An Adaptive Framework for Anomaly Detection in Time-Series Audio-Visual Data

PRATIBHA KUMARI AND MUKESH SAINI
Department of Computer Science and Engineering, IIT Ropar, Rupnagar 140001, India
Corresponding author: Pratibha Kumari (2017csz0006@iitrpr.ac.in)

This work was supported by the Department Of Science & Technology (DST), Government of India for the Technology Innovation Hub at the IIT Ropar in the framework of National Mission on Interdisciplinary Cyber-Physical Systems.

ABSTRACT Anomaly detection is an integral part of a number of surveillance applications. However, most of the existing anomaly detection models are statically trained on pre-recorded data from a single source, thus making multiple assumptions about the surrounding environment. As a result, their usefulness is limited to controlled scenarios. In this paper, we fuse information from live streams of audio and video data to detect anomalies in the captured environment. We train a deep learning-based teacher-student network using video, image, and audio information. The pre-trained visual network in the teacher model distills its information to the image and audio networks in the student model. Features from image and audio networks are combined and compressed using principal component analysis. Thus, the teacher-student network produces an image-audio-based light-weight joint representation of the data. The data dynamics are learned in a multivariate adaptive Gaussian mixture model. Empirical results from two audio-visual datasets demonstrate the effectiveness of joint representation over single modalities in the adaptive anomaly detection framework. The proposed framework outperforms the state-of-the-art methods by an average of 15.00 % and 14.52 % in AUC values for dataset 1 and dataset 2, respectively.

INDEX TERMS Adaptive learning, concept drift, long term surveillance, multimodal anomaly detection, unsupervised model.

I. INTRODUCTION Anomaly detection for automated video surveillance has grown remarkably owing to the widespread availability of economic closed-circuit television (CCTV) cameras and breakthrough of deep generative models such as convolutional autoencoder, Long Short Term Memory (LSTM) autoencoder, Generative Adversarial Network (GAN), etc. [1]–[3]. Various reconstruction-based deep learning frameworks [4], [5] have been explored in the past that aim at modeling the normality from the available dataset. State-of-the-art deep prediction-based frameworks [6], [7] have further mitigated the issue of narrow regularity gap in deep reconstruction-based models by achieving better normalcy models, resulting in improved anomaly detection and localization accuracy [8]. However, the applicability of such models is limited to anomaly detection in short pre-recorded offline videos of duration 2-5 minutes. With offline videos, the data distribution is stationary and known a priori. In real-life, we need to detect anomalies in a continuous stream of audio-visual data from a CCTV camera, called online videos. The data distribution may change over time in online videos, i.e., a normal event might change into abnormal and vice-versa. The change in the data distribution over time is also called concept drift [9]. An example of concept drift is variation in crowd density witnessed from morning to night in a school. More crowd appears normal during school hours, whereas it becomes abnormal if observed at night. With the above-mentioned frameworks, the samples which were regarded abnormal in training data will always remain abnormal. Thus, with one-time training, the model’s performance may degrade when applied to data having concept drift [10].

To account for such drifts, the model needs to be updated online, also referred to as adaptive learning [10], [11]. Adaptive learning techniques facilitate lifelong dynamic updation of the model at the arrival of any data sample. Thus, unlike static learning-based frameworks, adaptive learning frameworks adapt through the changes in the environment and hence do not show drastic degradation in performance.
The development of adaptive anomaly detection frameworks has received less attention. There have been a few attempts to use the Adaptive Gaussian Mixture Model (AGMM) in an online manner for anomaly detection [10], [12], [13] or foreground detection [14], [15]. The parameters of the modes in the mixture continuously adjust themselves in order to accommodate varying data distributions.

However, these frameworks utilize a single source of information for anomaly modeling. Relying solely on visual data accompanies multiple assumptions, such as a clear vision in the camera’s field of view, clear night vision, unhindered performance in adverse weather conditions, robustness toward sudden light changes and shadows, etc. [16], [17]. These assumptions do not fit well with real-life scenarios. In real-life scenarios, humans observe important physical entities and their interactions to understand a situation. In this paper, we call these physical entities as objects and observations made over time as a scene. Humans are more accurate at the scene or object analysis when additional sensory information such as pressure, video, audio, and temperature is also provided in complement [18], [19]. Audio information is linked to the events taking place in the scenario, and thus it can offer vital information for comprehending the scene.

Recently, a drastic shift from unimodal to multimodal frameworks has been observed. Joint audio-visual frameworks have proven to be more effective than video-only frameworks in a range of applications, including activity classification [20], [21], identity verification [22], [23], and video categorization [24], [25]. Additionally, deep learning-based hybrid fusion strategy has shown advantages over conventional fusion methods [26].

A detailed description of literature related to unsupervised anomaly detection and multimodal anomaly detection is presented in Section II.

Inspired by the efficacy of hybrid fusion-based multimodal frameworks, we propose a joint audio-visual adaptive framework for anomaly detection. It is comprised of two modules, namely ‘Representation’ and ‘Learning’. In the ‘Representation’ module, we employ a teacher-student network that facilitates a deep learning-based hybrid fusion of audio and visual modalities. The network hallucinates a video motion descriptor from audio and a representative video frame. The purpose of the network is to connect audio and visual modalities so that correlated audio and video cues can be combined. The trained network produces lightweight and semantically rich audio-visual descriptors. Further, Principal Component Analysis (PCA) [27] is used so as to reduce the dimensionality of the obtained deep features. The ‘Learning’ module learns about the distribution of conceptually distinct events occurring in the target scenario. For learning the scene dynamics, we use an AGMM-based framework [10]. Each time a sample (normal or abnormal) is received, the mixture parameters (mean, standard deviation, and weight) are updated to reflect changes in the environment. The distinct clusters in the AGMM (also referred to as modes) represent distinct events, some corresponding to normal, whereas the rest to abnormal events present in the scene. If the frequency of occurrence of an abnormal event increases over time, its corresponding mode gets strengthened in terms of its weight, hence exhibiting a smooth transition from abnormal to normal. On the other hand, if a normal event infrequently occurs in the future, its associated mode’s weight decreases, indicating a transition from normal to abnormal. Thus, the automated dynamic update of the mixture parameter facilitates adaptive scene learning.

