IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 27, 2020, accepted April 5, 2020, date of publication May 6, 2020, date of current version May 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992804

One-Class Classification in Images and Videos Using a Convolutional Autoencoder With Compact Embedding

MANASSÉS RIBEIRO[1,2], MATHEUS GUTOSKI[2], ANDRÉ E. LAZZARETTI[2], AND HEITOR S. LOPES[2]

[1]Catarinense Federal Institute of Education, Science and Technology (IFC), Videira 89560-000, Brazil
[2]Federal University of Technology Paraná–(UTFPR), Curitiba 3165, Brazil

Corresponding author: Manassés Ribeiro (manasses.ribeiro@ifc.edu.br)

This work was supported in part by IFC Campus Videira, in part by IFC/CAPES/Prodoutoral, and in part by CNPq under Grant 141983/2018-3, Grant 311778/2016-0, and Grant 423872/20168.

ABSTRACT In One-Class Classification (OCC) problems, the classifier is trained with samples of a class considered normal, such that exceptional patterns can be identified as anomalies. Indeed, for real-world problems, the representation of the normal class in the feature space is an important issue, considering that one or more clusters can describe different aspects of the normality. For classification purposes, it is important that these clusters be as compact (dense) as possible, for better discriminating anomalous patterns, which is a recurrent problem in OCC tasks. This work introduces a hybrid approach using deep learning and One-Class Support Vector Machine (OC-SVM) methods, named Convolutional Autoencoder with Compact Embedding (CAE-CE), for enhancing the compactness of clusters in the feature space. Such an approach is still underexplored in the literature, being restricted to models within the context of metric learning. Additionally, the absence of anomalous samples during training makes it difficult to determine when to interrupt the learning process, so as to avoid over-compression of the normal examples, thus resulting in overfitting of the model. In this work, we propose a novel sensitivity-based stop criterion, and its suitability for OCC problems was assessed. Using an OC-SVM for the classification task, several experiments were done using publicly available image and video datasets. We also introduce other two new benchmarks, specifically designed for video anomaly detection in highways. The final performance of the proposed method was compared with a baseline Convolutional Autoencoder (CAE). Overall results suggest that the enhanced compactness introduced by the CAE-CE improved the classification performance for most datasets. Also, the qualitative analysis of frames at the visual level indicated that features learned by CAE-CE are closely correlated to the anomalous events.

INDEX TERMS Anomaly detection, compact embedding, convolutional autoencoder, deep learning, feature extraction, one-class classification.

I. INTRODUCTION

Detecting anomalous behaviors is a recurrent subject in the pattern recognition field, especially because in real-world applications only one of the classes (that related to the normal behavior) is available during the training phase of the classifier [1]–[6]. When a single class is known, the usual approach for classifying patterns is the use of one-class classifiers. This sort of classifier is sometimes referred to as novelty or anomaly detector, because it is trained with previously known patterns that are arranged as one or more clusters (in the feature space) of normal concepts. Then, it is used to identify patterns, known as novelties or anomalies [7], which are somewhat different from those present in the original training dataset.

Three different approaches can be used for One-Class Classification (OCC) problems [7], [8]. The first one is based on the estimation of the probability density function (PDF) of the input patterns (density methods). From the PDF it is possible to establish if a given input pattern is an anomaly or not, based

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

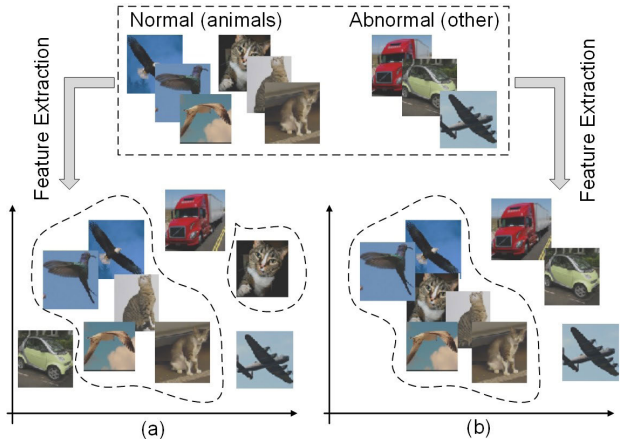

FIGURE 1. Representation of instances in feature space for OCC. The normal class comprises two types of animals (birds and cats), whilst all other examples (cars and aircraft) are considered anomalies. Standard hand-crafted feature descriptors for one-class problems are pictographically presented in (a). Such representation may lead to a high overlap between normal and abnormal concepts, impairing the OCC performance. The ideal feature extractor and the main objective of this work are presented in (b): a compact representation of the normal class leading to a large separation between normal and abnormal examples.

on its probability value. The second approach is concerned by reconstruction methods that use clustering to find out if a given input pattern is an anomaly or not, based on the distance from the input pattern to clusters previously defined in the training process. The last approach comprises models that impose boundaries upon the training dataset, assuming an unknown distribution. In this case, a boundary optimization problem is solved in order to represent the data. The most popular methods that use such an approach are One-Class Support Vector Machine (OC-SVM) and Support Vector Data Description (SVDD), which become identical under certain conditions [7].

The use of one-class classifiers for real-world problems, independently of their approach, encompasses a very important issue: the representation of the normal class in the feature space. Keeping in mind that it is possible to have more than one cluster (in the feature space) representing the normal class, the overall normal concept must be as compact (or dense) as possible, so that the classifier can better discriminate anomalies [9], [10]. Notice that the compact representation and dense representation terms are considered synonyms in this work. For the particular case of images and videos, it is known that the features extracted by standard methods may be inefficient when applied directly to OCC problems [11]–[14]. Hand-crafted image descriptors [15]–[19], and [20], such as the Histogram of Oriented Gradients (HOG), may perform well in particular problems where the gradient orientation is relevant for discriminating between the normal and abnormal classes. However, it may lead to poor classification performance in other sorts of problems where no *a priori* knowledge regarding such orientation is available. Consequently, the mapping performed by hand-crafted descriptors does not guarantee a compact representation of the normal class in the feature space.

Figure 1 highlights the importance of the compactness of the feature space. Suppose that, during the training process, the feature extractor is applied without attempting to achieve a compact representation in the feature space. Thus, normal and abnormal classes can be, eventually, overlapped in the feature space representation, see Figure 1 (a). This compromises the classification performance, regardless of the OCCs. On the other hand, Figure 1 (b) shows a feature extractor that creates an idealized dense representation. It groups in a compact form all the different normal concepts creating a wide separation margin between the normal and the abnormal concepts. Therefore, such representation may significantly improve the final classification performance, as suggested in [10], [21].

Currently, Deep Learning (DL) methods, such as the Convolutional Neural Networks (CNNs), are considered state-of-the-art for image classification problems. They have the advantage of learning, at the same time, the feature extractor and the classifier. The first is accomplished by many convolution filters at successive layers, and the latter, by adjusting the weights of the connections between neurons. This is done by minimizing the training error, considering that the class labels are given [22], [23]. Features learned this way can improve the intra-class separation since both the classifier and the feature extractor are optimized to increase the overall classification performance. However, CNNs are specially designed for supervised classification problems and are not directly applicable to OCC problems. However, recent works [12], [24], [25], and [26] have shown that both, Stacked Denoising Autoencoders (SDAE) and Convolutional Autoencoders (CAE) can be alternatives for OCC problems. This is possible because they are trained to minimize the Reconstruction Error (RE) of the normal class, and this error can be used as a classification score. However, in the bottleneck, i.e., the latent space where the number of neurons is the smallest, compactness is not always accomplished, since the patterns are mapped to minimize the RE of all instances. As a consequence, poor classification results were obtained in OCC problems for particular datasets, in the context of video anomaly detection, as shown in [24] and [12].

Similar results are observed when hand-crafted descriptors and DL methods are applied in the unsupervised context, i.e., clustering problems, such as [27]–[30], and [31]. A clustering algorithm aims at finding a set of objects in such a way that intra-cluster similarities are larger than those observed in inter-clusters. Usually, the similarity measure is defined in terms of pairwise distances. In other words, regions of the feature space that present the lower pairwise distances (compactness) may characterize a cluster. However, the compactness is highly dependent on the feature extraction and, in the unsupervised context, it is even more complex, since no class labels are available when clusters are being constructed. However, recent efforts have been done in learning representations from the clustering perspective – see, for instance,

in [31]. The general idea is to jointly learn representations and cluster assignments using deep neural networks, by reinforcing the compactness and increasing clusters separability. It follows the main purpose of CNNs, but in an unsupervised way. As a result, clustering accuracy in terms of assignment and the ground truth is highly increased when compared with the state-of-the-art feature extraction and clustering methods.

The compactness formulation presented in [31] was adapted to the one-class classification context in [32], resulting in relevant improvements for experiments with specific datasets (MNIST and STL-10). However, for datasets with large image sizes, fully-connected Autoencoders (AE) are no longer capable of capturing the 2D structure in both, images and video sequences. In other words, they use 1D vectors as input, removing local and correlated information that may characterize particular behaviors in the scene. This is a fundamental feature to detect anomalies. To cope with this issue, the CAE architecture seems to be more appropriate [24], [26], and [33], but it requires a modification of the original formulation presented in [31]. Another limitation discussed in [32] is the lack of a stop criterion for the cluster optimization process. This fact results in a very compact representation for the training set (in our case, examples of the normal class), but, on the other hand, a significant overlap between normal and anomalous events in the test set, due to the overfitting phenomenon [10], [34]. Such a limitation requires devising an adequate stop criterion, in order not only to improve the classification accuracy but, also, to improve the generalization capability of the model. The compact embedding and a regularization procedure in the anomaly detection context were initially proposed in a previous work of the authors of this paper [35], indicating the feasibility of this type of approach. The scope of the previous work (a Ph.D. thesis) was to present several aspects of anomaly detection in videos and images, using different approaches with deep learning methods.

