iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5257

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

A Review on Methods utilized for Audio based Detection of Violent Scenes in Videos

Mrunali D. Mahalle[1], Dr. Dinesh V. Rojatkar[2]
[1]*M. Tech. Student,* [2]*Associate Professor, Department of Electronics Engg., Government College of Engineering, Amravati, Maharashtra, INDIA*

Abstract: ***Internet video, movies have grown quite speedy in latest years with the success of multimedia social community as well as low cost of clever devices, accounting for 90% of the Internet traffic. Videos with detrimental content, such as horror videos, violent videos, and other detrimental videos are flooding. However, the increasing use of the associated technology via sensitive social groups creates the want for protection from harmful content material to hold the Internet video ecosystem, especially as the number of younger Internet users is growing rapidly. As the violent scene detection in videos has realistic significance in a number of applications, such as sensible surveillance, video retrieval, Internet filtering, film rating, toddler protection towards violent behaviour and so on. This paper details the distinct methods and techniques that are being used for the task of audio-based classification and detection of violent scenes in videos. This paper also contains our proposed method for audio based violent scene detection using extreme learning machine algorithm.***
Keywords: Violent Scene Detection, Visual Based System, Audio Based System, Extreme Learning Machine, Dataset.

I. INTRODUCTION

The wide variety of videos on the Internet, surveillance system, TV and other entertaining websites and mediums has been swiftly increasing in latest years, which allows to get entry to them easily, by means of giant parts of the population, including sensitive social groups, e.g., children, teenagers, etc. and has given their lives entertainment. This state of affairs additionally makes it viable for youth to easily attain violent contents at the equal time even though it needs to be filtered. Because of the big variety of them, however, it is nearly hard to give remarks on these videos manually to get rid of them. To prevent exposure of children and youngsters to beside the violent content, rankings are typically followed in maximum international locations where in an international rating is given to a film or web videos by a specific organizations, e.g., the Motion Picture Association of America (MPAA) in the USA, the Centre National du Cin´ema et de l'image anim ´ ee (CNC) in France, or the Motion Picture Code of Ethics Committee (Eiga Rinri Kanri Iinkai) in Japan. Primarily based on a hard and fast of standards, the score is provided globally for a given matter, commonly placing a minimal age for viewers. For example, ratings in japan are chosen consistent with four classes: all ages admitted (G), some matter may be inappropriate for youngsters underneath the age of 12 (PG-12), forbidden under 15 (R15+) or beneath 18 (R18+). International rankings however go through boundaries. Some nations, e.g., the People's republic of china, do not have a movie rating system as of nowadays. More importantly, ratings are highly dependent on time and nation. Rating rules have substantially modified through the years—you could without problems find a film rated by way of the MPAA as "no one 17 and under admitted" (NC-17) a few years back,
now rated as "parental guidance for children under 13" (PG-13)—and also are unique from one nation to any other, with however gender and violence as a constant preoccupation through all nations.
Subsequently, international ratings are poorly adapted to today net-primarily based content diffusion, wherein the range of video vendors have exploded at the side of the volume of matter to be had, in which country wide boundaries almost not exist and in which not all matter goes through ratings [1]. Many videos on social media are visually some time are not violent but by audio these web videos are violent as the phrase of abuse are used by way of person to supply the speech makes it a violent video. This makes it integral for the improvement of efficient, automatic, content-based violence detectors in digital content. Also, in video surveillance, to seriously guarantee public security thousands and thousands of surveillance cameras are deployed inside cities, however it is nearly not possible now a day to manually reveal all cameras to hold an eye on violent activities. Since violent conditions are usually accompanied via signs like arguments, shouts or an expand in the extent of the conversation, phrases of abuse used, gunshots, explosions and so on. So, a strength efficient system capable of acoustically detecting violence will be beneficial for detection of all violent situations. Therefore, for the prevention of violence in the digital content material and for the security surveillance gadget is important, and this can be finished by means of violent scene detection system (VDS).