We demonstrate the efficacy of the proposed framework on two collected audio-visual datasets from different scenarios.

The main contribution of the paper is the exploitation of an image and its accompanying audio for a joint scene representation. Further, we propose an adaptive framework for anomaly detection. In order to validate its performance, the proposed work is compared with various frameworks, viz., a unimodal-based adaptive anomaly detection framework [10], a non-adaptive framework [4], and a multimodal anomaly detection framework [17].

The rest of the paper is organized as follows. The literature survey is presented in Section II. We discuss the proposed work in Section III. The experimental results & analysis are presented in Section IV. We write the conclusions in Section V.
Further used to learn a refined normal model using One-class SVM (Support Vector Machine). Liu et al. [8] investigated that reconstruction-based models which tend to reconstruct whole frames from scratch may lead to over-fitting and can even reconstruct abnormal videos as well.

Another category of approaches, namely prediction-based networks, have also been explored for the detection of anomalies in videos. These networks aim at predicting the video frame immediately next to the current input video frame. Similar to reconstruction loss, the prediction loss is computed using the predicted future frame and ground truth future frame. The prediction loss is higher for anomalies. Future frame prediction [7], [30] or prediction along with the reconstruction-based frameworks [31], [32] proved better at minimizing the narrow regularity score gap between the two classes.

These works claim to consider the global temporal context. However, the global context is limited to the dataset available for training, i.e., up to 20-30 minutes of training data. Moreover, one-time training of the model limits its efficacy for datasets consisting of concept drift and hence constraining its applicability in real-world surveillance [10].

Contrary to deep learning-based approaches, an AGMM based framework incorporates global temporal context more effectively with its adaptive scene learning strategy. Cristani et al. [14] detected events in audio data by building feature-wise univariate AGMM. Moncrieff et al. [15] experimentally determined that multivariate AGMM performs better than univariate AGMM for audio surveillance. Kumari and Saini [10] used multivariate AGMM for anomaly detection in video data. A few semantic (object counts) and handcrafted (pixel-wise motion) features were used to represent the scene. These features are computed for each event in the streaming video. The average number of humans, vehicles, and animals in the event are taken as semantic features. In contrast to semantic features, low-level features are the fundamental features that can be extracted directly from an image without any object description [33]. For low-level features, authors have computed histograms on pixel-wise motion information.

However, for real-life cases, the video scenes are complex in nature, i.e., there can be occluded objects, varying background, camera motion, multiple events [34], etc. For example, a usual scenario at a traffic junction may encounter occluded people due to vehicles, background changes due to varying sunlight across the day, etc. Handcrafted features do not represent such a complex scene adequately. In contrast, deep learning has shown success in representing complex natural scenes. Additionally, multimodal information has not been utilized in an AGMM framework for anomaly modeling. Therefore, we intend to exploit the benefits of deep multimodal representation in the multivariate AGMM framework for anomaly modeling.

**B. MULTIMODAL ANOMALY DETECTION**

There have been very few works that employ multimodal information for anomaly detection. Kuklyte et al. [35] built a normalcy model from low-level features that were extracted from audio and visual modalities using a q-mode (q = 100) clustering technique. The normal data was clustered using the agglomerative clustering algorithm [36] to create q clusters. Outliers to these clusters were identified as anomalies. Park et al. [37] employed several hidden Markov models (HMMs), one for each defined normal behavior of a PR2 robot using haptic, visual, auditory, and kinetic data. The HMM model, which is probabilistic in nature, considers a Markov process with hidden or unobserved states [38]. It stores actions in the form of hidden states with their occurrence probabilities and probability of transition between actions. An action was treated as abnormal if its log-likelihood estimation was observed to be below a threshold. Further, Park et al. [39] used LSTM based autoencoder, which is an autoencoder implemented using the Encoder-Decoder LSTM architecture, to reconstruct the multimodal signal from a PR2 robot. An action with high RL was treated as an anomalous sample. Kitzler et al. [40] detected anomalies in atomic actions performed by a human. Two separate HMMs were built to model the actions in audio and video modalities. Class posterior probabilities by the two models were then used to estimate the incongruence between the models. Higher the incongruence; more was the disagreement between the modalities and hence an anomalous action. Rehman et al. [17] used late fusion-based audio-visual anomaly detection for generic monitoring. For video modality, optical flow combined with Particle Swarm Optimization (PSO) and Social Force Model (SFM) was used to detect abnormal video frames. The PSO technique iteratively regulated the swarm of data in order to give an optimal solution to the problem. PSO combined with the optical flow was then used to find the flow of moving objects/individuals in the crowd. Further, SFM is used to calculate interaction forces among individuals in the crowd so as to characterize the crowd’s behavior. For audio modality, the acoustic features-based SVM classifier is used to flag anomalies. Further, a late fusion is applied to give the final decision. The authors evaluated their approach on a self-curated dataset that contains one type of anomaly example, which is gun-shot.

The above-discussed approaches are developed and assessed mainly on datasets constituting fixed classes of simple activities. Additionally, these works follow static learning, and hence are not useful for anomaly detection under concept drift. Moreover, the early fusion or late fusion strategies have been considered in these works. A simplified assumption of conditional independence between the modalities is considered in early fusion-based models [41], which might not hold true in practice. Late fusion, on the other hand, might ignore the intra-class feature level relationship. As a result, researchers have favored an alternative strategy, namely intermediate or hybrid fusion.

