, Detection and classification of acoustic scenes and events, IEEE Transactions on Multimedia 17 (10) (2015) 1733–1746.

[21] X. Wang, E. E. Kerre, Reasonable properties for the ordering of fuzzy quantities (i), Fuzzy sets and systems 118 (3) (2001) 375–385.

[22] X. Wang, E. E. Kerre, Reasonable properties for the ordering of fuzzy quantities (ii), Fuzzy sets and systems 118 (3) (2001) 387–405.

[23] Y.-M. Wang, J.-B. Yang, D.-L. Xu, A preference aggregation method through the estimation of utility intervals, Computers & Operations Research 32 (8) (2005) 2027–2049.

[24] C.-H. Yeh, H. Deng, A practical approach to fuzzy utilities comparison in fuzzy multicriteria analysis, International Journal of Approximate Reasoning 35 (2) (2004) 179–194.