Violent Scene Detection System (VDS) can be divided into four primary parts:1) Visual based System (VBS), 2) Audio based System (ABS), 3) Audio-visual based VDS, 4) Audio-visual with text content based VDS. In visual primarily based strategy visual facts is extracted and represented as relevant features. Violent scenes can be realized by mean of decomposing the venture as action scene detection and bloody body detection. In audio-based strategy audio information is used to classify violence. On the basis of one-of-a-kind sounds such as screaming, arguments, phrases of abuse used, fights, shots, violent scenes can be detected. In audio visual based approach, emphasis is put on combining each visual and audio feature. Audio primarily based procedures are no longer as common as visual or audio-visual based tactics in spite of their performances. The audio content-based strategy commonly requires fewer computational sources than visible methods. There are specific elements which gives a compact representation of the given audio signal. Audio aspects can be labelled in two types: Time domain features and frequency domain features. These audio features can lead to three layers of audio understanding: low-level acoustics, such as the common frequency for a frame, midlevel sound objects, such as the audio signature of the sound a ball makes whilst bouncing, and high-level scene classes, such as background track taking part in in positive sorts of video scenes. All these features of audio can make a violent scene detection system environment friendly also with a low computational cost [9].

A. Violence Definition

The task of violent scene detection (VSD) has been carried out from earlier two decades, in the movies particularly in video surveillance field. Hence, as a result of numerous facets of violence, no dominant and generic satisfactory definition for violent event was ever proposed, even if proscribing ourselves to physical violence. From previous work, examples are taken for the definition of violence (in movies) are: “a series of human actions accompanied with bleeding” [13]; “scenes containing fights, regardless of context and number of people involved” [14]; “behaviour by persons against persons that intentionally threatens, attempts, or actually inflicts physical harm” [15]; “fast paced scenes which contain explosions, gunshots and person-on-person fighting” [16]. To deal with the variety of violence and offer a fairly objective definition with a view to ease the annotation method and maintain the dataset consistency, VSD annotations are constrained to physical violence in step with the subsequent definition, known as the ‘objective’ definition inside the sequel: “physical violence or accident resulting in human injury or pain” [17]. Some definitions only aim on action scenes, ignoring the distinction between actions and violence [18, 19].

B. Overview Of Elm Algorithm

ELM is a single layer feedforward neural network (SLFN). As it avoids multiple iteration the speed of learning of ELM is extremely fast. It is almost free from human intervention as compared to other ML and DL algorithms. It has homogenous models for compression, feature learning, clustering, regression and classification. ELM is easy for micro level real-time learning and control, up to thousand times faster, efficient for multichannel data fusion and potential for decision synchronization then DL. All these features make ELM algorithm capable for classification and detection [20]. Fig. [1] shows the structure of SLFN ELM.

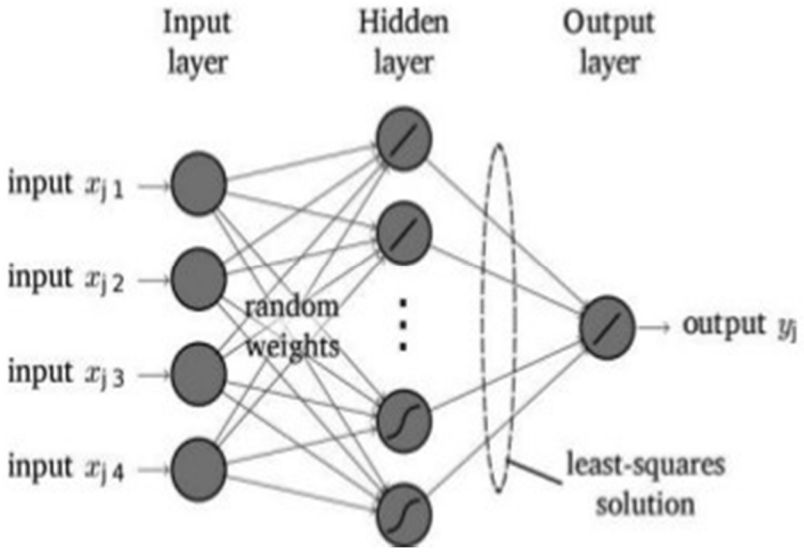


Fig. 1. Structure of Extreme Learning Machine.