The indisputable success of deep learning in modeling complex problems has attracted researchers to develop deep learning-based hybrid fusion frameworks for a variety of tasks such as activity detection [21], face recognition [23],...
III. PROPOSED WORK

In this section, first, we formulate the problem; then, we discuss the proposed framework. Figure 1 gives a high-level overview of the proposed framework. The ‘Representation’ module maps the target scene into an audio-visual descriptor, and the ‘Learning’ module employs an unsupervised online strategy for learning the target scene.

A. PROBLEM FORMULATION

Wrong or miss detection arising from a single source of information can be minimized by exploiting information from multiple sources. Hence, we propose a framework that utilizes audio and a characteristic image from audio-visual events to model the anomaly detection task effectively.

Given an online source of a continuous audio-visual stream, the goal is to assign an anomaly score in the range of [0,1] to the buffered data in each ΔT second without the need to store any past data. A larger score indicates a more anomalous situation. The ΔT seconds data at the 1th time-stamp is treated as an audio-visual event $E_i^{AV}$. The visual modality (1st RGB image) of $E_i^{AV}$ is represented as $E_i^I$, and the mel-spectrogram corresponding to the audio modality is represented as $E_i^A$. A joint representation $(\vec{Z}_i^{IA})$ is achieved by feeding $E_i^I$ and $E_i^A$ to the fusion network.

Only a single image and the corresponding audio information are utilized to map the scene into a semantically rich feature [21]. Such light representation saves the costly video processing, which is often desired in real-time surveillance applications. At the same time, the image captures most of the appearance information within $E_i^{AV}$, while the audio provides complimentary temporal information. The high-dimensional joint feature $(\vec{Z}_i^{IA})$ is further mapped to a lower-dimensional feature $\vec{p}_i^{IA}$ using a PCA model. We discuss this deep multimodal descriptor in detail in Section III-B.

B. DEEP MULTIMODAL SCENE REPRESENTATION

The ‘Representation’ module in the proposed framework consists of the IMGAUD2VID fusion network [21] followed by a PCA model. In this module, the fusion network follows teacher-student learning architecture that facilitates knowledge distillation [45]. The process of knowledge distillation helps in training a smaller model, known as the student model, by extracting knowledge from the large, cumbersome model, referred to as the teacher model.

The teacher model can be a single very large model or an ensemble of separately trained models trained with a very strong regularizer [46]. Once the teacher model is trained, its knowledge is then distilled into the student model for further processing. Typically, multiple networks, each for different modalities to be fused, are aggregated in the student model to leverage deep fusion under knowledge distillation. Since the student model used in the proposed work consists of more than one modality, hence, a feature fusion strategy is exploited in the student model [47] so as to facilitate efficient cross-modal learning. The fusion process is carried out by concatenating the feature vectors obtained from the audio and image modalities used in the network, followed by adding two fully connected layers in the student model, as explained later in this section.

The knowledge distillation process used in this framework helps in minimizing the loss, i.e., Kullback-Leibler divergence (KL-divergence), between the teacher and student models’ probabilistic outputs in the distilled model. The KL-divergence loss constraints the student model’s outputs in order to match the soft targets of the teacher model.

The proposed teacher-student architecture performs multimodal distillation from a video network to an image-audio network. The teacher model is represented by ‘Video-Net’, whereas the student model is represented by the combination of ‘Image-Net’ and ‘Audio-Net’, respectively. The teacher model is a video-based pre-trained network that distills...
The ‘Representation’ module in the proposed framework is followed by the ‘Learning’ module. The input to the ‘Learning’ module is the output of the ‘Representation’ module, i.e., $\tilde{Z}_t^V$. It consists of an AGMM that learns the scene dynamics in an adaptive manner. AGMM was originally used in [51] for foreground-background detection in a continuous video stream. The authors utilized separate univariate AGMM for modeling each pixel in the frame. The mixture can have a maximum of $K$ Gaussian at any time, also referred to as $K$ modes in the mixture. Some of the modes in the mixture correspond to foreground and whereas the rest for background. The decision for foreground-background is taken based on a predefined threshold. Similar to the foreground-background task, AGMM is also used for event detection [15], [18] as well as anomaly detection [10], [12], [13]. For multivariate cases, generally, the features are assumed to be independent, and hence a diagonal covariance matrix is considered for simplicity [15].

AGMM is well suited for multivariate scene modelling in an adaptive manner. The modes in the AGMM represent the distribution of the events present in the target scene based on the data observed so far. The variance and mean vectors of event clusters are stored in modes of the mixture. These values are dynamically updated on the arrival of new samples (events) in order to learn the current trend in the data. Additionally, the modes also contain weight information. Weights account for the frequency of occurrence of unique events. They are also used to determine abnormality associated with the mode, i.e., higher the weight, lower is the abnormality, and vice-versa.

When a new sample $P_t^A$ arrives, it is matched to the modes in the mixture using Mahalanobis distance [52]. If the distance with the best match (smallest distance) is found below a threshold (θ Mahal), then it is considered as a hit with that mode. Let the hit mode have index $h$.

The weights of all the modes in the mixture are updated according to the following equation:

$$\hat{w}_{i,t+1} = (1 - \alpha)w_{i,t} + \alpha (\hat{M}_{i,t})$$

(2)

here $w_{i,t}$ and $w_{i,t+1}$, denote the weights of $i^{th}$ mode at time $t$ and $t + 1$, respectively. $\hat{M}_{i,t} = 1$ if $i^{th}$ mode is matched; otherwise 0, $\alpha$ is the model learning rate and is kept between 0 to 1.

