Automatic Attendance Rating of Movie Content using Bag of Audio Words Representation

Avi Bleiweiss
Architecture Group, Intel Corporation, Santa Clara, U.S.A.

Keywords: MFCC, Vector Quantization, Bag of Words, Support Vector Machines, Ranked Information Retrieval.

Abstract: The sensory experience of watching a movie, links input from both sight and hearing modalities. Yet traditionally, the motion picture rating system largely relies on the visual content of the film, to make its informed decisions to parents. The current rating process is fairly elaborate. It requires a group of parents to attend a full screening, manually prepare and submit their opinions, and vote out the appropriate audience age for viewing. Rather, our work explores the feasibility of classifying age attendance of a movie automatically, resorting to solely analyzing the movie auditory data. Our high performance software records the audio content of the shorter movie trailer, and builds a labeled training set of original and artificially distorted clips. We use a bag of audio words to effectively represent the film sound track, and demonstrate robust and closely correlated classification accuracy, in exploiting boolean discrimination and ranked retrieval methods.

1 INTRODUCTION

Classifying multimedia scenes into semantic categories is considered a central information retrieval (IR) problem, and attracted great interest in both research and practice. This is primarily motivated by the ever increasing scale of online content-base data, both visual and auditory, becoming freely available. For movies, the Internet Movie Database (IMDb, 1990) is the most comprehensive and authoritative source, providing millions of titles in its repository, publicly accessible to users, in the forms of full length films, trailers and clips. Of greatest relevancy to our work, is the IMDb high quality, trailer gallery that is kept current with both upcoming and recent releases. Each trailer is believed to faithfully capture the essence of the fully featured film, by portraying a short plot preview with key scenes, and playing in a time limit of one to two minutes. Moreover, many of the trailer videos are already ground-truth labeled, and embed the content ratings for appropriate viewing audience. Contrary to the explicit approach that requires human perception experience, in attending an entire film screening, we propose a system that implicitly learns from the vast aggregate audio content, provided by the IMDb trailer sound tracks, to classify admissible content to an age group. Our claim is founded on the basis that the depiction of language and violence, two of the more principal content rating criteria, are more immediately identified with the trailer acoustic features. We clearly recognize the challenge involved in a learning system of no visual sensing, but contend the implicit approach is more scalable, has the potential to be less error prone, and respectfully more economically sound.

One of the more simple and very effective text retrieval models is the *bag of words* (Baeza-Yates and Ribeiro-Neto, 1999). Here, documents are represented as a histogram of words from a prescribed vocabulary, however, the model completely ignores any of word dependency and syntactic structure. Retrieval relevance is therefore reduced to evaluating similarity of words in a document with words in a query. Inspired by the prospect to produce superior classification performance, researchers have evolved the model to efficiently represent large scale, multimedia content. The *bag of visual words* formulation was successfully adapted to scene image classification (Yang et al., 2007), and image retrieval (Wu et al., 2009). Similarly, the *bag of audio words* modality emerged in content based audio retrieval (Chechik et al., 2008), video copy detection (Liu et al., 2010), and a higher level structure, *bag of system words*, for semantic annotation of musical pieces (Ellis et al., 2011). Our system leverages established results of the model convention, and expresses each of the movie trailers with an enhanced bag of audio words representation that fits both a discriminatory classifier like SVM, and

Bleiweiss A..
Automatic Attendance Rating of Movie Content using Bag of Audio Words Representation.
DOI: 10.5220/0004604601420150
In *Proceedings of the 10th International Conference on Signal Processing and Multimedia Applications and 10th International Conference on Wireless Information Networks and Systems* (SIGMAP-2013), pages 142-150
ISBN: 978-989-8565-74-7

follows similarity calculations directly from the well known, Vector Space Model (Salton et al., 1975).

The main contribution of this paper is a high performance software that retrieves movie trailer sound tracks from a large scale and continually growing, online audio archive, and automatically rates a film content to appropriate, viewing age group. Unlike the current manual and more involved rating process that studies every movie, individually. Our feature abstraction interface provides the flexibility to utilize both discriminative and ranked information retrieval processes for classification, subscribing to a unique and robust correlation based validation of system performance. Next, we provide a brief overview of the movie rating system, in Section 2. Followed by describing our methodology for acoustic feature extraction, end-to-end, in Section 3. Then, we detail algorithm and give theory to support the performance of our multi modal classification approach, in section 4. Proceeding with analyzing quantitative results of our experiments to demonstrate rating effectiveness, in Section 5, and conclude with a short summary and future prospect remarks, in Section 6.

