Audio-Based Semantic Concept Classification for Consumer Video

Keansub Lee, *Student Member, IEEE*, and Daniel P. W. Ellis, *Senior Member, IEEE*

***Abstract*—This paper presents a novel method for automatically classifying consumer video clips based on their soundtracks. We use a set of 25 overlapping semantic classes, chosen for their usefulness to users, viability of automatic detection and of annotator labeling, and sufficiency of representation in available video collections. A set of 1873 videos from real users has been annotated with these concepts. Starting with a basic representation of each video clip as a sequence of mel-frequency cepstral coefficient (MFCC) frames, we experiment with three clip-level representations: single Gaussian modeling, Gaussian mixture modeling, and probabilistic latent semantic analysis of a Gaussian component histogram. Using such summary features, we produce support vector machine (SVM) classifiers based on the Kullback–Leibler, Bhattacharyya, or Mahalanobis distance measures. Quantitative evaluation shows that our approaches are effective for detecting interesting concepts in a large collection of real-world consumer video clips.**

***Index Terms*—Audio classification, consumer video classification, semantic concept detection, soundtrack analysis.**

I. Introduction

MORE and more people are capturing everyday experiences using the video recording functions of small and inexpensive digital cameras and camcorders. These recordings are commonly shared with others through sites such as YouTube [1]. Such large consumer video archives contain copious information and consequently present many new opportunities for automatic extraction of information and the development of intelligent browsing systems. However, navigation and search within this kind of real-world material remain a considerable challenge. This paper addresses this challenge, looking in particular at the opportunities to exploit acoustic information—the soundtrack of a video—to see what useful descriptors can be reliably extracted from this modality. While the visual information in a video is clearly very rich, we believe that the soundtrack may offer a useful and complementary source of information.

Prior work on soundtrack analysis has typically focused on detecting or distinguishing between a small number of high level categories such as speech, music, silence, noise, or applause. The application domain has most often relatively carefully produced sources, such as broadcast audio or movie soundtracks. Saunders [2] presented a speech/music discrimination (SMD) based on simple features such as zero-crossing rate and short-time energy and a multivariate Gaussian classifier for use with radio broadcasts. This work reported an accuracy rate of 98% with 2.4-s segments. Scheirer *et al.* [3] tested 13 temporal and spectral features followed by a Gaussian mixture model (GMM) classifier, and reported an error rate of 1.4% in classifying 2.4-s segments from a database of randomly recorded radio broadcasts of speech and music. Williams *et al.* [4] approached SMD by estimating the posterior probability of around 50 phone classes on the same data, and achieved the same performance. Zhang *et al.* [5] proposed a system to segment and classify audio from movies or TV programs into more classes such as speech, music, song, environmental sound, speech with music background, environmental sound with music background, silence, etc. Energy, zero-crossing rate, pitch, and spectral peak tracks were used as features, and heuristic rule-based classifier achieved an accuracy rate of more than 90%. Ajmera *et al.* [6] used entropy and dynamism features based on posterior probabilities of speech phonetic classes [as obtained at the output of an artificial neural network (ANN)], and developed a SMD based on a hidden Markov model (HMM) classification framework. Lee *et al.* [7] developed a noise-robust musical pitch detector based on long-window autocorrelation for identifying the presence of music in the noisy, highly variable consumer audio collected by body-worn recorders. A support vector machine (SVM) classifier using both pitch and rhythm features achieved 92% average precision (AP) on 1873 YouTube videos.

For less constrained environmental sounds, research has considered problems such as content-based retrieval, surveillance applications, or context-awareness in mobile devices. A popular framework is to segment, cluster, and classify environmental recordings into relatively simple concepts such as “animal,” “machine,” “walking,” “reading,” “meeting,” and “restaurant,” with testing performed on a few hours of data. Wold *et al.* [8] presented a content-based audio retrieval system called “Muscle Fish.” This work analyzed sounds in terms of perceptual aspects such as loudness, pitch, brightness, bandwidth, and harmonicity, and adopted the nearest neighbor (NN) rule based on Mahalanobis distance measure to classify the query sound into one of predefined sound classes broadly categorized into animals, machines, musical instrument, speech, and nature. Foote [9] proposed a music and sound effects retrieval system

Manuscript received December 31, 2008; revised September 11, 2009. First published October 20, 2009; current version published July 14, 2010. This work was supported in part by the National Science Foundation under Grant IIS-0238301, in part by the Eastman Kodak Company, and in part by the EU project AMIDA (via the International Computer Science Institute). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

The authors are with the Laboratory for the Recognition and Organization of Speech and Audio (LabROSA), Electrical Engineering Department, Columbia University, New York, NY 10027 USA (e-mail: kslee@ee.columbia.edu; dpwe@ee.columbia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2034776

where 12 mel-frequency cepstral coefficients (MFCCs) plus energy were used as feature vectors. A tree-based vector quantizer (VQ) was applied on the feature vector space to partition it into regions. Sounds were classified by calculating the Euclidean or cosine distances between the histograms of VQ codeword usage within each sound. Guo *et al.* [10] used SVM classifiers with perceptual and cepstral features on the "Muscle Fish" data and roughly halved the errors in comparison to [8]. Ellis *et al.* [11] developed an automatic indexing mechanism at a coarse time scale (e.g., 60-s frames) using features such as average log energy and entropy deviation to identify the user's location based on nonspeech background ambience. Segmentation employed the Bayesian information criterion, and segments were then associated with one another via spectral clustering. This work gave frame-level accuracies of over 80% on a 62-h hand-labeled personal audio recordings. Malkin *et al.* [12] used linear autoencoding neural networks to achieve a lower error rate than a standard Gaussian mixture model (GMM) for classifying environments such as restaurant, office, and outdoor. A linear combination of autoencoders and GMMs yielded still better performance. Ma *et al.* [13] considered the problem of classifying the acoustic environment on a portable device, for instance to provide a record of daily activities. They used MFCC features classified by an adapted speech recognition HMM to achieve over 90% accuracy distinguishing 3-s excerpts of 11 environments; humans averaged only 35% correct on the same data. Chu *et al.* [14] investigate acoustic context recognition for an autonomous robot. They compared nearest-neighbor (NN), GMM, and SVM classifiers with a wide range of features on a five-way classification task, and found best performance using the SVM and a subset of features selected by a greedy scheme.

