Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5513582, 9 pages
https://doi.org/10.1155/2021/5513582

Research Article

A New Semantic and Statistical Distance-Based Anomaly Detection in Crowd Video Surveillance

Fariba Rezaei and Mehran Yazdi

School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

Correspondence should be addressed to Mehran Yazdi; yazdi@shirazu.ac.ir

Received 5 February 2021; Revised 9 April 2021; Accepted 7 May 2021; Published 17 May 2021

Academic Editor: Alireza Jolfaei

Recently, attention toward autonomous surveillance has been intensified and anomaly detection in crowded scenes is one of those significant surveillance tasks. Traditional approaches include the extraction of handcrafted features that need the subsequent task of model learning. They are mostly used to extract low-level spatiotemporal features of videos, neglecting the effect of semantic information. Recently, deep learning (DL) methods have been emerged in various domains, especially CNN for visual problems, with the ability to extract high-level information at higher layers of their architectures. On the other side, topic modeling-based approaches like NMF can extract more semantic representations. Here, we investigate a new hybrid visual embedding method based on deep features and a topic model for anomaly detection. Features per frame are computed hierarchically through a pretrained deep model, and in parallel, topic distributions are learned through multilayer nonnegative matrix factorization entangling information from extracted deep features. Training is accomplished through normal samples. Thereafter, K-means is applied to find typical normal clusters. At test time, after achieving feature representation through deep model and topic distribution for test frames, a statistical earth mover distance (EMD) metric is evaluated to measure the difference between normal cluster centroids and test topic distributions. High difference versus a threshold is detected as an anomaly. Experimental results on the benchmark Ped1 and Ped2 UCSD datasets demonstrate the effectiveness of our proposed method in anomaly detection.

1. Introduction

Automatic video surveillance has recently attracted the attention of researchers since a large number of cameras, installed in surrounding places, may not let human based-surveillance be error free. Thus, computer vision and machine learning come to help analyze the output videos for various tasks of automatic recognition and anomaly detection. Originally, raw signals are used to extract information through machine learning techniques [1]. However, the high dimensionality of video signals captured by high-resolution video cameras makes traditional methods computationally complex. Thereby, to combat the issue of curse of dimensionality, dimensionality reduction techniques have received more attention. Linear and nonlinear dimensionality reduction approaches can be applied as task-dependent techniques. PCA, MDS, LLE, and autoencoder are some to name a few. Generally speaking, all computer vision-based feature extraction methods like handcrafted features (SIFT, HOG, etc...) can also be considered a kind of dimensionality reduction.

New emerging embedding methods, basically introduced in natural language modeling/processing (NLP), map the original high-dimensional signals to embed spaces and consecutively capture high-level information, which besides the compression, the semantic relations of signals are also preserved [2, 3]. Embedding techniques in NLP are based on representing each word as a vector in a vector space model. Preliminary one hot encoding suffers from lack of preservation of semantic relations, since orthogonally between words neglects the probable coherence between them. Topic-based representations such as LSA, probabilistic LSA, LDA, and NMF try to capture semantics [3].

Embedding can also be applied to vision tasks to bridge the semantic gap in image or video analysis. Recently, deep learning architectures (CNN, RNN, AE, RBM, etc.) have been well studied for anomaly detection [4]. Diving into the high-level features, they have shown considerable results in comparison to handcrafted features. Supervised CNNs consist of both convolution and fully connected (FC) layer for feature extraction and classification/recognition, respectively. Ultraparameters in CNN are caused by those terminative FC layers, which may cause overfitting in limited dataset regimes when training from scratch. Therefore, attention is trended toward using only pretrained convolutional layers for feature extraction and powerful image representations, putting aside FC layers.

In most researches, anomaly detection is investigated based on defining a model(s) on normal samples and detecting anomalies as deviation from this normality. This deviation can be measured either by likelihood or similarity. In [5], an anomaly was defined based on interaction forces between pedestrians using the social force model (SFM), and LDA was used to compute likelihood for test set to evaluate deviation from a normal model in a probabilistic framework, whereas in [6, 7], normal training samples were used to create a dictionary model and deviation was calculated as high sparse reconstruction cost between an original test sample and its reconstruction through a linear combination of normal bases in the Euclidean space.

In this paper, we investigate a combination of the deep model, topic model, and statistical distance for anomaly detection. In contrast to previous methods which were based on either handcrafted or deep features, neglecting semantic and interpretable information, we analyze the combination of a deep model with a topic model hierarchically to produce semantic representation. We apply a pretrained deep model for hierarchical feature extraction from different layer levels, for each training image. Thereafter, we take the advantages of nonnegative matrix factorization (NMF) as a topic modeling approach in capturing semantic features. Specially, we applied a multilayer NMF, for hierarchical topic representation injecting information extracted from hierarchical layers of a deep model in hierarchical decompositions. After learning topic distribution per frame in the training stage, we apply K-means clustering to compute cluster centroids as typical normal topic-based representations. At test time, in a similar pipeline for feature extraction at the train stage, semantic representation for test frames is calculated and compared to typical normal topic distributions through a statistical distance metric. Here, the earth mover distance (EMD) metric is chosen as a distance metric since it has shown efficient performance in comparing distributions.