Based on the above-mentioned issues, this paper extends the idea presented by [31] and [35], and introduces a CAE-CE (Convolutional Autoencoder with Compact Embedding) as a feature learning stage specifically suited for OCC problems. The main idea is to increase the compactness of the normal class during the training step and, next, to achieve an improved separation between normal examples and anomalies in the test set. The proposed method has three main steps. First, features are learned with the compact representation, using our proposed formulation and architecture for the CAE-CE. Second, a one-class classifier is used to establish boundaries around the compact (dense) normal concepts. In the last step, OCC is accomplished using the test set (which includes normal examples and anomalies). In short, this work aims at filling some of the gaps in current OCC problems, as follows:

1) The Deep Embedding Clustering proposed in [31] is extended to the anomaly detection context with a convolutional architecture, thus allowing to capture the 2D structure of image and video sequences and perform anomaly detection in different datasets.
2) A dense representation for image and video anomaly detection is learned and such representation ends up significantly increasing the final classification performance on real-world datasets, especially when compared with CAE-based approaches [24], [26].
3) A stop criterion is proposed for the CAE-CE optimization, providing to the proposed method a regularization procedure and increasing the performance on the test set.

This paper is organized as follows. Section II presents some related works found in recent literature, mainly focused on compact representations for OCC. Section III addresses the fundamental topics related to CAEs, the proposed CAE-CE, and the OCC trained with the compact learned features. Section IV describes in detail the proposed method. Section V presents how the experiments were done, their results, and a short discussion. Finally, Section VI reports the general conclusions drawn and suggests future research directions.

II. RELATED WORK

A common approach for reducing the effect of the feature descriptor in the OCC performance is to search or build the most appropriate similarity measure (metric) between pairs of examples in the normal dataset. An increase in classification performance is expected by minimizing the distance between those data pairs that belong to the same class (reducing intra-class distances and increasing compactness), subject to the constraint that the data pairs that belong to different classes are well separated (increasing inter-class distances) [36]. In OCC, the learning process of a metric is ill-posed, mainly because only the normal set is available during the training stage [37]. A possible way to overcome such limitation is to minimize distances in the normal set, including a constraint to prevent the solution from being achieved by shrinking the entire space to a single point [38]. The most common solution to this problem is presented in the Relevant Component Analysis (RCA) [36]. However, the main limitation of RCA and its kernel version (kernel-RCA) is that they preserve most of the pairwise distance characteristics in the mapped space, that is, they are not locally adaptive. Such limitation implies that the anomaly detector that is fitted in one part of the feature space will be suboptimal for other parts, even with a preprocessing stage using RCA and kernel-RCA [39].

In order to circumvent the limitation presented in RCA-based solutions, [10] proposed an approach that eliminates the constraint that prevents the convergence to the single point solution. The authors present a kernel null-space for anomaly detection in the multi-class classification context, which makes all known classes to have zero intra-class variance. Results using Caltech-256 and ImageNet datasets show that the proposed method outperforms standard models for multi-class anomaly detection problems. Similar results were obtained by [40], who applied the kernel null-space method to five person re-identification benchmarks. The method significantly overcame state-of-the-art alternatives in some cases.

An extension for local learning in the context of multi-class anomaly detection is presented in [21]. The idea is to use the nearest neighbors of a test sample in the training set to learn a local model that improves the final classification. The limitation of the kernel null-space method and its variants is that the feature extraction and anomaly detection are performed at different stages, so that the features extracted in a preprocessing stage may not be the most appropriate in the anomaly detection context, even with a classifier based on the null-space representation, as briefly suggested in [21]. DL methods, on the other hand, have been investigated for several classification problems in the recent literature, and have already achieved state-of-the-art performance for object recognition in supervised classification of images and videos [22], [23], and [24]. As a matter of fact, this high performance is associated to the features that are learned automatically in the training process, especially when compared with hand-crafted image or video descriptors [15] that, in general, are not designed for particular classification problems. However, in the context of OCC, DL methods are still in the early stages of development. Autoencoders and its variants are, to date, one of the most frequently used models [24].

In [12], an appearance and motion SDAE was proposed to extract features of video surveillance datasets. Based on the features learned, multiple OC-SVM models were used to predict the anomaly scores and classify each frame as normal or anomalous. A similar procedure was presented in [26], where two AEs (SDAE and CAE) were used to learn regular motion patterns from video sequences. The main advantage of this approach is the possibility of jointly capturing regularities (degrees of normality) from multiple datasets. Nevertheless, the anomalies may be characterized by motion and appearance features, thus requiring that the input of the CAE includes such sort of features. In [41], a performance comparison of a deep AE with the proposed hybrid model for different anomaly detection problems was presented. The hybrid model is composed of an unsupervised deep belief network (to extract generic underlying features) and a linear OC-SVM, leading to a scalable and computationally efficient model. More recently, [24] proposed a CAE in the anomaly detection context for automated video surveillance, by using the RE of each frame as the anomaly score. In that work, a method for aggregating high-level spatial and temporal features was also introduced, leading to increased performance in anomaly detection using public-domain datasets.

The application of an SDAE in the context of unsupervised classification was proposed by [31]. In this method, a set of data points is clustered by simultaneously learning representations and cluster assignments. The optimization was based on the Kullback-Leibler (KL) divergence between a centroid-based probability distribution and an auxiliary target distribution. By minimizing the KL divergence, the clusters become denser at each iteration, increasing cluster cohesion and separation, and producing semantically meaningful and well-separated representations. In terms of unsupervised clustering accuracy, the proposed method outperformed the state-of-the-art clustering methods for MNIST, STL, and REUTERS datasets and it is significantly less sensitive to the choice of the hyperparameters.

An extension of this method was presented by [32] in the context of one-class image classification. Relevant results were achieved for MNIST and STL datasets, but the main limitation of the proposed AE is that it does not capture the 2D local structure in image and video sequences since it is based on a fully-connected architecture. Such characteristic results in significant redundancy in the network's parameters, and removes the local information that can be extracted from the images [42]. Of course, this can be particularly relevant in the anomaly detection context, since anomalies can be characterized by different low and high-level local features in a scene. Additionally, the authors do not address the stop criterion when training the SDAE, which may result in overfitting of the model, causing an excessive compression of the clusters. Good results for the training set can be achieved, but at the expense of a huge overlap among examples in the test set.

Inspired by the relevant results obtained for multi-class anomaly detection with kernel null-space representations [10], [21], the recent advances with CAEs in image and video anomaly detection, and the Deep Embedding Clustering presented in [31], we propose here a Convolutional Autoencoder with Compact Embedding (CAE-CE) for OCC in images and videos. The CAE-CE simultaneously learns a set of cluster centers that represent the normal concept and a set of the parameters (weights) that map data points into the bottleneck with dense representation (w.r.t. each center), in order to increase the discrimination between normal examples and anomalies. We extend the method present introduced before in [31] by including a stop criterion and a convolutional approach, maintaining the compactness constraint. This results in a relevant improvement of the final classification performance for real-world datasets, particularly when compared to other CAE-based approaches [24], [26].

III. DEEP LEARNING WITH AUTOENCODERS

A. CONVOLUTIONAL AUTOENCODER

The AE was introduced by [43] and is an unsupervised fully connected one-hidden-layer neural network that learns from unlabeled datasets. The idea is that the AE is trained to reconstruct the input pattern at the output of the network. An AE takes an input $\mathbf{x} \in \mathbb{R}^d$ and first maps it to the latent representation (hidden layer) $\mathbf{h} \in \mathbb{R}^{d'}$. This is done using the mapping function $\mathbf{h} = f_\Theta = \sigma(\mathbf{W}\mathbf{x} + b)$ with weights $(\mathbf{W})$ and bias (b). The set of parameters are represented as $\Theta = \{\mathbf{W}, b\}$, and σ is the activation function. For reconstructing the input, a reverse mapping $f : \mathbf{y} = f_{\Theta'}(h) = \sigma(\mathbf{W}'\mathbf{h} + b')$ is used, such that $\Theta' = \{\mathbf{W}', b'\}$. The parameters $\mathbf{W}$ learned from the input layer to the hidden layer define the encoder, and the parameters $\mathbf{W}'$ learned from the hidden layer to the output layer define the decoder. Optionally, the decoder

parameters $\mathbf{W}'$ may be constrained by $\mathbf{W}' = \mathbf{W}^T$, which is known as *tied weights* [42].

CAEs, as proposed in [42], are similar to the ordinary AE, but the difference between them is the fact that the weights in the CAE are shared among the inputs, preserving the spatial locality, similarly to a CNN [44]. The loss function is given by:

$$e(\mathbf{x}, \mathbf{y}, \mathbf{W}) = \frac{1}{2N} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{y}_i\|_2^2 + \lambda \|\mathbf{W}\|_2^2 , \tag{1}$$

where λ is the regularization parameter for the term $\|\mathbf{W}\|_2^2$, used during the training procedure of the CAE. Similarly to CNNs, CAEs architectures contain convolutional, deconvolutional, pooling, and unpooling layers.

The convolutional layer abstracts the information of a filter into a scalar value parameterizing the number and size of maps, as well as the kernels' size. It connects multiple input activations within the fixed receptive field of a filter to a single activation output in the feature map. For the input $\mathbf{x}$, the hidden layer mapping (latent representation) of the $k-th$ feature map is given by Equation 2:

$$\mathbf{h}_k = \sigma \left(\mathbf{x} * \mathbf{W}_k + b_k\right), \tag{2}$$

where b is the bias, σ is the hyperbolic tangent, and the symbol $*$ corresponds to the 2D-convolution. The reconstruction is obtained using Equation 3:

$$\mathbf{y} = \sigma \left(\sum_{k \in H} \mathbf{h}_k * \tilde{\mathbf{W}}_k + b \right), \tag{3}$$

where there is one bias b per input channel and H identifies the group of latent feature maps. $\tilde{\mathbf{W}}$ corresponds to the flip operation over both dimensions of the weights $\mathbf{W}$.

