electronics

MDPI

Article

Violence Recognition Based on Auditory-Visual Fusion of Autoencoder Mapping

Jiu Lou, Decheng Zuo *, Zhan Zhang and Hongwei Liu

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; loujiu@hit.edu.cn (J.L.); zhangzhan@hit.edu.cn (Z.Z.); liuhw@hit.edu.cn (H.L.)
* Correspondence: zuodc@hit.edu.cn

Abstract: In the process of violence recognition, accuracy is reduced due to problems related to time axis misalignment and the semantic deviation of multimedia visual auditory information. Therefore, this paper proposes a method for auditory-visual information fusion based on autoencoder mapping. First, a feature extraction model based on the CNN-LSTM framework is established, and multimedia segments are used as whole input to solve the problem of time axis misalignment of visual and auditory information. Then, a shared semantic subspace is constructed based on an autoencoder mapping model and is optimized by semantic correspondence, which solves the problem of audiovisual semantic deviation and realizes the fusion of visual and auditory information on segment level features. Finally, the whole network is used to identify violence. The experimental results show that the method can make good use of the complementarity between modes. Compared with single-mode information, the multimodal method can achieve better results.

Keywords: violence recognition; auditory-visual fusion; autoencoder mapping; shared semantic subspaces; CNN-LSTM

Citation: Lou, J.; Zuo, D.; Zhang, Z.; Liu, H. Violence Recognition Based on Auditory-Visual Fusion of Autoencoder Mapping. *Electronics* **2021**, *10*, 2654. https://doi.org/10.3390/electronics10212654

Academic Editors: Xiaofeng Liu, Harry Yang, Zhenhua Guo and Jane You

Received: 27 August 2021
Accepted: 26 October 2021
Published: 29 October 2021

1. Introduction

The wide application of high-definition multimedia data acquisition equipment has guaranteed public social security and greatly protected the safety of people and property. The exploration and research of violent behavior detection based on multimedia data streams are an important direction in intelligent signal processing [1].

The process of violence recognition mainly consists of feature extraction and recognition model design. Earlier work mainly focused on the design of violent features. Studies proved that appearance and motion features in visual information could effectively describe violent behaviors [2], such as RIMOC (rotation-Invariant feature modeling MOtion Coherence) [3] and GRADIENT direction histogram HOG [4], the latter of which described appearance. STIP (space–time interest points) [5], iDT (improved dense trajectories) [6], and GMOF (Gaussian model of optical flow) have also been described in relation to motion [7]. Along with continuous research, scholars have found that the auditory channel also plays an important role in the detection of violent scenes, such as those often accompanied by shooting, explosions, roaring, and screaming, which are typical sounds related to the breaking of objects; however, nonviolent scenario sounds tend to be relatively slow. The auditory feature is thus represented by the classic Mel-frequency cepstral and LPC (linear predictive coding) used to identify violence [8]. After feature extraction, the final recognition results are often obtained through classifiers, such as SVM [9], bag-of-words mode [10], etc.

With the development of deep network technology, the deep network model represented by a convolutional neural network (CNN) [11] can realize end-to-end task processing by integrating feature extraction and recognition models. A long- and short-term memory network (LSTM) [12] realizes the acquisition of a sequence context, and these models have made breakthroughs in the field of computer vision [13]. The unprecedented prosperity of deep learning has prompted scholars to try to use deep networks to identify violence. In

terms of visual features, Accattoli et al. used a 3D convolutional network to detect violent behaviors [14] and Tripathi et al. used a convolutional neural network to extract multi-level video features [15]. Additionally, Deniz et al. proposed fast motion detection based on an extreme acceleration mode [16]. Sharma et al. realized violent video scene detection based on a deep neural network [17]. In terms of auditory, Garcia-Gomez et al. realized the recognition of violent events based on auditory feature screening [18], while Chen L B et al. explored the classification of violent audio using a temporal tensor feature [19]. Moreover, Wang Y et al. examined audio time localization based on a continuous time model [20].

As can be seen from the above analysis, visual and auditory information in violent videos contain different features related to violence, and these features from different modes are characterized by information complementarity. In the process of violence recognition, these different modal features need to be fused together to improve the accuracy of violence recognition. Scholars have tried to fuse multimodal features at different granularities, which can be divided into two categories: a late fusion method [21], which integrates the classification results of classifiers based on different modal features at the decision level, and an early fusion method [22], which combines and splices different modal features. However, due to the particularity of violent behavior, there are common problems that are difficult to solve in the process of auditory-visual information fusion for the identification of violence. First, auditory-visual data are often misaligned on the time axis, for instance, an explosion might be heard first, followed by a crowd rioting, or vice versa. Second, the semantic expression bias of visual and auditory information, such as normal behavior, is shown in a video accompanied by an explosion, or, alternatively, violent behavior is shown but without any abnormal background sound. Both of these are problems that need to be solved in the process of multi-modal feature fusion.

At present, multimodal fusion for the recognition of violence mostly adopts a late fusion method at the decision level [23,24], primarily. This is mainly because the information fusion at the decision level is equivalent to the fusion of semantically similar features in the same feature space (i.e., decision scores), which has less risk and is relatively easy to achieve. However, the effect of a decision-level fusion method on the improvement of violent video recognition performance is limited, as only the scores after each mode decision can be used in decision-level fusion, and the semantic consistency of each mode of information is not taken into account. Compared with decision-level fusion, an advantage of the feature-level fusion method is that it can more intuitively fuse more modal information and better capture the relationship between various modes. Good feature-level fusion methods can significantly improve the performance of video classification. However, the difficulty of this method lies in the different semantic meanings of various modal features and the difficulty of establishing feature subspaces with uniform semantic representation.