The work most directly comparable to the current paper is that by Eronen *et al.* [15]. Similar to [13], they investigated the classification of 24 contexts such as restaurant, office, street, kitchen with a view to applications in portable devices that could alter their behavior to best match an inferred situation. They compared a variety of features and classifiers, and achieved best performance with a simple approach of training a five-component GMM on the MFCCs for each class, then classifying a test sample according to the GMM under which it achieves the highest likelihood. We take this as our baseline comparison system in the results below.

None of this prior work has directly addressed the classification of consumer videos by their soundtracks, and this domain raises a number of novel issues that are addressed for the first time in this paper. First, we are dealing with the relatively large number of 25 concepts, comparable only to the 24 contexts in [15]; other systems used only between 2 and 12 concepts. Second, our concepts are drawn from a user study of photography consumers [16], and thus reflect actual types of queries that users would wish to make rather than simply the distinctions that we expect to be evident in the data. Third, in all previous work there has been exactly one ground-truth label for each clip example (i.e., the data were exclusively arranged into a certain number of examples of each category). Consumer-relevant concepts cannot be so cleanly divided, and in our data most clips bear multiple labels, requiring a different approach to classification; our approach is inspired by similar work in music clip tagging, which has a similarly unpredictable number of relevant tags per item [17]. Finally, our data set is larger than any previously reported in environmental sounds, consisting of the soundtracks from 1873 distinct videos obtained from YouTube (as described below). These soundtracks are typically rather poor quality, often contain high levels of noise, and frequently have only sparse instances of "useful" (i.e., category-relevant) sounds. Thus, this is a much more demanding task than has been addressed in earlier work.

In addition to the novelty of the problem, this paper makes a number of specific technical contributions. First, we illustrate the viability of classifying, based only on soundtrack data, concepts like "beach" or "night" that on first sight seem unrelated to audio. Second, we show how to address the problem of overlapping concepts through the use of multiple, independent classifiers. Finally, we introduce a novel technique based on probabilistic latent semantic analysis (pLSA) which outperforms our baseline Gaussian-SVM classifiers.

Our concepts have diverse characteristics in terms of consistency, frequency, and interrelationships. For example, the labels "music" and "crowd" typically persist over most or all of any clip to which they apply, and hence should be well represented in the global feature patterns (e.g., mean and covariance of the clip's frame-level features). However, the concept "cheer" manifests as a relatively small segment within a clip (at most a few seconds), which means that the global patterns of an entire clip may fail to distinguish it from others. This points to the need for methods that can emphasize local patterns embedded in a global background, such as the pLSA approach described below.

We briefly review the selection, definition, and annotation of semantic concepts for consumer videos in Section II-A. Audio-based detectors are described in Section III. The evaluation and discussion of experimental results are given in Sections IV and V, respectively.

II. Data and Labels

A. Semantic Concepts

Our goal is to provide classification that is relevant to users browsing personal video collections, thus our concepts must reflect the actual needs of this target group. In previous work [18], we defined the set of 25 concepts used here by starting from a full ontology of over 100 concepts obtained through user studies conducted by the Eastman Kodak Company [18]. This set was pared down based on criteria of usefulness and viability (i.e., whether the concept was suitably unambiguous to be labeled by humans, and whether it should be detectable in either video or audio). These selected concepts fall into several broad categories including activities, occasions, locations, or particular objects in the scene, as shown in Table I. Most concepts are intrinsically visual, although some concepts, such as music and cheering, are primarily acoustic. Since most of the selected concepts are dominated by the visual cues, using visual cues achieved higher accuracy for most concepts than using audio cues. However, audio models provided significant benefits. For

TABLE I
DEFINITION OF THE 25 CONCEPTS, AND COUNTS OF MANUALLY LABELED EXAMPLES OF EACH CONCEPT FROM 1873 VIDEOS

Category	Concept	Definition	Examples
Activities	dancing	one or people dancing	189
	singing	singers visible and audible	345
	ski	one or people skiing	68
Locations	beach	sand and water visible	130
	museum	exhibitions of arts, antiques	45
	park	some greenery in view	118
	playground	swings, slides in view	96
Occasions	birthday	birthday cake, caps, songs	68
	graduation	caps and gowns visible	72
	picnic	people and food outdoors	54
	parade	people or vehicles moving	91
	show	concerts, plays, recitals	211
	sports	soccer, basketball, football, baseball, volleyball, ping-pong	84
	wedding	bride and groom in view	57
Objects	animal	dogs, cats, birds, wild animals	61
	baby	infant, 12 months or younger	112
	boat	boats in the water	89
	group of 3+	three or more people	1126
	group of 2	two people	252
	one person	single person	316
Scenes	crowd	many people in the distance	533
	night	outdoors at night	300
	sunset	the sun in view	68
Sounds	cheer	acclamation, hurrah	388
	music	clearly audible professional or amateur music	653
Total	25 concepts		1873

example, by their nature, concepts like "music," "singing," and "cheer" can primarily be detected in the acoustic domain. Even for some visually dominated concepts (like "museum" and "animal"), audio methods were found to be more reliable than visual counterparts, implying that the soundtracks of video clips from these concepts provide rather consistent audio features for classification. By combining visual baseline detectors and audio baseline detectors through context fusion, the proposed Audio–Visual Boosted Conditional Random Field (AVBCRF) method algorithm improves the performance by more than 10% compared with the visual baseline. The improvements over many concepts are significant, e.g., 40% for "animal," 51% for "baby," 228% for "museum," 35% for "dancing," and 21% for "parade." This paper describes for the first time the detail of the audio-based detectors used in that work.

B. Video Data

We downloaded 4539 videos (about 200 videos for each concept) from YouTube [1] by using the most relevant keyword queries associated with the definition of the 25 concepts. For these downloaded videos, we first manually filtered them to discard videos not consistent with the consumer video genre, or low-quality videos (e.g., those with particularly bad sound quality).