The deconvolutional layer performs the inverse operation of the convolution layer, and reconstructs the input taking into account the required shape of the output [26]. Convolutional and deconvolutional layers can be stacked to build deep CAE architectures. The filters in the first convolutional layers extract low-level features, whilst the middle layers extract high-level features from the input frames. In this work, the high-level features are basically appearance features.

Pooling layers were originally intended for fully supervised feed-forward architectures for downsampling the latent representation by a constant factor. The idea of the pooling layer is to obtain translation-invariant representations, allowing more complex representations when combined with convolutional layers. It also reduces the spatial size of the representation, reducing the number of parameters and computation in the network, by using operations such as the maximum value over non-overlapping rectangular sub-regions (patches). On the other hand, unpooling layer performs the reverse operation, reconstructing the original size of each rectangular sub-region.

B. CONVOLUTIONAL AUTOENCODER WITH COMPACT EMBEDDING

The CAE-CE introduced in this work follows the idea presented in [31]. Clusters are constructed using data mapped by the bottleneck (latent representation or feature space) of a deep CAE. The proposed approach simultaneously learns a set of K cluster centers $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ in the feature space and the parameters (weights) of the deep CAE that maps data points into the bottleneck. Notice that, in our work, the clusters represent the normal concepts of the OCC problem.

Given the initial mapping provided by the AE and the initial K cluster centers, the idea is to alternate, iteratively, between two main steps: (1) compute a soft assignment between the embedded points and the cluster centroids; (2) update the deep mapping (weights of the AE) and refine the cluster centroids by learning from current high confidence assignments (an auxiliary target distribution is used for this purpose). To do so, the optimization is performed by minimizing the KL divergence loss between soft assignments q_{ij} and the auxiliary distribution p_{ij} [45]:

$$D_{KL} = KL\left(P \| Q\right) = \sum_{i} \sum_{j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{4}$$

The soft assignment is defined as the probability of assigning a mapped sample $\mathbf{z}_i$ (in the bottleneck) to cluster $\boldsymbol{\mu}_k$, using the Student's t-distribution as a kernel to measure such similarity, i.e.:

$$q_{ij} = \frac{\sum_{k} \left(1 + \left\|\mathbf{z}_i - \boldsymbol{\mu}_k\right\|^2\right)}{1 + \left\|\mathbf{z}_i - \boldsymbol{\mu}_j\right\|^2}. \tag{5}$$

On the other hand, the auxiliary distribution is calculated using soft assignments with the following relationship:

$$p_{ij} = \frac{q_{ij}^2 / \sum_{m} q_{mj}^2}{\sum_{k} \left(q_{ik}^2 / \sum_{m} q_{mk}^2 \right)}. \tag{6}$$

The choice of the distribution P is the most important step to achieve compactness for each cluster centroid. According to [31], this distribution must present the following properties: strengthen predictions (improve normal class purity), put more emphasis on data points assigned with high confidence, and normalize loss contribution of each centroid to prevent large clusters from distorting the bottleneck embedding.

Notice that, in our case, the degree of freedom of the Student's t-distribution was set to 1, following the recommendations of [45] and [31]. By using this auxiliary distribution, it is possible to improve cluster purity, putting more emphasis on data points assigned with high confidence, and normalize the loss contribution of each centroid to prevent large clusters from distorting the latent representation [31].

The cluster centers are then optimized jointly with the CAE parameters using the Stochastic Gradient Descent method

with momentum, and the standard backpropagation to compute parameters' gradients, which are defined as:

$$\frac{\partial L}{\partial \mathbf{z}_i} = 2\sum_k \frac{(p_{ij} - q_{ij})\,(\mathbf{z}_i - \boldsymbol{\mu}_k)}{\left(1 + \|\mathbf{z}_i - \boldsymbol{\mu}_k\|^2\right)}, \tag{7}$$

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = -2\sum_i \frac{(p_{ij} - q_{ij})\,(\mathbf{z}_i - \boldsymbol{\mu}_k)}{\left(1 + \|\mathbf{z}_i - \boldsymbol{\mu}_k\|^2\right)}. \tag{8}$$

The optimization continues iteratively until the stop criterion is met. This is discussed in Section IV-C.

C. ONE-CLASS SUPPORT VECTOR MACHINE

In this work, we use the formulation proposed in [7] for the OC-SVM, i.e., SVDD. For a given input class, with N examples and features $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, we assume that there is a closed surface (hypersphere) that surrounds it. The hypersphere is characterized by its center $\mathbf{a}$ and radius R. In the original formulation, the SVDD model contains two terms. The first term (R^2) is related to the structural risk and the second term penalizes objects located at a large distance from the edge of the hypersphere, keeping the trade-off between empirical and structural risks. The minimization problem can be defined as:

$$\varepsilon(R, \mathbf{a}, \vec{\xi}) = R^2 + C_1 \sum_i \xi_i, \tag{9}$$

grouping almost all patterns within the hypersphere:

$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i,\ \xi_i \geq 0, \quad \forall i, \tag{10}$$

in which C_1 gives the trade-off between the volume of the description and the errors that are represented by the distance between outliers and the edge of the hypersphere, ξ_i. The parameter C_1 is related to the ν of the standard one-class SVM presented in [46]. The solutions of the SVDD and the one-class SVM are identical when the Gaussian kernel is used, and $C_1 = 1/\nu N$. The SVDD optimization problem is usually solved through its Lagrangian dual problem, which consists of maximizing:

$$L = \sum_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{11}$$

with respect to the Lagrangian multiplier $\boldsymbol{\alpha}$, subject to the following constraint:

$$0 \leq \alpha_i \leq C_1, \quad \forall i. \tag{12}$$

The main feature of the SVDD model is the representation of the input data in a high-dimensional space without the need of large additional computational effort [7]. This representation allows more flexible descriptors of the input data, following the same general idea of Support Vector Machines. The RBF kernel (used in this work), is given by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\upsilon^2}\right), \tag{13}$$

where υ represents the kernel parameter (width). In this representation, a new pattern $\mathbf{z}$ is classified as an anomaly if:

$$AS = \sum_i \alpha_i \exp\left(\frac{-\|\mathbf{z} - \mathbf{x}_i\|^2}{\upsilon^2}\right) - \frac{1}{2}\left[1 + \sum_{i,j} \alpha_i \alpha_j \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\upsilon^2}\right) - R^2\right] < 0, \tag{14}$$

where AS is the anomaly score of the pattern $\mathbf{z}$. With the RBF kernel, the formulation of the SVDD model is equivalent to the OC-SVM proposed in [47], and discussed in [7].

IV. THE PROPOSED METHOD

The proposed method for anomaly detection has four main steps. The first step, defined as data preparation aims at arranging the datasets for OCC, as presented in Subsection IV-A. Next, the pretraining step is dedicated to training a CAE using examples of the normal class. The optimization is done by minimizing the RE between the input and output images, as in standard AEs. Once the CAE is optimized, the decoder part of the network is discarded. The third step consists of fine-tuning the encoder using CAE-CE for increasing the compactness of the normal class, as discussed in Subsection III-B. The fine-tuning is repeated until the stop criterion is met. Finally, an OC-SVM is trained and used for classifying. Figure 2 presents a high-level overview of the proposed approach, which will be detailed in the next sections.

A. DATA PREPARATION

All datasets used in this work are publicly available, including those first introduced in this work (UTFPR-HSD1 and UTFPR-HSD2). Each dataset is composed of a number of video clips, and each video frame was previously labeled as normal or abnormal. Frames were extracted from video clips to produce the training and test sets. Notice that the training set is composed of normal samples whilst the test set has both, normal and anomaly samples. See Section V-B for further details.

B. CAE PRETRAINING

The CAE training method uses the backpropagation algorithm [48] to minimize the RE, as mentioned in Section III-A. To optimize the loss function, we use the adaptive subgradient method AdaGrad. It computes a dimension-wise learning rate that adapts the rate of gradients as a function of all previous updates in each dimension. AdaGrad is widely used due to its theoretical guarantee of convergence and empirical success [49]. The weights are initialized using the Xavier algorithm, which automatically determines the scale of initialization based on the number of input and output neurons [50].

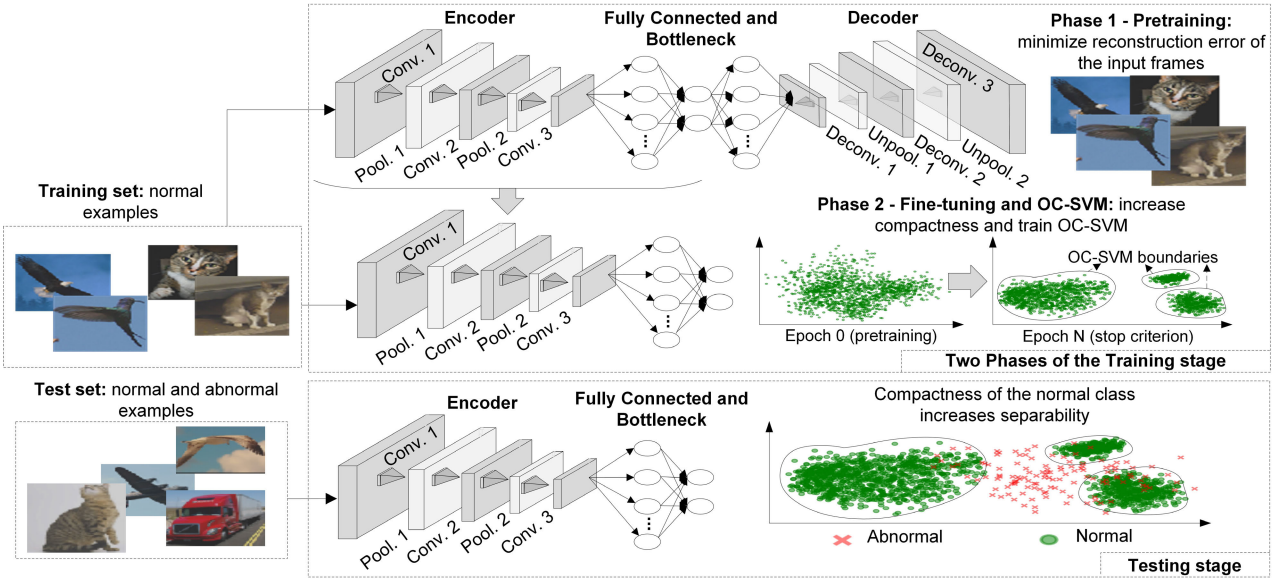

FIGURE 2. Overview of the proposed method.

