ISSN:0975-9646

Effective News Video Classification Based On Audio Content: A Multiple Instance Learning Approach

Vivek P[#1], Kumar Rajamani[*2], Lajish V L[#3]

[#]*Department of Computer Science, University of Calicut, Kerala, India- 673 635*

[*]*Robert Bosch Engineering and Business Solutions, Bangalore, India -560 095*

Abstract— **This paper introduces a novel method for binary classification of news videos based on audio content using Multiple Instance Learning (MIL) approach. In this work violent incident videos are classified from news video archives using MIL methods. News audio sequences are segmented into instances and features have been extracted from each instances. Features of the instances of the same audio files are grouped together (called bag). Bags and instances are properly labelled and fed into MIL classifier for classification. A good performance of MIL methods on news video classification are observed and initial experimental results are found to be promising. The classification system provides over 90% accuracy witch indicate that MIL offers an effective paradigm for multimedia classification based on its audio modality.**

Keywords—**Video classification, Multiple Instance Learning (MIL), Feature extraction, mi-Graph, mi-SVM.**

I. INTRODUCTION

As videos form a major portion of information disseminated in the world every day, research has begun on automatically classifying it. Because of the huge amount of video to categorize, automatic classification of video feeds has been gaining importance recently, especially with the wide acceptance of information and communication technology (ICT) and systems such as the Carnegie Mellon University (CMU) Informedia system [1]. At this time people have access to a tremendous amount of news video, both on television and the Internet. The amount of video that a viewer has to cull from is now so immense that it is infeasible for a human to go through it and distinguish violent incident news videos among immensely colossal volume of broadcast feeds.

For the purpose of video classification, features are drawn mainly from text, audio, and visual modalities. Usually multimedia approaches are found scarcely more often in the literature than text-only approaches. The audio content based approach usually require fewer computational resources than visual methods [2]. There are different features which provides a compact representation of the given audio signal. Among them Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear prediction (PLP) coefficients are widely accepted and used features [3].

Different supervised and unsupervised learning algorithms are available in machine learning approaches. Multiple Instance Learning (MIL) is proposed as a variation of supervised learning for problems with incomplete knowledge about labels of training examples. Melih Kandeir et. al [24] conducted a benchmark study over the performance of different MIL methods. In their study they found that mi-Graph and mi-SVM gives considerably better result compared to other MIL methods.

In this work, we propose a novel approach for audio content based video classification using MIL methods. The results obtained using these methods are evaluated using different performance metrics. The rest of this paper is organized as follows. Section II reviews the literature on video classification and Multiple Instance Learning. Section III describes the proposed classification methodology. Section IV describes the experimental results and section V concludes the work.

II. REVIEW ON VIDEO CLASSIFICATION AND MULTIPLE INSTANCE LEARNING

Most of the research on video classification has the intent of classifying an entire video, some others have focused on classifying segments of video such as identifying violent [4] or scary [5] scenes in a movie or distinguishing between different news segments within an entire news broadcast [6]. Video classification experiments often try to classify video into different broad categories like politics, entertainment, cultural etc. but some other works focused their efforts on more specific tasks such as identifying sports videos [7], informational videos [8] etc. among all video.

In video classification, features are drawn mainly from three modalities: visual, text and audio. For video classification purpose, we found many of the standard classifiers, such as Bayesian, Support Vector Machines (SVM), Artificial Neural Network (ANN), Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) [9]. The TREC Video Retrieval Evaluation (TRECVid) was an international benchmarking activity which encouraged research in video classification, indexing and retrieval [10]. Liu et al. [18] considered the problem of discriminating five types of videos namely commercial, basketball game, football game, news report, and weather forecast. They have designed an ergodic HMM using the clip based features as observation vectors.