C. *Proposed Methodology*

Violent scene detection is a challenging task, researchers had proposed lots of different techniques for VSD system. All the traditional machine learning algorithms (ML) and deep learning algorithms (DL) take lots of time for training the model. Extreme Learning Machine (ELM) algorithm takes less time as compared to ML and DL as ELM is single hidden feedforward neural network and avoids multiple iterations and also it can be used to solve reversion and classification problems [11]. [12] The work done is using kernel ELM and three-dimensional histograms of gradient orientation for VSD which is only based on visual features only. The accuracy obtained using KELM with hockey dataset is 95.05% and with movies dataset is 99.95% which shows that approach using ELM will be efficient for higher accuracy. This work was based visual based but accuracy given by approach using ELM is very high. There was very less research work on audio-based violent concept detection approach, so the proposed work contains the audio based VSD system using ELM algorithm.

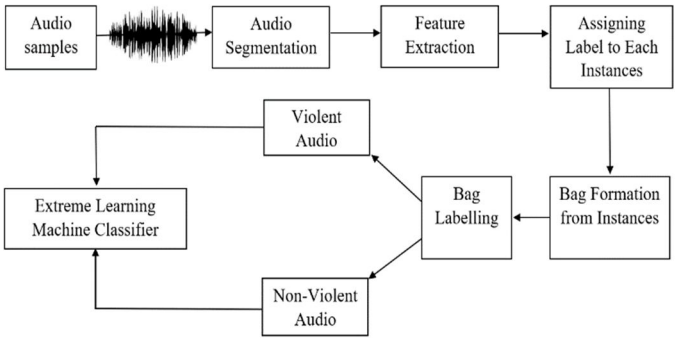


Fig. 2.: Block diagram of proposed work

The block diagram in fig. [2]. gives a general view of the proposed work:

1) *Audio Samples Dataset:* In the first step of the process some audio samples are used for training the model and some samples are used for testing. The Media Eval, movies, hockey or self-recorded video and audio recordings from movies and some real films can be used as a dataset to train the model. This is given to the next step that is pre-processing.
2) *Audio Segmentation:* This block is used to process the data, which will be given by training dataset to model or test the data that is input data. Audio segmentation is a pre-processing step in which audios clips are split into particular time length overlapped segments. This data is then given as input to the feature extraction block for feature selection.
3) *Feature Extraction:* In the feature extraction process, the feature will be extracted on the basis of some rules which are decided for extraction of feature. Instance will be created from each audio segments by extracting features in order to separate different types of sound, for example speech, music, environmental sounds, screaming, silence, and combination of these sounds. These instances are then given to next step which will be used for assigning label.
4) *Assigning Label to each Instances:* All the feature set i.e. group of instances belonging to the same audio sounds files are grouped into a bag. Labels will be assigned for the bag in which instances are present. All the bags along with their labels are assigned into a classifier for classification.
5) *Output:* After classification, it gives output that detects whether the audio clips are violent or non-violent type. The extreme learning machine algorithm is used for the detection of violent scene.

II. EXISTING AUDIO BASED METHODS USED FOR VSD:

During the last few years, the popularity of video sharing websites and applications through mobiles, tablets and smart TVs has given rise to a huge increase in the number of videos that can be uploaded and accessed by large portions of the population, including sensitive social groups, e.g., children, teenagers, etc. The necessity to protect such sensitive groups from accessing offensive content, along with the inherent difficulty in manually annotating huge volumes of video data, highlight the need for the development of efficient, automatic, content-based violence detectors. Also, violence detection has a practical significance in intelligent surveillance system. This has made violent scene detection to gain lot of attention over the past few decades. After a detailed literature survey, it was found that violence detection can be classified using different parameters such as shown below:

A. Approaches used for Detection

1) Audio based approach
2) Visual based approach
3) Audio-visual based approach
4) Audio-visual with text content-based approach

B. Features used for Detection

1) Audio Features

a) Time domain features
b) Frequency Domain Features

2) Visual Features

Depends upon, for which situations the visual features are used to detect violent scenes for example:

a) Global Features
b) Local Features

Many techniques have been proposed for the purpose of detection of violence in movies, real time videos, audio clips and surveillance system and it involve the utilization of different features of audio and visual cues.