However, semantic consistency is important in multimodal fusion, especially in visual and auditory information fusion. When the semantics of multimodal information are consistent, the information is complementary; otherwise, they may interfere with each other (such as the famous "McGurk effect" [25], which is a perceptual cognitive phenomenon manifested primarily in the interaction between auditory and visual perception in the process of speech perception). Sometimes, human hearing is so clearly affected by vision that it can lead to mishearing. When one sound does not match another, people mysteriously perceive a third sound, and merging them can even have the opposite effect. Therefore, in the case of semantic inconsistencies between multimodal forms of information, feature fusion between modes without any measures cannot achieve information complementarity between modes, and it may also lead to the degradation of an algorithm's performance [26].

In violent videos, semantic consistency can be understood as the featuring of violent audio and violent video; this consistency means that either both audio and visuals feature violent scene descriptions or neither of them do. Violence in multimedia data analyses can be found due to the particularity of violence, and semantic inconsistencies in audiovisual information are embodied by two notions. First, audiovisual data are not aligned on a timeline. Second, there are semantic expression deviations between visual and auditory

information. Both of these are problems that need to be solved in the process of multi-modal feature fusion.

Violence as captured by existing recognition algorithms combines audiovisual features that do not consider semantic consistency problems. As such, this paper proposes a recognition model of violence, which uses a CNN-LSTM architecture for fragment levels feature extraction and uses the autoencoder [27] model to represent the shared semantic subspace mapping for audiovisual information fusion. Through this approach, we seek to circumvent problems related to audiovisual information time axis misalignment. Then, the segment-level visual and auditory features are integrated into the same shared subspace using an autoencoder model, and semantically corresponding labels are introduced to optimize the autoencoder model to solve the problem of semantic consistency. Our experimental results show that this method can improve the performance of violent behavior recognition methods.

In Section 2 of this paper, a feature extraction method for visual and auditory channels based on the CNN-LSTM model is introduced. In Section 3, the construction method for a visual and auditory feature shared subspace and fusion detection model is introduced. In Section 4, the MediaEval VSD2015 dataset [28] is used to verify the validity of the proposed method.

2. Auditory and Visual Feature Extraction Method

Violent behavior shows a certain persistence in the time axis, so it is necessary to extract the identification features of violent behavior within a certain time span. In this paper, CNN-LSTM architecture is used to extract the auditory and visual features of violence on the temporal axis; in other words, the AlexNet [29] structure in the CNN network is used to extract the frame-level features, and the LSTM is used to integrate the frame-level features to obtain the segment-level features. This section details the specific feature extraction methods.

2.1. Auditory Feature Extraction Based on CNN-LSTM

Auditory information is a key element in the recognition of violence. There are two approaches to auditory feature extraction. One is the filter-based acoustic feature extraction method, the typical features of which include Mel-frequency cepstral features and LPC features. These features have achieved good effects in the field of speech recognition. However, this feature extraction method does not consider the differences between different tasks, which may lead to a lack of key information related to a given task and affect the results. The other approach is the end-to-end feature extraction method based on a depth network, which directly takes an audio signal as input and uses the depth network for feature extraction. When deep network is used for end-to-end feature learning, a large number of data sets with uniform distribution of the occurrence frequency of classified events are needed [30]. However, the current violent audio data sets are difficult to meet this requirement. This is because violent sounds are mostly sudden, such as gunfire and shouting, and their occurrence time and frequency are unfixed, random, and unpredictable. As a result, the distribution of violent audio events and non-violent audio events in data sets is uneven, and the network cannot fully learn the features of violent audio. Therefore, this paper does not use the end-to-end feature extraction model for feature extraction of violent audio.

An audio signal can also be used contain all audio-related information of a spectrum diagram or an audio waveform envelope for characterization. In this paper, CNN net, as the best image feature extraction method, is used to extract audio features, and the original audio waveform is mapped to the two-dimensional field as the network input, achieving end-to-end audio feature extraction. This method not only improves the representative accuracy of violent behavior but also solves problems related to the form and scale of audio and video features regarding their inconsistent processing of visual and auditory information fusion.

In order to avoid unnecessary network training and parameter tuning overhead, this paper uses the pre-trained AlexNet structure as the feature extraction model of the CNN network to extract frame-level features. The AlexNet structure contains five convolutional layers, three pooling layers, and three fully connected layers. The output of the last full connection layer passes through the SoftMax layer of 1000 neurons, which generates a probability distribution of the 1000 class label output results. At the same time, due to the continuity of violent behavior on the time axis, this paper selects the LSTM network to process the temporal relationships between audio frame-level features and to obtain segment-level features. The specific structure of the auditory feature extraction module is shown in Figure 1.

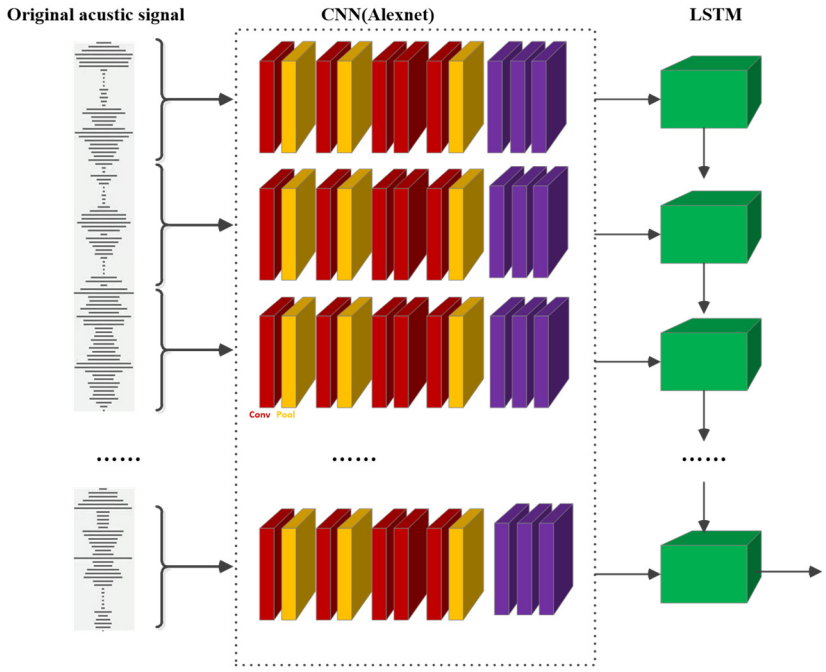

Figure 1. The module of violent auditory feature extraction.