Non-consumer videos fall mainly into two categories: broadcasting content, and user-edited videos. The "sports" videos downloaded with using keywords like soccer, basketball, football, baseball, volleyball, and ping-pong, contain many commercial videos captured from TV. Some consumer videos are also extensively edited, e.g., the highlights of a field trip can have many abrupt changes of locations in single video clip. Some clips look like music videos, with the original soundtrack largely or completely replaced by added music. These types are also excluded.

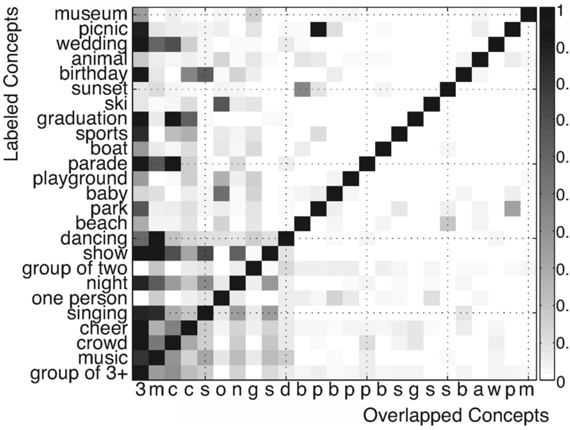


Fig. 1. Co-occurrence matrix for the 25 manually annotated labels within the 1873 video set. Co-occurrence counts within each row are normalized by the total number of instances of that row's concept to give the conditional probability of observing the overlapped concept given the labeled concept.

In consequence, the 4539 YouTube videos were reduced to 1873 (41%) relevant, consumer-style clips, 1261 (28%) irrelevant (non-consumer), and 1405 (31%) poor-quality videos whose soundtracks had bandwidth less than 8 kHz. We used only the 1873 relevant videos with adequate sound quality as our experimental data set. The average duration of a clip from this set was 145 s.

Videos were downloaded based on the tags and description provided by their owners. However, people will generally tag their videos according to subjective definitions (e.g., labeling an entire ski trip as relating to the concept "skiing"). To ensure accurate labels, we manually reviewed every one of the 1873 videos, and tagged it with the concepts that it contained, as defined in Table I.[1] On average, each video ended up with three concept labels, and some labels were very likely to co-occur with others, as illustrated in Fig. 1. For example, "group of three or more," "music," "crowd," and "cheer," are all highly overlapped with other concepts. More details on the video collections and labels are provided in [16].

III. AUDIO CONCEPT DETECTION ALGORITHMS

Our fundamental frame-level feature is the MFCCs commonly used in speech recognition and other acoustic classification tasks. The single-channel (mono) soundtrack of a video is first resampled to 8 kHz, and then a short-time Fourier magnitude spectrum is calculated over 25-ms windows every 10 ms. The spectrum of each window is warped to the Mel frequency scale, and the log of these auditory spectra is decorrelated into MFCCs via a discrete cosine transform.

After the initial MFCC analysis, each video's soundtrack is represented as a set of $d = 21$-dimensional MFCC feature vectors, where the total number of frames depends on the

[1] These labels, along with the references to the YouTube videos, are available on our website, http://www.labrosa.ee.columbia.edu/projects/consumervideo/.

duration of the original video. (21 dimensions were chosen based on results from our earlier experiments [11]; general audio classification usually benefits from using more MFCC dimensions than are commonly used for speech recognition.) To reduce this set of MFCC frames, regardless of its original size, to a single fixed-dimension clip-level feature vector, we experimented with three different techniques: Single Gaussian modeling (1G), Gaussian mixture modeling (GMM), and probabilistic latent semantic analysis of a Gaussian component histogram (pLSA). Each of these is discussed in more detail below.

These fixed-size representations are then compared to one another by several distance measures: the Kullback–Leibler divergence (KL), Bhattacharyya distance (Bha), and Mahalanobis distance (Mah). The distances between all clips form the input to a support vector machine classifier as described in the next subsection.

A. Support Vector Machines (SVMs)

The SVM is a supervised learning method used for classification and regression that has many desirable properties [19]. Data items are projected into a high-dimensional feature space, and the SVM finds a separating hyperplane in that space that maximizes the margin between sets of positive and negative training examples. Instead of working in the high-dimensional space directly, the SVM requires only the matrix of inner products between all training points in that space, also known as the kernel or gram matrix. With a method similar to [20], we exponentiate the matrix of distances between examples $D(f, g)$ to create a gram matrix $K(f, g)$

$$K(f, g) = \exp\left(-\gamma \cdot D(f, g)\right) \tag{1}$$

where $\gamma = \{2^{10}, 2^{9}, \ldots, 2^{-10}\}$, and f and g index the video clips. We use the so-called slack-SVM that allows a trade-off between imperfect separation of training examples and smoothness of the classification boundary, controlled by a constant C that we vary in the range $\{10^{1}, 10^{2}, \ldots 10^{10}\}$. Both tunable parameters γ and C are chosen to maximize classification accuracy over a held-out set of validation data. After training an independent SVM model for each concept, we apply the classifiers to summary features derived from the test video clips. The resulting distance-to-boundary is a real value that indicates how strongly the video is classified as reflecting the concept. The test videos are then ranked according to this value. Following conventions in information retrieval, we evaluate classifiers by calculating their average precision (AP), which is the proportion of true results in the ranked list when truncated at the nth true item, averaged over all n.

B. Single Gaussian Modeling (1G)

The basic assumption of single Gaussian modeling is that different activities (or concepts) are associated with different sounds whose average spectral shape and variation, as calculated by the cepstral feature statistics, will be sufficient to discriminate categories. This approach is based on common practice in speaker recognition and music genre identification, where the distribution of cepstral features, collapsed across time, is found to be a good basis for classification [21], [22]. Specifically, to describe a clip's sequence of MFCC features as a single feature vector, we ignore the time dimension and treat the set as a "bag of the frames" in MFCC feature space, which we then model as a single, full-covariance Gaussian distribution. This Gaussian is parameterized by its 21-dimensional mean vector μ and 21×21-dimensional (full) covariance matrix Σ. The overall process of the single Gaussian modeling is illustrated in Fig. 2.