C. Audio Based Approach

The audio-based tactic uses time and frequency domain audio features in order to categorize violence. In [2] they had explained applications, different acoustic features and methods used for audio-based event detection. They had explained different types of features which are temporal features whose processing time is slower as compared to frequency features which is categorised into physical features and perceptual features. They had also explained various methods such Hidden Markov Models (HMM), Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) which had been used for audio-based event detection. C. Clavel. [3] built an audio-based surveillance system using GMM in order to detect violent events in crowded places which were considered as environment full of noise. They had used a typical event such as cries, shots and explosions for detection of abnormal situations. The approach was based on binary classification and had conducted many experiments to reduce the false rejection (FR) and false detection (FD) rates which can be calculated has:

$$FR = \frac{number\ of\ failed\ detections}{number\ of\ events\ to\ detect}$$

$$Fd = \frac{number\ of\ false\ detected\ windows}{number\ of\ windows}$$

Giannakopoulos T. [4] proposed a violence content classification system using audio features with SVM as a classifier by extracting audio segments from real movies. They had used six segments level time and frequency domain audio features. Three features were from time domain they are energy entropy in order to calculate abrupt changes in the energy level of the acoustic signal, short time energy and zero crossing rate (ZCR) used to the number of time-domain zero-crossings, divided by the number of samples in the frame which is the broadly used time domain feature. Three features from frequency domain are spectral flux, spectral rolloff, and signal amplitude. The data was classified on audio-basis with an accuracy of 85.5% and correctly detected the violent segments with an accuracy of 90.5%.

Giannakopoulos T. [5] further recommended his work by using multi-class classification algorithm for audio segments recorded from movies for violence detection. The binary classification task was accomplished by the usage of Bayesian Network in order to classify audio segment into six classes 3 for violent which contains scenes such as shots, fights, screaming and 3 for non-violent which are music speech and other non-violent sounds, particularly by using 12D audio features. These audio features are ZCR, MFCC, Pitch, Energy Entropy, Spectral Rolloff, Chroma Vector Feature, and Spectrogram Feature. The accuracy obtained by using the proposed method is 73.2%. The equation for ZCR can be given as: $ZCR = \frac{1}{N}\sum_{n=1}^{N-1}\frac{|sgn\{x(n)\}-sgn\{x(n-1)\}|}{2}$, where sgn(.) stands for the sign function, i.e., sgn{x(n)} = +1if x(n) ≥ 0 and −1 if x(n) < 0.

Esra Acar. [6] proposed a VSD system for Hollywood movies by using midlevel audio features with Bag of audio words (BoAW) representation based on MFCC which were constructed by using two different methods, namely the vector quantization-based (VQ-based) method and the sparse coding-based (SC-based) method.

The result shown by the method based on SC method was slightly better than VQ based method. An SVM classifier were used as a classifier. The training data set used contains higher number of non-violent shots then violent once to overcome the problem random undersampling were used. The accuracy obtained using VQ based is 72.2% and by SC based is 79.1%.

Md. ZaighamZaheer et al.[7] proposed a deep learning based system for scream sound detection in surveillance system. The polynomial Deep Boltzmann Machines (DBM) algorithm were used for different screaming sounds which are produced in different situations. MFCC feature were used as an input to the system. An 100% accuracy is achieved by the proposed system by using its self-recorded scream dataset.

Vivek P. [8] proposed a multiple instance learning (MIL) approach for binary classification of news based on audio features. News based on audio signals were segmented into particular instances and acoustic features like MFCC and Perceptual Linear Prediction (PLP) were extracted from each instance. Features of the instances having same audio files are grouped together called bags. These bags and instances were properly labelled and were fed to classifiers. mi-Graph and mi-SVM were two methods which were used as a MIL classifier. About 120 news recordings were used as dataset among which 60 had found violent. mi-Graph method using MFCC showed the best performance with accuracy of 90.0% as compared to mi-SVM with accuracy of 80.3%.