2 MOVIE RATING SYSTEM

The movie rating system is governed by an independent body, comprised of parents, with the purpose of giving advance warnings to parents, so that they can make informed decision about which films their children see. The movie rating system evolved both as a useful and valued tool for parents, but also become an essential guardian of the freedom of artistic, creative and political expression. In the United States, the Motion Picture Association of America (MPAA, 1922), through the Classification and Rating Administration (CARA, 1968), issues ratings for the movies. The modern system was instituted in late 1968 and is entirely voluntary. However, most major studios have agreed to submit all titles for rating prior to theatrical release. Respectfully, most movie theater chains avoid showing unrated domestic films. After screening films, the personal opinions of the group of parents in attendance, are used to arrive at one of five film ratings (Table 1). For some ratings, the MPAA adds a brief explanation as to why a particular film received certain rating. By convention, most film trailers will have the MPAA rating right at the beginning, and a fully featured film will have the MPAA logo at the end of the closing credits.

The motion picture rating system, classifies the content of a film primarily in terms of its depiction of matured theme, language, violence, and adult activities. For a General Audiences (G) rated movie, all ages are admitted. The movie contains nothing to offend parents for viewing by children. A Parental Guidance Suggested (PG) film class identifies some material that is inappropriate for children. Parents are urged to give guidance to their children, before letting them view the film. Parents Strongly Cautioned (PG-13) rating implies some material is unsuitable for children under 13. A PG-13 rank holds more of a serious warning to parents to determine whether their children should attend this motion picture. The Restricted (R) rating indicates that children under 17 are not allowed to attend, unless they are accompanied by a parent or an adult guardian. Finally, NC-17 rating is considered by most parents too adult oriented and no child under the age of 17 is admitted. NC-17 rated movies are rare and have little to no success in the box office. Often, studios and distributors edit the NC-17 film to qualify for an R rating. Accordingly, the NC-17 movie rating is outside the scope of our study. If a film is not submitted for rating or is an uncut version of a film that was submitted, the label NR (Not Rated) is often used.

3 ACOUSTIC FEATURES

Our basic framework for extracting the bag of audio words from a movie trailer, proceeds in several stages. First, the sound track of its original digital audio is converted into a high quality, WAV file. Next, the signal captured is divided into overlapping, short time segments, and primitive, audio feature vectors are derived from each. This follows a vector quantization (VQ) (Gersho and Gray, 1992) process that maps the original, acoustic feature vectors into a set of k clusters. The collection of clusters constitutes a vocabulary, and each cluster corresponds then to a unique audio word, or a term. Effectively, reducing the trailer representation to a highly compressed, single word vector of k dimensionality, with each element retaining a count of audio word occurrences in a WAV file.

Table 1: United States motion picture attendance ratings.

Rating	Description
G	General Audiences
PG	Parental Guidance Suggested
PG-13	Parents Strongly Cautioned
R	Restricted
NC-17	No One 17 and Under Admitted

3.1 Audio Feature Extraction

Models for human perception of sound, are based upon frequency analysis performed in the inner ear (Rabiner and Schafer, 2007). The cepstrum, the power spectrum of the logarithm spectrum, of a speech signal found to be an effective indicator of pitch detection (Noll, 1964). On this basis, Davis and Mermelstein (Davis and Marmelstien, 1980) formulated the mel-frequency cepstrum coefficients, MFCC, that represents the short term power spectrum of a sound, and captures the nonlinearity of human hearing. MFCC is one of the more compact acoustic feature representation, and is widely used in application domains that include automatic speech recognition, speaker identification, audio-video misalignment detection (Perelygin and Jones, 2011), and multimedia event classification (Pancoast and Akbacak, 2012). In our work, we use MFCC features with the parametric representation of the Fourier transformed cepstrum, derived based on a mel scale. The subjectively, nonlinear mel scale is further defined in equation 1, where f is the linear frequency in Hz.

$$f_{mel} = 1125 \ln(1 + \frac{f}{700}) \quad (1)$$

The computational flow for extracting the MFCC feature vector, from a discrete time speech signal, is depicted in Algorithm 1. First, the signal for each movie trailer is pre-emphasized with $alpha = 0.95$. Then, we segment the time data into frames at a rate of 100Hz, or 10ms duration, using a hamming window with 50% overlap. Our movie clips are consistently recorded in 16-bit mono, using a 44,100Hz sampling rate that yields 512 zero padded samples, per frame. Power-of-2 frame padding, warrants an efficient successive computation of the Fourier power spectrum, using the discrete Fourier transform (DFT) algorithm. Then, we apply a set of 48 triangular filters, spaced linearly in mel scale, on the spectrum obtained previously, and compute the logarithmic filter bank, energy coefficients. Finally, mel cepstral coefficients are derived, by employing the discrete cosine transform (DCT) on the mel bank, energy spectra. The features extracted for each frame, consist of the standard 12 MFCCs, as well as the log energy of the frame, resulting in a 13-dimensional feature vector. In our experiments, trailer sound tracks are recorded for approximately one minute, generating about 12,000 MFCC vectors on average, per WAV file.