Also, the mean and variance vector of only the hit mode are updated as follows:

$$\hat{\mu}_{h,t+1} = (1 - \beta_t)\hat{\mu}_{h,t} + \beta_t P_t^A$$

$$\hat{\sigma}_{h,t+1}^2 = (1 - \beta_t)\hat{\sigma}_{h,t}^2 + \beta_t (P_t^A - \hat{\mu}_{h,t+1})^T (P_t^A - \hat{\mu}_{h,t+1})$$

(3)

(4)

here, $\hat{\mu}_{h,t}$ and $\hat{\sigma}_{h,t}^2$ represent the mean and variance vector of $h^{th}$ mode at time $t$; and $\hat{\mu}_{h,t+1}$ and $\hat{\sigma}_{h,t+1}^2$ at time $t + 1$: $\beta_t$ is the distribution update factor inversely proportional to the distance between $\hat{\mu}_{h,t}$ and $P_t^A$. The coefficient $\beta_t$ is computed as follows:

$$\beta_t = \left(1 - \frac{\text{dist}(\hat{\mu}_{h,t}, P_t^A)}{\theta_{\text{Mahal}}}\right) \times \alpha$$

(5)

here, $\text{dist}(\hat{\mu}_{h,t}, P_t^A)$ is the Mahalanobis distance between $\hat{\mu}_{h,t}$ and $P_t^A$.

In case of a miss, i.e., in case no match is found in the mixture, then a new mode for the new sample is created. If the number of modes has reached the upper limit ($K$), then the weakest mode is deleted and a new mode corresponding to the new sample is added to the mixture. Let the newly created mode in case of a miss have an index $m$. It is initialized with
a small weight ($W_0$), a high variance ($z$), and the mean vector as the current sample, i.e., $\vec{P}_{IA}$.

In order to avoid the overflow of weights in online learning, weight normalization is performed after processing a sample. Hence, following the weight update, the weights are normalized such that their sum is 1.

An anomaly score is assigned to $\vec{P}_{IA}$ on the basis of the relative weight of the matched mode ($h$th mode in case of a hit) or the newly created mode ($m$th mode in case of a miss). For simplicity, both modes are denoted with index $hm$.

Thus, the anomaly score for the event $\vec{P}_{IA}$ can be assigned as $(1 - w_{hm,t+1})$.

It is to be noted that during the test/application time, the ‘Representation’ module is static while the ‘Learning’ module updates its parameters in real-time.

**IV. EXPERIMENTS AND RESULT ANALYSIS**

**A. DATASETS AND ANNOTATION PROCEDURE**

There is a lack of publicly available audio-visual datasets for anomaly detection. Hence, we collect two long-duration audio-visual datasets, namely ‘PhD-Lab’ and ‘Reception’, for evaluation of the proposed work. The datasets have been recorded using a Sony 64MP Quad Camera of ‘Realme 7’ smartphone. The audio data is sampled at 22.5 kHz, and the video at 20 frames per second. Fixed surveillance location (single scene) and untrimmed footage are the two primary conditions for any dataset to be used for the evaluation of adaptive anomaly detection frameworks [10]. Therefore, each dataset is captured by placing the smartphone in a fixed spot. Rather than trimming the data for a specific set of events, the entire audio-visual clip is considered as one complete dataset. Rare events correspond to anomalies in our datasets. Some less probable events that are considered anomalous in our datasets might not be a direct security threat. However, such events can be of interest to monitoring personnel; thus, considered anomalous. Also, it might be worth noting that the probabilities of events are not static due to concept drift. Hence, the ground truth of an event might change over time.

The ‘PhD-Lab’ dataset is recorded in a research lab. The dataset is recorded continuously for a duration of 82 minutes. This dataset is accompanied by various types of audio-visual events performed intentionally. Some of the abnormal activities include changes from sitting to walking or dancing, leaving the scene, more people entering into the scene, music, laughter, etc. A few activities were carried out on purpose.

The second dataset, i.e., ‘Reception’, is recorded in the reception area of an educational institute. The dataset is recorded continuously for a duration of 189 minutes. This dataset is primarily composed of natural anomalies. Additionally, these long-duration data capture the outdoor light change, i.e., the gradual transition from evening to night. Several natural anomalies were observed, including shouting, running, intense door closings, a student switching lights on/off, and crowds. A few activities were carried out on purpose.
such as playing music and engaging in some sort of fight using umbrellas.

Table 1 gives detail of example anomalies and the fraction of anomalies for each of the datasets.

The datasets are annotated manually by three people separately, using a common set of instructions. A time segment is treated as anomalous if it catches one’s attention. The annotators were given the audio-visual datasets to listen and watch using the ‘vlc’ player on separate laptops. Starting and ending times of the time segments, which seemed interesting to them, and hence anomalous, were noted. No predefined category of events was provided to them. Rather, the annotators were allowed to use their own judgment to define a situation and determine whether it is an anomaly or not. Also, it was instructed that when an anomalous event is observed for a sufficient duration, then all its future occurrences will be treated as normal by them.

The final annotated anomalous segments are then mapped to per second labels. In this mapping, all the seconds falling in the time-segments noted by the annotators are labeled as 0, indicating them as normal. Thus, for each annotator, a vector representing per second labels is achieved. Inter-annotator agreement measure using Cohen’s kappa score between the annotator’s pair on different datasets. In order to minimize the biases produced from an individual’s opinion, majority voting is opted. Thus, if at least two people refer to an event as anomalous, then it is considered as an anomaly, else normal. In this manner, we achieve the final per second ground truth (\(\hat{s}\)).

### Table 1. The table lists the example anomalies and the percentage of the anomaly in the audio-visual datasets.

| Dataset       | Duration (minutes) | % Anomaly | Example anomalies                                                                 |
|---------------|--------------------|-----------|-----------------------------------------------------------------------------------|
| PhD-Lab       | 82                 | 10.43%    | video play on pc, gathering, music, laugh, changing color of light, etc.          |
| Reception     | 189                | 3.81%     | telephone ringing, siren, light on-off, music, gathering, shouting, etc.          |

### Table 2. Cohen’s kappa score between the annotator’s pair on different datasets.

| Annotator’s Pair | PhD-Lab | Reception |
|------------------|---------|-----------|
| 1st - 2nd        | 0.5320  | 0.6287    |
| 2nd - 3rd        | 0.5187  | 0.3067    |
| 3rd - 4th        | 0.7371  | 0.2592    |