S. Sarman. [10] proposed and audio based violent scene classification using ensemble learning approach. Technicolcor violent scene detection dataset were used which contains 31 enormously violent and non-violent movies, in order to detect audio based violent scenes. MFCC and ZCR are two audio features, i.e. one from time domain and one from frequency domain were used. Data were classified using three classifiers such as SVM, Random Forest, and Bagging in which SVM with ZCR showed better performance as compared to others. This approach obtained a superior result with the official metric MAP@100 of 66%.

III. SUMMARIZED WORK DONE

Table1. Shows the comparative study of various methods used in the violent scene detection in machine learning and also in deep learning.

Table 1: Summary Of Methods Used For Violent Scene Detection

Researchers	Methodology		Scenes	Datasets	Results
	Features	Classification			
Theodoros Giannakopoulos, 2006	Energy Entropy, ZCR, Short Time Energy, Spectral Flux, Spectral Rolloff, and Signal Amplitude.	Support Vector Machine classifier	Violent audio segments are shots, explosions, fights and screams, non-violent audio segments are music and speech	audio segments, extracted from several movie genres	Classified 85.5% Detected 90.5%
Theodoros Giannakopoulos, 2007	12D Audio Features are: ZCR, MFCC(4), Spectrogram Features(2), Chroma Vector Features(2), Energy Entropy, Spectral Rolloff, Pitch	Multi Class Classification Using Bayesian Networks	6 Audio classes: 3 Violent are Shots, Fights Screams. 3 Non-Violent are Music, Speech, others non-violent sounds	6 Datasets consisting of 200 minutes of movie recordings.	73.2%
Esra Acar, 2013	Mid-level audio features which are BoAW representations based on MFCCs	Two Class SVM with RBF kernel model is constructed using VQ and SC based mid-level audio features.	Boundaries between scenes are automatically generated.	32,708 video shots from 18 Hollywood movies	VQ based is 72.2% SC based is 79.1%.
Md. Zaigham Zaheer, 2015	Mel-Frequency Cepstral Coefficient (MFCC)	Deep Boltzmann Machine (DBM) a Deep Learning Algorithm.	All screaming sounds produced at different situations	Self-Recorded screams dataset.	100%
Vivek P, 2016	MFCC and Perceptual Linear prediction (PLP)	mi-Graph and mi-SVM were used as a MIL classifier	Violent incident news videos from large news video archive.	120 News Recordings	mi-Graph with 90.0% mi-SVM with 80.3%
Sercan Sarman, 2018	MFCC and ZCR	SVM, Random Forest, Bagging used as Ensemble Learning	Gunshot, explosion, and scream.	Technicolor Violent Scene Detection dataset	Official metric MAP@100 of 66%