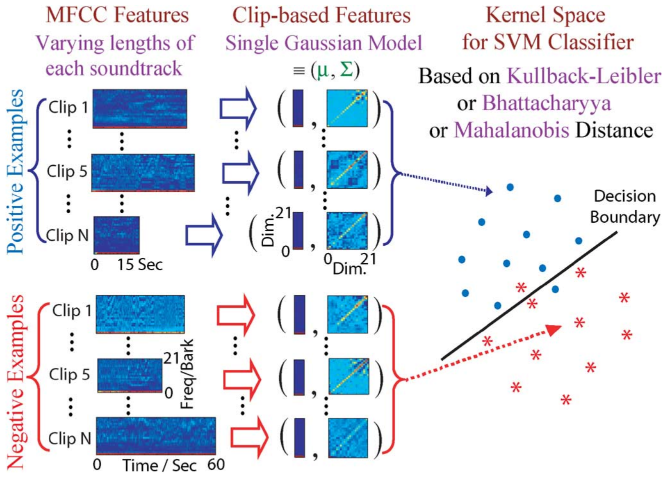


Fig. 2. Process of calculating clip-level features via a single Gaussian model per clip, and using them within an SVM classifier.

To calculate the distance between two Gaussians, as required for the gram-matrix input (or kernel matrix) for the SVM, we have experimented with three different distance measures. First is the KLdivergence: If two clips f and g are modeled by single Gaussians as

$$f(x) = \mathcal{N}(\mu_f, \Sigma_f), g(x) = \mathcal{N}(\mu_g, \Sigma_g) \tag{2}$$

respectively, then the distance between the clips is taken as the KL divergence between Gaussians $f(x)$ and $g(x)$, i.e.,

$$\begin{aligned} D_{KL}(f, g) = {} & (\mu_f - \mu_g)^T \left(\Sigma_f^{-1} + \Sigma_g^{-1}\right)(\mu_f - \mu_g) \\ & + trace\left(\Sigma_f^{-1}\Sigma_g + \Sigma_g^{-1}\Sigma_f\right) - 2d. \end{aligned} \tag{3}$$

The second distance measure is the Bha distance, defined by

$$\begin{aligned} D_B(f, g) = {} & \frac{1}{4}(\mu_f - \mu_g)^T(\Sigma_f + \Sigma_g)^{-1}(\mu_f - \mu_g) \\ & + \frac{1}{2} log\left|\frac{\Sigma_f + \Sigma_g}{2}\right| - \frac{1}{4} log|\Sigma_f \Sigma_g|. \end{aligned} \tag{4}$$

The final approach simply treats the d-dimensional mean vector μ concatenated with the $d(d+1)/2$ independent values (diagonal and upper triangular elements) of the covariance matrix Σ as a point in a new $21 + 231$-dimensional feature space describing the clip. These 252-dimensional features, denoted by h_f and h_g for videos f and g, are compared to one another using the Mahalanobis (i.e., covariance-normalized Euclidean) distance to build the gram matrix

$$D_M(f, g) = (h_f - h_g)^T \Sigma_h^{-1}(h_f - h_g) \tag{5}$$

where Σ_h is the covariance of these features taken across the entire training set. We assumed Σ_h to be diagonal, i.e., consisting only of the variance of each dimension.

C. Gaussian Mixture Models (GMMs)

In order to capture details of feature distributions that may not be well fit by a single Gaussian, we also experimented with using a mixture of diagonal-covariance Gaussians, estimated via the EM algorithm, to describe the bag-of-frames distribution. To compare GMMs, we use just one distance measure, an approximation to the Bhattacharyya distance that was shown to give good performance in tasks requiring the comparison of GMMs [23]: Assume that the distributions of two clips, $f(x)$ and $g(x)$, are represented by two different GMMs

$$f(x) = \sum_a \pi_a \mathcal{N}(\mu_a, \Sigma_a), \quad g(x) = \sum_b \pi_b \mathcal{N}(\mu_b, \Sigma_b) \quad (6)$$

where π_a, μ_a, and Σ_a are the prior weight, mean, and covariance of each Gaussian mixture component used to approximate clip f, and the b-subscripted values are for clip g. To simplify notation, we call $f_a = \mathcal{N}(\mu_a, \Sigma_a)$ and $g_b = \mathcal{N}(\mu_b, \Sigma_b)$ henceforth.

Although there is no closed-from expression for the Bhattacharyya divergence between two GMMs, it can be approximated by variational methods [23]. The Bhattacharyya similarity between two distributions $f(x)$ and $g(x)$ is

$$\begin{aligned} B(f,g) &\equiv \frac{1}{2}\int \sqrt{f(x)g(x)}dx \\ &\geq \sqrt{\sum_{ab} \pi_a \pi_b B^2(f_a, g_b)} \equiv \hat{B}^v(f,g) \end{aligned} \quad (7)$$

where $B(f_a, g_b)$ is the Bhattacharyya distance between a particular pair of single Gaussians, one from each mixture. To preserve the identity property that $\hat{B}(f,g) = 1/2$ if and only if $f = g$, the variational Bhattacharyya similarity $\hat{B}^v$ is normalized using the geometric mean of $B(f,f)$ and $B(g,g)$

$$\hat{B}_{\mathrm{norm}}(f,g) = \frac{\hat{B}^v(f,g)}{\sqrt{\hat{B}^v(f,f)\hat{B}^v(g,g)}}. \quad (8)$$

With this normalized Bhattacharyya approximation, the corresponding Bhattacharyya divergence is defined as $D_B(f,g) = -\log(2\hat{B}_{\mathrm{norm}}(f,g))$.

D. Probabilistic Latent Semantic Analysis (pLSA)

Unlike the Gaussian models' assumption that each concept is distinguished by the global distribution of all short-time feature vectors, this approach recognizes that each soundtrack will consist of many different sounds that may occur in different proportions even for the same category, leading to variations in the global statistics. If, however, we could decompose the soundtrack into separate descriptions of those specific sounds, we might find that the particular palette of sounds, but not necessarily their exact proportions, would be a more useful indicator of the content. Some kinds of sounds (e.g., background noise) may be common to all classes, whereas some sound classes (e.g., a baby's cry) might be very specific to a particular class of videos.

To build a model better able to capture this idea, we first construct the vocabulary (or palette) of sounds by constructing a large GMM, composed of M Gaussian components; we experimented with M in the range 256 to 1024. This large GMM was trained on MFCC frames subsampled from all videos from the training set, regardless of label. (We observed a marginally better performance after training the GMM on a set of frames selected as the central points of about 100 groups, clustered by the K-means algorithm on each clip, instead of a random sampling method). The resulting M Gaussians are then considered as anonymous sound classes from which each individual soundtrack is assembled—the analogs of words in document modeling. We assign every MFCC frame in a given soundtrack to the most likely mixture component from this "vocabulary" GMM, and describe the overall soundtrack with a histogram of how often each of the M Gaussians was chosen when quantizing the original clip's frames.

Suppose that we have given a collection of training clips $C = \{c_1, c_2, \ldots c_N\}$ and an M-mixture of Gaussians $G = \{g_1, g_2, \ldots g_M\}$. We summarize the training data as a $N \times M$ co-occurrence matrix of counts O with elements $o_{ij} = o(c_i, g_j)$, the number of times mixture component g_j occurred in clip c_i. Normalizing this within each clip gives an empirical conditional distribution $P(g|c)$. Note that this representation also ignores temporal structure, but it is able to distinguish between nearby points in cepstral space provided they were represented by different Gaussians in the vocabulary model. The idea of using histograms of acoustic tokens to represent the entire soundtrack is also similar to that of using visual token histograms for image representation [24], [25].

We could use this histogram $P(g|c)$ directly, but to remove redundant structure and to give a more compact description, we go on to decompose the histogram with pLSA [26]. This approach, originally developed to generalize the distributions of individual words in documents on different topics $Z = \{z_1, z_2, \ldots z_K\}$, models the histogram as a mixture of a smaller number of "topic" histograms, giving each document a compact representation in terms of a small number of topic weights. The individual topics are defined automatically to maximize the ability of the reduced-dimension model to match the original set of histograms. (This technique has been used successfully in an audio application by Arenas–García *et al.* [27], who use pLSA as a way to integrate and condense different features of music recordings for applications in similarity and retrieval.)

Specifically, the histogram-derived probability $P(g|c)$ that a particular component g will be used in clip c is approximated as the sum of contributions from topics z, $p(g|z)$, weighted by the specific contributions of each topic to the clip, $p(z|c)$, i.e.,

$$P(g|c) = \sum_{z \in Z} P(g|z)P(z|c) \quad (9)$$

which embodies the assumption that conditioning on a topic z makes clip c and component g independent. During training, the topic profiles $P(g|z)$ (which are shared between all clips), and the per-clip topic weights $P(z|c)$, are optimized by using the

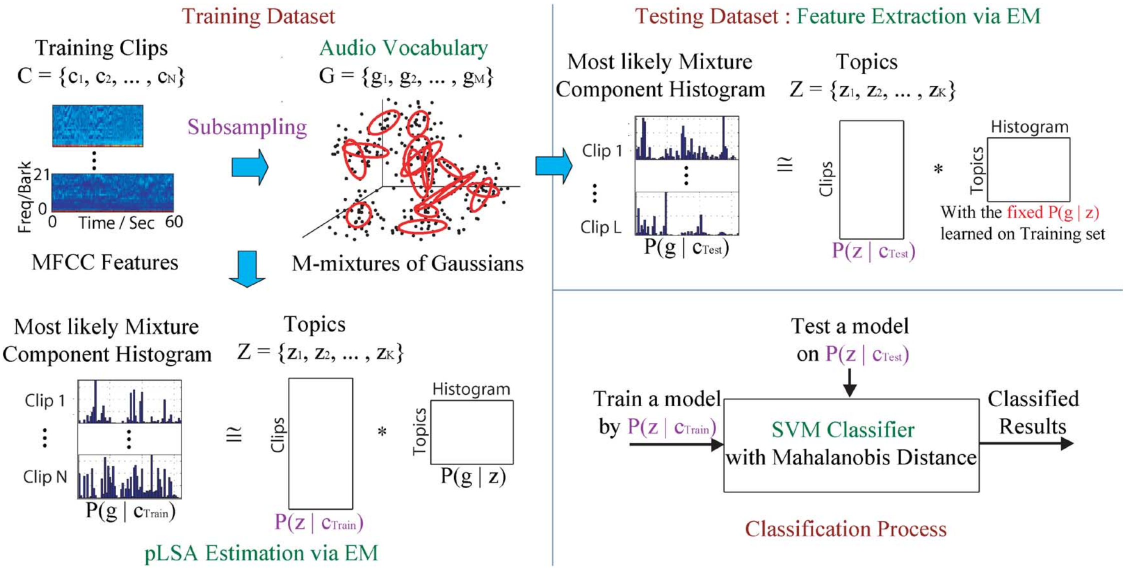


Fig. 3. Illustration of the calculation of pLSA-based features and clip-level comparisons, based on GMM component histograms. Top left shows the formation of the global GMM; bottom left shows the formation of the topic profiles, $P(g|z)$ and topic weights, $P(z|c_{\mathtt{train}})$ in training data; top right shows the analysis of each testing clip into topic weights, $P(z|c_{test})$ by matching each histogram to a combination of topic profiles estimated by training data, and bottom right shows the final classification by an SVM.

expectation–maximization (EM) algorithm. In the Expectation (E) step, posterior probabilities are computed for the latent variables

$$P(z|c,g) = \frac{P(z)P(c|z)P(g|z)}{\sum_{z' \in Z} P(z')P(c|z')P(g|z')}. \tag{10}$$

Parameters are updated in the maximization (M) step

$$\begin{aligned} P(g|z) &\propto \sum_{c} o(c,g)P(z|c,g) \\ P(c|z) &\propto \sum_{g} o(c,g)P(z|c,g) \\ P(z) &\propto \sum_{c}\sum_{g} o(c,g)P(z|c,g). \end{aligned} \tag{11}$$

The number of distinct topics determines how accurately the individual distributions can be matched, but also provides a way to smooth over irrelevant minor variations in the use of certain Gaussians. We tuned it empirically on the development data, as described in Section IV. Representing a test item similarly involves finding the best set of weights to match the observed histogram as a (nonnegative) combination of the topic profiles; we minimizing the KL distance via an iterative solution, i.e., the per-clip topic weights $P(z|c)$ of testing data sets are optimized by using the EM algorithm with fixed the topic profiles $P(g|z)$ that is already estimated on training set.