Algorithm 1: MFCC.

```
Input: time data $s$, sample rate $f_s$, filter bank $f_n$
$x = \text{pre-emphasize}(s)$
$F = \text{frames}(x)$
for $f$ in $F$ do
    $w = \text{hamming}(f)$
    $spectrum = \mathcal{DFT}(w)$
    $melspectra = \log_{10}(\text{melbank}(spectrum, f_s, f_n))$
    $melcepstral = \mathcal{DCT}(melspectra)$
    $melcepstral = melcepstral \cup \text{energy}(f)$
end for
```

3.2 Bag of Audio Words

The VQ step performs clustering in the 13-dimensional, MFCC space. We apply K-means (Lloyd, 1982) to a set of MFCC features, extracted from each movie trailer, and identify k dense regions that collectively constitute the audio words. For every MFCC vector, K-means computes the nearest Euclidean distance to an iterated cluster centroid, and assigns to each MFCC feature, a cluster index in $\{1, 2, ..., k\}$. Then, a sound track is mapped to a k-dimensional vector $[f_1, f_2, ..., f_k]$ that encodes the frequency of each audio word, or a term, f_t, in the trailer. This histogram representation makes equivalent, segments that are acoustically similar, but their MFCC vector varies slightly. Figure 1 depicts audio word histogram of four movie trailers, one for each of the MPAA ratings. The choice of the parameter k is an important system performance trade-off. Large k increases computation time and results in a more discriminative model. On the other hand, while more efficient and of higher vocabulary consistency, a small set of clusters is less separable. In our evaluation, we explore the parameter k in a wide range of 100 to 3000, and study the vocabulary size impact on movie rating accuracy.

In a bag of audio words model, the order of the features in a movie sound track m is ignored. Rather, each trailer audio content is represented as a count vector in $\mathbb{N}^{|V|}$, where $|V|$ is the total number of words in the vocabulary. Borrowing from IR, we define the term frequency, $tf_{t,m}$, of term t in a movie feature vector m, as the number that t occurs in m. Whereas the relevance of a movie to a query, does not increase proportionally with the term frequency. Hence, to ensure less than linear, matching score growth, IR introduces instead a log-frequency weighting entity, we label $w_{t,m}$:

$$w_{t,m} = \begin{cases} 1 + \log_{10} tf_{t,m}, & \text{if } tf_{t,m} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Similarly to stop words in IR, frequent MFCC features in a trailer auditory sequence, are less informative than rare terms. To capture this notion, we intro-

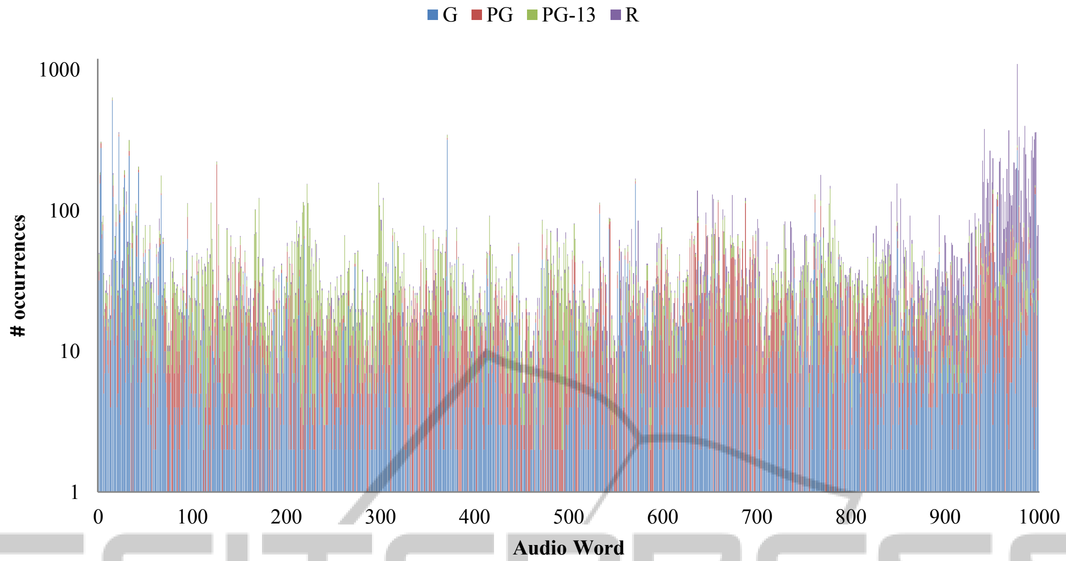

Figure 1: Audio word histogram for a sample trailer from each movie rating class. G (Aladdin), PG (Animals United) and PG-13 (10 Years) are fairly distinct, while R (A Late Quartet) is more observant on high word indices. Shown for a 1000 clusters configuration, with number of occurrences plotted on a logarithmic scale.