Training a CAE does not require label information of the input data. However, as usual in OCC problems [24], [26], we use indirect labeling, i.e, all training instances belong to the group of videos without anomalies.

The CAE architecture used was similar to the model recently proposed by Hasan *et al.* [26] and also used in our previous work [24]. It is composed of three convolutional layers and two pooling layers in the encoder side, and the same mirrored structure in the decoder side. However, in this work, three fully connected layers were added to the architecture, besides the bottleneck layer. Table 1 summarizes the structure of the layers. The column *Size* presents the filter size for convolution and deconvolution operations, as well as the kernel size for max pooling and max unpooling operations. The column *Dimension* presents the shape of the output data after each layer. A stride of 4 is used in the first convolutional layer, followed by a stride of 1 in the remaining convolutional layers. Each fully connected layer is followed by a dropout layer with 50% drop probability. All convolutional and fully connected layers use the hyperbolic tangent activation function. The decoder part follows the exact inverse structure of the encoder but uses deconvolutional and unpooling operations.

The CAE-CE architecture is inherited from the original CAE used in the pretraining step. The decoder part of the CAE is discarded, and the bottleneck becomes the output layer of the CAE-CE. Hence, the CAE-CE architecture is equivalent to the encoder part of the CAE.

C. CAE-CE OPTIMIZATION AND STOP CRITERION

The CAE-CE is initialized with the weights learned by the CAE. Recall that the goal now is to simultaneously optimize the representations and the cluster centers through an unsupervised and iterative process. The centers are initialized using the k-means algorithm, where k is a user-defined parameter. In this work, experiments were done with $\{2, 3, 4, 5, 10\}$ clusters, and the best-performing value was used in further experiments. The model was trained using SGD with momentum in a standard backpropagation procedure. For more details refer to Subsection III-B.

TABLE 1. Convolutional autoencoder layers and output sizes.

Encoder		
Layer	Size	Dimension
Input	-	1 x 235 x 155
Conv. 1	11x11	256 x 57 x 37
Pool. 1	3x3	256 x 28 x 18
Conv. 2	5x5	128 x 28 x 18
Pool. 2	3x3	128 x 14 x 9
Conv. 3	3x3	64 x 14 x 9
Fully 4	-	2016
Fully 5	-	504
Fully 6	-	168
Bottleneck	-	50

In real-world OCC problems, samples of the abnormal class are often difficult or even impossible to obtain, but samples of the normal class are abundant. Therefore, using only the normal class samples, we propose a stop criterion based on sensitivity to estimate the best epoch to stop the optimization process.

The sensitivity, or true positive rate (TPR), measures the rate of positive examples that are correctly classified as positive. In order to define a stop criterion, we use a validation set composed exclusively of normal examples extracted from the test set, meaning that no instance of the anomalous class was used for validation. Sensitivity is evaluated after every training epoch, and the best stopping epoch is the one in

which the highest sensitivity value is achieved. This is done by retaining the network parameters every time TPR reaches a new maximum value. The stop criterion is, then, defined by a quality metric (TPR $= 1$) or by the stagnation of the optimization (no improvement after a number of epochs).

To evaluate the TPR as a stop criterion, it was compared with the product of sensitivity and specificity (TPR$\times$TNR), where TNR is the true negative rate (specificity). This product indicates the best performance obtained by the classifier considering the balance between positive and negative classes. Therefore, we can evaluate whether the TPR stop criterion is a good approximation to the best performance that could be achieved, measured by TPR$\times$TNR. The analysis of the effectiveness of our proposed stop criterion is shown in Subsection V-C.

D. CLASSIFICATION AND EVALUATION

The next step of the proposed method is to perform OCC. The optimized CAE-CE is used for extracting features from the training and test sets. This is done by forwarding every instance throughout the network and capturing its embedded representation at the bottleneck. For the classification task, we employ the OC-SVM, which was described in Section III-C.

Since the OC-SVM depends on parameter tuning, we use a combinatorial search strategy [12]. Experiments were done to find the most appropriate values for the main parameters, kernel and regularization (C_1), and the following ranges were tested for them: $\{0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.9\}$ and $\{0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0\}$, respectively. Once the parameters are defined and the optimization has converged, we used Equation 14 to compute the distance, in the hyperspace, of each data point (of the test set) to the OC-SVM's decision border. We represent this distance as either positive or negative. A positive distance indicates that the given data point is within the hypersphere, whilst a negative distance indicates that it is outside the decision border.

At first, the distance sign could be used as the classification result itself, considering zero as the classification threshold. However, this method can be misleading, since it prioritizes the overall accuracy, and does not consider a balance between sensitivity and specificity. Therefore, another approach based on the Area Under the ROC Curve (AUC) and Equal Error Rate (EER) is proposed.

The EER selects the threshold that leads to equal sensitivity and specificity, i.e., the ideal balance between TPR and TNR. This measure is independent of a classification threshold and provides direct analysis of the classification performance. Moreover, it enables the comparison of results with other approaches in the literature [12], [26], [51], [52].

Based on the anomaly score of Equation 14, a Normalized Anomaly Score (NAS) was proposed. Both, NAS and AS, are normalized in the range [0, 1]. Then, NAS is smoothed by using a moving average filter and taken its one's complement, yielding the score S_{NAS}, according to Equation 15:

$$S_{NAS}(n) = 1 - \frac{1}{N}\sum_{j=0}^{N-1} NAS(n+j), \tag{15}$$

where N is the number of samples of the moving average and S_{NAS} is the smoothed normalized AS of the frame n.

V. EXPERIMENTS, RESULTS AND DISCUSSION

The CAE-CE model proposed in this work was built using Theano [53] and Lasagne [54], running in the Ubuntu 14.04.3 LTS operating system. Theano is written in Python and used for computing mathematical expressions, and Lasagne is a library to build neural networks in Theano. All experiments were run in a dedicated GPU server with Intel i7-5820K CPU at 3.3 GHz, 32GB of RAM, and two NVIDIA Titan XP GPUs.

The experiments are presented in the following order: in Section V-A we perform a preliminary experiment to evaluate the viability of the compact representation idea proposed in this work. In Section V-B, the datasets used in this work are presented in detail, including two novel video anomaly detection datasets. Section V-C presents the main experiment and a discussion about the results obtained. Next, in Section V-D, we present experiments and analysis of the stop criterion discussed in Section IV-C. Finally, in Section V-E, we provide a visual analysis of the results for two datasets.

A. PRELIMINARY ANALYSIS OF THE CAE-CE OPTIMIZATION

Our working hypothesis is that a dense feature space increases the separability between normal and abnormal classes. Therefore, a Proof of Concept (PoC) was conducted in this preliminary experiment. For this purpose, we used two well-known datasets: MNIST [55] and notMNIST.[1] The first one (only images of numbers) was considered as the normal class, and the former (only images of letters), as the anomaly class.

A CAE is optimized by minimizing the RE between input and output. This process maps similar features near to each other in the bottleneck feature space. However, in the traditional CAE, there is no attempt to achieve a dense bottleneck representation. This effect is shown in Figure 3 (a), where it is observed a significant overlap between normal and abnormal samples mapped by the CAE's bottleneck. This overlap hinders the performance of the classifier.

In contrast, when the CAE-CE introduces compactness in the representation, a significant reduction of the overlapped areas takes place, as shown in Figure 3 (b). Observe that the representation of normal samples has been segmented into three highly compact clusters when compared with the mapping provided by the traditional CAE. This suggests that the ideal feature space for the normal class should be dense to improve the classification performance. Notice that both Figures are aimed at illustrating the principle behind the method,

[1] Available at yaroslavvb.blogspot.com.br/2011/09/notmnist-dataset.html

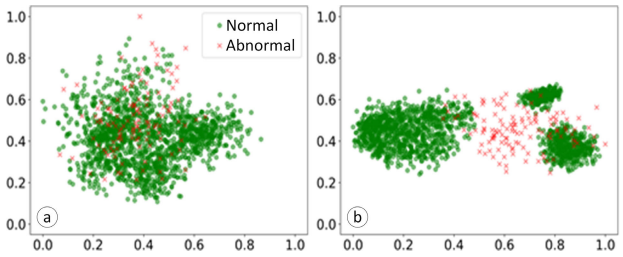

FIGURE 3. Examples of two feature space representations. The normal class comprises only numbers (MNIST dataset), whilst the anomaly class is comprised of letters (notMNIST dataset). The CAE's feature space representation is presented in (a), showing a high overlap between both classes. In (b) the compact representation of the normal class is shown, leading to a higher separation between classes, when compared with (a).

since axes x and y are the two first principal components of the Principal Component Analysis (PCA) of the data.