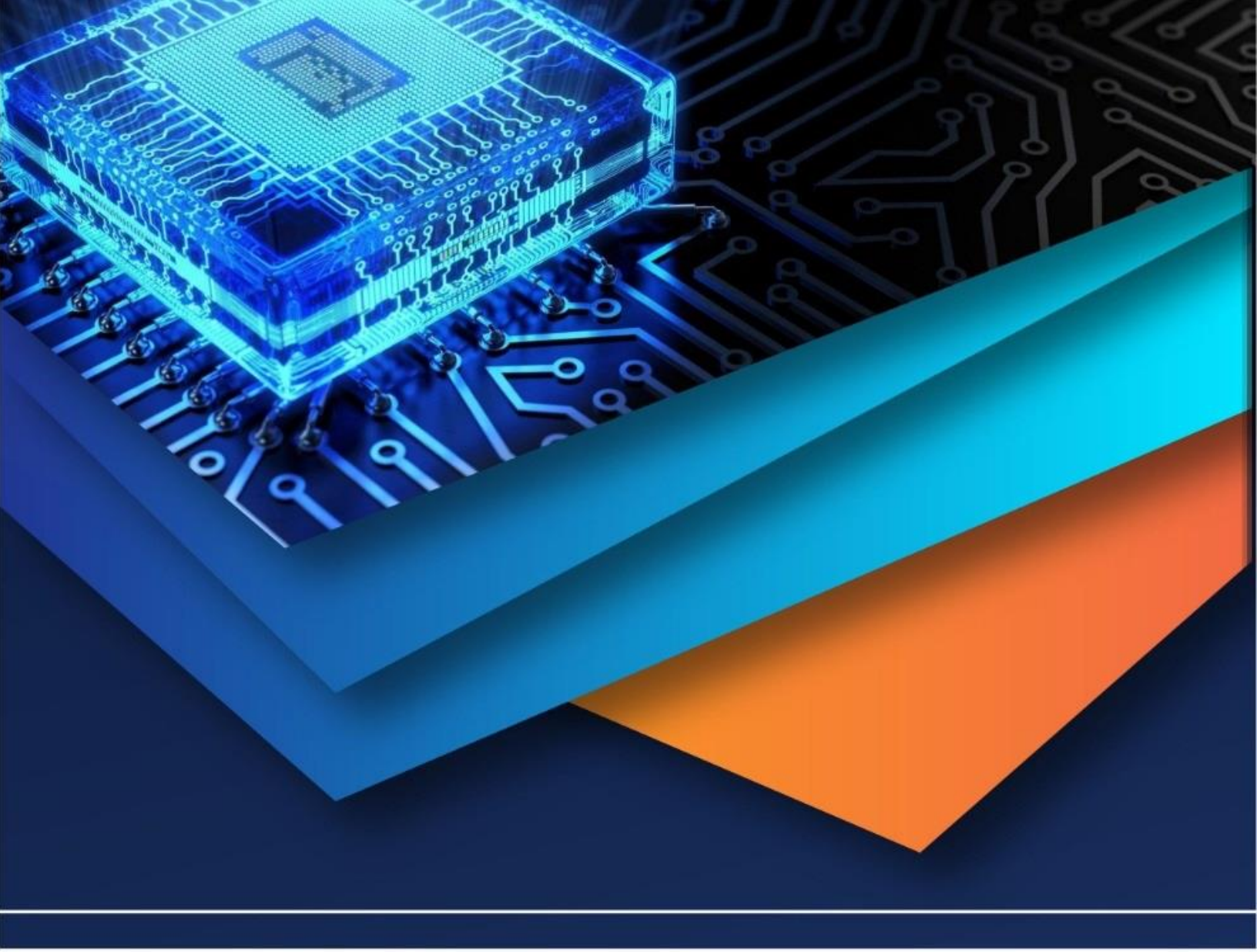
IV. CONCLUSION

This paper presents a review on audio based violent scene detection methodology and its development. To develop the productive abnormal scenes/ VSD system or to increase the accuracy of the system, the Violent identification/recognition is most needed. Researchers use various features from time domain and frequency domain (like ZCR from time domain and MFCC from frequency domain are most widely used features) and classifiers (such as SVM, Random Forest, Deep learning classifier, etc) to differentiate the various violent scenes at a different type of segmentation i.e. gun shots, explosions, fights, screams, etc.

In between audio or visual approach, lots of research/work is done in visual approach. The reason behind is the lag in the technology of audio based violent scene detection, where the complexity or difficulty occurs during misconceptions created. Due to this their forms a large research gap between audio and visual based violent analysis. Also audio based approach requires less computational sources than visual methods. Because of this there is a huge opportunity for research on audio-based approach. The approach audio based violent scene detection using extreme learning machine algorithm has not been used for violence detection. Expanding the system further by adding extreme learning machine algorithm using MATLAB will add more versatility to the system for the detection of violence in movies, videos which can be used for internet filtering, film ratings and also it can be used for surveillance system.