Video classification literature is the common area where audio based approaches are found rather than in texts and video approach. Advantages of audio approaches are that they typically require fewer computational resources than visual methods and more dependable than text. In the earlier works time domain features are used widely [11-15] and later, researches are started using combined features

from both time and frequency domains for better recognition accuracy [14-16]. Among these features MFCC is identified as the most used and trustable one [17, 18].

The use of weakly supervised machine learning technique can reduce the computational cost in a large manner. So far no work has been reported on the use of Multiple Instance Learning (MIL) approach on news video classification. The MIL problem was first formulated for the task of digit recognition [19], in which a neural network was trained with the information of whether a given digit was present, but not where it was present. Another early application of MIL was to the problem of drug discovery [20], in which the bags were molecules of the drug and the instances were conformations of those molecules. MIL has also been applied to object detection in images [21], video classification to match names and faces [22], and to text classification [23], in which the bags were documents and the instances were sentences or paragraphs. Many approaches have been introduced for MIL, including mi-Graph, Gaussian process multiple instance learning (GPMIL), MILBoost, mi-SVM and Bag key instance SVM (B-KI-SVM) [24].

III. Audio Content Based Video Classification Using MIL

The goal of this work is to enable automatic learning of news categories. Given a list of news videos of interest, the proposed method will produce discriminative model to distinguish them.

In the first section, MIL and its procedures has been discussed, followed by sub-section on feature extraction, where different feature extraction techniques used in this work are discussed. The last sub-section explains the MIL classifiers, mi-Graph and mi-SVM which have been used for classification.

In MIL, extracted news audio is split into 25ms length overlapped segments. Instances are created from each audio segments by extracting features from it. All the feature sets (group of instances) belonging to the same news files are grouped into a bag. Labels are assigned for instances and for the bags as a whole, assuming the bag label to be the maximum of the instance labels within the bag. Finally, these bags along with their labels are fed into a classifier for classification. A bag with a positive label indicates that there is at least one positively labelled instance and for a negatively labelled bag, all instances are known to have negative labels. Thus as shown in figure 1 violent incident news videos are represented by positive bag and other news sets by negative bag. The schematic diagram of the proposed work is shown in figure 2. The following sessions describe the methods used for extraction of audio features and MIL based classification.

A. Extraction of audio features

To classify the videos, we extracted audio part from it and split it into overlapping audio segments. The signals are segmented into constant-time segmentation, like 25ms blocks [25]. Speech signal analysis is generally performed over short-time frames with a Fixed Frame Length (FFL) and a fixed frame rate (FFR), based on the assumption that these signals are non-stationary and, exhibit quasi-stationary behaviour in short durations [26]. This method benefits from simplicity of implementation and the ease of comparing blocks of the same length. Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear prediction (PLP) coefficients are extracted as features and further used for classification purpose. The algorithms used for MFCC and PLP based feature extraction techniques are described below.