Finally, each clip is represented by its vector of topic weights and the SVM's gram matrix is calculated as the Mahalanobis distance in that topic weight vector space. (Again, we assumed the feature covariance matrix was diagonal.) We compared several different variants of the topic weight vector: unmodified $P(z|c)$, log-transformed $\log(P(z|c))$, and log-normalized $\log(P(z|c)/P(z))$, which normalizes the topic weight by the prior of topics and then takes the logarithm. The process of pLSA feature extraction is illustrated in Fig. 3.

IV. Evaluation

We evaluate our approaches using fivefold cross validation on our labeled collection of 1873 videos: At each fold, SVM classifiers for each concept are trained on 40% of the data, tuned on 20%, and then tested on the remaining 40%, selected at random.

We then evaluated all our approaches in terms of the AP for detecting the 25 concepts across the 1873 consumer-style videos. Fig. 4 shows the results of the 1G with the three different distance measures, KL, Mahalanobis, and Bhattacharyya. 1G+KL gives better performance for location-related concepts such as "park," "playground," and "ski"; by contrast, audio-dominated concepts such as "music," "cheer," and "singing" are best with the 1G+Mah. Concepts "group of 3+," "crowd," and "baby" are well detected by 1G+Bha, possibly because human speech plays an important role in discriminating them from other concepts. On average, 1G+KL performs the best among the three distance measures.

Fig. 5 shows the results for GMMs with between 2 and 16 Gaussian components per model. Between-model distance is calculated by the approximated Bhattacharyya divergence. Although the optimal number of Gaussian is strongly dependent on the total duration of positive examples of the class, the 8-GMM is a good compromise (the best AP), able to capture detail across all the classes.

The performance of the pLSA of the GMM histogram is shown in Figs. 6 and 7. To build the gram matrix for the SVM, we tested various summary features, including the raw histogram counts $P(g|c)$ (i.e., without decomposition into pLSA topics), the per-clip topic weights $P(z|c)$, log-topic weights $\log(P(z|c))$, and log-normalized topic weights $\log(P(z|c)/P(z))$. In each case, the gram matrix contained Mahalanobis distances, i.e., normalized by the variance of the features across the entire training set. By comparing the three curves for 1024-GMM histograms in Fig. 7, we see that log-normalized topic weights perform significantly better than

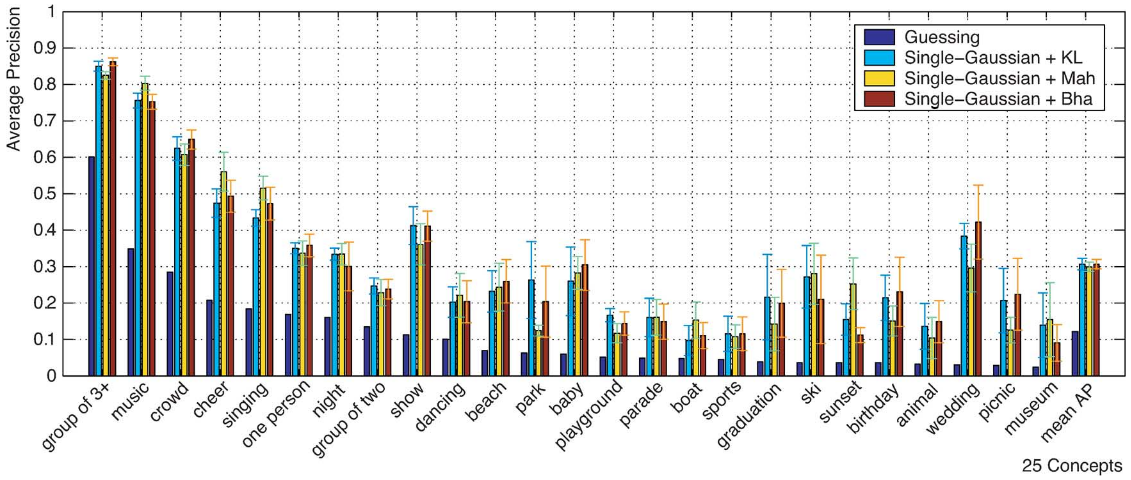


Fig. 4. Average precision (AP) across all 25 classes for the 1Gs, using each of the three distance measures, KL, Mahalanobis, and Bhattacharyya. Labels are sorted by the guessing baseline performance (shown). Bars and error-bars indicate the mean and standard deviation over fivefold cross-validation testing respectively.