duce the movie frequency weight, mf_t, that amounts to the number of trailers that contain the term t. mf_t is therefore the inverse measure of the informativeness of t. We then formally follow to define the inverse movie frequency, imf of t as

$$imf_t = \log_{10}(N/mf_t), \tag{3}$$

with N the size of our trailer training set. Finally, analogous to the IR $tf.idf$ scheme, the term and inverse movie frequency measures are combined to yield $tf.imf$ weighting, for a given term, that is the product of the term tf and imf weights:

$$w_{t,m} = (1 + \log_{10} tf_{t,m}) \log_{10}(N/mf_t). \tag{4}$$

$tf.imf$ increases with the number of feature occurrences within a trailer acoustic model, and also with the rarity of a feature in the trailer training collection. Each movie trailer auditory content, is now represented by a real-valued vector of $tf.imf$ weights $\in \mathbb{R}^{|V|}$. Either an unnormalized or normalized weight vector version, is passed on to a classifier of choice.

4 CROSS CLASSIFICATION

The task of rating a movie content constitutes a multi class, classification problem. Decomposing the problem into a series of two-class, binary subproblems, using a one-against-all discriminative process, is one intuitive path of implementation. On the other hand, with a generic and orthogonal bag of audio words representation, the problem can be viewed as a search into a dataset of trailers that returns a ranked list of the most relevant, auditory content, to match a sound track query. Our design incorporates both boolean and indexed classification modalities, using support vector machines (SVM) and ranked information retrieval (RIR) methods, respectively. This cross classification formulation, merits a unique correlation dimension that ensures a more robust validation of our system rating performance.

4.1 Support Vector Machines

The bag of words feature representation, lends itself well to a discriminative classifier, such as support vector machines (SVM). SVM effectively models nonlinear decision boundaries, by using a kernel function. Respectively, we pass the vectors of weights, extracted from the auditory content of the movie trailers, for both training, and testing the learned model with new examples. For our classifier, we selected SVM-Light (Joachims, 1999), owing to its robust, large scale SVM training, and implemented a C++ wrapper on top, to seamlessly communicate with our movie rating, software components. Both the linear and radial basis function (RBF) kernels are studied in our work, specifically to compare rating performance impact, as a function of varying vocabulary size.

With four movie rating classes, we train four SVM models, each separating one group of trailers from the rest. The i-th SVM trains one of G, PG, PG-

13, and R class of trailers, all labeled as ground-truth true, and the remaining movie rating classes are labeled false. At the classification step, an unlabeled test trailer is assigned to a rating class that produces the largest value of hyper-plane distance, in feature space. Implementation wise, our software controls the one-against-all outer loop, invoking SVM-Light repeatedly four times, once for each rating class vs. rest configuration.

4.2 Ranked Information Retrieval

The Vector Space Model is the most commonly used in IR. This model ranks more relevant documents higher than less relevant ones, with respect to a query that is comprised of a collection of terms. Both the query and the documents are represented as vectors in the space, and documents are ranked based on their proximity to the query. Proximity is the similarity measure of two vectors, and is roughly a function of the inverse of the distance between them. The notion adapted by the IR community, to rank documents in increasing order of the cosine of the angle between the query and the document vectors, stems from the cosine function property, for monotonically decreasing in the interval $[0^\circ, 180^\circ]$.

A vector is length normalized by dividing each of its weighting components, by its length. For normalization, IR uses the L2-norm, expressed as $\|\vec{x}\|_2 = \sqrt{\sum_i x_i^2}$. Dividing a vector by its L2-norm, makes it a unit vector, and thus short and long trailer sequences of scaled terms, now have comparable weights. We define cosine similarity as the dot product of the query sound track vector, q, and a training trailer audio vector, m, both length normalized:

$$\cos(\vec{q}, \vec{m}) = \vec{q} \cdot \vec{m} = \sum_{i=1}^{|V|} q_i m_i. \tag{5}$$

We compute the cosine similarity score for the query trailer and each of the training trailers in our dataset. Training trailers, with respect to the query, are ranked for relevancy by their score, and the top $M(M = 10)$ are returned for further analysis.