B. IMPLEMENTATION DETAILS

In order to train the fusion network, the Kinetics dataset [48] is used. The dataset consists of 400 classes of activities and about 0.2 million samples, each of approximately 10 seconds. Each sample is further divided into 1-second segments, thus resulting in 2 million audio-visual samples for training the network. I3D [55] model is used as the teacher model in the fusion network. It is a 3D convolution-based neural network pre-trained on the Kinetics dataset. It takes an RGB video of 1 second and produces a video feature of 2048 dimensions. The student model consists of two networks, viz., ‘Image Net’ and ‘Audio Net’ for processing images and audios, respectively. Two VGG-16 [56] models in the student model are used, one for the ‘Image Net’ and the other for the ‘Audio Net’. Both the models take an input of \(112 \times 112\) and \(187 \times 42\), respectively. The sampling rate of audio for all datasets is kept at 22 kHz. For a 1-second audio data sampled at 22 kHz, a mel-spectrogram is computed using 512 fft points, 128 hop length, and 40 mel filterbanks. The fusion network is trained using Adam optimizer with an initial learning rate of 1e-03 for 60 epochs and batch size of 64. The learning rate is decayed after 30 epochs by 10. The deep features extracted from the fusion network, i.e., \(\tilde{Z}_d\) are fed further to train the PCA model for each dataset. The number of components in the PCA model is chosen to be 10 for the ‘PhD-Lab’ dataset and 20 for the ‘Reception’ dataset. Thus, we get a \(1 \times 10\) and \(1 \times 20\) dimensional feature representation for an audio-visual event in ‘PhD-Lab’ and ‘Reception’ datasets, respectively. These choices for the number of components in the PCA model are discussed in the next section.

For each dataset, audio-visual events of 1 second without overlap are extracted. Since shuffling of the dataset can destroy the temporal context and the notion of concept drift, hence it is kept unshuffled. Each dataset is split into a 50:50 ratio for training and testing purposes. The split ratio is chosen such that the test data has sufficient anomaly samples since, with standard splits like 70:30 or 80:20, test data may get a few or no samples of anomalies due to class imbalance. The training data is used to train the PCA model and learn the optimal set of hyper-parameters of AGMM.

The hyper-parameter \(W_0\) is kept fixed for all the experiments at 0.005. Other hyper-parameters, i.e., \(\alpha\) and \(K\) are varied in sets \{0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\} and \{2, 4, 6, 8, 10, 12, 14, 16, 18\}, respectively. A new sample is said to match a mode only if it lies in \(\pm 1\sigma\), i.e., within 68% population, \(\sigma\) being the standard deviation. However, in multidimensional cases, the sigma rule changes, i.e., a higher times sigma is required to accommodate the same 68% population [57]. Following the sigma rules from [57], for feature sizes of 10 for ‘PhD-Lab’, and 20 for ‘Reception’, \(\theta_{Mahal}\) is set to 3.3 and 4.8, respectively.

Hyper-parameter tuning is carried out using Grid-search. The hyper-parameters are varied in the specified set, and model performance on the training data is observed. The set of hyper-parameters for which the model achieves the best performance is selected and further used to report the performance on test data.

The ‘Learning’ module is not started from scratch for testing purposes. Instead, the same model that has already seen the entire training data is continued. For performance, ROC (Receiver Operating Characteristic curve) and AUC performance is selected and further used to report the performance on test data.
C. SELECTION OF SUITABLE FEATURE SIZE & HYPER-PARAMETERS OF AGMM

Selecting a suitable feature size for dimensionality reduction is important. Extremely small feature size after dimensionality reduction might discard important information, whereas large feature size results in reduced performance of proximity-based outlier detection. Hence, suitable values of feature size for PCA and \((K, \alpha)\) pairs for AGMM are selected experimentally. For the same, the effect of feature size \((s_{\text{IA}})\) on model performance is studied. Different PCA models are trained for different feature sizes in \(\text{PCA}_{\text{size}} = \{4, 6, 8, 10, 20, \text{ and } 30\}\). \(\theta_{\text{Mahal}}\) for each of the feature dimensionality is chosen according to the sigma rule in [57]. It is set to 2.2 for a feature size of 4, 2.6 for a feature size of 6, 3 for a feature size of 8, 3.3 for a feature size of 10, 4.8 for a feature size of 20, and 5.8 for a feature size of 30.

For each feature size considered, AUCs on different \((K, \alpha)\) pairs for the training data are observed. For each feature size, the final values of \((K, \alpha)\) are selected, for which training AUC is maximum. The finalized \((K, \alpha)\) pair, along with the training AUCs for each feature size, are reported in Table 3. From the table, it can be seen that for the ‘PhD-Lab’ dataset, the model attains the best performance for feature size 10 and \((K, \alpha)\) pair of \((10, 0.05)\). In the case of ‘Reception’ dataset, the best performance is attained for feature size 20 and \((K, \alpha)\) pair of \((6, 0.05)\). These finalized values are chosen for all further experiments involving the proposed framework, as mentioned in the previous section.

D. MODEL EVALUATION

In order to validate the proposed framework, it is compared with its unimodal variants, its fusion variants, a unimodal feature-based AGMM framework [10], a deep learning-based non-adaptive framework [4], and a handcrafted feature-based multimodal anomaly detection framework [17].

1) COMPARISON OF THE PROPOSED FRAMEWORK WITH ITS UNIMODAL VARIANTS

In this section, the proposed framework is compared with its unimodal variants, i.e., image and audio-based models. For the same, an image \((\vec{Z}_I)\), audio \((\vec{Z}_A)\), and a joint \((\vec{Z}_{\text{IA}})\) feature from the fusion network are considered, and the PCA model is trained separately for each variant. Further, scene learning is performed using the ‘Learning’ module.