B. DATASETS

In this work, we used three benchmark video datasets (*Ped1*, *Ped2*, and *Avenue*) frequently used for anomaly detection problems. Additionally, we introduce two video anomaly datasets, namely UTFPR-HSD1 and UTFPR-HSD2. Finally, an image dataset (STL-10) is used in the context of OCC, with the purpose of showing that the proposed method works for both, videos and images, as discussed in the following subsections.

1) STL-10

The STL-10 dataset [56] has 10 classes and contains a large number of unlabeled image data (not used in this work) as well as a small set of labeled images. The dataset was originally devised for unsupervised learning and classification. However, in this work, the dataset was adapted to fit the OCC context using the labeled part of the dataset. We used the classes with more training examples, as follows: classes containing images of animals were considered the normal class, whilst anomalies were composed of the remaining data. The training set had 3000 images and test set 8000 images (4800 belong to the normal class and 3200 are anomalies).

2) UTFPR-HSD1 AND UTFPR-HSD2

These two video datasets, first introduced in this work, aimed at detecting anomalies in highways, traffic monitoring, vehicle identification, and vehicle counting. They were acquired at Curitiba, Brazil, on a busy highway that crosses the city in which traffic control is very important. To avoid traffic jams, heavy truck traffic is not allowed at given times of the day on this highway. Hence, the automatic detection of unauthorized vehicles in this route is extremely relevant for traffic control. Video surveillance is used to monitor the highway, and videos were captured in both normal (without trucks) and abnormal (with trucks) scenarios. Videos with normal traffic include not only cars, but, also vans, pickups (and small trucks), motorcycles, and buses. Anomalies are found when at least one (large) truck is in the scene. The UTFPR-HSD1 dataset is composed of 6602 frames in the training set, and 1660 frames in the test set. UTFPR-HSD2, in its turn, contains 5640 frames in the training set and 1986 frames in the test set. The difference between those datasets is the position of the camera regarding the highway. In the UTFPR-HSD1 dataset, videos were acquired almost at the ground level, transversely to the highway. On the other hand, UTFPR-HSD2 videos were acquired at three different scenarios, from an elevated level above the highway, including the incoming and outgoing traffics. All datasets were annotated by a human expert considering the specific anomaly detection context described here.

The UTFPR-HSD1 dataset contains 23 training clips and 6 test clips, whilst the UTFPR-HSD2 training set contains 25 clips and the test set contains 7 clips. The videos were captured with a resolution of 1920×1080 pixels at 25 fps. Both datasets are publicly available,[2] including the ground truth and training/test splits used in this work. It is important to emphasize that similar datasets are not available in the related literature, and this was the main motivation to introduce those new datasets.

3) AVENUE

The Avenue dataset [57] was captured in an avenue of the Chinese University of Hong Kong campus, and was also devised as an anomaly detection dataset. The training set contains approximately 15328 frames, and the test set about 15324 frames. The normal class contains people normally walking from/to different directions, and the abnormal class contains people running, throwing objects, and loitering.

4) UCSD Ped1 AND Ped2

The UCSD Pedestrian [52] is a video anomaly detection dataset captured by a stationary camera in a pedestrian walkway. The normal class includes only pedestrians, whilst the anomalies include vehicles, bicycles, skateboarders, and wheelchairs passing throughout the pedestrians in the walkway. The dataset is divided into two subsets, Ped1 and Ped2. Ped1 has about 5500 frames of the normal class and 3400 frames of anomalies. Ped2 subset is smaller and contains 346 frames of the normal class and 1652 frames of anomalies.

C. CLASSIFICATION PERFORMANCE

This is the main experiment and its objective is to evaluate the hypothesis that the dense representation achieved by the CAE-CE increases the classification performance when compared to a regular CAE by increasing the separability between normal events and anomalies.

The baseline for comparisons is the classification results achieved by the CAE's bottleneck representations. We also included, when available, results obtained by the state-of-the-art methods for each dataset. Notice that the computational cost of these experiments was very high. For instance, the set

[2]https://github.com/bioinfolabic/UTFPR-HSD

TABLE 2. AUC/EER results for all datasets using CAE, CAE-CE, and the current state-of-the art found for four video datasets and one image dataset.

Datasets	CAE		CAE-CE		State-of-the-art		
	AUC	EER	AUC	EER	AUC	EER	Reference
STL-10	0.681	0.357	**0.716**	**0.331**	0.689	-	[32]
UTFPR-HSD1	0.876	0.195	**0.912**	**0.148**	-	-	-
UTFPR-HSD2	0.909	0.174	**0.916**	**0.146**	-	-	-
Avenue	0.816	0.269	**0.828**	**0.247**	0.772	0.27	[24]
UCSD Ped 1	0.565	0.458	0.652	0.362	**0.927**	**0.160**	[51]
UCSD Ped 2	0.700	0.365	0.768	0.306	**0.908**	**0.170**	[12]

TABLE 3. Confusion matrices of the classification results using the EER for all datasets. "N" represents the normal class, "A" represents the abnormal class, and "Pred." means the predicted class.

UTFPR-HSD1

	CAE Pred. N	CAE Pred. A
N	1029	74
A	251	306

	CAE-CE Pred. N	CAE-CE Pred. A
N	1090	56
A	190	324

UTFPR-HSD2

	CAE Pred. N	CAE Pred. A
N	431	253
A	91	1211

	CAE-CE Pred. N	CAE-CE Pred. A
N	446	215
A	76	1249

Avenue

	CAE Pred. N	CAE Pred. A
N	8486	998
A	3126	2714

	CAE-CE Pred. N	CAE-CE Pred. A
N	8742	917
A	2870	2795

UCSD Ped1

	CAE Pred. N	CAE Pred. A
N	415	566
A	350	669

	CAE-CE Pred. N	CAE-CE Pred. A
N	488	447
A	277	788

UCSD Ped2

	CAE Pred. N	CAE Pred. A
N	230	603
A	132	1045

	CAE-CE Pred. N	CAE-CE Pred. A
N	251	504
A	111	1144

of experiments only for the Avenue dataset required about 45 hours running.

Results are shown in Table 2. In this table, it is observed that, when the feature space is optimized by the CAE-CE, the classification tends to be better when compared with the baseline CAE, for both images and videos. In some cases, such as UCSD Ped1, the AUC is about 15% better. A similar result is observed in the image dataset (STL-10), in which an improvement over 5% was achieved. Notice that STL-10 is a challenging dataset since it contains images with different appearance characteristics (background, sizes, and shapes) [32].

Table 2 allows a broad comparison between the state-of-the-art results and the proposed CAE-CE for video datasets. In the Avenue dataset, our results are better than those achieved by [26], but inferior compared to [14]. However, it is worth mentioning that in [14], different preprocessing stages are applied to the image frames, mainly including optical flow. Also, for UCSD Ped2, we achieved satisfactory results, but not as good as those obtained by [12] and [58]. For the UCSD Ped1 dataset, our results are worse than those of [51]. It is important to mention that the above-mentioned state-of-the-art approaches are very elaborate, including special schemes not used in our work, such as data-augmentation, the division of frames into patches, and large CAE architectures. However, as a matter of fact, no method achieves the state-of-the-art for all datasets.

Table 3 presents the confusion matrix obtained with the EER threshold at the event level, i.e., the classification result of each frame of all videos of the test set. Using the table, TPR and TNR can be computed. For all datasets, the table shows that the CAE-CE achieved better results when compared with the CAE. For instance, the TPR and TNR gains for UTFPR-HSD1 were 5.93% and 5.88%, respectively; 9.13% and 9.47% for the UCSD Ped2 dataset; and 17.59% and 17.79% for UCSD Ped1 dataset. Furthermore, both TPR and TNR increased for all datasets.

D. STOP CRITERION

This experiment aims at analyzing the effectiveness of the stop criterion proposed in Section IV-C. Recall that the working hypothesis is that the TPR can be used as a stop criterion for CAE-CE with a similar effectiveness of the TPR $\times$ TNR approach. Figure 4 shows the behavior of the KL-divergence minimization (KL loss) represented by a blue line, and both TPR (green line) and TPR$\times$TNR (red line) along 1000 training epochs. For this study, a small learning rate value was used, and all plotted values are normalized in the range of [0, 1].

Figure 4 (top) shows that, in the early epochs, the KL loss increases abruptly because of the latent space geometry is being modified to assign samples with respect to their centers. However, after this initial transient, the KL loss decreases and tends to converge along time. Furthermore, improvements in TPR and TPR $\times$ TNR are observed until they achieve their maximum values (around epoch 280), highlighted by the dashed vertical red line. After that, they tend to decrease abruptly despite the convergence observed for the KL loss. Moreover, after the best point is achieved, TPR and TPR $\times$ TNR tend to diverge, thus suggesting the overfitting of the model to the normal samples. For this reason, the best point for stopping the optimization process is when the TPR achieves the maximum value. Notice that this does not necessarily occur when the KL loss achieves its minimum. However, proceeding beyond the suggested stopping point will lead the model to be overfitted to the normal samples, impairing the classification performance.