5 EMPIRICAL EVALUATION

To evaluate our system in practice, we have implemented a Direct2D audio application that reads raw WAV files, splits each to a property header and data sections, and loads the normalized time signal and its sampling rate parameter to our movie rating, C++ library. Our library operates on the raw audio, commences feature extraction followed by vector quantization, and performs discrimination and similarity calculations. We use the hold out method with cross validation to rank the performance of our system. Formally, our library sets up one of random and 10-fold resampling modes, and each rating class becomes a two-way data split of trailers, with train and test sets, owning 80/20 percent shares, respectively.

5.1 Experimental Setup

Disposed at matching movie content to audience age, the productivity of bag of audio words representation is assessed. We build a labeled base set of 25 trailers, for each the G, PG, PG-13, and R ratings. The base set is drawn randomly from previously rated, movie trailers, off IMDb (IMDb, 1990). Our collection incorporates more recent and modern titles that are readily accessible online, but also subscribes to a fair share of productions that span over two decades of movie making. The high quality, 16-bit mono, 44.1KHz WAV files, each produced of a minute long recording, is about 5MB in size, setting the base set to total 100-minute footage of combined 0.5GB size. With an average of nearly 12,000 MFCC feature vectors, extracted from the audio sequence of a base trailer, our system processes a total of 1.2 million features. We then apply signal distortion, artificially to each of the labeled recording, and augment our learning data set by a factor of ten, to yield an aggregate of 250 auditory samples, per rating class. Leading to a combined 1000 system trailer set that subscribes to a 5GB footage of both physical and virtual audio. Rather than deforming the source signal (Riley et al., 2008), a step that is computationally intensive, we obtain an identical effect by simply perturbing the histogram of the base word vector, and randomly modulating the term frequency of a word in the interval $\{-5\%, +5\%\}$. This slightly warped version of a word vector, conforms to both discriminatory and similarity margins, to rule rating class inclusion.

Scalability to a dynamically grown, synthetic auditory data set, for attaining higher classification performance, is a principal guidance in the design of our proposed system. Here we discuss three implementation considerations that ensure the efficiency of major computational steps. First, the task of processing over a million feature vectors, is a critical compute section in our implementation. We find the execution of feature extraction and vector quantization, in sequentially iterating the base set, prohibitively inefficient. Rather, we exploit concurrency, using the latest C++11 futures and asynchronous launching methodology. In parallel processing independent au-

Table 2: Vocabulary statistical data, collected for a discrete set of cluster counts, and for each movie rating.

Number Clusters	Rating	Average MFCC Vectors	Max $tf_{t,m}$	Median $tf_{t,m}$	Mean $tf_{t,m}$	Standard Deviation $tf_{t,m}$	Average Single Word Clusters
100	G	11,679	4292	53.20	116.79	192.82	3.04
	PG	12,045	3687	49.60	120.45	201.26	3.48
	PG-13	11,507	2953	49.48	115.08	198.84	3.00
	R	11,955	3740	51.84	119.56	219.08	1.60
500	G	11,679	2256	12.00	23.36	39.82	42.88
	PG	12,045	2916	12.00	24.09	42.48	43.48
	PG-13	11,507	754	12.16	23.02	34.21	31.68
	R	11,955	1860	11.28	23.91	43.84	40.52
1000	G	11,679	1579	6.80	11.68	19.26	129.76
	PG	12,045	1162	6.60	12.05	19.72	131.80
	PG-13	11,507	435	6.88	11.51	15.60	109.00
	R	11,955	809	5.84	11.96	21.73	154.56
1500	G	11,679	626	4.80	7.79	11.10	248.60
	PG	12,045	1162	4.64	8.03	13.00	266.04
	PG-13	11,507	357	4.84	7.67	9.90	238.80
	R	11,955	450	3.84	7.97	14.71	331.40
2000	G	11,679	554	3.76	5.84	7.63	408.16
	PG	12,045	1159	3.68	6.02	9.47	442.00
	PG-13	11,507	357	3.56	5.75	7.32	408.32
	R	11,955	440	3.12	5.98	10.65	552.20
2500	G	11,679	554	3.08	4.67	5.92	624.16
	PG	12,045	579	2.92	4.81	6.35	649.08
	PG-13	11,507	357	3.04	4.60	5.68	621.88
	R	11,955	359	2.52	4.78	7.96	792.32
3000	G	11,679	554	2.60	3.89	4.90	886.92
	PG	12,045	578	2.64	4.02	5.13	876.96
	PG-13	11,507	357	2.56	3.83	4.60	884.60
	R	11,955	359	2.32	3.99	6.03	1038.96

dio streams, and generating MFCC vectors, followed by constructing histogram of words for each, we achieved a close to linear performance gain of about 3.5X, compared to the serial implementation, on a four cores, 2^{nd} generation, 2.8GHz Intel Core processor. Second, the automatic construction of synthetically distorted histogram vectors at runtime, not only bypasses the extensive MFCC processing of a source audio signal, it furthermore strips down memory footprint for temporary buffers, substantially. More importantly, the automatic augmenting of our auditory data set, from a relatively small, manual trailer production set, adds flexibility to gracefully enhance the performance of our rating system. Note the data amplification factor, set in our study to ten, is a system controlled parameter, supplied by the user. Thirdly, the computation of the inverse movie frequency, imf weighting, is independent of any query vector, and is therefore computed only once in our system, at training time. We use the well established and more efficient, standard IR weighting scheme, $lnc.ltc$, with imf computed for the test vectors, but is set to one, for the entire training partition.

5.2 Experimental Results

To understand how terms are distributed across our collection of base and distorted auditory content, we use the Zipf law. The law states that the collection frequency cf_i of the i_{th} most common term, is proportional to $1/i$. For vocabularies of different sizes, we plot the frequency of an audio word against its frequency rank in Figure 2, in a log-log scale. A straight line with a slope of -1, corresponding to the Zipf ideal function, is also depicted for reference (Ref). Our data shown to consistently fit the law, with the exception of the extremely low frequency terms. This is likely a side effect of our K-means implementation that produces rare words, we attribute to sparse, noisy audio frames. The slopes of the vocabulary curves, are however less steeper than the line predicted by the law, indicating a more evenly distribution of audio words. Table 2 further provides broader quantitative statistics to term weights, for each vocabulary size we use in our study, per movie rating. An interesting observation is the outstanding larger count of average single word clusters, indicative of content sparseness, for the movie rating R, specifically for richer vocabularies.

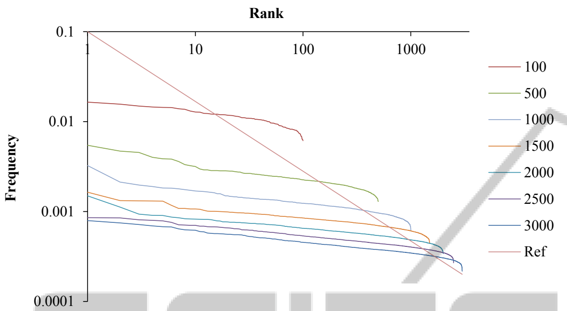

Figure 2: Zipf law distribution as a function of increased vocabulary size, plotted in a log-log scale.

We then study the impact of parametrically varying the bag of audio word representation, on our movie rating performance. Our multi modal classification methodology, evaluates system performance, for each of G, PG, PG-13, and R attendance classes, as a function of increased vocabulary size. For SVM, we use the F1-Score measure, and compare linear to RBF kernels, reporting our best results in Figures 3 and 4, respectively. Clearly, the vocabulary size affects greatly the classification score, however, the behavior is not necessarily monotonic. Performance rises first, might reach an optimal peak, then either remains flat or declines and rises again to reach a similar, or an extended local maximum. G, PG, and PG-13 ratings, are strongly grouped and follow a similar path trend, to closely intersect at an F1-Score of 0.67, for an optimal vocabulary size of 2000. However, the R class performs distinctively different, with an average classification rate of 0.575, peaking at a lexicon size of 3000 words, for a 0.64 score. We contend this one-from-rest scoring disparity is fairly explicable. The R sound track contains only sparsely spread, unique audio content to be classified as a restricted theme, a claim further supported statistically in Table 2. But the remainder of its content is a mixture of audio, relevant in the other ratings, thus making R less discriminatory. For the RBF kernel, we observe better performance for small vocabularies of highly correlated words, when compared to the linear kernel. But for larger vocabularies of linearly separable words, the RBF advantage is less obvious. More compelling for RBF, is the even tighter lumping of the G, PG, and PG-13 rating functions, extended to small vocabularies, and more importantly, the R class is further aligned with the rest.

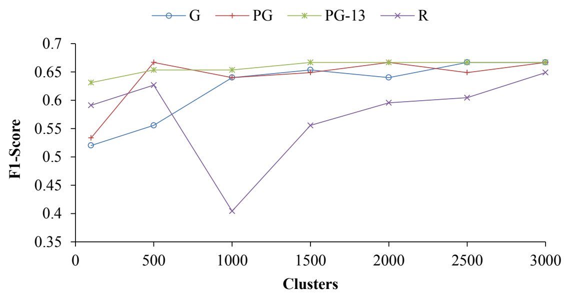

Figure 3: Linear SVM: F1-Score classification performance as a function of increased vocabulary size, shown for each movie rating class.