For each dataset, modality-wise AUC % on the test data is reported in Table 4. From the table, it can be observed that the AUC measure improves significantly after performing the hybrid fusion for both datasets. Also, it can be seen that the image modality performs better than the audio modality. This is because the number of visual anomalies was more compared to audio-based anomalies in our datasets. Overall, it can be observed that the model performs better when the joint representation is used instead of unimodal representations. We see an increment of 0.82% and 34.08% in AUC values when compared with image and video modalities for the ‘PhD-Lab’ dataset. For the ‘Reception’ dataset, there is an increment of 12.37% and 41.93% in AUC values when compared with image and audio modalities, respectively. These improvements signify the importance of multimodal information over unimodal information-based modeling for the task of anomaly detection.
2) COMPARISON OF FUSION VARIANTS OF THE PROPOSED FRAMEWORK

In this section, early, late, and hybrid fusion variants of our proposed framework are evaluated and compared. In the case of early fusion, the image and audio features extracted from ‘Image-Net’ and ‘Audio-Net’ of the untrained fusion network are concatenated. Let the concatenation be represented as $\mathbf{C}_{IA}$. Further, PCA models are trained using the training volume of these concatenated features with sizes 10 and 20 for ‘PhD-Lab’ and ‘Reception’ datasets, respectively. For each dataset, its trained PCA model is used to transform the whole data into a lower dimension. AGMM is applied to these descriptors, and performance on test data is reported. For the late fusion variant, we take anomaly scores from unimodal variants achieved in the previous section and compute the average anomaly scores. AUC is computed on these averaged anomaly scores and binary ground truth ($\mathbf{s}$). The hybrid fusion variant refers to the proposed framework.

AUC for each dataset and fusion variants are reported in Table 5. It can be observed that hybrid fusion is outperforming the early and late fusion variants of the proposed framework. For the ‘PhD-Lab’ dataset, late fusion performs the second best. The hybrid fusion outperforms early and late fusion by 12.26% and 0.14% in AUC values. In the case of the ‘Reception’ dataset, early fusion performs the second best. The hybrid fusion shows an improvement from early and late fusion by 2.34% and 12.05% in terms of AUC. The associated ROC curves are plotted in Figure 4. Results with both datasets in Table 5 and Figure 4 demonstrate the efficacy of hybrid fusion over early and late fusion.

3) COMPARISON WITH OTHER APPROACHES

As mentioned in Section II, adaptive scene learning-based anomaly detection in multimedia data has not yet been exploited except in work by Kumari and Saini [10]. Therefore we compare the proposed framework with [10]. Further, in order to show the efficacy of adaptive learning over static learning (one-time training), a comparison of proposed work with a recent deep learning-based anomaly detection framework by Wang et al. [4] has been performed. Additionally, we compare with a state-of-the-art audio-visual information-based anomaly detection framework proposed for generic surveillance by Rehman et al. [17].

Kumari and Saini [10] used object counts and spatial-temporal motion-based histogram features from video data for anomaly detection. Hence, for a fair comparison, the ‘Representation’ module in the proposed work is replaced with the handcrafted video feature as in [10], and the ‘Learning’ module is kept the same. For comparison, the event size is kept as 1 second. $\theta_{Mahal}$ is kept as 5 as in [10]. For each dataset, suitable $(K, \alpha)$ pair is selected using the training data. The hyper-parameters resulting in the best performance on the training data are selected as mentioned in Section IV-B. Selected $(K, \alpha)$ pairs for the approach Kumari and Saini [10] are (8, 0.05) and (6, 0.001) for ‘PhD-Lab’ and ‘Reception’ datasets, respectively.

The approach by Wang et al. [4] was originally employed on video data. For comparison, the same joint features as applied in the proposed approach are used to train their two-stage architecture. This non-adaptive approach uses an autoencoder-based normality estimation in the first stage and...
TABLE 6. The table compares the performance of the proposed framework with a non-adaptive deep learning-based framework (Wang et al. [4]) on our joint feature, the low-level+semantic video feature based AGMM (Kumari and Saini [10]), and a handcrafted feature based multimodal approach (Rehman et al. [17]).

| Datasets → Approach ↓ | Modality | Learning | Feature | PhD-Lab test AUC | Reception test AUC |
|-----------------------|----------|----------|---------|------------------|-------------------|
| Wang et al. [4]       | Image-audio | unsupervised & non-adaptive | deep | 65.73 | 73.59 |
|                       |          |          |         | 53.59 | 76.23 |
|                       |          |          |         | 65.70 | 74.21 |
|                       |          |          |         | 60.94 | 64.18 |
|                       |          |          |         | 66.96 | 72.82 |
|                       |          |          |         | 64.98 | 66.75 |
| Kumari and Saini [10] | Video    | unsupervised & adaptive | handcrafted | 65.16 | 65.20 |
| Rehman et al. [17]    | Video-audio | audio: supervised & non-adaptive video: unsupervised & non-adaptive | handcrafted | 67.32 | 73.88 |
| Proposed              | Image-audio | unsupervised & adaptive | deep | 76.44 | 81.82 |

FIGURE 5. ROC based comparison of the proposed framework with the state-of-the-art works.

In Table 6, AUCs of different approaches along with other information, viz., type of modality, type of learning employed, and type of features used, is reported. It can be observed from the table that the proposed framework achieves a gain of 17.31% and 25.49% in AUC values as compared to the handcrafted video feature-based AGMM framework [10] for the ‘PhD-Lab’ and ‘Reception’ datasets, respectively. It signifies the importance of deep joint representation over a handcrafted unimodal feature-based anomaly detection.

The multimodal anomaly detection framework by Rehman et al. [17] uses optical flow-based PSO and SFM for anomaly score estimation on video frames. We replicate their algorithm for per frame anomaly score computation with the same parameter settings. For each 1-second of event, the anomaly score of constituent frames is averaged, and a single score is obtained. For audio modality, first, the acoustic features as used in [17] are extracted for each 1-second of data. It is followed by training an SVM classifier using the training data (50% of total data). Since the range of scores can vary for audio and video modalities, they are normalized in order to bring them in the same range. Finally, late fusion (decision fusion) is applied, and the per-second average anomaly score is computed using the individual anomaly scores of audio and video modalities. AUC and ROC curves are computed on these scores and binary ground truth (s) corresponding to the test data.