Our main contributions are as follows:

(1) We take the advantages of both the deep model (pretrained VGG-Net) and the topic model (multilayer NMF), hierarchically and in combination to reach high-level and semantic frame representation

(2) Since topic distributions are extracted at the final level as the frame representations, after K-means clustering, some normal representative topic distributions for normality are achieved, and then, EMD statistical distance metric is applied in clustering-based anomaly detection framework

The organization of the rest of this paper is as follows: literature review in three domains of anomaly detection, topic modeling, and statistical learning methods are provided in Section 2. Section 3 introduces our proposed pipeline for crowd anomaly detection. Experimental results are reported in Section 4. Finally, Section 5 concludes this paper.

2. Literature Review

In this section, we review researches in anomaly detection, topic modeling, and statistical distance separately.

2.1. Anomaly Detection. Video surveillance studies for anomaly detection was started by using traditional handcrafted feature extraction and model learning and improved over the years by applying end-to-end deep architectures. Formerly, low-level features like color, texture, and its variants, like mixture of dynamic texture (MDT), SIFT, SURF, optical flow, and trajectories, were extracted either from appearance, motion, or both, depending on the anomaly definition. At model learning stages, binary classifiers like SVM, decision tree, and NN have been applied for supervised scenarios [1]. However, in semisupervised and unsupervised scenarios, given only normal videos at the training stage, a model for normal behavior is created and an anomaly is detected as a deviation from this model. This has been done for instance by one-class SVM (OCSVM) or fitting a Gaussian model on normal samples. Some researchers took the idea of the inherent sparsity of vision. A dictionary was learned from normal samples, and at the test time, a large reconstruction error was interpreted as an anomaly. Reconstruction was done as a linear combination of dictionary bases which are representative of all normal samples. Dictionary can be learned offline through codebook generation or online through updating along with observing new normal samples [8].

Recently, deep learning methods have commenced entering to the practical realm like vision, lexical, and speech. The intermediate image representations learned through CNN, especially when trained on large-scale datasets like ImageNet, have been proven to be powerful image descriptors.

In [9], anomalous behaviors were captured through a novel concept of aggregation of ensembles (AOE), based on fine-tuning different pretrained ConvNets and a pool of classifiers. They assumed that different CNN architectures learn different levels of representation from crowd videos, and thus, an ensemble of CNNs will enable enriched feature sets to be extracted. Autoencoder-based architectures were also studied where a large reconstruction error was considered a sign of anomaly score. The autoencoder can reduce dimensionality and is vastly used in unsupervised learning problems or as the preliminary stage of supervised task [10]. In particular, after training an AE or sparse AE on normal samples, the bottleneck layer can be considered feature extraction layers for any test samples. Some researchers tried to incorporate both handcrafted and deep features in a unified configuration. In [11], a trajectory-pooled deep convolutional descriptor was introduced combining dense trajectories and convolutional feature maps which results in high discriminative features. Convolutional networks outperform both traditional low-level features and their compositional forms like BoW, Fisher Kernel, and VLAD, [12] although sometimes are used cooperatively. In [12], features extracted from within layers of a convolutional network were used in VLAD to compress the data and subsequently feed to SVM for classification. Wimmer et al. [13] applied Fisher vector encoding to the output feature maps of CNN to find fixed-length representation for image classification.

Sabokrou et al. investigated video anomaly detection through different deep architectures [14–21]. Autoencoder-based anomaly detection and localization using sparsity was introduced in [14, 15]. An architecture based on deep 3D autoencoder, deeper 3D convolutional neural network (CNN), and cascade of two cascaded classifiers was proposed in [16] for anomaly detection. High speed and accurate detection and localization of anomalies were achieved in [18] using fully convolutional neural networks (FCNs) and cascaded outlier detection. Some researches applied generative adversarial networks and its variants for image anomaly detection [17, 19, 22]. Semisupervised anomaly detection was analyzed in [23] based on information theory. A novel self-supervised representation learning based on integration of a neighbourhood-relational encoding (NRE) among the training data and an encoder-decoder structure was proposed in [20]. In [21], they propose an adversarial training approach to detect out-of-distribution samples in an end-to-end model through jointly training two deep neural networks which collaborate at test time to detect novelties.