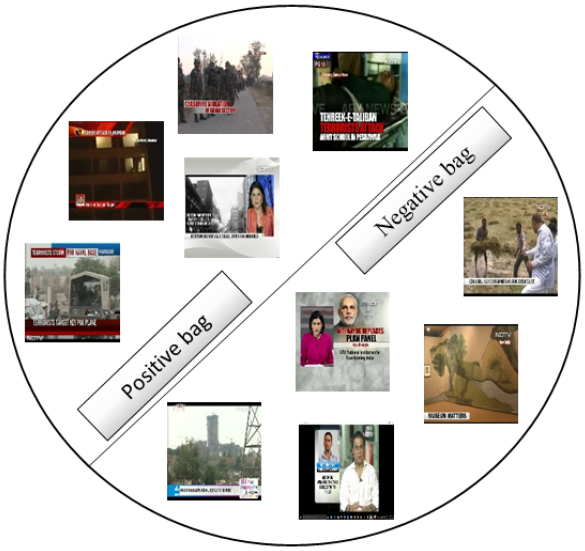

Fig. 1 MIL for news video classification

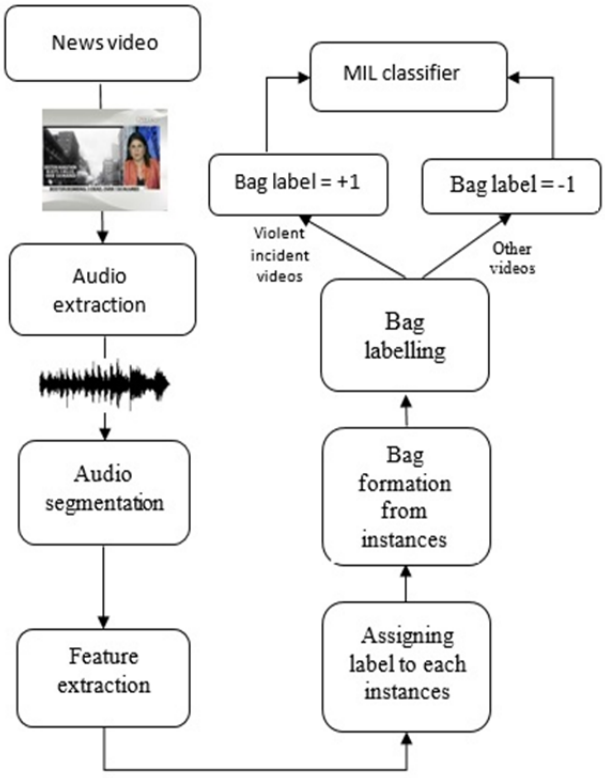

Fig. 2 Schematic diagram of proposed methodology.

1) *Mel Frequency Cepstral Coefficients (MFCC):* The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC. MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for audio recognition [27]. The procedure to determine MFCC is described in the following algorithm.

Algorithm: MFCC feature extraction

Step 1: Segmentation of voiced speech signal into 25 ms-length frames.

Step 2: Calculate the periodogram estimate of the power spectrum for each frame.

Step 3: Apply the mel filterbank to the power spectra, sum the energy in each filter.

Step 4: Take the logarithm of all filterbank energies.

Step 5: Take the DCT of the log filterbank energies.

Step 6: MFCCs are the amplitudes of the resulting spectrum.

The mel-scale frequency mapping is formulated as:

$$m(f) = 1125\ (1 + \frac{f}{700}) \quad (1)$$

2) Perceptual Linear prediction (PLP): The Perceptual Linear Prediction (PLP) model is developed by Hermansky [28]. PLP models the human speech based on the concept of psychophysics of hearing [29]. PLP discards irrelevant information of the speech and thus improves speech recognition rate. The procedure to determine PLP coefficients are described as follows:

Algorithm: PLP feature extraction

Step 1: The N- point DFT is applied on the segmented input signal *x(n).*

Step 2: The critical-band power spectrum is computed through discrete convolution of the power spectrum with the piece-wise approximation of the critical-band curve.

Step 3: Equal loudness pre-emphasis is applied on the down-sampled *θ(B)* and then intensity-loudness compression is performed.

Step 4: Inverse DFT is performed for getting the equivalent autocorrelation function.

Step 5: PLP coefficients are computed after autoregressive modelling and conversion of the autoregressive coefficients to cepstral coefficients.

B. MIL based Classification

No more than 3 levels of headings should be used. All headings must be in 10pt font. Every word in a heading must be capitalized except for short minor words as listed in Section III-B. Multiple Instance Learning (MIL) is a variation of supervised learning for problems with incomplete knowledge about labels of training examples. In MIL, the labels are assigned to bags of instances. The binary classifier labels a bag positive if no less than one instance in that bag is positive, otherwise bag is labelled as negative [30]. That is the MIL training set consists of bags $\{X_1, X_2 \ldots, X_n\}$ and bag labels $\{y_1, y_2 \ldots, y_n\}$, where $X_i = \{x_{i1}, x_{i2} \ldots, x_{im}\}$, $x_{ij} \in X$ and $y_i \in \{-1, 1\}$. The goal of MIL is to either train an instance classifier h(X): $X \rightarrow Y$ or a bag classifier H(X): $X^m \rightarrow Y$.