In Table 6, AUCs of different approaches along with other information, viz., type of modality, type of learning employed, and type of features used, is reported. It can be observed from the table that the proposed framework achieves a gain of 17.31% and 25.49% in AUC values as compared to the handcrafted video feature-based AGMM framework [10] for the ‘PhD-Lab’ and ‘Reception’ datasets, respectively. It signifies the importance of deep joint representation over a handcrafted unimodal feature-based anomaly detection. The approach by Wang et al. [4] shows degradation of 12.40% and 6.83% in AUC values compared to the proposed approach for the ‘PhD-Lab’ and the ‘Reception’ datasets, respectively. ‘PhD-Lab’ has a higher concept drift (added by acted events) than the ‘Reception’ dataset. Therefore we observe a larger
V. CONCLUSION

In this paper, the efficacy of using joint information over single modalities for anomaly detection is shown. An end-to-end joint feature-based adaptive anomaly detection framework is presented. For hybrid fusion, a deep learning-based teacher-student network is used, followed by a PCA model to reduce the dimensionality of the features. The proposed joint feature-based adaptive framework shows an average increment of 21.40%, 10.74%, and 12.15% in AUC values when compared to a unimodal information-based AGMM framework, a deep learning-based non-adaptive framework, and a handcrafted feature-based multimodal anomaly detection framework, respectively. Hence, the proposed framework is well suited for anomaly detection in multimedia data having concept drift.

Like all other works, we also assume events to be of fixed duration. However, the assumption does not hold true in real-life. In our future work, we plan to incorporate adaptive event size for modeling.

REFERENCES

[1] S. Szymanowicz, J. Charles, and R. Cipolla, “X-MAN: Explaining multiple sources of anomalies in video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshop, Jun. 2021, pp. 3224–3232.

[2] K. K. Santhosh, D. P. Dogra, and P. P. Roy, “Anomaly detection in road traffic using visual surveillance: A survey,” ACM Comput. Surv., vol. 53, no. 6, pp. 1–26, Nov. 2021.

[3] R. Nayak, U. C. Pati, and S. K. Das, “A comprehensive review on deep learning-based methods for video anomaly detection,” Image. Vis. Comput., vol. 106, Feb. 2021, Art. no. 104078.

[4] S. Wang, Y. Zeng, Q. Liu, C. Zhu, E. Zhu, and J. Yin, “Detecting abnormality without knowing normality: A two-stage approach for unsupervised video abnormal event detection,” in Proc. 25th ACM Int. Conf. Multimedia, Seoul, South Korea, Oct. 2018, pp. 636–644.

[5] J. Hu, E. Zhu, S. Wang, X. Liu, X. Guo, and J. Yin, “An efficient and robust unsupervised anomaly detection method using ensemble random projection in surveillance videos,” Sensors, vol. 19, no. 19, p. 4145, Sep. 2019.

[6] C. Park, M. Cho, M. Lee, and S. Lee, “FastAno: Fast anomaly detection via spatio-temporal patch transformation,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Waikoloa, HI, USA, Jan. 2022, pp. 2249–2259.

[7] X. Wang, Z. Che, B. Jiang, N. Xiao, K. Yang, J. Tang, J. Ye, J. Wang, and Q. Qi, “Robust unsupervised video anomaly detection by multi-path frame prediction,” 2020, arXiv:2011.02763.

[8] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection—A new baseline,” in Proc. CVPR, Salt Lake City, UT, USA, Jun. 2018, pp. 6536–6545.

[9] G. Widmer and M. Kubat, “Learning in the presence of concept drift and hidden contexts,” Mach. Learn., vol. 23, no. 1, pp. 69–101, 1996.

[10] P. Kumari and M. Saini, “Multivariate adaptive Gaussian mixture for scene level anomaly modeling,” in Proc. IEEE Conf. Multimedia Big Data (BigMM), New Delhi, India, Sep. 2020, pp. 54–62.

[11] 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.

[12] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, “Deep autoencoding Gaussian mixture model for unsupervised anomaly detection in Proc. Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, Apr./May 2018, pp. 1–19.

[13] Z. Chen and T. Ellis, “A self-adaptive Gaussian mixture model,” Comput. Vis. Image Understanding, vol. 122, pp. 35–46, May 2014.

[14] M. Cristani, M. Bicego, and V. Murino, “On-line adaptive background modelling for audio surveillance,” in Proc. 17th Int. Conf. Pattern Recognit., Cambridge, U.K., 2004, pp. 399–402.

[15] S. Moncrieff, S. Venkatesh, and G. West, “Online audio background determination for complex audio environments,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 3, no. 2, p. 8, May 2007.

[16] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, “Deep audio-visual learning: A survey,” Int. J. Autom. Comput., vol. 18, no. 3, pp. 351–376, Apr. 2021.

[17] A.-U. Rehman, H. S. Ullah, H. Farooq, M. S. Khan, T. Mahmood, and H. O. A. Khan, “Multi-modal anomaly detection by using audio and visual cues,” Proc. 34th Conf. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, Dec. 2020, pp. 9758–9770.

[18] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, “Listen to look: Action recognition by previewing audio,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2020, pp. 10457–10467.

[19] S. Soleymani, A. Babaei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, “Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification,” in Proc. 24th Int. Conf. Pattern Recognit., Beijing, China, Aug. 2018, pp. 3469–3476.

[20] A. R. Lejbbole, B. Krogh, K. Nasrollahi, and T. B. Moeslund, “Attention in multimodal neural networks for person re-identification,” in Proc. CVPR Workshops, Salt Lake City, UT, USA, Jun. 2018, pp. 179–187.