2.2. Topic Modeling. Topic modeling is an unsupervised method, originally introduced for text analysis, but has been also noticed in vision. It is based on the idea that documents containing similar contents will likely use a similar set of words that are indicated by topics. Topic modeling discovers patterns as low-dimensional latent representation given unlabeled collection of documents constituted of words. pLSA, LDA, and NMF are among the most common probabilistic topic modeling approaches [24–26]. Topic models take as input a set of documents J, a set of words V, and in a cooccurrence matrix of words and documents $F = \|n_{wj}\|_{w\epsilon V.j\epsilon J}$ (or BoVW representation, and produce a set of topic T, or more especially $P(w \mid k)$ and $p(k \mid j)$, for $w \in V.j \in J.k \in T$, as word distribution per topic and topic distribution per document, respectively. Consider n_{wj} as the number of times the word w appears in document j, then documents can be represented as mixtures of topics.

F can be decomposed into two matrices $F = \Phi\Theta$, where $\Phi = \{\phi_{wk}\}_{w\epsilon V.k\epsilon K}$ is a word-topic matrix with $\phi_{wk} = p(w \mid k)$ and $\phi_k = \{\phi_{wk}\}_{w\epsilon V}$, and $\Theta = \{\theta_{kj}\}_{k\epsilon K.j\epsilon J}$ is a topic-document matrix with $\theta_{kj} = p(k \mid j)$ and $\theta_j = \{\theta_{kj}\}_{k\epsilon K}$. The decomposition can be solved through the various topic model algorithms with a different assumption. For instance, LDA uses a predefined number of topics, whereas hierarchical Dirichlet process (HDP) [27] estimates the best number of topics based on the training dataset.

In [28], Niebles et al. studied the application of latent topic models, namely, pLSA and LDA, for action categorization. Especially, they extract spatiotemporal interest points along the input volumes followed by codebook generation. In an unsupervised fashion, they succeeded in detecting and localizing actions, which were considered latent topics. New learning algorithms based on EM and variational Bayes inference were proposed in [29] for activity analysis in videos where the description of activities and behaviors was made by the dynamic topic model. The activities and behaviors were described by a dynamic topic model. They also evaluated anomaly localization procedures in the topic modeling framework. In [30], scene classification was made by discovering objects per image in an unsupervised fashion using pLSA. They subsequently used object distribution in each image for scene classification using supervised kNN. Topic modeling-based abnormal behavior recognition has been previously investigated in [5, 31]. In almost all cases, low likelihood corresponds to abnormal test samples. An unsupervised topic model (pLSA) anomaly detection and localization were studied in [32] based on extra information of location and size beside quantized spatiotemporal gradient descriptors to create a more informative vocabulary over visual clips. Each document (frame) is fully described by a corresponding distribution over topics.

2.3. Statistical Distance. Statistical distances try to find the distance between two statistical objects, and when accompanied with a symmetric property, they are known as a metric. In the anomaly detection area, distance measures such as Jensen Shannon divergence or Z score value were applied for comparing query observation to those extracted patterns from normal samples [33]. According to the evaluation of this distance concerning the threshold, the anomaly can be detected. As a powerful statistical distance, earth mover distance (EMD), also known as the Wasserstein metric, was applied in the image domain [34, 35] to compare two probability distributions, mainly based on low-level features like color or texture. It is based on computing statistical distance between two signatures. The typical signature consists of a list of pairs:

$$S = \{(x_1.m_1).(x_2.m_2) \cdots (x_n.m_n)\}, \tag{1}$$

where each x_i is a certain feature, and m_n is its mass (how many times that feature occurs in the record). Considering two signatures P and Q which contain m and n clusters, respectively,

$$P = \{(p_1.w_{p1}).(p_2.w_{p2}) \cdots (p_m.w_{pm})\}, \tag{2}$$

$$Q = \{(q_1.w_{q1}).(q_2.w_{q2}) \cdots (q_n.w_{qn})\}, \tag{3}$$

and $p_i(q_i)$ is the cluster representative and $w_{pi}(w_{qi})$ is the weight of cluster i. Also, consider $D = [d_{i.j}]$ as the ground distance between clusters p_i and q_j. It can be chosen or learned according to the problem at hand. The aim is to find flow matrix $F = [f_{i.j}]$, where $f_{i.j}$ is the flow between p_i and q_j, such that the below overall cost is minimized with its related constraints.

$$\begin{gathered}\min \sum_{i=1}^{m}\sum_{j=1}^{n} f_{i.j} d_{i.j},\\ f_{i.j} \geq 0\ 1 \leq i \leq m.1 \leq j \leq n,\\ \sum_{j} f_{i.j} \leq w_{pi} 1 \leq i \leq m,\\ \sum_{i} f_{i.j} \leq w_{qi}.1 \leq j \leq n,\\ \sum_{i=1}^{m}\sum_{j=1}^{n} f_{i.j} = \min \left\{ \sum_{j=1}^{m} W_{pi}. \sum_{j=1}^{n} W_{qj} \right\}.\end{gathered} \tag{4}$$

This optimization can be solved via linear programming. It is based on solving a kind of transportation problem. Once the flow F is calculated, then the EMD is defined as the work normalized by the total flow:

$$\mathrm{EMD}(P.Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i.j} d_{i.j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i.j}}. \tag{5}$$

EMD suffers from high computational complexity $O(N^3 \log N)$. Wavelet EMD was proposed in [36] to reach a linear time algorithm for approximating the EMD for low-dimensional histograms using the sum of absolute values of the weighted wavelet coefficients of the difference histogram.