The brief description of two MIL based classification methods, mi-Graph and mi-SVM, used in our study are given below.

1) *mi-Graph:* mi-Graph is a simple but effective method represents each bag by a similarity graph [31]. First, the cross-similarities of bag instances are calculated by an instance-level kernel function $k_{inst}(xi,xj)$. A graph is then constructed by placing a node per each instance within a bag and each node pair is connected by an edge if the two corresponding instances are more similar to each other than a threshold δ. Let W_b be the affinity matrix of bag b, whose entry is $w^b_{nm} = 1$, if there is an edge between the nodes of instances n and m, and $w^b_{nm} = 0$ otherwise. Consequently, similarity between bags b and c are computed by the following kernel function:

$$K_{bag}(X_b, X_c) = \frac{\sum_{n=1}^{N_b}\sum_{m=1}^{N_c} v_{bn} v_{cm} k_{inst}(X_{bn}, X_{cm})}{\sum_{n=1}^{N_b}\sum_{m=1}^{N_c} v_{cm}} \quad (2)$$

Where, $v_{bn} = 1/\sum_{u=1}^{N_b} W^b_{nu}$, $v_{bn} = 1/\sum_{u=1}^{N_c} W^c_{mu}$ are the sum of the weights of the edges incident to nodes (instances) n and m of bags b and c, respectively. Based on the resultant bag-level Gram matrix, the arbitrary kernel learner is trained. The intuition behind this kernel is that for instances that are similar to a large number of other instances within the bag, W_{ia} has a smaller value, and for instances different from the rest of the bag, W_{ia} is large. Hence, the influence of odd instances within bags are enhanced, and others are down weighted.

2) *mi-SVM:* This method approaches MIL as a semi-supervised learning problem, treating the labels of positive bag instances as latent variables [32]. These latent variables are added to the optimization problem and inferred from data.

$$\begin{aligned} &\min_y \min_{w,b,\xi} \frac{1}{2}\|W^2\| + C\sum_{i=1}^{N}\xi_b, \quad (3)\\ &\text{s.t } y_i\left(W^T\phi(X_i)\right) \geq 1 - \xi_i, \forall i,\\ &\substack{\xi_i \geq 0, \forall i,\\ \max(y_b)} = Y_b, \forall b. \end{aligned}$$

where w is the vector of model parameters defining the planar decision boundary, C is the regularization constant, ξ_b are slack variables, and $\phi(.)$ is a function that maps an instance from the original feature space to a Reproducing Kernel Hilbert Space (RKHS) [33]. At each iteration, the approximate solution can be found as follows: trains an instance-level standard SVM based on the current assignments of the latent variables, then update these variables by making predictions with the learned SVM.

IV. EXPERIMENTAL RESULTS

We evaluated MIL on news video dataset. Above discussed MFCC and PLP features and two MIL techniques have been used in the experiment part. The details of the dataset and performance evaluations are discussed in this section.

A. News video dataset and pre-processing

120 English news recordings have been gathered from news chronicles for the implementation of the proposed method. Among them 60 news videos have been found violent incident news videos, which mainly reports about emergency situations like terror attack, bomb blast, murder etc. and the remaining entries are randomly selected from different categories. In all news videos the reporter narrates the news briefly for the first 10-15 seconds and goes for the detailed reporting. In the proposed experiment, we extracted narration audio portion, and the audio signals are segmented into 25ms frames.

B. News video classification using MIL techniques

Frames are considered as the instances of the audio signal. MFCC and PLP features have been extracted from each frame. We report the following four performance metrics for video classification evaluation.