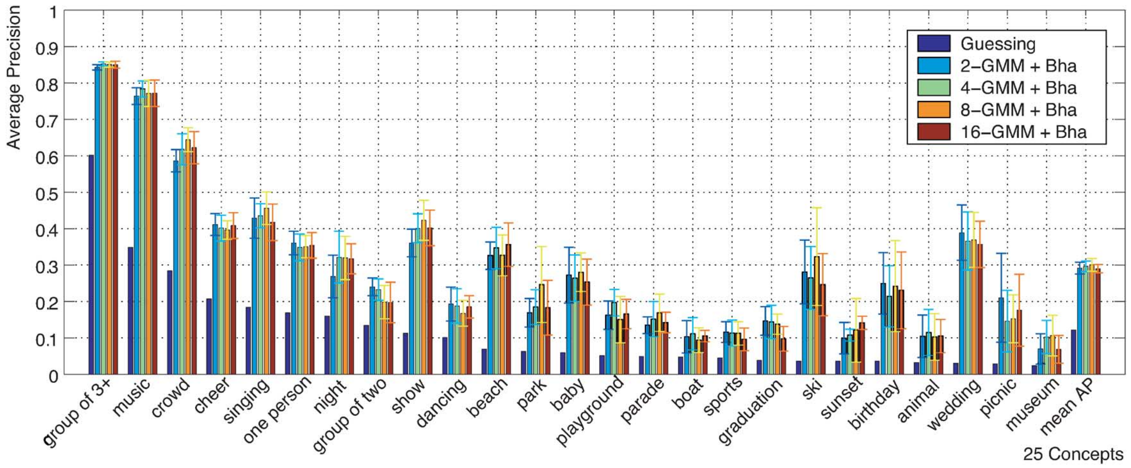


Fig. 5. As Fig. 4, but using GMMs with 2, 4, 8, and 16 components, and approximated Bhattacharyya distance.

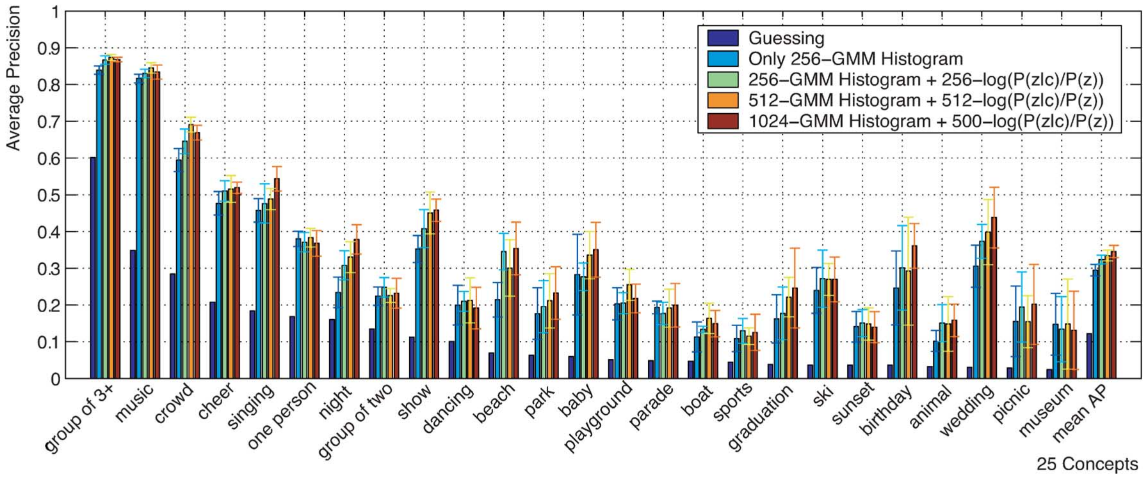


Fig. 6. As Fig. 4, but using pLSA modeling of component-use histograms for GMMs of 256, 512, and 1024 components. Also shown is performance using the 256 component histogram directly, without pLSA modeling.

the raw histogram or unnormalized weights. As we increase the number of Gaussian components used to build the histograms, we see increasing benefits for the less-frequent (lower-prior) concepts. The best performance is obtained by using around 500 topics to model component use within a 1024-mixture GMM.

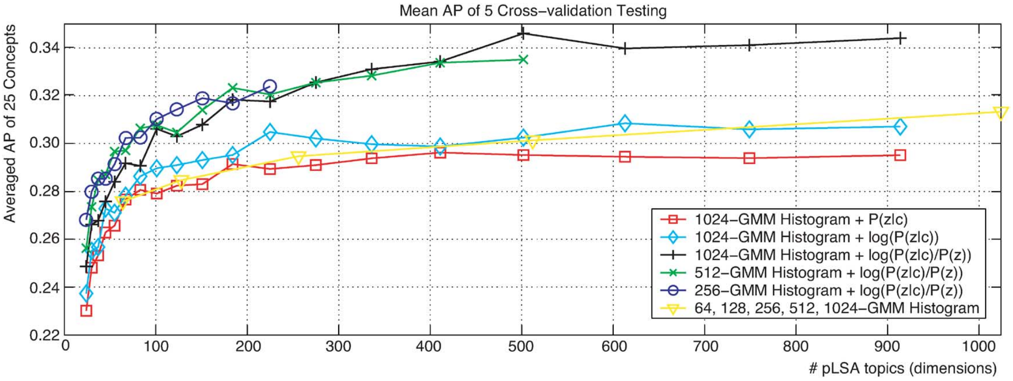


Fig. 7. AP averaged across all classes for pLSA models using different numbers of "topics" (latent dimensions) and different treatments for the inferred per-clip topic strengths, $p(z|c)$.