Rare studies have gained from EMD in anomaly detection. To the best of our knowledge, only in [7], wavelet EMD was applied in conjunction with sparse representation for anomaly detection instead of the Euclidean distance, for its robustness. In this paper, we investigate wavelet EMD on our proposed clustering-based anomaly detection.

3. Proposed Method

In this paper, we analyze anomaly detection at frame level in crowded scenes. Our proposed architecture is shown in Figure 1. The pipeline consists of two stages: (1) feature extraction and (2) anomaly detection. The feature extraction stage itself consists of two parts entangled with each other: (1) hierarchical feature extraction through pretrained VGG-Net [37] and (2) hierarchical latent representation from multilayer NMF. Both architectures start from low-level features and increase in depth to high-level information resulting in ultimate representation.

In the second stage, we applied clustering-based anomaly detection. Precisely, K-means is applied to all processed training samples' ultimate representations, to create typical normal clusters. Since the training dataset consists of only normal samples, thus, cluster centroids are normal frame representatives. At test time, test frames are processed to be represented in learned topic space from the training stage and compared to each cluster centroids. A large statistical distance from all centroids is detected as an anomaly. In the following, we explain each part in more detail.

3.1. Preprocessing and Feature Extraction

3.1. Preprocessing and Feature Extraction. The dataset is separated into two subsets as train and test set. Let $X_{\mathrm{train}} = [x_1.x_2 \cdots x_{n_{\mathrm{Train}}}]^T \in R^{n_{\mathrm{Train}} \times B_0}$, where n_{Train} is the number of frames in the train dataset, $B_0 = m \times n \times c$ and m, n, and c are the width, height, and number of channel, respectively, for the original captured image.

3.1.1. Deep Representation. Pretrained model is applied for feature extraction in problems encountering scarcity of training datasets, since training from scratch may result in overfitting. As higher layer feature maps are task specific, we extract more general features from lower layers. We resized each frame to be in a compatible size as the input for VGG-Net model $(m_0 \times n_0 \times c_0)$ and extract features hierarchically from different depths of the architecture. Let $a^0 = x(\in R^{m_0 \times n_0 \times c_0})$ be a typical train image in compatible size with VGG input layer. Then,

$$a^l = f\left(w^{l-1} a^{l-1} + b^{l-1}\right) \in R^{m_l \times n_l \times c_l}, \tag{6}$$

is the output feature map from layer l. w^{l-1} and b^{l-1} are VGG weights and biases pretrained, respectively, for layer l. $m_l \times n_l$ is the spatial size of the feature map, and c_l is the feature map's depth at layer l. We extract feature maps from L different depths $(l = 1 \cdots .L)$; then, feature maps at each layer $l\,(l = 1.2 \cdots .L)$ are separately feed to the global average pooling (GAP) layer to get representations in vector format. GAP layers take input volumes of size $m_l \times n_l \times c_l$ and create $1 \times c_l$ dimensional vector by spatial averaging. Therefore, for each frame x, now, we have L vector representations, $f_{Dl} \in R^{c_l}(l = 1.2 \cdots .L)$. Considering all training samples, now we have L different size matrices, $M_l \in R^{n_{\mathrm{Train}} \times f_{Dl}}$.

3.1.2. Topic-Based Representation. In parallel, we try to capture semantic information based on the topic model. Specially, we applied multilayer NMF since multilayer has been shown to improve performance by capturing more semantic features [38]. We adopt a similar approach to [39] by considering a frame as a document and trying to extract topic distribution per document. However, we apply multilayer NMF for hierarchical topic modeling. Single-layer NMF decomposes a nonnegative matrix V into two low-rank nonnegative basis and coefficient matrices W and H.

$$\mathrm{V} = \mathrm{WH}'.\mathrm{V} \in R^{m \times n}.\mathrm{W} \in R^{m \times k}.\mathrm{H} \in R^{n \times k}, \tag{7}$$

where H is the new low-dimensional representation for V . The decomposition is solved as an optimization problem through a multiplicative update approach. In multilayer NMF, computed latent representation in preceding layers is decomposed hierarchically in subsequent layers. Consider $X_{\mathrm{train-pca}} = \mathrm{PCA}(X_{\mathrm{train-vec}})$ and $X_{\mathrm{train-pca}} = [x_1.x_2 \cdots x_{n_{\mathrm{Train}}}]^T \in R^{n_{\mathrm{Train}} \times D_0}$, where PCA applied to each vectorized frame to decrease dimensionality from $m_0 \times n_0$ to $D_0 < m_0 \times n_0$ per

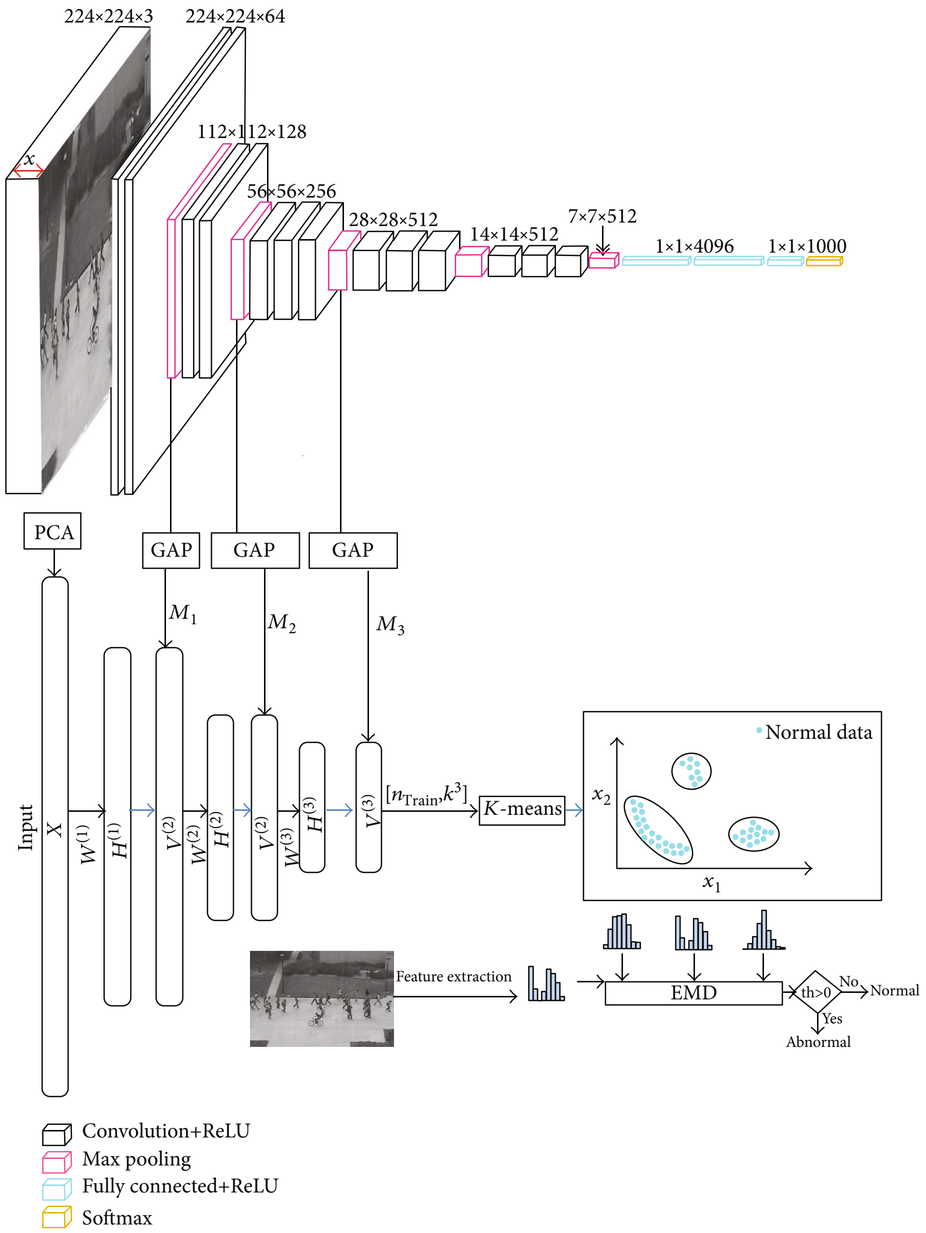

Figure 1: Our proposed architecture for anomaly detection. It consists of two stages of hierarchical feature representation and cluster-based anomaly detection.

frame and standardized to stay in range [0-1] . Let $H_0 = X_{\text{train-pca}}$ as input to the first stage of multilayer NMF. Then, it can be decomposed as $H_0 = W_1 H_1$. Instead of directly applying the second NMF to H_1, as the new low-dimensional representation, H_1 is processed to V_1 before being introduced to the next layer. V_l is computed as $V_l = f(H_l.M_l).l = 1 \cdots L$ where $f(.)$ is the nonlinear function, like softmax, and M_l is feature representation from pretrained VGG-Net at layer l .

$$V_l = f(H_l.M_l) = W_{l+1}H'_{l+1}.W_{l+1} \in R^{D_{l-1}\times D_l}.H_{l+1} \in R^{n_{\text{Train}}\times D_l}. \quad (8)$$

Table 1: UCSD dataset in detail.