In Figure 4 (a), normal samples are represented with two clusters (red and green) plotted in a 2D space after the CAE optimization. Figure 4 (b) shows the compactness of the

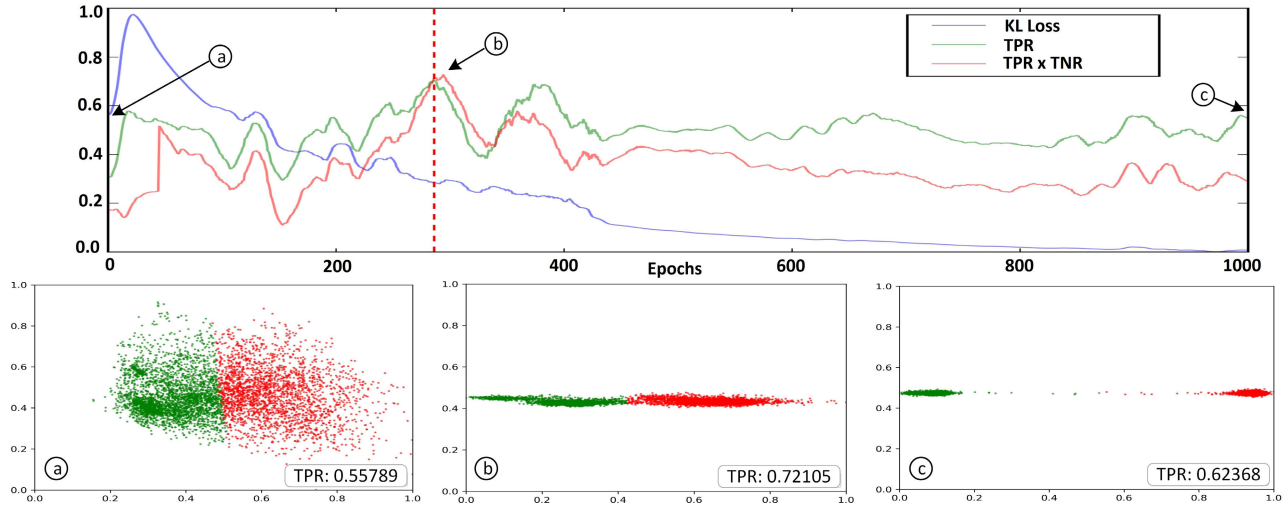

FIGURE 4. (top) KL loss convergence, and both TPR and TPR×TNR curves for a video clip from the UTFPR-HSD1 dataset. (bottom) Figures in the boxes are a compact representation accomplished by using Principal Component Analysis: (a) shows the bottleneck representation after the CAE optimization is done; (b) shows the CAE-CE bottleneck representation at the best optimization point; (c) shows the CAE-CE bottleneck representation when the training process converged w.r.t. KL loss.

two clusters imposed by the CAE-CE approach when the optimization reaches the suggested stopping point. Since the KL loss has not yet converged, the optimization could go further, reaching the point shown in Figure 4 (c), which is the best possible compactness obtainable by the CAE-CE approach. However, at this point, the compactness level is not adequate for anomaly detection problems, since the model is overfitted to normal samples and leads to a poor classification performance (observed in TPR and TPR×TNR curves).

Concerning the generality of the TPR approach, Figure 5 shows the behavior of both TPR (dashed blue line) and TPR×TNR (red line) over 75 CAE-CE training epochs, for all the datasets. Values were individually normalized in the [0, 1] range, so that the plots can be observed on the same scale, as well as to see if their peaks overlap. The peaks are pointed in blue for TPR and red for TPR×TNR.

In the figure, it is observed that in most cases the TPR peak matches with the TPR × TNR peak in the same epoch or very close to it. Also, notice that, in general, the plots follow similar orientation on both TPR and TPR×TNR. These facts reinforce our hypothesis that TPR can have a similar effectiveness as the TPR×TNR approach to find the stop point.

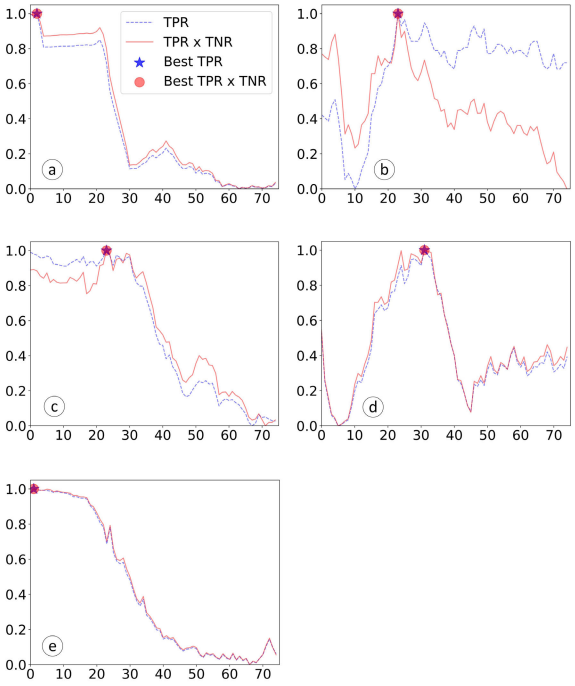

FIGURE 5. Stop criterion based on sensitivity for all datasets: (a) Avenue, (b) UTFPR-HSD1, (c) UTFPR-HSD2, (d) USCS Ped1, and (e) UCSD Ped2. Axes *x* and *y* represent the training epochs and the normalized values, respectively.

E. QUALITATIVE ANALYSIS

In order to better understand the practical differences between our method and the baseline CAE, we perform a visual analysis of some video fragments extracted from the datasets. To accomplish this, we plot the distance of each frame to the OC-SVM decision border in the hypersphere over time, as explained in Section IV-D. This analysis is performed for two different situations, the first when our method showed improvements over the baseline and, the other when no significant improvement was achieved.

Figure 6 (top) shows the plot of the S_{NAS} for both the CAE (green line) and the CAE-CE (blue line), computed as described in Section IV-D, and using a fragment of the UTFPR-HSD1 dataset. The ground truth, annotated by a human expert, and the EER threshold for anomaly detection are shown as red and dashed black lines, respectively. Recall that values above the EER threshold are

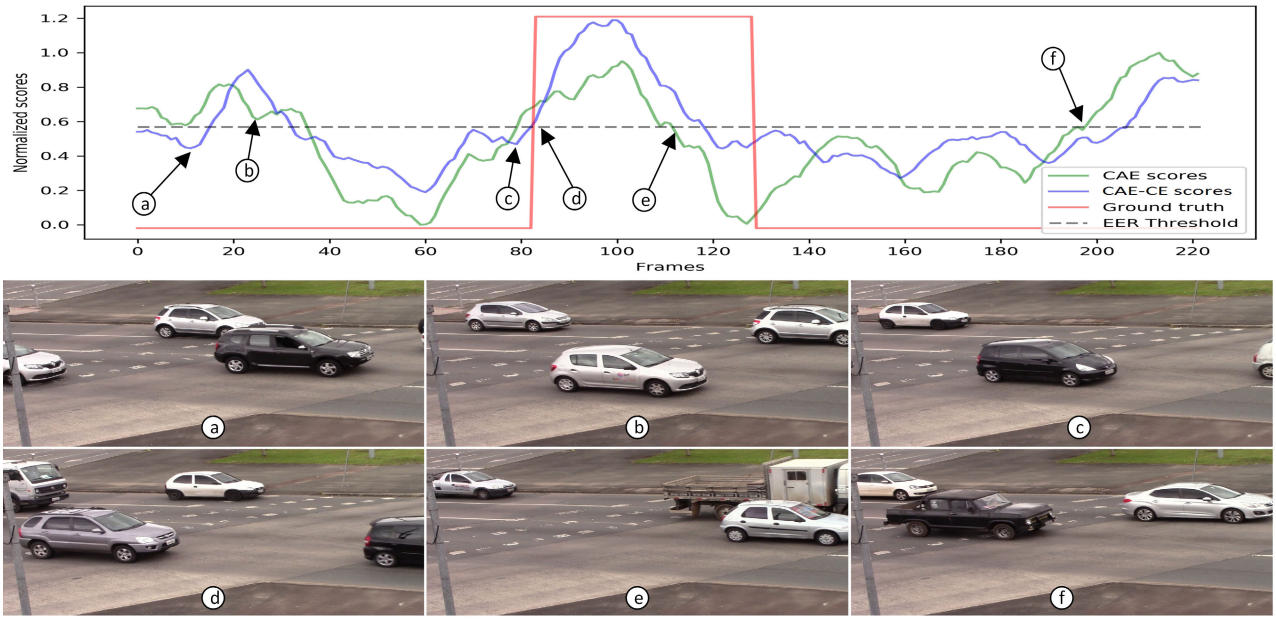

FIGURE 6. S_{NAS} scores computed on a fragment of the UTFPR-HSD1 dataset.

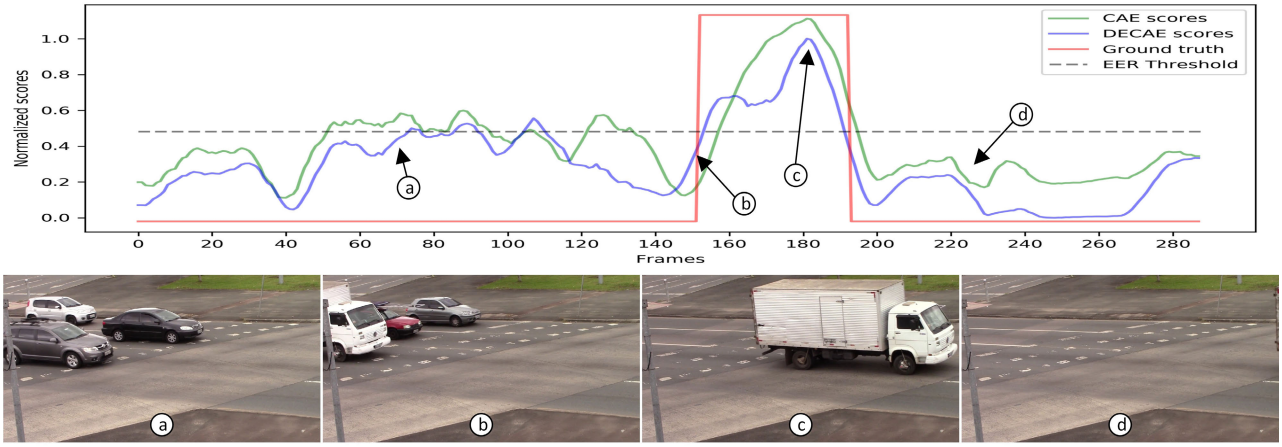

FIGURE 7. Another example of S_{NAS} scores for the UTFPR-HSD1 dataset.

considered anomalies. Basically, this video shows cars crossing the highway background. In a given moment a truck (the anomaly) appears, crosses the highway, and leaves the scene.