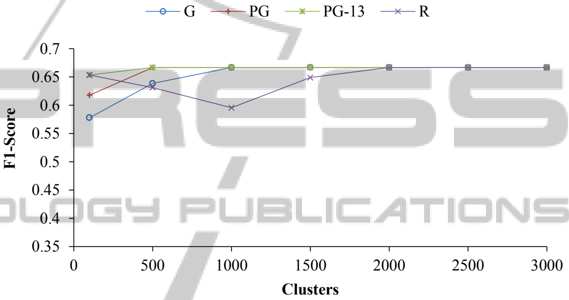

Figure 4: RBF SVM: F1-Score classification performance as a function of increased vocabulary size, shown for each movie rating class.

For RIR, a baseline method of finding nearest neighbors in using cosine distance, depicted non commensurate and inconsistent spiky performance. Rather, we use the mean average precision (MAP) measure that is susceptible to the entire ranking of a set of queries. A query is a trailer member of our test held out, data partition. First, we compute the average precision (AP) score, for an individual query. Each time, one of the top ranked training trailers, by way of similarity, is relevant and hence matches the query label, we accumulate its precision score at the current, non interpolated recall, and average out the scores. Seeking both the performance of common and rare terms in our bag of audio words representation, we weigh each query equally, amid computing the arithmetic average of all the query APs, to yield the desired MAP measure. Figure 5 depicts our system MAP as a function of increased vocabulary size. The G rating function rises almost monotonically, and flattens out at an optimal lexicon size of 2000 words, with an exceptionally high MAP score of 0.995. For the remaining rating classes, the score changes more mildly. PG peaks at a MAP of 0.93 and a 1000 word vocabulary, whereas PG-13 and R follow an almost identical performance path, tracing along a 0.92 MAP score. Figure 6 further illustrates the precision-recall curve for each rating, at a fixed vocabulary size of 100. For each of the precision-recall curves, we report commensurate area-under-curve (AUC) measures, in Figure 7, that closely parallel our MAP figures.

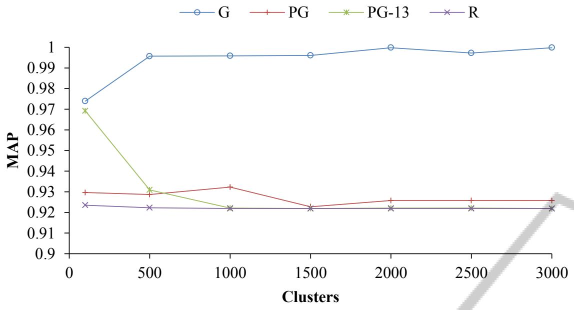

Figure 5: Ranked Retrieval: Mean average precision (MAP) classification performance as a function of increased vocabulary size, shown for each movie rating class.

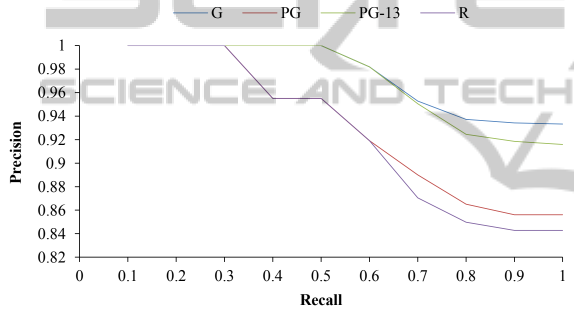

Figure 6: Ranked Retrieval: Precision-Recall curves for vocabulary size of 100 audio words, shown for each movie rating class.

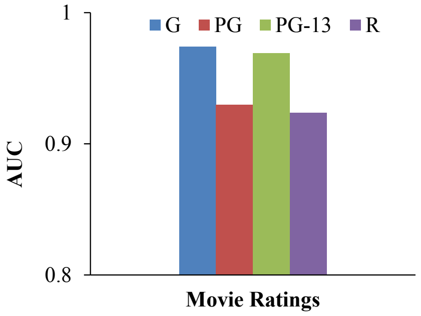

Figure 7: Ranked Retrieval: area-under-curve (AUC) measures, drawn from Precision-Recall curves (Figure 6), shown for each movie rating class.

Owing to a common and an effective weight vector representation, our research benefited markedly from comparing classification performance of a boolean discriminatory process against a discretized, ranked search. While classification modalities largely concur on performance results, there are however differences that warrant a brief review. Notably is the tendency for each model to distinguish one boundary rating class from the rest, but in a seemingly non-explicable, opposing directions of performance scores. For SVM, the R rating stands relatively less optimal, potentially implying bounded generalization, due to over-fitting. We suspect that by increasing the amplification factor of our distorted, audio data set, over-fitting can be greatly mitigated. On the other hand, RIR attributes higher MAP performance to the G rated movies, compared to the rest of the rating classes. Subscribing to the kids gamut of movie content, the G theme signifies more consistent and relatively flat acoustic features. Hence, similarity based RIR, is more successful in assigning relevancy to the majority of the top ranked, query results.