Accuracy : Percentage of correctly classified test points.
F1 score : Harmonic mean of precision and recall.
AUC-ROC : Area under Receiver Operating Characteristics (ROC) curve.
AUC-PR : Area under precision–recall curve.

We have conducted video classification experiments using MFCC and PLP features separately based on MIL techniques and the detailed experiment results are listed in Table 1.

TABLE I
VIDEO CLASSIFICATION RESULTS AND PERFORMANCE MATRICES

MIL method	Feature	Accuracy (%)	F1 score	AUC-ROC	AUC-PR
mi-Graph	MFCC	90.0	0.91	0.89	0.81
	PLP	85.0	0.86	0.94	0.96
mi-svm	MFCC	80.3	0.86	0.89	0.93
	PLP	79.4	0.87	0.88	0.94

In this work, we have performed news video classification experiment based on audio content detection using two different MIL methods. There were two types of features used for audio content detection, namely MFCC and PLP. Two MIL techniques mi-Graph and mi-svm are used for classification purpose. From the experiment result it is evident that the MIL works effectively on the audio classification. It is also noted that mi-Graph with MFCC feature gives slightly better result compared to others.

V. CONCLUSIONS

In this study, we tried to classify violent incident news videos from large news video archive. mi-Graph and mi-SVM techniques are used for this study as mi-Graph directly models In this study, we tried to classify violent incident news videos from large news video archive. mi-Graph and mi-SVM techniques are used for this study as mi-Graph directly models within bag instance relationships and mi-SVM is semi-supervised in its nature. mi-Graph using MFCC features appears as the best-performing method with 90.0% classification accuracy and 0.91 F1 score. Other results are also in a considerable range. The good performance of the selected features and MIL methods motivates for further research in this direction. A detailed analysis of other MIL methods can also be taken as a future work. The outcome of the work can also be used to classify different categories of video from large data archive.

REFERENCES

[1] Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., and Wactlar, H., "Informedia, "Digital Video Library", *Communications of the ACM*, vol. 38(4), pp. 57-58, 1995.

[2] Brezeale, D., & Cook, D. J., "Automatic video classification: A survey of the literature". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38(3), pp. 416-430, 2008.

[3] Mporas, Iosif, Todor Ganchev, Mihalis Siafarikas, and Nikos Fakotakis. "Comparison of speech features on the speech recognition task." *Journal of Computer Science*, vol. 3(8), pp. 608-616, 2007.

[4] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in *Proc. International Conference on Image Processing (ICIP '98)*, 1998, vol. 1, pp. 353–357.

[5] Moncrieff, Simon, Svetha Venkatesh, and Chitra Dorai. "Horror film genre typing and scene labeling via audio analysis.", in *Proc. Multimedia and Expo, 2003 IEEE. ICME'03*, 2003, vol. 2, pp. II-193.

[6] W. Zhu, C. Toklu, and S.-P. Liou, "Automatic news video segmentation and categorization based on closed-captioned text," in *Proc. IEEE International Conference on Multimedia and Expo (ICME 2001)*, 2001, pp. 829–832.

[7] V. Kobla, D. DeMenthon, and D. Doermann, "Identifying sports videos using replay, text, and camera motion features," in Proc. *SPIE conference on Storage and Retrieval for Media Databases*, 2000.

[8] J. Fan, H. Luo, J. Xiao, and L. Wu, "Semantic video classification and feature subset selection under context and concept uncertainty," in *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, 2004, pp. 192–201.

[9] Brezeale, Darin, and Diane J. Cook. "Automatic video classification: A survey of the literature." *in Proc. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 2008, vol. 38(3) pp. 416-430.

[10] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in MIR '06: in *Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA:ACM Press, 2006, pp. 321–330.

[11] P. Q. Dinh, C. Dorai, and S. Venkatesh, "Video genre categorization using audio wavelet coefficients," in *Proc. Fifth Asian Conference on Computer Vision*, 2002.