Figure 6 (bottom) shows six frames extracted from a video fragment from the UTFPR-HSD1 dataset. An analysis of the S_{NAS} curve and related events in the video are provided for each frame:

1) In this frame, there are only cars crossing the highway, and, therefore, this is a normal event. The S_{NAS} line of the CAE-CE is under the EER threshold whilst the CAE S_{NAS} line is above. In this case, the CAE leads to a misclassification, however, the CAE-CE corrects the mistake.
2) Around frame 20, some false-positives for abnormal behavior happen when using the EER threshold. Both the CAE and the CAE-CE S_{NAS}'s are oscillating, thus indicating that the events are located near the normal-abnormal decision border. Thus, if the event is outside the decision border (beyond normal concept), the CAE-CE optimization enforces it to be a pronounced abnormal event. Considering that the S_{NAS} wrongly classified this event in both CAE and CAE-CE, it is suggested that the training dataset does not contain a sufficient representative amount of this particular pattern to recognize it as a normal event.
3) Between frames 35 and 82 of the clip, there are some fluctuations in the S_{NAS} under the EER threshold. Once more, the CAE-CE has corrected the classification error.
4) The anomaly appears around frame 83 when a truck enters the scene. Soon after this point, the CAE-CE's S_{NAS} scores became higher than those of the CAE, meaning that CAE-CE detected the anomaly

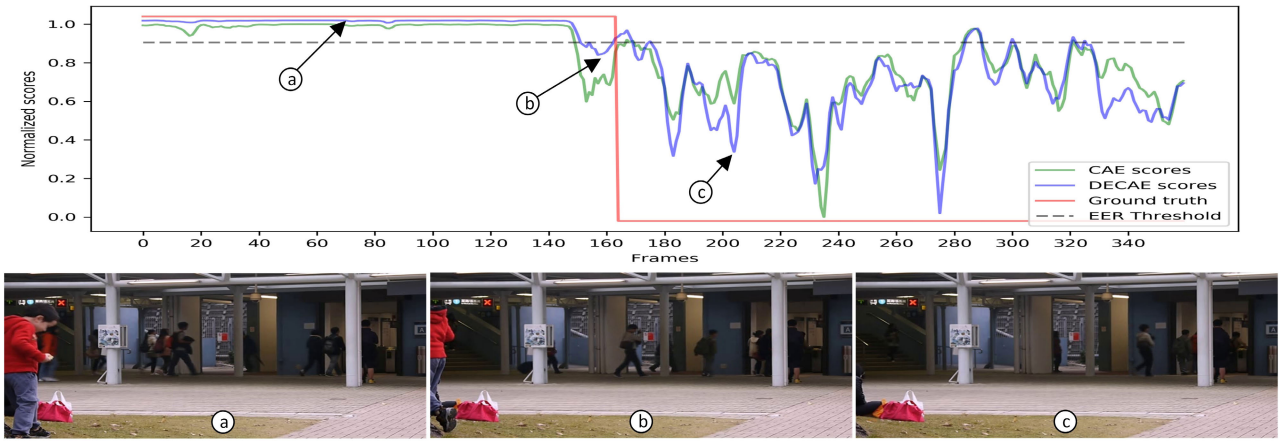

FIGURE 8. S_{NAS} scores computed on a fragment of the Avenue dataset.

with more confidence. Also, it is observed that the CAE-CE detected the anomaly precisely at the time that it started, whilst the CAE wrongly classified some previous frames as abnormal.

5) The truck is partially occluded by a car. This happens from around frame 88 of the clip all the way to frame 123. Whilst the truck stays in the scene, the CAE-CE enforces the S_{NAS} scores to be higher than the CAE scores. Around frame 120, a transition from an abnormal to a normal event starts, where both the CAE and the CAE-CE present some fluctuations. At this point, we can observe once again the effect of compactness since, despite the classification error, the CAE-CE was able to keep the S_{NAS} score near the EER threshold. The truck completely leaves the scene around frame 127.
6) A black pickup truck enters the scene around frame 200 of the clip. Both the CAE and the CAE-CE classify this vehicle as a false-positive anomaly. This is explained due to the visual similarity between this pickup and small trucks, thus leading to false-positives. Even so, the CAE-CE has better performance when compared to the CAE for keeping the S_{NAS} scores closer to the threshold. Moreover, the CAE-CE successfully corrected the classification error for a few frames just after starting the event.

Furthermore, Figure 7 (top) shows the S_{NAS} scores from another fragment of the UTFPR-HSD1 dataset. It is observed that, in most cases, the CAE-CE reinforced the S_{NAS} scores to better approximate the ground truth, thus classifying events with higher confidence than the CAE. In the frame shown in Figure 7 (a), the CAE-CE was able to correct the misclassification caused by the CAE. In the frame shown in Figure 7 (b), the CAE-CE was able to detect the abnormal event (the white truck) from the very beginning, when it is entering the scene, whilst the CAE can do this only some frames later. In Figure 7 (c), both methods successfully had their highest S_{NAS} scores when the anomaly was in the middle of the scene. For this particular moment, CAE displayed a higher value than CAE-CE. In Figure 7 (d), the CAE-CE showed the lowest S_{NAS} score when the scene became completely empty, meaning that it detected normality with high confidence. In general, the analysis of results suggests that, for this kind of dataset, the compactness introduced by the CAE-CE reduces S_{NAS} fluctuations, improving the overall classification performance.

On the other hand, in some cases, the classification performance does not increase significantly after introducing more compactness by CAE-CE. This is the case with the Avenue dataset, shown in Figure 8. In this video fragment, a boy (the abnormal event) slowly enters the scene from the left. He stands for a moment and then leaves. The first part of the video does not cause any significant fluctuations of S_{NAS} scores for both methods, as shown in the frame of Figure 8 (a). Just before Figure 8 (b), the boy starts to leave the scene, causing some fluctuations in the S_{NAS} scores. At this time, the CAE-CE presents a better performance, keeping the S_{NAS} closer to the classification threshold. From this point on, only normal events occur in the video, as illustrated by the frame shown in Figure 8 (c). No significant difference is noticeable regarding the performance of the methods.

VI. CONCLUSIONS

In this work, we proposed the CAE-CE, a novel approach for anomaly detection problems in images and videos. CAE-CE is based on the Kullback-Leibler divergence for learning compact (dense) representations. We showed experimentally that such compactness is able to increase the separability between normal examples and anomalies.

The qualitative analysis of frames at the visual level indicated that the fluctuations of S_{NAS} follow the events occurring in the frames. Despite the need for semantic interpretation of the frames' contents, the CAE-CE was shown to be more related to the anomalous events

than the CAE. Therefore, our experimental results support the initial hypothesis that compacting the representation of the normal concept in anomaly detection problems can be valuable for increasing the classification performance.

Throughout the experiments, the need to establish an effective stop criterion for the training phase of the CAE-CE has emerged. We showed that the proposed sensitivity-based criterion is a plausible alternative, considering the nature of OCC problems, where the anomalous events are unknown. Results suggest that, even though the sensitivity method is not perfect, it can be seen as an approximation to the optimal stop point. On the other hand, experiments showed that training the CAE-CE model beyond the best TPR epoch compromises the classification performance. Experimental results suggest that the TPR-based approach may be useful for other similar OCC problems. Future experiments will assess its generalization capability for other datasets.

Besides, all experiments to validate the proposed approach were performed using publicly available datasets, including the new datasets introduced, UTFPR-HSD1 and UTFPR-HSD2, fully-annotated and designed for several tasks, including anomaly detection in highways. It is important to emphasize that there are no similar datasets available in the related literature, and an annotated dataset may be welcome to foster other research in this area.

One of the hardest tasks in automatic video analysis is to identify anomalies, since most of the time only normal events take place. Anomalies are ill-defined events that happen unexpectedly in a given context. Consequently, the development of computer vision methods has been the subject of growing research in recent years, on both, theoretical and practical grounds. In this sense, this work raised important issues for anomaly detection in videos from the point of view of an OCC problem. We believe that the methodological contributions of this work can be promptly applied to real-world problems and, so, further research is encouraged. Additionally, future research directions shall focus on extending the proposed CAE-CE to automatically estimate the ideal number of clusters, as well as to investigate the usefulness of temporal information to improve the classification performance.

ACKNOWLEDGMENT

M. Ribeiro thanks the Catarinense Federal Institute of Education, Science and Technology (IFC) *Campus* Videira and IFC / CAPES / Prodoutoral for both support and scholarship; M. Gutoski thanks CNPq for the scholarship number 141983/2018-3; H.S.Lopes thanks CNPq for the research grants no. 311778/2016-0 and 423872/20168. Special thanks to NVIDIA Corporation for the donation of the Titan-Xp GPUs used in this work and Lucas A. Albini for the support with the UTFPR datasets.

REFERENCES

[1] W. J. Scheirer, A. D. R. Rocha, A. Sapkota, and T. E. Boult, "Towards open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013.

[2] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, Nov. 2017.

[3] M. Amorim, F. D. Bortoloti, P. M. Ciarelli, E. O. T. Salles, and D. C. Cavalieri, "Novelty detection in social media by fusing text and image into a single structure," *IEEE Access*, vol. 7, pp. 132786–132802, 2019.