In this module, the red rectangle represents each convolution layer, the yellow represents the pooling layer, the blue represents the fully connected layer, and the green represents the LSTM structure. The convolution layer contains the ReLU activation function, which makes the activation mode of the network sparser. The pool layer contains a local response normalization operation to avoid gradient disappearance and improve the network's training speed.

The two-dimensional spectrum features of the auditory signal as input in Figure 1 can be spectral or audio waveform envelopes. In order to verify the effectiveness of the two inputs, this paper uses the model shown in Figure 1 to extract auditory features. Two fully connected layers of 2048 $\times$ 512 and 512 $\times$ 2 are added to the output end of the model as classifiers. Through verification on the MediaEval 2015 training set [26], the experimental results are shown in Figure 2. The horizontal axis in Figure 2 is the number of iterations, while the vertical axis is the recognition accuracy. As can be seen from Figure 2, in terms of the ability to distinguish violent from nonviolent audio, the recognition accuracy of the original audio envelope map is better than that of the spectrogram, at least in most cases. Therefore, this paper selects the audio waveform envelope as input information for the feature extraction of the auditory channel.

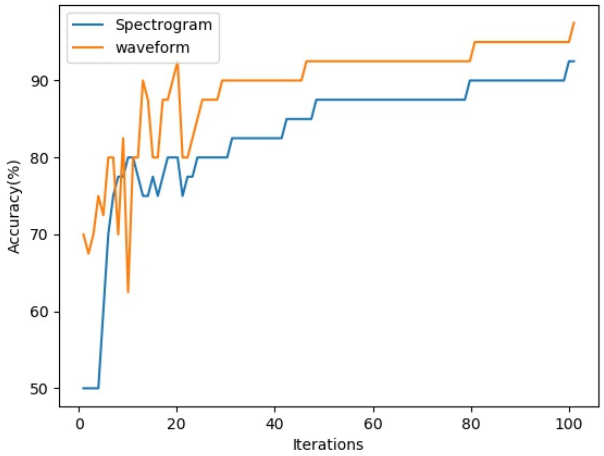

Figure 2. Iterative training results of violent behavior based on auditory features.

2.2. Visual Feature Based on CNN-ConvLSTM

Visual information plays a key role in the detection of violence. Violent behavior recognition primarily detects violent continuous actions in a video; these need to be processed by images of the video frame. Considering that the object of violent behavior recognition is violent action, the inter-frame differences in a video can theoretically extract the required information more accurately than the video frame itself [7]. Therefore, the difference between the adjacent frames of a video is selected as the input of the network model in this paper, and the same AlexNet structure is used for visual frame-level feature extraction.

Considering that this paper uses the frame-level features extracted by the difference between video images, meaning higher requirements for local spatial features, this paper selects the ConvLSTM network [31] to capture the temporal relationship between visual frame-level features. As such, this process is realized by the ConvLSTM network in this paper.

Therefore, regarding the feature extraction of video violent behavior, this paper adopts feature extraction architecture that is consistent with the structure of an audio feature extraction module. In other words, AlexNet is used as the CNN subject to extract image features, but the classic LSTM module is replaced by ConvLSTM, and the original input signal becomes the difference between frames of the image.

3. The Deep Network for Auditory Visual Information Fusion

In Section 2, the features of different modes are acquired. Next, the features need to be fused. Sharing a subspace can eliminate feature heterogeneity among different modes, and then capture complementary information and high-level semantics among different modes, thus realizing feature fusion at a semantic level. However, the semantic inconsistencies in violent videos pose a challenge to the design of shared subspace models. In order to solve the modal characteristics of different semantic inconsistency problems, we designed a shared subspace based on the autoencoder-mapping model, its labels, and introduced a semantic relation between semantic equivalent labels to optimize the learning sharing subspace. In so doing, we sought to solve semantic consistency issues through audiovisual information fusion in terms of the given recognition model framework and the implementation of the algorithm that recognizes violence.

3.1. Shared Semantic Subspace Based on Autoencoder

3.1.1. Shared Semantic Subspace

Spatial learning aims to obtain isomorphic subspaces shared by multiple modalities to capture complementary information between different modalities. Suppose the auditory feature extracted in Section 2.1 is f_{audio}, the visual feature extracted in Section 2.2 is f_{visual}, the feature mapping functions from auditory, visual from the shared semantic subspace are

$h()$ and $g()$, and the mapping functions from the shared semantic subspace for auditory and visual are $H()$ and $G()$, respectively. Then, the mapping relationship from visual features to auditory features is expressed as Equation (1), and the mapping relationship from auditory features to visual features is expressed as Equation (2).

$$f'_{audio} = H(g(f_{visual})) \tag{1}$$

$$f'_{visual} = G(h(f_{audio})) \tag{2}$$

After integrating visual and auditory features into the same subspace by Equations (1) and (2), the shared semantic features f'_{audio} and f'_{visual} are obtained. At this time, they have the same semantic properties and can be fused in different ways. In this paper, the CONCAT method is adopted to directly combine the visual and auditory features in the shared subspace with the input feature vector of the final violence event detector, which is shown in Formula (3).

$$f_{fusion} = CONCAT(f_{visual}, f'_{visual}, f_{audio}, f'_{audio}) \tag{3}$$

It can be seen from the above analysis that the process of obtaining the shared subspace actually computes the mapping function of isomorphic subspaces with different modality characteristics. The mapping function can be projection calculation, matrix decomposition, multi-label learning, discrete hash optimization [32], distance, etc. In this paper, an autoencoder mapping model is used to calculate the isomorphic subspace of auditory-visual features.

3.1.2. Shared Semantic Subspace Based on Autoencoder

An autoencoder is an unsupervised neural network model that can learn the deep representation of input data. Therefore, this function can be used to obtain the isomorphic shared semantic subspace of auditory-visual features, as shown in Figure 3. It can be seen from the figure that the model consists of an encoder and a decoder. The auditory-visual features share the same encoder, and each has its own decoder. The ideal output of the decoder should be equal to the corresponding input.

When the inputs are the visual and auditory features with semantic consistency, the error of the autoencoder model includes two parts. One is the error of the auditory decoder y_{audio} and the other is the error of the visual decoder y_{visual}. The sum of the two as the total error can be backpropagated to update the weights of the autoencoder. The encoder can map audiovisual features to the common coding space, which is equivalent to mapping functions $g()$ and $h()$. Then, the decoder is used to map the features to different modality spaces, and the compensation features of other modalities are obtained. Finally, these features are spliced using Equation (3) and used as input for the classification model to identify violent behaviors.

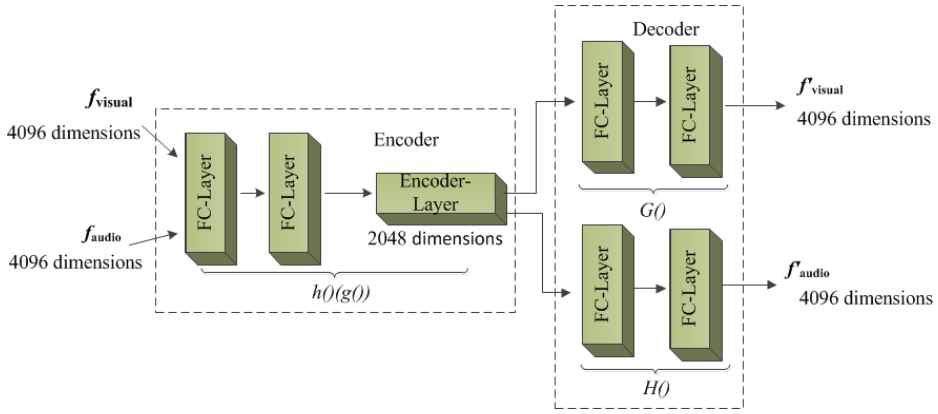

Figure 3. Autoencoder model.

3.1.3. Model Optimization Based on Semantic Correspondence

In our analysis of the VSD2015 dataset, we found that, for the same video, visual and auditory information showed semantic inconsistencies in relation to time axis misalignment and semantic deviation; this poses a challenge for visual information fusion. To address this problem, we suggest a new label called "semantic mapping" for the dataset using L_{corr}. This label is used to describe whether the audiovisual data of the same video contain the same semantic information. Video data containing blood, weapons, physical violence, etc. are considered visual violence. Audio that contains gunshots, screams, and explosions is considered auditory violence. Audio and video data are marked separately to prevent interference with each other. If the visual semantic label of the video is the same as the audio semantic label, the audio and video are considered to have a semantic correspondence $L_{corr} = 1$; otherwise, there is no semantic correspondence $L_{corr} = -1$. Semantic tags provide metrics for constructing shared subspaces with different modal features. In this paper, semantic tags are introduced into the calculation of loss function for the training of an autoencoder model. When there is semantic correspondence between visual and auditory information in a video, a loss function is the absolute error of visual and auditory coding information. The loss function is shown in Formula (4).

$$y_{autocoder} = \begin{cases} \left(\frac{1}{N} \sum_{i=1}^{N} |f_{audio} - f'_{audio}| + \frac{1}{N} \sum_{i=1}^{N} |f_{visual} - f'_{visual}| \right), L_{corr} = 1 \\ 1 - \left(\frac{1}{N} \sum_{i=1}^{N} |f_{audio} - f'_{audio}| + \frac{1}{N} \sum_{i=1}^{N} |f_{visual} - f'_{visual}| \right), L_{corr} = -1 \end{cases} \quad (4)$$

The loss function is designed to reduce the interference of blind splicing features. In this sense, the discriminative ability of the self-encoding mapping model for the semantic correspondence of violent videos is enhanced, which is more conducive to eliminating the interference between noncorresponding features. In addition, semantic-embedded learning can be regarded as a form of regularization, which helps to enhance the generalization ability of models and prevent overfitting.

3.2. Violent Behavior Recognition Model Based on Visual and Auditory Fusion

3.2.1. Network Structure

According to Sections 2 and 3.1, this paper designed a violent behavior recognition model based on the auditory-visual information fusion of an autoencoder. The model structure is shown in Figure 4. The model comprises four parts: visual feature extraction, auditory feature extraction, the autoencoder model, and the full connection recognition model. Regarding visual and auditory feature extraction, a two-channel feature extraction method is adopted, and the network structure adopts the classic AlexNet network in CNN. In relation to visual features, the interframe differences in terms of video are used as original input, and the segment-level visual features are extracted by the AlexNet–ConvLSTM network. In terms of auditory features, audio waveform is used as network input and the AlexNet–LSTM network is used to extract segment-level auditory features. Then, the autoencoder model is used to construct the shared semantic subspace to eliminate the semantic biases of visual and auditory features, and the CONCAT method is used to achieve the combination of visual and auditory features. Finally, the full connection model is used to identify violent behavior.