[12] J. Fan, H. Luo, J. Xiao, and L. Wu, "Semantic video classification and feature subset selection under context and concept uncertainty," in *Proc. of the 4th ACM/IEEE-CS joint conference on Digital libraries*, 2004, pp. 192–201.

[13] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *Proc. Proceedings of the third ACM international conference on Multimedia*, 1995, pp. 295–304.

[14] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong, "Integration of multimodal features for video scene classification based on HMM," in *Proc. Third IEEE Workshop on Multimedia Signal Processing, 1999*, pp.53–58.

[15] R. S. Jasinschi and J. Louie, "Automatic tv program genre classification based on audio patterns," in *Proc. IEEE 27th Euromicro Conference*, 2001, pp. 370–375.

[16] M. Roach, J. Mason, and L.-Q. Xu, "Video genre verification using both acoustic and visual modes," in *Proc. International Workshop on Multimedia Signal Processing*, 2002.

[17] R. S. Jasinschi and J. Louie, "Automatic tv program genre classification based on audio patterns," in *Proc. of IEEE 27th Euromicro Conference*, 2001, pp. 370–375.

[18] Z. Liu, J. Huang, and Y. Wang, "Classification of TV programs based on audio information using hidden markov model," in *Proc. IEEE Signal Processing Society Workshop on Multimedia Signal Processing*, 1998,pp. 27–32.

[19] J. D. Keeler, D. E. Rumelhart, and W. K. Leow. *Integrated segmentation and recognition of hand-printed numerals*. In Advances in Neural Information Processing Systems 3, pages 557–563, San Francisco, CA, Morgan Kaufmann Publishers, Inc. 1990.

[20] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. "Solving the multiple instance problem with axis-parallel rectangles". *Artificial Intelligence*, vol. 89(1-2), pp. 31–71, January 1997.

[21] Y. Chen and J. Z. Wang. "Image categorization by learning and reasoning with regions". *Journal of Machine Learning Research*, 5:913–939, 2004.

[22] J. Yang, R. Yan, and A. G. Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *Proc. ACM Intl. Conf. on Multimedia*, 2005, pages 31–40, New York, NY, USA.

[23] S. Andrews, I. Tsochantaridis, and T. Hofmann. "Support vector machines for multiple-instance learning". In *Advances in Neural Information Processing Systems* 15, pages 561–568. MIT Press, Cambridge, MA, 2003.

[24] Kandemir, Melih, and Fred A. Hamprecht. "Computer-aided diagnosis from weak supervision: A benchmarking study." *Computerized Medical Imaging and Graphics,* 2015, vol. 42(1), pp. 44-50.

[25] S. Young. Large vocabulary continuous speech recognition: a review. IEEE Signal Processing Magazine, 13(5):45–57, 1996.

[26] Tan, Zheng-Hua, and Ivan Kraljevski. "Joint variable frame rate and length analysis for speech recognition under adverse conditions." *Computers & Electrical Engineering,* 2014, vol. 40(7) pp. 2139-2149.

[27] Vergin, Rivarol, Douglas O'shaughnessy, and Azarshid Farhat. "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition." *IEEE Trans. on Speech and Audio Processing*, 1999, vol. 7(5), pp. 525-532.

[28] Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." *Journal of the Acoustical Society of America* 87.4 (1990): 1738-1752.

[29] Mporas, Iosif, et al. "Comparison of speech features on the speech recognition task." *Journal of Computer Science,* 2007, vol. 3(8), pp. 608-616.

[30] Yang, Jun., Review of multi-instance learning and its applications, *Tech. Rep*, 2005.

[31] Zhou ZH, Sun YY, Li YF. Multi-instance learning by treating instances as non-iidsamples. In *Proc. ICML.* 2009. p. 1249–56.

[32] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning. In *Advances in NIPS,* 2003.

[33] Schölkopf, Bernhard, and Alexander J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.