Dataset	Resolution	Number of training sequences	Number of test sequences
Ped1	158 × 238	34~200 images	36~200 images
Ped2	240 × 360	16120~200 images	12120~200 images

Here, we use softmax as a nonlinear function to have a distribution-like representation. Since the ReLu activation function has been applied in deep architecture, nonnegativity is preserved. Bringing in M_ls in multilayer NMF decomposition results in both high-level and semantic information, which can improve the performance of the subsequent tasks.

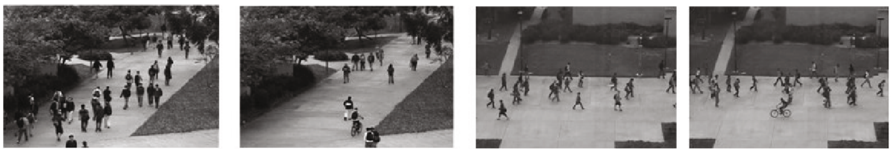

Figure 2: Typical normal and abnormal samples of the UCSD dataset. Left to right: normal frame and abnormal frame for Ped1 and normal frame and abnormal frame for Ped2.

By decomposing V_l in the next layer, we force the architecture to learn how to combine information from the previous layer; therefore, $D_l < D_{l-1}$. Training separately each NMF layer, to learn W_l and H_l, ultimate data representation V_L is acquired. Finally, V_L integrates features throughout the deep model and topic model.

3.2. Anomaly Detection. Upon training completion, $V_L \in R^{n_{\text{Train}} \times D_L}$ is acquired from normal frames in the training set. We apply K-means algorithm to V_L to find K cluster centroids as normality representatives. Therefore, now, we have K cluster centroids $s_i.i = 1 \cdots .K$ which are used in cluster-based anomaly detection. Each test frame x_{test} is fed to our learned feature extraction block from the training phase, and ultimate representation $V_{L.test}$ is acquired. $V_{L.test}$ can be considered as the final topic distribution for x_{test}. $V_{L.\text{test}}$ is compared to each s_i and exceedance of statistical wavelet EMD distance from threshold th is detected as an anomaly.

$$\underset{i=1\,:\,K}{\min}\left(d_{\text{EMD}.i}\left(V_{L.\text{test}}.s_i\right)\right) > \text{th} \rightarrow V_{L.\text{test}}, \quad (9)$$

is an abnomal frame.

4. Results and Discussion

We conducted experimental analysis on UCSD dataset as one of the benchmark datasets in crowd anomaly detection introduced in [40], recorded with a static camera at 10 fps. This dataset contains two scenes as Ped1 and Ped2, each of which is split into train and test sequences. The nonpedestrian objects, like bikers, skaters, and small carts, are considered anomalies. More details about this dataset are provided in Table 1. Typical normal and abnormal sample frames for Ped1 and Ped2 datasets are also shown in Figure 2.

When originally introduced, VGG [37] was trained on the ImageNet dataset which only consists of object classes; however, recently, pretrained VGG on both the ImageNet and Places dataset is provided which consider scene classes, as well. 1000 classes from the ImageNet and the 365 classes from the Places365Standard [41] were merged to train a VGG16-based model (Hybrid1365-VGG [42]). We use VGG model pretrained both on the ImageNet and Places datasets to improve the capability of our deep feature extraction block in capturing both objects and scenes features. For this paper, our algorithms have been implemented in Python and run on a PC with 2.9 GHz Core i5 GPU, with GTX1080 GPU, and 16G RAM. Original frames are resized to be compatible with VGG, as VGG accepts input of size 224 × 224 × 3 . Feature maps from different depths, namely, block2 – pool, block3 – pool, and block4 – pool of VGG architecture, were extracted and resulted in $(56 \times 56 \times 128)$, $(28 \times 28 \times 256)$, and $(14 \times 14 \times 512)$ feature maps, respectively. Then, we applied global average pooling to each feature map separately which results in $f_{D1} : 128D, f_{D2} : 256D$, and $f_{D3} : 512D$ representation vectors in hierarchical order. On the other hand, we applied multilayer NMF with $L = 3$ on our train set with reduced dimensionality by PCA (2000D vector each frame). W_0 , W_1, and W_2 are learned separately with a multiplicative updates. D_1, D_2, and D_3 are chosen as 512, 256, and 128, respectively. K-means clustering with $K = 50$ is applied to the final representation $V_L \in R^{n_{\text{Train}} \times D_L}$ to generate typical representative centroids. In the UCSD dataset, there are $n_{\text{Train}} = 6800$ for Ped1 and $n_{\text{Train}} = 2550$ for Ped2 datasets.

Table 2: Fixed parameter used in the proposed algorithm.