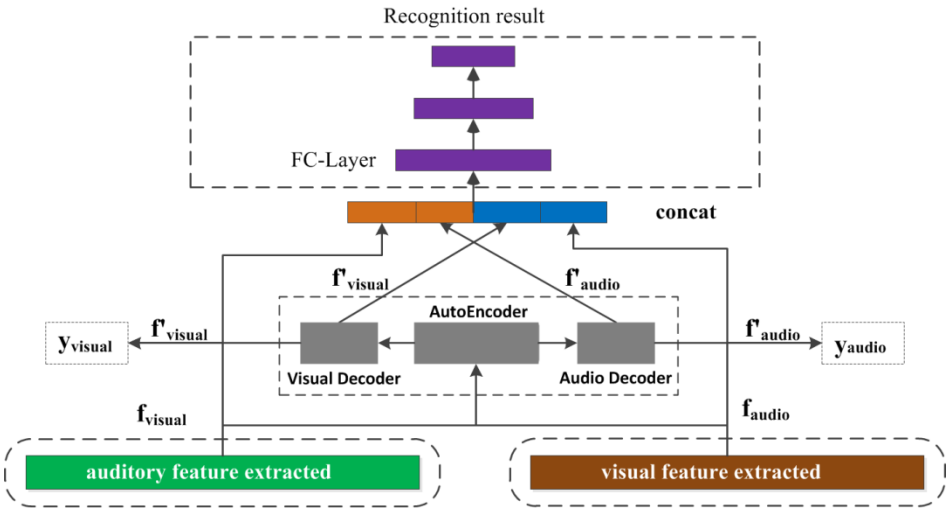

Figure 4. Violence recognition network based on auditory visual fusion of Autoencoder mapping.

In the method used in this article, timing information is summarized through the LSTM network in the final stage of visual and auditory feature processing; this approach can cover the entire multimedia paragraph, so there is no rigidity in terms of the length and sampling rate of the audio and video, etc. Therefore, the feature time axis alignment problem is solved. On the other hand, this method also greatly reduces the complexity of visual and auditory feature fusion and improves the stability of the model. Furthermore, in the output module of autoencoder mapping, and in addition to the CONCAT method for visual and auditory feature splicing, this paper also tries to use the Add method for feature combination. Experimental results show that the CONCAT method is better than the Add method at the feature level.

3.2.2. Algorithm Realization

According to the model structure in Figure 4, the back propagation (BP) mechanism is used for updating parameters. The autoencoder model is used to calculate the shared semantic subspace, two errors need to be considered in the process of model training. These are the error of the autoencoder model $y_{autocoder}$ and the violence recognition error y_{recog}, y_{recog}, which is calculated by cross entropy loss function. Thus, the error function can be written as:

$$y_{total} = y_{autocoder} - log\frac{e^{x_{label}}}{\sum_{j=1}^{N} e^{x_j}} \quad (5)$$

where N represents the number of input samples x. The AlexNet–ConvLSTM and AlexNet–LSTM networks are used for auditory-visual feature extraction. Since AlexNet in these two networks has been pretrained on the ImageNet dataset, the AlexNet parameters are frozen during training, and only the parameters of ConvLSTM, LSTM, the autoencoder model, and the fully connected classifier are updated. The specific Algorithm 1 for this is as follows.

Algorithm 1. Algorithm of Auditory Visual Fusion of Autoencoder Mapping

Input: Video frame sequence, Audio frame waveform image, Label x_{label}, Iteration number T
Output: Weights of Network model

1: Initialize the network weights, freeze some parameters of AlexNet, t = 1;
2: for t = 1:T do
3: Compute network model output: f_{audio}, $f\prime_{audio}$, f_{visual}, $f\prime_{visual}$ and label
4: Calculate the error value y_{total} at time t according to formula (10)
5: Calculate the error gradient δ_k of hidden layer element k at time t
6: Calculate the error gradient δ_{ct} of state C_t at time t
7: Update network weight vectors W

4. Experiments and Results Analysis

4.1. The Experimental Setup

4.1.1. Dataset

In this paper, the movie dataset MediaEval 2015 [28] was used to identify violent behaviors in videos. The specific information from this dataset is shown in Table 1. The data are derived from 199 Hollywood movies and include visual and auditory clips and violence-labeling information. A violent video is defined as a video clip with an R8 X-rated content, and it includes explosions, screaming or fighting, shooting, knife crime, and a variety of other forms of violence. We specified 6144 samples for the training set and 4756 samples for the test set. The training set included 272 samples of violence as well as 5872 samples of nonviolence; the test set included 230 samples of violence and 4526 nonviolent samples. Due to the unbalanced number of samples in this dataset the use of violent and nonviolent video in the training process was characterized by the label-shift confrontational and unsupervised domain-adaptive method to enhance data processing [33], and we added random noise with all kinds of data, as well as rotation or transition to make the size of the two classes of samples consistent. At the same time, in order to optimize the self-organizing mapping model using modal semantics, the frame-level semantic corresponding label was re-labeled, which was used for model training together with the violence label of the dataset itself.

Table 1. Experimental dataset.

Dataset Name	Type	Data Scale	Length/Clips (sec)	Scenario	Annotation
MediaEval2015	Violence	502	8~12	Movie	Frame-Level

The MediaEval 2015 competition officially provided the MAP (mean average precision) indicator as a performance evaluation indicator for the recognition of violence.