[4] S. Park, M. Kim, and S. Lee, "Anomaly detection for HTTP using convolutional autoencoders," *IEEE Access*, vol. 6, pp. 70884–70901, 2018.

[5] S. Bhakat and G. Ramakrishnan, "Anomaly detection in surveillance videos," in *Proc. 26th Int. Conf. High Perform. Comput., Data Analytics Workshop (HiPCW)*, New York, NY, USA: Association Computing Machinery, Dec. 2019, pp. 252–255.

[6] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[7] D. M. J. Tax, "One-class classification," Ph.D. dissertation, Adv. School Comput. Imag., Tech. Univ. Delft, Delft, The Netherlands, 2001.

[8] J. B. Camiña, M. A. Medina-Pérez, R. Monroy, O. Loyola-González, L. A. P. Villanueva, and L. C. G. Gurrola, "Bagging-RandomMiner: A one-class classifier for file access-based masquerade detection," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 959–974, Jul. 2019.

[9] D. Xu, R. Song, X. Wu, N. Li, W. Feng, and H. Qian, "Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts," *Neurocomputing*, vol. 143, pp. 144–152, Nov. 2014.

[10] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler, "Kernel null space methods for novelty detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, Jun. 2013, pp. 3374–3381.

[11] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, Jan. 2017.

[12] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understand.*, vol. 156, pp. 117–127, Mar. 2017.

[13] S. Naseer, Y. Saleem, S. Khalid, M. K. Bashir, J. Han, M. M. Iqbal, and K. Han, "Enhanced network anomaly detection based on deep neural networks," *IEEE Access*, vol. 6, pp. 48231–48246, 2018.

[14] E. Duman and O. A. Erdem, "Anomaly detection in videos using optical flow and convolutional autoencoder," *IEEE Access*, vol. 7, pp. 183914–183923, 2019.

[15] H. A. Perlin and H. S. Lopes, "Extracting human attributes using a convolutional neural network approach," *Pattern Recognit. Lett.*, vol. 68, pp. 250–259, Dec. 2015.

[16] S. Biswas and R. Venkatesh Babu, "Anomaly detection via short local trajectories," *Neurocomputing*, vol. 242, pp. 63–72, Jun. 2017.

[17] G. H. F. de Carvalho, L. A. Thomaz, A. F. da Silva, E. A. B. da Silva, and S. L. Netto, "Anomaly detection with a moving camera using multiscale video analysis," *Multidimensional Syst. Signal Process.*, vol. 30, no. 1, pp. 311–342, Jan. 2019.

[18] T. Qasim and N. Bhatti, "A low dimensional descriptor for detection of anomalies in crowd videos," *Math. Comput. Simul.*, vol. 166, pp. 245–252, Dec. 2019.

[19] S. Biswas and V. Gupta, "Abnormality detection in crowd videos by tracking sparse components," *Mach. Vis. Appl.*, vol. 28, nos. 1–2, pp. 35–48, Feb. 2017.

[20] A. S. Hassanein, M. E. Hussein, W. Gomaa, Y. Makihara, and Y. Yagi, "Identifying motion pathways in highly crowded scenes: A non-parametric tracklet clustering approach," *Comput. Vis. Image Understand.*, vol. 191, Feb. 2020, Art. no. 102710.

[21] P. Bodesheim, A. Freytag, E. Rodner, and J. Denzler, "Local novelty detection in multi-class recognition problems," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.* Piscataway, NJ, USA: IEEE press, Jan. 2015, pp. 813–820.

[22] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Understand.*, vol. 172, pp. 88–97, Jul. 2018.

[23] V. Murino, S. Gong, C. C. Loy, and L. Bazzani, "Image and video understanding in big data," *Comput. Vis. Image Understand.*, vol. 156, pp. 1–3, Mar. 2017.

[24] M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes, "A study of deep convolutional auto-encoders for anomaly detection in videos," *Pattern Recognit. Lett.*, vol. 105, pp. 13–22, Apr. 2018.
[25] Y. Xu, L. Lu, Z. Xu, J. He, J. Zhou, and C. Zhang, "Dual-channel CNN for efficient abnormal behavior identification through crowd feature engineering," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 945–958, Jul. 2019.
[26] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Piscataway, NJ, USA: IEEE Press, Jun. 2016, pp. 733–742.
[27] Y. Wang, X. Li, and X. Ding, "Probabilistic framework of visual anomaly detection for unbalanced data," *Neurocomputing*, vol. 201, pp. 12–18, Aug. 2016.
[28] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Piscataway, NJ, USA: IEEE Press, Jun. 2016, pp. 5147–5156.
[29] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 132–149.
[30] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, vol. 80, Jul. 2018, pp. 4393–4402.
[31] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. 33rd Int. Conf. Mach. Learn. (JMLR)*, 2016, pp. 478–487.
[32] M. Gutoski, M. Ribeiro, N. M. Romero Aquino, A. E. Lazzaretti, and H. S. Lopes, "A clustering-based deep autoencoder for one-class image classification," in *Proc. IEEE Latin Amer. Conf. Comput. Intell. (LA-CCI)*. Piscataway, NJ, USA: IEEE Press, Nov. 2017, pp. 1–6.
[33] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, Apr. 2017.
[34] S. Kim, Y. Choi, and M. Lee, "Deep learning with support vector data description," *Neurocomputing*, vol. 165, pp. 111–117, Oct. 2015.
[35] M. Ribeiro, "Deep learning methods for detecting anomalies in videos: Theoretical and methodological contributions," Ph.D. dissertation, Federal Univ. Technol., Akure, Nigeria, Feb. 2018.
[36] N. Shental, T. Hertz, D. Weinshall, and M. Pavel, "Adjustment learning and relevant component analysis," in *Proc. 7th Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2002, pp. 776–792.
[37] D.-Y. Yeung, H. Chang, and G. Dai, "A scalable kernel-based algorithm for semi-supervised metric learning," in *Proc. 20th Int. Joint Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2007, pp. 1138–1143.
[38] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," *J. Mach. Learn. Res.*, vol. 6, no. 1, pp. 937–965, 2006.
[39] A. E. Lazzaretti and D. M. J. Tax, "An adaptive radial basis function kernel for support vector data description," in *Proc. 3rd Int. Workshop Similarity-Based Pattern Recognit.* Cham, Switzerland: Springer, 2015, pp. 103–116.
[40] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Piscataway, NJ, USA: IEEE Press, Jun. 2016, pp. 1239–1248.
[41] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognit.*, vol. 58, pp. 121–134, Oct. 2016.
[42] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. 21st Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer-Verlag, 2011, pp. 52–59.
[43] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biol. Cybern.*, vol. 59, nos. 4–5, pp. 291–294, Sep. 1988.
[44] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *Proc. 19th Eur. Symp. Artif. Neural Netw.*, 2011, pp. 489–494.
[45] L. V. D. Maaten, "Learning a parametric embedding by preserving local structure," in *Proc. 12th Int. Conf. Artif. Intell. Statist., (AISTATS)*, vol. 5, 2009, pp. 384–391.
[46] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, May 2000.
[47] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
[48] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[49] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.
[50] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[51] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* Piscataway, NJ, USA: IEEE Press, Jun. 2012, pp. 2112–2119.
[52] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* Piscataway, NJ, USA: IEEE Press, Jun. 2010, pp. 1975–1981.
[53] The Theano Development Team *et al.*, "Theano: A Python framework for fast computation of mathematical expressions," 2016, *arXiv:1605.02688*. [Online]. Available: http://arxiv.org/abs/1605.02688
[54] S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sønderby, and D. Nouri, "Lasagne: First release," Zenodo, Geneva, Switzerland, Tech. Rep. 01, 2015.
[55] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[56] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.* Piscataway, NJ, USA: IEEE Press, Jun. 2011, pp. 215–223.
[57] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Piscataway, NJ, USA: IEEE Press, Dec. 2013, pp. 2720–2727.
[58] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* Piscataway, NJ, USA: IEEE Press, vol. 2, Jun. 2009, pp. 935–942.

MANASSÉS RIBEIRO received the Ph.D. degree in computer engineering from the Federal University of Technology–Paraná, Brazil, in 2018. Since 2010, he has been a Professor of computer science with the Catarinense Federal Institute of Education, Science and Technology (IFC), Videira, Brazil. He is currently a Charter Member of the Brazilian Association of Computational Intelligence (ABRICOM). His research interests include machine learning, pattern recognition, and image processing.

MATHEUS GUTOSKI received the bachelor's degree in information technology from Santa Catarina State University, Brazil, in 2015, and the M.Sc. degree in computer engineering from the Federal University of Technology–Paraná (UTFPR), Brazil, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include deep learning, pattern recognition, and computer vision.

ANDRÉ E. LAZZARETTI received the B.Sc., M.Sc., and D.Sc. degrees in electrical engineering from the Federal University of Technology–Paraná, in 2007, 2010, and 2015, respectively. He is currently a Professor with the Department of Electronics, Federal University of Technology–Paraná. His research interests include machine learning, instrumentation, and digital signal processing.

HEITOR S. LOPES received the Ph.D. degree in electrical engineering from the Department of Electronics, Federal University of Technology–Paraná, Curitiba, Brazil, in 1996. Since 2008, he has been a Researcher of the Brazilian National Research Council (CNPq). In 2014, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, USA. He is currently a Titular (Full) Professor with the Department of Electronics, Federal University of Technology–Paraná. He is also the Founder and the Head of the Bioinformatics and Computational Intelligence Laboratory (LABIC), and the current elected President of the Brazilian Association of Computational Intelligence (ABRICOM). He has advised 50 M.Sc. and Ph.D. students and published around 300 articles in conferences and journals.