Dataset/parameters	Ped1	Ped2
Number of training samples	2550	6800
L (number of levels for feature hierarchies)	3	
K (K-means clustering)	50	
Threshold (for WEMD distance comparison)	0.33	0.24

In our experiment, there are some parameters that we investigate their values and fixed after evaluation. These parameters are shown in Table 2.

VGG16 consists of several layers (C11-C12-P1-C21-C22-P2-C31-C32-C33-P3-C41-C42-C43-P4-C51-C52-C53-P5-FC1-FC2-FC3). Convolutions and fully connected layers have trainable parameters. Three last fully connected layers provide task specific features. So, we focus on first 5 convolution layers. We chose $L = 3$ to achieve a trade-off between accuracy and complexity. The number of clusters in K -means clustering was also evaluated for $K = 30{,}40{,}50{,}60$ and chosen as $K = 50$ based on accuracy evaluation. We decided on the value of threshold for WEMD comparison based on average distance from training samples representations, since the training dataset consists only of normal samples.

For Ped 1, we compare our proposed approach both to traditional methods (SRC [6], MPPCA [43], and MDT [40]) and high-level deep learning-based methods (AVID [19], Sabokrou [8], and deep cascade [16]). As introduced and calculated in [26], evaluation metrics such as equal error rate (EER) and area under curve (AUC) are computed at frame level and compared to the state-of-the-art methods. EER indicates the point where false positive rate equals to false negative rate. The lower the EER is, the higher accuracy can be achieved. A comparison of EER of our proposed

Table 3: Comparison of AUC performance for the UCSD Ped1 dataset at frame level.

Method	SRC [6]	MPPCA [43]	MDT [40]	AVID [19]	Sabokrou [8]	Deep cascade [16]	Proposed approach
EER	19	40	25	12.3	8.4	9.1	8.1
AUC	86	59	81.8	—	93.2	—	93.9

Table 4: Comparison of EER performance for the UCSD Ped2 dataset at the frame level.

Method	SF [5]	MPPCA [43]	MDT [40]	Conv-AE [44]	AVID [19]	Deep anomaly [18]	Deep cascade [16]	ALOCC [17]	ST-AE [45]	Proposed approach
EER	42	36.0	24.0	21.7	14.	13.5	9.	13	12.0	6.1
AUC	63	71	85	90	—	—	—	—	87.4	97.3

Table 5: Accuracy criteria for the Ped 1 and Ped 2 datasets.

Dataset/criteria	Ped1	Ped2
Accuracy	90.3	95.4

approach to the previous method is shown in Table 3 for Ped1. Results show the comparable performance for our proposed method. Besides, AUC as the area under ROC curve is computed and compared to the state-of-the-art. Results show the outperformance of our proposed approach in AUC, as well.

For Ped 2, The Ped1 dataset suffers from the perspective problem. For this reason, most researches have been conducted on Ped2. We compare our proposed approach both to traditional methods (SF [5], MPPCA [43], and MDT [40]) and high-level deep learning-based methods (Conv-AE [44], AVID [19], deep anomaly [18], deep cascade [16], ALOCC [17], and ST-AE [45]). A comparison of EER of our proposed approach to the previous method is shown in Table 4 for Ped2. Results show the comparable performance for our proposed method. Besides, AUC is computed and compared to the state-of-the-art. Results show the outperformance of our proposed approach in AUC.

Moreover, we evaluated accuracy as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}. \quad (10)$$

The results, shown in Table 5 for the Ped1 and Ped2 datasets, indicate the high performance of our proposed method.

5. Conclusions

In this paper, we discussed a new semantic and statistical distance-based crowd anomaly detection at the frame level. In particular, inspired by the earth mover distance metric applied previously on low-level vision features, we applied this statistical distance to hierarchically learned features, through pretrained deep convolutional neural network and topic model, for anomaly detection. Features from VGG-Net, pretrained on hybrid dataset (Places dataset and ImageNet dataset) and multilayered NMF as semantic interpretable features, were computed in combination as hierarchical representation and used in clustering-based anomaly detection using wavelet EMD statistical distance. Experimental results show the outperformance of our proposed approach. In the future, we will investigate anomaly localization by patch analysis through the kernel convolutional network (CKN) [46] and EMD in a similar framework to localize anomalies.

Data Availability

The readers can access the UCSD Ped1 and Ped2 datasets in http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems the MIT press:neurIPS proceedings*, pp. 3111–3119, 2013.

[3] P. Wiriyathammabhum, D. Summers-Stay, C. Fermuller, and Y. Aloimonos, "Computer vision and natural language processing: recent approaches in multimedia and robotics," *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, pp. 1–44, 2016.

[4] R. Wang, K. Nie, T. Wang, Y. Yang, and B. Long, "Deep learning for anomaly detection," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 894–896, Houston, TX, USA, 2020.

[5] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 935–942, Miami, FL, USA, 2009.

[6] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR 2011*, pp. 3449–3456, Colorado Springs, CO, USA, 2011.