4.1.2. Experimental Parameters Config

(1) Model module parameter Settings.

In order to verify the recognition ability of the proposed method in a real scene, we carried out violent behavior recognition experiments of a single channel and a visual and auditory fusion channel. The network structure is shown in Figure 4, and the specific network configurations are shown in Table 2.

Table 2. Parameters of violence recognition network based on auditory visual fusion of autoencoder mapping.

Module Name	Type	Input/Output Data Dimension	Repeat Times
Auditory feature extraction	AlexNet	(227 × 227 × 3, 4096)	1
	LSTM	(4096, 4096)	1
Video feature extraction	Substract	(227 × 227 × 3 × 2, 227 × 227 × 3)	1
	AlexNet	(227 × 227 × 3, 4096)	1
	Conv-LSTM	(4096, 4096)	1
Autoencoder	FC+ReLu	(4096, 2048)	1
	FC+ReLu	(2048, 4096)	2
	CONCAT	(4 × 4096, 4 × 4096)	1
Classifier	FC+ReLu	(4 × 4096, 4096)	1
	FC+ReLu	(4096, 2)	1

The network was constructed according to the network settings shown in Table 2, and the experimental dataset with frame level annotation was used for model training and testing. The hyperparameters in the training process are shown in Table 3.

Table 3. Hyperparameters of network training.

Hyperparameters of Network	Default Value
Learning Rate	10^{-5}
LR decay rate	0.5
Batch	16
Hidden size	128
Loss function	Cross Entropy
Penalty coefficient ratio	1:16
Optimized	A dam

(2) Evaluation indicators.

MediaEval 2015 provides a performance evaluation for video violence detection using the mean average precision (MAP) metric. In addition, the commonly used accuracy, P, recall, R, and F1 values were also used to evaluate the results of this method. Finally, experiments are carried out according to Algorithm 1.

4.2. Experimental Results

4.2.1. Validation of Feature Combination Method

As can be seen from Figure 4, common features of audiovisual modes can be obtained through a shared subspace, and the combination of these features also affects the effect of violence detection. In this paper, two combination methods—CONCAT and Add—were tested, and the baseline system in this paper was compared with several late-fusion methods. The experimental results are shown in Table 4. As can be seen from Table 4, the feature fusion method is superior to the late fusion method, and the CONCAT combination is superior to the Add combination. This shows that the feature fusion method can observe more information and make full use of the complementarity between multi-modal features compared with the later fusion method that uses decision-level scoring. In terms of feature fusion, compared with the Add method, the CONCAT combination can save information from different modes better. Therefore, the CONCAT method was used for feature splicing in the following experiments.

Table 4. Comparison of experimental results of different feature combination methods.

Method		*P*	*R*	*F*1	MAP
Late fusion	SVM	0.29	0.70	0.41	17.4%
	Average Fusion	0.34	0.75	0.46	19.1%
	3-layer Perceptron	0.33	0.73	0.45	18.6%
Feature fusion	Add	0.42	0.79	0.54	29.2%
	Concat	0.51	0.84	0.63	31.54%

4.2.2. Visual and Auditory Information Fusion Visualization Based on Autoencoder Mapping

To verify the effectiveness of the autoencoder mapping model, 230 violent videos were selected in this paper, while 230 nonviolent videos were randomly selected for feature visualization. Due to their high final feature dimensions, PCA (principal component analysis) and tSNE [26]. First, the PCA method is used to calculate the important components of visual and auditory features, and then these components are combined according to the violent and non-violent labels. Finally, the tSNE method is used to reduce the high-dimensional features to two-dimensional space.

Visualization of feature distribution in ultimate feature space, presenting the feature distribution of non-violence (red-circles) and violence (green-triangles) before autoencoder (left) and after autoencoder (right), such as in Figure 5. Note that the composition of the MediaEval 2015 dataset is complex. Whether or not the data are self-encoded, the distribution of the data is chaotic. However, it can still be seen from the figure that, after the encoding and decoding of the self-encoding model, the distribution of the two types of features in the right figure is obviously concentrated, which is more orderly than in the left

figure; this helps us to establish a more effective high-dimensional classification model and proves that self-encoding mapping can realize the complementarity of different pieces of modal information, at least to a certain extent.

Figure 5. Visualization of feature distribution in ultimate feature space.

4.2.3. Violence Test Results

The hyperparameters set in Table 3 were used for training and testing, and the experimental results are shown in Table 5. It can be seen from the experimental results that the auditory and visual channel fusion has greater significance for reducing the false detection rate. The fusion method proposed in this paper effectively combines the effective information of the two channels, which improves the performance of the model under all evaluation indicators, especially the *F1* value and map value, both of which represent a significant improvement in comprehensive recognition ability.

Table 5. Fusion modality feature recognition results.

	P	*R*	*F*1	MAP
Auditory feature	0.46	0.73	0.56	16.47%
Visual feature	0.36	0.82	0.50	20.21%
Fusion modality feature	0.51	0.84	0.63	31.54%

In order to further prove the effectiveness of the method proposed in this paper, the experimental results obtained by this method are compared with the results of other teams in MediaEval 2015, as shown in Table 6. From the experimental results, it can be seen that the auditory-visual information fusion method proposed in this paper based on an autoencoder achieves the best recognition effect under this dataset; the auditory feature MAP value is increased by 5.04% compared with the best result, and the visual feature is also improved. The MAP value of the fusion audio and video increased by 1.94%, which fully proved the effectiveness of the method proposed in this paper.

Table 6. Comparison of recognition results of violent behavior.

	Team	MAP
Auditory feature	ICL-TUM-PASSAU [28]	14.9%
	TCS-ILAB [34]	6.38%
	RECOD [35]	11.43%
	Proposed	16.47%
Visual feature	KIT [36]	12.9%
	Umons [37]	9.67%
	Proposed	20.21%
Fusion modality feature	RUCMM [38]	21.6%
	NII-UIT [39]	26.8%
	Fudan-Huawei [40]	29.6%
	Proposed	31.54%

5. Conclusions

This paper proposes an auditory-visual information fusion model based on an autoencoder for violent behavior recognition. The model is divided into three parts. First, an audiovisual feature extraction framework based on CNN-LSTM is proposed; this can be used to obtain the overall feature of the segment level that helps to solve the problem of misalignment on the time axis. Then, a shared semantic subspace based on an autoencoder is constructed to fuse visual and auditory features on the basis of ensuring the consistency of semantic information. Finally, the fully connected model is used to obtain the results of violent behavior recognition. The shared semantic subspace based on the autoencoder realized the complementarity of different modalities, and, after feature fusion, the model obtained better recognition results and improved identification accuracy and reduced the rate of missing detection. This shows that the feature extraction and multimodality feature fusion method proposed in this paper can effectively utilize the information related to violent events in visual and auditory features, make up for the inherent shortcomings of the visual and auditory channels, and effectively improve the accuracy of violent behavior recognition. However, the work conducted in this paper still needs to be improved. For example, the accuracy of visual and auditory feature representation by different convolution models is not considered in our experiment, while only AlexNet is used in our experiments is used in experiments. Therefore, in future work, we will focus on the impact of visual and auditory feature expression on the fusion effect, and will further explore the design of a shared subspace by the loss-based attention [41] for various Convolutional Neural Networks.

Author Contributions: J.L. and D.Z. conceived of the key idea and designed the study. Z.Z. performed the simulations. H.L. analyzed the raw data and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China [grant number 62171155], the Natural Science Foundation of Jilin Province, China [grant number YDZJ202101ZYTS191].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are from the MediaEval 2015 effective impact of movies task database used in the MediaEval challenge. The database acquisition address: https://liris-accede.ec-lyon.fr/ (accessed on 27 August 2021). Please note, that if you want to acquire access to the database, you need to apply first and then download it.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ramzan, M.; Abid, A.; Khan, H.U.; Awan, S.M.; Ismail, A.; Ahmed, M.; Ilyas, M.; Mahmood, M. A review on state-of-the-art violence detection techniques. *IEEE Access* **2019**, *7*, 107560–107575. [CrossRef]
2. Nayak, R.; Pati, U.C.; Das, S.K. A comprehensive review on deep learning-based methods for video anomaly detection. *Image Vis. Comput.* **2021**, *106*, 104078. [CrossRef]
3. Ribeiro, P.C.; Audigier, R.; Pham, Q.C. RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. *Comput. Vis. Image Underst.* **2016**, *144*, 121–143. [CrossRef]
4. Dhiman, C.; Vishwakarma, D.K. High dimensional abnormal human activity recognition using histogram oriented gradients and Zernike moments. In Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, 14–16 December 2017; Springer: Berlin/Heidelberg, Germany, 2017; p. 5.
5. Senst, T.; Eiselein, V.; Kuhn, A.; Sikora, T. Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation. *IEEE Trans. Inform. Forensics Secur.* **2017**, *12*, 2945–2956. [CrossRef]
6. Bilinski, P.; Bremond, F. Human violence recognition and detection in surveillance videos. In Proceedings of the 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, CO, USA, 23–26 August 2016; IEEE: Piscataway, NJ, USA, 2016; Volume 7, pp. 30–36.
7. Zhang, T.; Yang, Z.; Jia, W.; Yang, B.; Yang, J.; He, X. A new method for violence detection in surveillance scenes. *Multimed. Tools Appl.* **2016**, *75*, 7327–7349. [CrossRef]