[21] H. Tian, Y. Tao, S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, “Multimodal deep representation learning for video classification,” World Wide Web, vol. 22, no. 3, pp. 1325–1341, Mar. 2019.

[22] S. Kang, J. Kim, H. Choi, S. Kim, and C. D. Yoo, “Pivot correlational neural network for multimodal video categorization,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 2018, pp. 386–401.

[23] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: A survey,” Multimedia Syst., vol. 16, no. 6, pp. 3605–3607, 2010.

[24] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” J. Roy. Statist. Soc. B, vol. 61, no. 3, pp. 611–622, 1999.

[25] A. Del Giorno, J. A. Bagnell, and M. Hebert, “A discriminative framework for anomaly detection in large videos,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Venice, Italy, Oct. 2017, pp. 2895–2903.
X. Liu, L. Jiao, L. Li, X. Tang, and Y. Guo, “Deep multi-level fusion detection via deep predictive coding network,” in Proc. 27th ACM Int. Conf. Multimedia, Nice, France, Oct. 2019, pp. 1805–1813.

Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, “Spatio-temporal AutoEncoder for video anomaly detection,” in Proc. 25th ACM Int. Conf. Multimedia, Mountain View, CA, USA, Oct. 2017, pp. 1933–1941.

Y. Lai, R. Liu, and Y. Han, “Video anomaly detection via predictive autoencoder with gradient-based attention,” in Proc. IEEE Int. Conf. Multimedia Expo. (ICME), London, U.K., Jul. 2020, pp. 1–6.

C. Zhang and T. Chen, From Low Level Features to High Level Semantics. Boca Raton, FL, USA: CRC Press, 2003.

E. Ricci, G. Zen, N. Sebe, and S. Messelodi, “A prototype learning framework using emd: Application to complex scenes analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 3, pp. 513–526, Jun. 2012.

J. Kuklyte, P. Kelly, C. O. Conaire, N. E. O’connor, and L.-Q. Xu, “Anti-social behavior detection in audio-visual surveillance systems,” in Proc. Pattern Recog. Artif. Intell. Human Behaviour Anal. Workshop, Reggio Emilia, Italy, Dec. 2009, pp. 9–11.

W. H. E. Day and H. Edelsbrunner, “Efficient algorithms for agglomerative hierarchical clustering methods,” J. Classification, vol. 1, no. 1, pp. 7–24, 1984.

D. Park, Z. Erickson, T. Bhattacharjee, and C. C. Kemp, “Multimodal execution monitoring for anomaly detection during robot manipulation,” in Proc. IEEE Int. Conf. Robot. Automat. (ICRA), Stockholm, Sweden, May 2016, pp. 407–414.

L. Rahbini and B. Juang, “An introduction to hidden Markov models,” IEEE ASSP Mag., vol. 3, no. 1, pp. 4–16, Jan. 1986.

D. Park, Y. Hoshi, and C. C. Kemp, “A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder,” IEEE Robot. Automat. Lett., vol. 3, no. 3, pp. 1544–1551, Jul. 2018.

J. Kittler, I. Kalouskampis, C. Zor, Y. Xu, Y. Hicks, and W. Wang, “Intelligent signal processing mechanisms for nuanced anomaly detection in action audio-visual data streams,” in Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP), Calgary, AB, Canada, Apr. 2018, pp. 6563–6567.

D. Ramachandram and G. W. Taylor, “Deep multimodal learning: A survey on recent advances and trends,” IEEE Signal Process. Mag., vol. 34, no. 6, pp. 96–108, Nov. 2017.

X. Liu, L. Jiao, L. Li, X. Tang, and Y. Guo, “Deep multi-level fusion network for multi-source image pixel-wise classification,” Knowl. Based Syst., vol. 221, Jun. 2021, Art. no. 106921.

X. Liu, L. Jiao, L. Li, C. Cheng, F. Liu, S. Yang, and B. Hou, “Deep multiview union learning network for multisource image classification,” IEEE Trans. Cybern., early access, Nov. 5, 2020, doi: 10.1109/TCYB.2020.3029787.

X. Liu, L. Jiao, J. Zhao, J. Zhao, and D. Zhang, “Deep multiple instance learning-based spatial–spectral classification for PAN and MS imagery,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 1, pp. 461–473, Sep. 2017.

G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.

T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, “A comprehensive survey on model compression and acceleration,” Artif. Intell. Rev., vol. 53, no. 7, pp. 5113–5155, Feb. 2020.

M. Zhao, T. Li, M. A. Alsheikh, and Y. Tian, “Through-wall human pose estimation using radio signals,” in Proc. IEEE/ACVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 7356–7365.

W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” 2017, arXiv:1705.06950.

C. C. Aggarwal, “An introduction to outlier analysis,” in Proc. Outlier Anal., Cham, Switzerland, Dec. 2017, pp. 1–34.

D. Bollegala, “Dynamic feature scaling for online learning of binary classifiers,” Knowl. Based Syst., vol. 129, pp. 97–105, Aug. 2017.

C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Fort Collins, CO, USA, Dec. 1999, pp. 246–252.

R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, “The Mahalanobis distance,” Chemometrics Intell. Lab. Syst., vol. 50, no. 1, pp. 1–18, Jan. 2000.

J. Cohen, “A coefficient of agreement for nominal scales,” Educ. Psychol. Meas., vol. 20, pp. 37–46, Apr. 1960.

J. R. Landsis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, vol. 33, no. 1, pp. 159–174, Mar. 1977.

J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in Proc. IEEE Int. Conf. Comput. Vis., Honolulu, HI, USA, Jul. 2017, pp. 6299–6308.

K. Simonany and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

G. Gallego, C. Cuevas, R. Mohedano, and N. García, “On the Mahalanobis distance classification criterion for multidimensional normal distributions,” IEEE Trans. Signal Process., vol. 61, no. 17, pp. 4387–4396, Sep. 2013.

T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, Jun. 2006.

A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” Pattern Recognit., vol. 30, pp. 1145–1159, Jul. 1997.