[7] X. Zhu, J. Liu, J. Wang, C. Li, and H. Lu, "Sparse representation for robust abnormality detection in crowded scenes," *Pattern Recognition*, vol. 47, no. 5, pp. 1791–1799, 2014.

[8] M. Sabokrou, M. Fathy, Z. Moayed, and R. Klette, "Fast and accurate detection and localization of abnormal behavior in crowded scenes," *Machine Vision and Applications*, vol. 28, no. 8, pp. 965–985, 2017.

[9] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned ConvNets," *Neurocomputing*, vol. 371, pp. 188–198, 2020.

[10] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semisupervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[11] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deepconvolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4305–4314, Boston, MA, USA, 2015.

[12] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1798–1807, Boston, MA, USA, 2015.

[13] G. Wimmer, A. Vécsei, M. Häfner, and A. Uhl, "Fisher encoding of convolutional neural network features for endoscopic image classification," *Journal of Medical Imaging*, vol. 5, no. 3, article 034504, 2018.

[14] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 56–62, Boston, MA, USA, 2015.

[15] M. Sabokrou, M. Fathy, and M. Hoseini, "Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder," *Electronics Letters*, vol. 52, no. 13, pp. 1122–1124, 2016.

[16] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, 2017.

[17] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3379–3388, Salt Lake City, UT, USA, 2018.

[18] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, vol. 172, pp. 88–97, 2018.

[19] M. Sabokrou, M. Pourreza, M. Fayyaz et al., "Avid: adversarial visual irregularity detection," in *Asian Conference on Computer Vision*, pp. 488–505, Springer, Cham, 2018.

[20] M. Sabokrou, M. Khalooei, and E. Adeli, "Self-supervised representation learning via neighborhoodrelational encoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8010–8019, Seoul, Korea (South), 2019.

[21] M. Sabokrou, M. Fathy, G. Zhao, and E. Adeli, "Deep end-to-end one-class classifier," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 675–684, 2021.

[22] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft, "Image anomaly detection with generative adversarial networks," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 3–17, Springer, Cham, 2018.

[23] L. Ruff, R. A. Vandermeulen, N. Görnitz et al., "Deep semi-supervised anomaly detection," 2019, arXiv preprint arXiv:1906.02694.

[24] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147–153, 2015.

[25] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception by hierarchical Bayesian models," in *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, Minneapolis, MN, USA, 2007.

[26] X. Wang and E. Grimson, "Spatial latent Dirichlet allocation," in *Advances in neural information processing systems. MIT Press: Cambridge*, pp. 1577–1584, MA, USA, London, UK, 2008.

[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[28] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[29] O. Isupova, D. Kuzin, and L. Mihaylova, "Learning methods for dynamic topic modeling in automated behavior analysis," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 9, pp. 3980–3993, 2018.

[30] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *European Conference on Computer Vision*, pp. 517–530, Springer, Berlin, Heidelberg, 2006.

[31] O. P. Popoola and Kejun Wang, "Video-based abnormal human behavior recognition| a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, 2012.

[32] D. Pathak, A. Sharang, and A. Mukerjee, "Anomaly localization in topic based analysis of surveillance videos," in *2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 389–395, Waikoloa, HI, USA, 2015.

[33] Bo du and Liangpei Zhang, "A discriminative metric learning based anomaly detection method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 6844–6857, 2014.

[34] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[35] M. A. Ruzon and C. Tomasi, "Edge, junction, and corner detection using color distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1281–1295, 2001.

[36] S. Shirdhonkar and D. W. Jacobs, "Approximate earth mover's distance in linear time," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, AK, USA, 2008.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv preprint arXiv:1409.1556.

[38] H. A. Song, B. K. Kim, T. L. Xuan, and S. Y. Lee, "Hierarchical feature extraction by multi-layer non-negative matrix factorization network for classification task," *Neurocomputing*, vol. 165, pp. 63–74, 2015.

[39] X. Wan, "A novel document similarity measure based on earth mover's distance," *Information Sciences*, vol. 177, no. 18, pp. 3718–3730, 2007.

[40] Weixin Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, 2014.

[41] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: a 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[42] L. Wang, Z. Wang, W. Du, and Y. Qiao, "Objectscene convolutional neural networks for event recognition in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 30–35, Boston, MA, USA, 2015.

[43] J. Kim and K. Grauman, "Observe locally, infer globally: a spacetime MRF for detecting abnormal activities with incremental updates," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 2921–2928, Miami, Fla, USA, 2009.

[44] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, Las Vegas, NV, USA, 2016.

[45] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X. S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proceedings of the 25th ACM international conference on multimedia, Mountain View*, pp. 1933–1941, California, USA, 2017.

[46] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," *Advances in Neural Information Processing Systems the MIT press:neurIPS proceedings*, pp. 2627–2635, 2014.