8. Mu, G.; Cao, H.; Jin, Q. Violent scene detection using convolutional neural networks and deep audio features. *Commun. Comput. Inform. Sci. CCPR* **2016**, *663*, 451–461.
9. Xie, J.; Yan, W.; Mu, C.; Liu, T.; Li, P.; Yan, S. Recognizing violent activity without decoding video streams. *Optik* **2016**, *127*, 795–801. [CrossRef]
10. Peixoto, B.M.; Avila, S.; Dias, Z.; Rocha, A. Breaking down violence: A deep-learning strategy to model and classify violence in videos. In Proceedings of the 13th International Conference on Availability, Reliability and Security, Hamburg, Germany, 27–30 August 2018; ACM Library: New York, NY, USA, 2018; Volume 50, pp. 1–7.
11. Manzo, M.; Pellino, S. Voting in transfer learning system for ground-based cloud classification. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 542–553. [CrossRef]
12. Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S.W. Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access* **2018**, *6*, 1155–1166. [CrossRef]
13. Sreenu, G.; Saleem Durai, M.A. Intelligent video surveillance: A review through deep learning techniques for crowd analysis. *J. Big Data* **2019**, *6*, 1–27. [CrossRef]
14. Accattoli, S.; Sernani, P.; Falcionelli, N.; Mekuria, D.N.; Dragoni, A.F. Violence detection in videos by combining 3D convolutional neural networks and support vector machines. *Appl. Artif. Intell.* **2020**, *34*, 329–344. [CrossRef]
15. Tripathi, G.; Singh, K.; Vishwakarma, D.K. Violence recognition using convolutional neural network: A survey. *J. Intell. Fuzzy Syst.* **2020**, *39*, 7931–7952. [CrossRef]
16. Oscar, D.; Ismael, S.; Gloria, B.; Tae-Kyun, K. Fast Violence Detection in Video. In Proceedings of the 9th International Conference on Computer Vision Theory and Application (VISAPP), Lisbon, Portugal, 5–8 January 2015; Volume 2, pp. 478–485.
17. Sharma, M.; Baghel, R. Video Surveillance for violence detection using deep learning. *Lect. Notes Data Eng. Commun. Technol.* **2020**, *37*, 411–420.
18. García-Gómez, J.; Bautista-Durán, M.; Gil-Pita, R.; Mohino-Herranz, I.; Rosa-Zurera, M. Violence Detection in Real Environments for Smart Cities. In Proceedings of the 10th International Conference of Ubiquitous Computing and Ambient Intelligence (UCAmI), San Bartolomé de Tirajana, Spain, 29 November 2016; Volume 10070, pp. 482–494.
19. Chen, L.; Jakubowicz, J.; Yang, D.; Zhang, D.; Pan, G. Fine-Grained urban event detection and characterization based on tensor cofactorization. *IEEE Trans. Hum.-Mach. Syst.* **2017**, *47*, 380–391. [CrossRef]
20. Wang, Y.; Neves, L.; Metze, F. Audio-Based Multimedia Event Detection Using Deep Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2742–2746.
21. Lejmi, W.; Khalifa, A.B.; Mahjoub, M.A. Fusion Strategies for Recognition of Violence Actions. In Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, Hammamet, Tunis, 30 October–3 November 2017; pp. 178–183.
22. Asad, M.; Yang, J.; He, J.; Shamsolmoali, P.; He, X. Multi-frame feature-fusion-based model for violence detection. *Vis. Comput.* **2021**, *37*, 1415–1431. [CrossRef]
23. Song, D.; Kim, C.; Park, S.-K. A multi-temporal framework for high-level activity analysis: Violent event detection in visual surveillance. *Inform. Sci.* **2018**, *447*, 83–103. [CrossRef]
24. Xia, Q.; Zhang, P.; Wang, J.; Tian, M.; Fei, C. Real Time Violence Detection Based on Deep Spatio-Temporal Features. In Proceedings of the 13th Chinese Conference on Biometric Recognition, Zhuzhou, China, 12–13 October 2018; Volume 10996, pp. 157–165.
25. Michael, S.B. Chapter 42-Audiovisual speech integration: Neural substrates and behavior. In *Neurobiology of Language*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 515–526.
26. Gu, C.; Wu, X.; Wang, S. Violent video detection based on semantic correspondence. *IEEE Access* **2020**, *8*, 85958–85967. [CrossRef]
27. Ivanovic, B.; Leung, K.; Schmerling, E.; Pavone, M. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *IEEE Robot. Autom. Lett.* **2021**, *6*, 295–302. [CrossRef]
28. Sjöberg, M.; Baveye, Y.; Wang, H.; Quang, V.L.; Ionescu, B.; Dellandréa, E.; Schedl, M.; Demarty, C.; Chen, L. The MediaEval 2015 Affective Impact of Movies Task. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 14–15 September 2015.
29. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
30. Cramer, J.; Wu, H.-H.; Salamon, J.; Bello, J.P. Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3852–3856.
31. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.; Wong, W.; Woo, W. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inform. Process. Syst.* **2015**, 802–810.
32. Shi, X.; Xing, F.; Zhang, Z.; Sapkota, M.; Guo, Z.; Yang, L. A scalable optimization mechanism for pairwise based discrete hashing. *IEEE Trans. Image Process.* **2021**, *30*, 1130–1142. [CrossRef] [PubMed]
33. Liu, X.; Guo, Z.; Li, S.; Xing, F.; You, J.; Jay Kuo, C.-C.; Fakhri, G.; Woo, J. Adversarial unsupervised domain adaptation with conditional and label shift: Infer, Align and Iterate. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, ON, Canada, 11–17 October 2021.

34. Chakraborty, R.; Maurya, A.K.; Pandharipande, M.; Hassan, E.; Ghosh, H.; Kopparapu, S.K. TCS-ILAB-MediaEval 2015: Affective Impact of Movies and Violent Scene Detection. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 14–15 September 2015.
35. Moreira, D.; Avila, S.; Perez, M.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; Rocha, A. RECOD at MediaEval 2015: Affective Impact of Movies Task. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 14–15 September 2015.
36. Vlastelica, M.P.; Hayrapetyan, S.; Tapaswi, M.; Stiefelhagen, R. KIT at MediaEval 2015-Evaluating Visual Cues for Affective Impact of Movies Task. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 14–15 September 2015.
37. Seddati, O.; Kulah, E.; Pironkov, G.; Dupont, S.; Mahmoudi, S.; Dutoit, T. UMons at MediaEval 2015 Affective Impact of Movies Task Including Violent Scenes Detection. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 14–15 September 2015.
38. Jin, Q.; Li, X.; Cao, H.; Huo, Y.; Liao, S.; Yang, G.; Xu, J. RUCMM at MediaEval 2015 Affective Impact of Movies Task: Fusion of Audio and Visual Cues. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 14–15 September 2015.
39. Vu, L.; Sang, P.; Duy-Dinh, L.; Shinichi, S.; Duc-Anh, D. NII-UIT at MediaEval 2015 Affective Impact of Movies Task. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 14–15 September 2015.
40. Dai, Q.; Zhao, R.; Wu, Z.; Wang, X.; Gu, Z.; Wu, W.; Jiang, Y. Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. In Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 14–15 September 2015.
41. Shi, X.; Xing, F.; Xu, K.; Chen, P.; Liang, Y.; Lu, Z.; Guo, Z. Loss-based Attention for Interpreting Image-level Prediction of Convolutional Neural Networks. *IEEE Trans. Image Process.* **2021**, *30*, 1662–1675. [CrossRef] [PubMed]