REFERENCES

[1] Claire-H´ el`ene Demarty, C´edric Penet, Mohammad Soleymani and Guillaume Gravier. “VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation”, © Springer Science+Business Media New York 2014.
[2] Rajeswari Natarajan and Chandrakala.S. “Audio-Based Event Detection in Videos - a Comprehensive Survey”, International Journal of Engineering and Technology (IJET) Vol 6 No 4 Aug-Sep 2014.
[3] C. Clavel, T. Ehrette and G. Richard, "Events Detection for an Audio-Based Surveillance System”, 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, 2005, pp. 1306-1309 (2005).
[4] Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. “Violence Content Classification Using Audio Features”, in Hellenic Conference on Artificial Intelligence, 2006, pp. 502–507.
[5] Theodoros Giannakopoulos, Aggelos Pikrakis and Sergios Theodoridis. “A Multi-Class Audio Classification Method With Respect To Violent Content In Movies Using Bayesian Networks”, IEEE 9th Workshop on Multimedia Signal Processing (2007).
[6] Esra Acar, Frank Hopfgartner, and Sahin Albayrak. “Detecting Violent Content in Hollywood Movies by Mid-level Audio Representations”, IEEE 11th International Workshop on Content-Based Multimedia Indexing (CBMI) (2013).
[7] Md. Zaigham Zaheer, Jin Young Kim, Hyoung-Gook Kim, Seung You Na “A Preliminary Study on Deep-Leaning Based Screaming Sound Detection”, International Conference on IT Convergence and Security (ICITCS) 978-1-4673-6537-6/15/$31.00 ©2015 IEEE.
[8] Vivek P, Kumar Rajamani, and Lajish V L. “Effective News Video Classification Based On Audio Content: A Multiple Instance Learning Approach”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (6), 2016.
[9] Marta Bautista Duran, Joaquin Garcia-Gomez, Roberto Gil-Pita, Inma Mohino-Herranz, and Manue lRosa-Zurera. “Energy Efficient Acoustic Violent Detector For Smart City”, International Journal of Computational Intelligence Systems, Vol. 10 (2017).
[10] S. Sarman and M. Sert, "Audio based violent scene classification using ensemble learning”, *2018* 6th International Symposium on Digital Forensic and Security (ISDFS), Antalya, 2018, pp. 1-5.
[11] Shifei Ding, Han Zhao, Yanan Zhang, Xinzheng Xu and Ru Nie. “Extreme learning machine: algorithm, theory and applications”, © Springer Science + Business Media Dordrecht 2013.
[12] Jing Yu & Wei Song & Guozhu Zhou & Jian-jun Hou. “Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation”, In Multimedia Tools and Applications volume 78, pages8497–8512 Springer 2019.
[13] ChenL-H, HsuH-W, WangL-Y, SuC-W (2011) “Violence detection in movies”, In: 2011 8th international conference on computer graphics, imaging and visualization (CGIV), pp 119–124 (2011).
[14] De Souza FDM, Chavez GC, do Valle Jr EA, de Araujo AA (2010) “Violence detection in video using spatio-temporal features”, In: Proceedings of the 2010 23rd SIBGRAPI conference on graphics, patterns and images. IEEE Computer Society, Washington, DC, pp 224–230
[15] Giannakopoulos T, Makris A, Kosmopoulos D, Perantonis S, Theodoridis S (2010) “Audio-visual fusion for detecting violent scenes in videos”, In: Konstantopoulos S et al (eds) Artificial intelligence: theories, models and applications, LNCS, vol 6040. Springer, pp 91–100 (2010).
[16] Gong Y, Wang W, Jiang S, Huang Q, Gao W (2008) “Detecting violent scenes in movies by auditory and visual cues”, In: Huang Y-M et al (eds) Advances in multimedia information processing - PCM 2008, LNCS, vol 5353. Springer, pp 317–326 (2008).
[17] M. Sj¨ oberg, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C.-H. Demarty. “The MediaEval 2014 affect task: Violent scenes detection”, In Working Notes Proceedings of the MediaEval 2014 Workshop, October 2014.
[18] ChenL-H,SuC-W,WengC-F, LiaoH-YM (2009) “Action scene detection with support vector machines”, J Multimed 4:248–253 (2009).
[19] Wang S, Jiang S, Huang Q, Gao W (2008) “Shot classification for action movies based on motion characteristics”, In: Proceedings of the IEEE international conference on image processing, pp 2508–2511 (2008).
[20] Extreme Learning Machines (ELM) Tutorial https://www.ntu.edu.sg/home/egbhuang/pdf/ELM-Tutorial.pdf.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY