PSYCHOLOGY AND EDUCATION (2019) 56(1): 28-32
ISSN: 1553-6939

Multiple Instance Learning for Automatic Content-Based Classification of Speech Audio

B. Bhaskar Reddy[1], P. Imran Khan[2], Dr. B. Dhananjaya[3]
Associate Professor[1,3], Assistant Professor[2]
Department of Electronics and Communication Engineering
Bheema Institute of Technology and Sciences, Adoni-518301.[1,2,3]

Abstract*: Speech analytics researchers are working to improve their ability to decipher audio material. This research presents a new method for classifying news audio clips based on their content, called the Multiple Instance Learning (MIL) approach. Audio classification and segmentation benefit from content-based analysis. As a starting point, a classifier that can predict the category of an audio sample has been proposed. Perceptual Linear Prediction (PLP) coefficients and Mel-Frequency Cepstral Coefficients (MFCC) are two kinds of features used for audio content identification (MFCC). For classification, two MIL approaches, mi-Graph and mi-SVM, are used. Different performance matrices are used to assess the outcomes gained via the use of various approaches. The results of the experiments clearly show that the MIL has great audio categorization capacity.

***Keywords* :Audio classification, Multiple Instance Learning(MIL); Featureextraction; mi-Graph;mi-SVM.**

I. INTRODUCTION

Because audio constitutes a significant amount of the information distributed in the globe on a daily basis, several scholars are trying to categorise it using different criteria [1]. People in today's digital environment have easy access to a wealth of news audio and video through radio, television, and the internet. The quantity of multimedia data accessible is now so vast that it is impossible for a person to go through it all and identify essential files among them. Automatic content-based analysis provides relevant data for audio classification and segmentation. The audio content information may be utilised to categorise the file. As a result, audio content comprehension is a current study topic in speech analytics. A classifier that can predict the category of the input audio is an important step in this approach. A unique audio categorization approach based on Multiple Instance Learning is presented in this research (MIL).

In Machine Learning (ML), there are a variety of supervised and unsupervised learning methods. Multiple

MIL, a variation on supervised approaches for dealing with patrial information on the labels of training samples, has been presented. MIL techniques have been compared in terms of performance by MelihKandeir and his colleagues [2]. They found that mi-Graph and mi-SVM performed much better than other MIL approaches in their investigation.

Text, audio, and visual modalities are the primary sources of characteristics used in multimedia categorization. As a rule, multimedia techniques are more common than text-only approaches in academic publishing. When compared to visual solutions, audio content-based approaches often use less computer resources [3]. A compact representation of an audio stream may be achieved via a variety of aspects. [4] The PLP coefficients and MFCC are two of the most often utilised features [4] in the field of perceptual linear prediction (PLP). Vivek P et al. suggested a news video classification approach that uses the MIL algorithm to classify violent event footage from news video archives [5]. This paper proposes a new technique for classifying speech audio based on content utilising MIL approaches. To perform the trials, researchers used a custom-built audio database of recent news broadcasts. Different performance criteria are used to assess the quality of the data generated by various approaches.

II. REVIEWONAUDIOCLASSIFICATIONUSING MULTIPLEINSTANCELEARNING

The field of automatic audio analysis and categorization is one that is only now gaining traction in the world of multimedia. One of the first engines for content-based categorization, search and retrieval of audio data was developed by ErlingWold et al in 1996. Using the new approach, sounds can be categorised based on the type of audio they contain. Any acoustical characteristic or a combination of features might be the basis for a query. One way to do the experiment is to use the engine to get sounds that are similar to (or different from) previously taught classes depending on these qualities. Using acoustic similarities, Jonathan T. Foote presented a technique for retrieving audio documents in 1997 [7]. A supervised vector quantizer is utilised to produce the statistical similarity measure in this study.

Simple classification situations like speech-music classification or speech-silent classification were used in the beginning. Autonomous audio content analysis based on perceptual characteristics was suggested by Pfeiffer et al. [8]. Audio recordings were categorised as speech, quiet, laughing, and non-speech noises by D. Kimber et al. to separate conversation records from extraneous sounds during meetings [9].

Zhang and Kuo developed a heuristic-based methodology to categorise audio recordings into several categories, such as songs and talks over music [10].

In video classification literature, audio-based techniques are likewise more prevalent than text and video approaches. As opposed to visual techniques, audio approaches use less processing resources and are more trustworthy than text. Early studies relied on time-domain characteristics, but subsequent studies incorporated features from the time- and frequency-domains to improve identification accuracy [11, 12]. It has been determined that MFCC is the most widely utilised and reliable of these options [13]. In 1998, Liu et al. investigated the difficulty of differentiating five categories of news: commercials, basketball games, football games, news reports, and weather forecasts [14]. As observation vectors, they've created an ergodic HMM based on the clip-based features. As of 2017, Visser et al. [15] have developed an audio filter predictor for event categorization and extraction. From unlabeled training data, deep neural networks (DNN) are commonly utilised to extract the necessary target audio recordings.

Initial MIL research began with the digit recognition challenge, where a neural network was trained with just the existence of a particular digit without defining its location. It was also used to identify a medicine, in which the bags were drug molecules and examples were their conformations. [17] Text classification [20], where documents are treated as bags and sentences as instances, is another use of MIL, as are the detection of objects in photos [18], the classification of videos to match names and faces [19], and the classification of photographs [20]. Multi-instance learning has seen a slew of new methodologies published in recent years, including the mi-Graph and the Gaussian Process Multiple Instance Learning (GPMIL) algorithms. Using a poorly supervised machine learning approach may have a significant impact on computing costs. Speech audio categorization using the Multiple Instance Learning (MIL) technique has remained mostly unexplored till now.

III. FEATUREEXTRACTIONFROMNEWSAUDIOSFORCLASSIFICATION

Extraction of audio features by the MIL classifier is described in detail in this section. Prior to processing, audio is extracted from the news stream and chunked together. Signals are divided into 25ms segments using constant-time blocks [22]. Speech, according to a standard speech signal analysis, is non-stationary and only exhibits quasi-stationary behaviour for brief periods of time. Using a fixed frame length (FFL) and frame rate (FFR), it's normally done in brief periods of time. (FFR). This method is simple to develop since blocks of the same length may be compared. MFCC and PLP coefficients are extracted as features and utilised for classification purposes. MFCC and PLP coefficients are used as features. The algorithms for MFCC and PLP feature extraction are outlined here.

Mel Frequency Cepstral Coefficient (MFCC) Feature Mel There are several prominent spectral features used in speech recognition, including Frequency Cepstral Coefficients (MFCC). MFCC takes into account the sensitivity of the human ear to frequencies in order to improve audio recognition. The following method outlines the steps necessary to calculate the MFCC.

Algorithm

Step1:Segmentationofvoicedspeechsignalinto25ms-lengthframes.

Step2:Calculatetheperiodogramestimateofthepowerspectrumfor eachframe.

Step3:Apply themelfilterbanktothepowerspectraandtakethesumofenergyineachfilter.

Step4:Calculatethelogarithmvalueforallfilterbankenergies.

Step5:FindtheDCT ofthefilterbankenergies(log).

Step6:FindtheMFCCastheamplitudesofthesubsequentspectrum.

Themel-scalefrequencymappingisformulatedas:

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

A. PerceptualLinearPrediction(PLP)Feature

ThePLPmodel,proposedbyHermansky[25],isbased ontheconceptofpsychophysicsofhumanhearing.PLPthrow-outs irrelevant information in the speech and makessignificantimprovementinspeechrecognitionrate.Theprocedure to determine PLP coefficients are described asfollows:

Algorithm

Step 1: The segmented input signal x is subjected to the N-point DFT (n).

Using the piece-wise approximation of the critical-band curve, the power spectrum is convolved with the critical-band power spectrum in Step 2.

This is the third and last step in the process of equalising the down-sampled (B) before performing intensity-loudness compression.

For the analogous autocorrelation function, an inverse DFT is used in Step 4.

For Step 5: Autoregressive modelling, followed by cepstral coefficients to calculate the PLP coefficients for autoregressive models.

IV. CONTENT BASED AUDIOCLASSIFICATION USING MIL

This project seeks to automatically classify news audios depending on the content. Analytical applications will have lower search costs as a result of this classification. The suggested technique will generate a discriminative model to discriminate between a list of relevant news audios.

The implementation specifics of the MIL technique, as well as an explanation of the mi-Graph and mi-SVM algorithms used for classification, are covered in depth in the following sections.

A. Preparation of Malayalam News Audio Corpus(MNAC)

In the beginning, the news text corpus is produced from the online news portals of the most prominent Malayalam daily.

These news text sentences are then divided into five categories: state news, national news, international news, sports news, and news with cultural significance.. State news, national news, and international news all fall under the first three categories; however, sports and culture are not included in this list. Five thousand five hundred twenty-five For the construction of the dataset, sentences belonging to a variety of different categories are gathered.

35 speakers (female and male) of various ages each contributed news audio samples for the creation of a Malayalam News Audio Corpus (MNAC). There were 150 sentences from each of the five news categories (MNSTC) selected at random for each speaker to utter. We've assigned a unique identifier to each one of these 5,250 spoken utterances so that we know which category they belong in: "News." There are 5.35 seconds of news audio clips included in the dataset.

B. MILforNewsAudio Classification

For feature extraction, the input news audio files from the MNAC audio corpus are first separated into 25 ms-long overlapping chunks. Extracting features from each audio segment is used to generate a new instance. A single bag is used to organise a collection of features (also known as instances) culled from the same set of news files. Bags and cases have been tagged correctly. As a result, the bag label is larger than all of the other labels in the bag. Then, the MIL classifier receives these bags and their matching labels. To name a bag positive, there must be at least one positively labelled instance inside it; to label it negative, all of the examples must be negatively labelled. As can be seen in Figure 1, the positive bag represents newsgroups of interest, whereas the negative bag depicts uninterested newsgroups. Figure 2 depicts a schematic representation for the audio categorization approach based on MIL.

MIL differs from supervised learning in that it often addresses issues involving only partial knowledge of the labels of training samples. In the case of MIL, a bag is labelled as negative if it contains only negative instances, and as positive if it contains only positive examples.. That is the MILtraining set consists of bags $\{X_1, X_2, \ldots X_n\}$and bag labels $\{y_1, y_2, \cdots y_n\}$,where$X_i=\{x_{i1},x_{i2}...,x_{im}\}$,$x_{ij}\in Xandy_i\in\{-1,1\}$.ThegoalofMIListoeithertrainaninstanceclassifierh(X):X$\rightarrow$YorabagclassifierH(X):$X^m\rightarrow Y$.

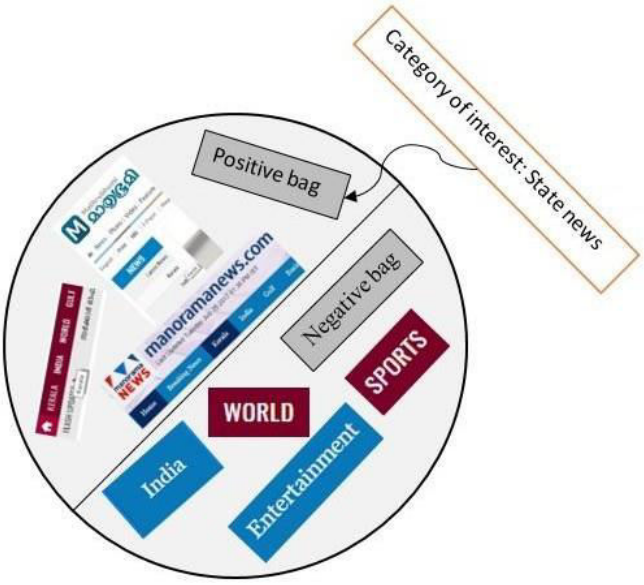


Fig. 1.MIL approach for news audio classificationconsideringstatenewsasthearea ofinterest

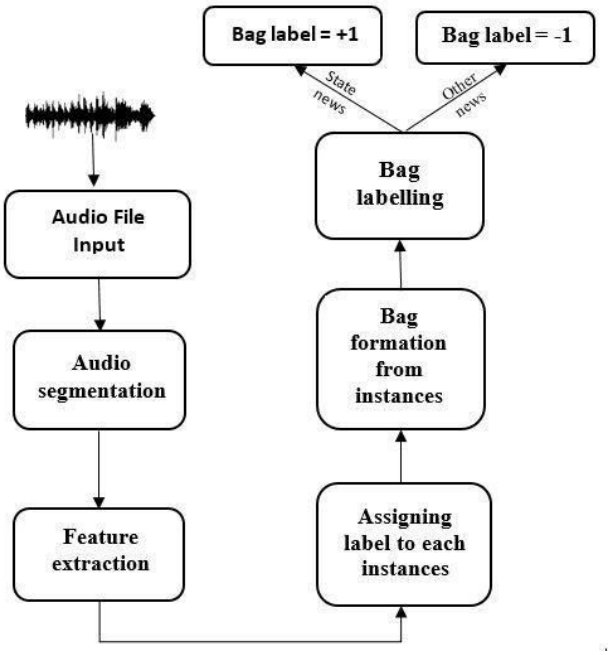


Fig.2.Schematicdiagramofproposednewsaudiocla ssificationmethodology

The next subsections provide an overview of the mi-Graph and mi-SVM classification algorithms, both of which are based on MIL.

1) mi-Graph: The mi-Graph MIL approach uses a similarity graph to describe each bag, making it both efficient and easy [26]. Kernel functions are used to calculate the cross-similarity of bag instances in this technique..

Eachinstancesarerepresentedasnodesandnodepairsareconnectedonlyiftheirexistasimilaritybetweenthemoverathresholdδ.Let$W_b$betheaffinitymatrixofbag$b$,whoseentryisw$^b_{nm}$=1,ifthereisanedgebetweenthenodesofinstances$n$and$m$,andw$^b_{nm}$=0otherwise.Accordingly,similarityamongtwobags$b$and$c$arecalculatedbythefollowingkernel 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, $v_{bn} = 1/\sum_{u=1}^{N_b} W_{nu}^{b}$, $v_{bn} = 1/\sum_{u=1}^{N_c} W_{mu}^{c}$ add up the weights at nodes n and m of the bags in bag b and bag c, respectively, to arrive at the total weight. The bag-level Gram matrix is then computed after training an arbitrary kernel learner. Bags with more identical instances have a lower value, but bags with more unique instances have a higher value. As a result, unusual events in bags have a greater impact, while others are diminished..

1) mi-SVM: Consider the positive bag instances as latent variables in this semi-supervised learning problem [27]. The optimization problem is fed with these latent variables, which are then inferred from the data.

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

where$w$isthevectorofmodelparameters,Cistheregularization constant, $\xi_b$are slack variables, and $\phi(.)$There are latent variables in this semi-supervised learning problem, which are the positive bag instances. Latent variables, which are subsequently derived from the data, are provided into the optimization problem as input.

V. SIMULATIONEXPERIMENTSANDRESULTS

The assessment of the proposed MIL based audio categorization is done using MNAC news audio archive. The MFCC and PLP characteristics and two MIL approaches viz. mi-Graph and mi-SVM have been employed for the tests. After conducting keyword spotting studies, the experiment is carried out on the audio samples that were generated as a consequence. The assessment of the suggested technique is undertaken by treating the news audio samples included in the dataset as two types viz. state news audio and non-state news audio. Non-state news, on the other hand, may be further classified into several binary classifications, such as national and non-national, sports and non-sports, and news with cultural and non-cultural significance. Fig. 3 shows a block diagram of the proposed MIL-based news audio categorization assessment model. The MIL classifier takes as input the audio recordings that have been keyword detected. Negative and positive bags are created by the MIL classifier for each audio file. The material is classified as state news if it is categorised as positive. The specifics of the performance assessments are described in this section.

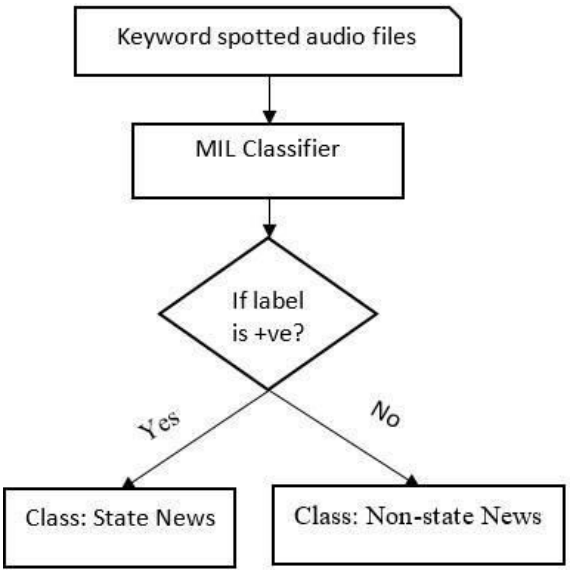


Fig.3.EvaluationmodelfortheMILbasednewsaudio classifier

The audio signals are divided into 25 ms frames as the initial step in the processing process. The audio signal is represented as a series of frames, each of which represents an individual occurrence. Each frame has been analysed for MFCC and PLP characteristics. The proposed MIL classifier's audio classification performance is measured using the four performance measures listed below. information

Accuracy:measurement(%)ofhowclosearesultcomesto the true value.

F1score: Functionofprecisionandrecall.

AUC-ROC: Area under Receiver OperatingCharacteristics(ROC) curve.

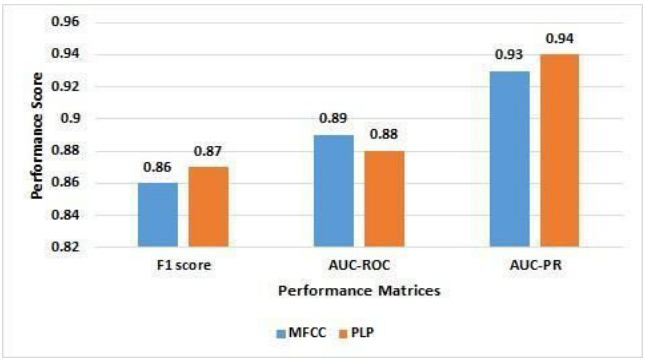
AUC-PR: Areaunderprecision–recallcurve.

MFCC and PLP features based on two distinct MIL approaches, viz. mi-Graph and mi-SVM, are used in the news audio classification tests. Table I shows the results of the audio categorization using performance matrices derived by using state news as positive bags.

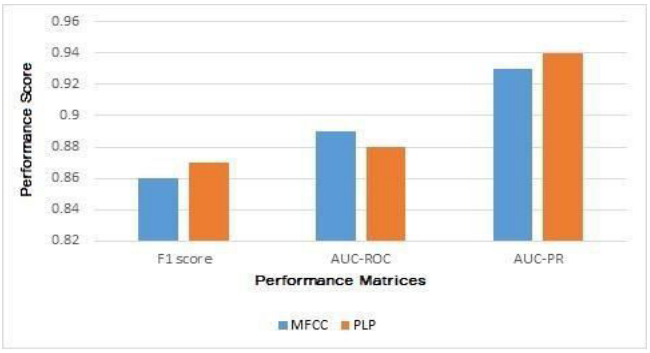
Table- I: MIL based news audio classification resultsandperformancematrices

MIL method	Feature	Accuracy (%)	F1 score	AUC-ROC	AUC-PR
mi-Graph	MFCC	95.8	0.96	0.98	0.99
	PLP	93.2	0.93	0.97	0.97
mi-svm	MFCC	85.0	0.86	0.94	0.96
	PLP	80.3	0.86	0.89	0.93

The results of the experiments show that the MIL classification approach is successful in classifying speech recordings. Furthermore, it is clear that mi-Graph with the MFCC feature outperforms other approaches. Mi-graph and mi-SVM based audio categorization performance scores are shown graphically in Figure



a. mi-Graph



b. mi-SVM

Fig.4.Performancescoresfor(a)mi-Graph(b)mi-SVMbasednewsaudio classification

VI. CONCLUSION

The MIL methodology is used to develop a new way of content-based audio categorization. Mi-Graph and mi-SVM algorithms are used to classify the news audio recordings extracted from the indigenous MNAC audio collection. mi-Graph represents the link between the bag and the instances directly, whereas mi-SVM is semi-supervised. MFCC and PLP characteristics are used in the news audio classification tests. The suggested mi-Graph and mi-SVM algorithms are also evaluated utilising MFCC and PLP parameters. When compared to other methods of audio classification, mi-Graph employing MFCC features comes out on top, with an F1 score of 0.96 and a classification accuracy of 95.8%. A wide range of audio, multimedia, and speech analytics applications might benefit from the capacity of the MIL-based audio class
ifier to categorise and retrieve audio samples according to their content.