We compare our movie rating performance to the conceptually resembling task of Multimedia Event Detection (MED). Alike, MED searches multimedia corpora to identify user defined events, based on precomputed metadata. In their state-of-the-art work, Wang et al. (Wang et al., 2012) present their MED approach, employing bag of words representation for an enhanced combination of both static and dynamic visual descriptors, and MFCC audio features. They report an event specific MAP performance that ranges from 0.2 to 0.7, showing a moderate rate increase in augmenting the feature set selection. While the possible extension of our feature set merits further evaluation, at first order, our system MAP scores appear to justify a design with the sole use of audio features.

6 CONCLUSIONS

We have demonstrated the apparent potential in automating the rating process of age admittance to movie viewing, by classifying the auditory content of the movie sound track. Our training data set is comprised of an extremely small seed of high quality, hand recorded trailers, that is artificially augmented by a dynamically grown distorted set, computed off our low dimensionality, histogram vectors. Thus leading to an efficient and scalable software that executes on a compact memory footprint. The bag of audio words representation prove effective for both discriminative and retrieval classification modalities. Whereas exploiting term weighting to achieve plausible classification accuracy, serves a sound empirical basis for other challenging applications in real world, audio content understanding. Optimizing our system performance, as a function of the manual to synthetic data division, is one vital area for future research. We perceive our current unigram representation to evolve into a more generic n-gram model, by building bigram and trigram constructs, to better capture adjacent, temporal relations of audio features, at a trailer level.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their constructive and helpful feedback.

REFERENCES

Baeza-Yates, R. and Ribeiro-Neto, B., editors (1999). *Modern Information Retrieval*. ACM Press Series/Addison Wesley, Essex, UK.

CARA (1968). Classification and Rating Administration. http://www.filmratings.com/.

Chechik, G., Ie, E., Rehn, M., Bengio, S., and Lyon, R. F. (2008). Large scale content-based audio retrieval from text queries. In *ACM International Conference on Multimedia Information Retrieval (MIR)*, Vancouver, Canada.

Davis, S. B. and Marmelstien, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Ellis, K., Coviello, E., and Lanckriet, G. R. (2011). Semantic annotation and retrieval of music using a bag of systems representation. In *International Society for Music Information and Retrieval Conference (ISMIR)*, pages 723–728, Miami, FL.

Gersho, A. and Gray, R. M., editors (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, MA.

IMDb (1990). Internet Movie Database. http://www.imdb.com/.

Joachims, T. (1999). Making large-scale svm learning practical. In *Advances in Kernel Methods: Support Vector Learning*, pages 169–184. MIT-Press.

Liu, Y., Zhao, W., Ngo, C., Xu, C., and Lu, H. (2010). Coherent bag-of audio words model for efficient large-scale video copy detection. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, pages 89–96, Xi'an, China.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

MPAA (1922). Motion Picture Association of America. http://www.mpaa.org/.

Noll, A. M. (1964). Short-time spectrum and cepstrum techniques for vocal-pitch detection. *Acoustical Sociecty of America*, 36(2):296–302.

Pancoast, S. and Akbacak, M. (2012). Bag-of-audio-words approach for multimedia event classification. In *Conference of the International Speech Communication Association*, Portland, OR.

Perelygin, A. and Jones, M. R. (2011). Detecting audio-video asynchrony. Machine Learning , Stanford, http://cs229.stanford.edu/projects2011.html.

Rabiner, L. R. and Schafer, R. W., editors (2007). *Introduction to Digital Speech Processing*. Now Publishers Inc., Hanover, MA.

Riley, M., Heinen, E., and Ghosh, J. (2008). A text retrieval approach to content-based audio retrieval. In *International Society for Music Information and Retrieval Conference (ISMIR)*, pages 295–300, Philadelphia, PA.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communication of the ACM*, 18(11):613–620.

Wang, F., Sun, Z., Zhang, D., and Ngo, C. (2012). Semantic indexing and media event detection: ECNU at TRECVID 2012. In *TREC Video Retrieval Evaluation Workshop (TRECVID)*, Gaithersburg, MD.

Wu, Z., Ke, Q., Sun, J., and Shum, H. Y. (2009). A multi-sample, multi-tree approach to bag-of-words image representation for image retrieval. In *IEEE International Conference on Computer Vision, (ICCV)*, pages 1992–1999, Kyoto, Japan.

Yang, J., Jiang, Y., Hauptmann, A. G., and Ngo, C. (2007). Evaluating a bag-of-visual-words representations in scene classification. In *ACM International Workshop on Multimedia Information Retrieval (MIR)*, pages 197–206, Bavaria, Germany.