REFERENCES

1. Christel, Michael, Scott Stevens, and Howard Wactlar. "Informediadigital video library." In Proceedings of the second ACM internationalconferenceonMultimedia,pp.480-481.ACM,1994.
2. Kandemir, Melih, and Fred A. Hamprecht. "Computer-aided diagnosisfromweaksupervision:Abenchmarkingstudy."ComputerizedMedicalImagingandGraphics 42(2015):44-50.
3. Kandemir, Melih, and Fred A. Hamprecht. "Computer-aided diagnosisfromweaksupervision:Abenchmarkingstudy."ComputerizedMedicalImagingandGraphics 42(2015):44-50.
4. Mporas,Iosif,TodorGanchev,MihalisSiafarikas,andNikosFakotakis."Comparisonofspeechfeaturesonthespeechrecognitiontask."JournalofComputerScience3,no.8(2007):608-616.
5. VivekP,KumarRajamani,LajishVL,"EffectiveNewsVideoClassification Based On Audio Content: A Multiple Instance LearningApproach", International Journal of Computer Science and InformationTechnologies(IJCSIT),Vol.7(6),page:2556-2560,ISSN:0975-9646,2016
6. Wold,Erling,ThomBlum,DouglasKeislar,andJamesWheaten."Content-basedclassification,search, and retrieval of audio." IEEEmultimedia3,no.3(1996):27-36.
7. Foote, Jonathan T. "Content-based retrieval of music and audio." InMultimedia Storage and Archiving Systems II, vol. 3229, pp. 138-148.InternationalSocietyforOptics and Photonics,1997.
8. Pfeiffer, Silvia, Stephan Fischer, and Wolfgang Effelsberg. "Automaticaudiocontentanalysis."InProceedingsofthefourthACMinternationalconferenceonMultimedia,pp.21-30.ACM,1997.
9. Kimber, Don, and Lynn Wilcox. "Acoustic segmentation for audiobrowsers." Computing ScienceandStatistics (1997): 295-304.
10. Zhang, Tong, and C-C. Jay Kuo. "Video content parsing based oncombined audio and visual information." In Multimedia Storage andArchiving Systems IV, vol. 3846, pp. 78-90. International Society forOptics andPhotonics,1999.
11. Dinh, PhungQuoc, ChitraDorai, and SvethaVenkatesh. "Video genrecategorizationusingaudiowavelet coefficients."ACCV2002 (2002).
12. Jasinschi, Radu S., and Jennifer Louie. "Automatic tv program genreclassificationbasedonaudiopatterns."InEuromicroConference,2001.Proceedings.27th,pp.370-375.IEEE,2001.
13. Thomas F Quatieri. Discrete-time speech signal processing: principlesand practice.PearsonEducationIndia,2006.
14. Liu,Zhu,JinchengHuang,andYaoWang."ClassificationTVprogramsbasedonaudioinformationusinghiddenMarkovmodel."InMultimedia Signal Processing, 1998 IEEE Second Workshop on, pp.27-32.IEEE,1998.
15. Visser, Erik, YinyiGuo, Lae-Hoon Kim, RaghuveerPeri, and ShuhuaZhang."Deepneuralnetbasedfilterpredictionforaudioeventclassification and extraction." U.S. Patent 9,666,183, issued May 30,2017.
16. James D Keeler, David E Rumelhart, and Wee KhengLeow. Integratedsegmentationandrecognitionofhand-printednumerals.InAdvancesinneuralinformationprocessingsystems, pages 557–563,1991.
17. Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez.Solving the multiple instance problem with axis-parallelrectangles.Artificialintelligence,89(1):31–71,1997.
18. Yixin Chen and James Z Wang. Image categorization by learning andreasoningwithregions.JournalofMachineLearningResearch,5(Aug): 913–939,2004.
19. Jun Yang, Rong Yan, and Alexander G Hauptmann. Multiple instancelearningforlabelingfacesinbroadcastingnewsvideo.InProceedingsofthe 13th annual ACM international conference on Multimedia, pages31–40.ACM,2005.