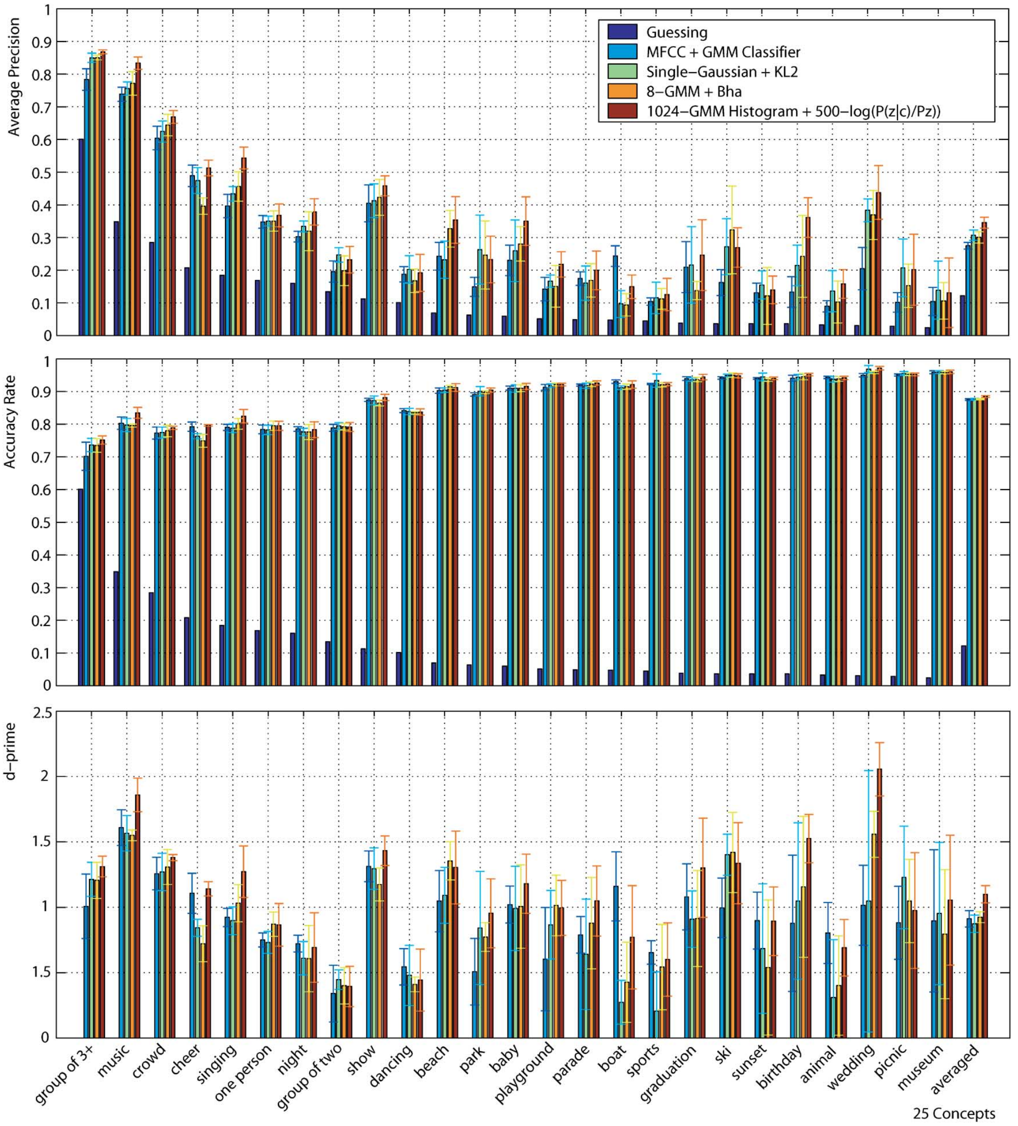


Fig. 8. Best results from Figs. 4, 5, and 6, illustrating the relative performance of each representation. Concepts are evaluated with average precision (AP), accuracy and d'.

V. Discussion

Fig. 8 compares the best results for each of the three modeling approaches, (1G+KL, 8-GMM+Bha, and pLSA-500+lognorm) along with a comparison system based on [15]. The comparison system builds an eight-component diagonal-covariance GMM for the MFCC features of clips bearing each label, and ranks items based on the likelihood under that GMM, i.e., it lacks the final SVM stage of the other systems. The figure compares the systems in terms of average precision (AP), accuracy rate, and

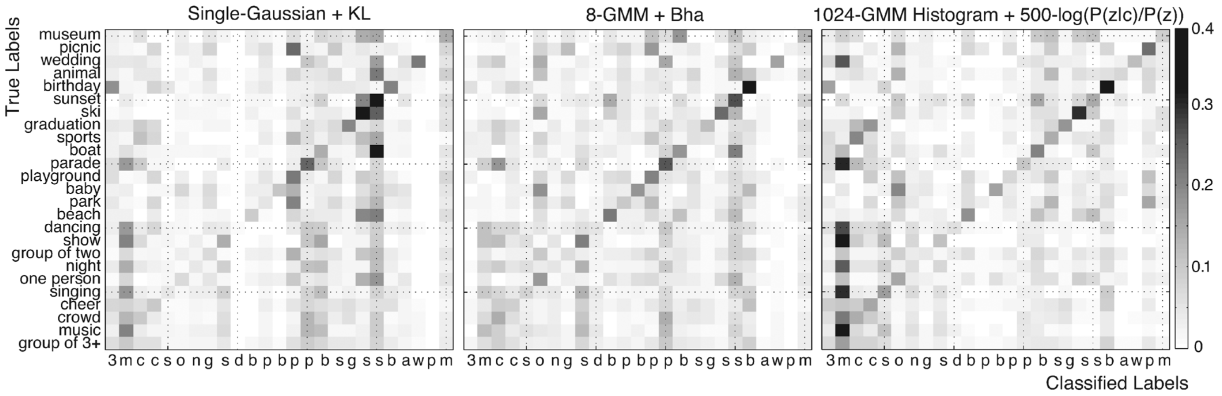


Fig. 9. Confusion matrix of classified labels within 750 testing clips according to three approaches.

d'. The AP is the Average of Precisions calculated separately for each true clip. The accuracy rate is the proportion of clips correctly labeled. d' is a threshold-independent measure of the separation between the two classes (presence and absence of the label) when mapped to two unit-variance Gaussian distributions. Note that accuracies can be made very high for concepts with small prior probabilities simply by labeling all clips as negative; d' and AP are less vulnerable to this bias. To obtain a hard classification (for accuracy and d' calculation) from our SVM-based rankings, we need to choose a threshold for the distance-to-boundary values. We set this threshold independently for each class at the point at which the number of positive classifications matched the prior of the class.

Most striking is the wide variation in performance by concept, which is to be expected since different labels will be more or less evident in the soundtrack as well as being supported by widely differing amounts of training data. Indeed, the main determinant of performance of these classifiers appears to be the prior likelihood of that label, suggesting that a large amount of training data is the most important ingredient for a successful classifier. This is, however, confounded by the correspondingly higher baseline. In some cases these factors may be distinguished: a less frequent concept "ski" has AP similar to that of the more frequent concept "beach," suggesting that it is more easily detected from the audio. However, the error bars, showing the range of variation across the fivefold cross validation, reveal that "ski" gives less consistent results, presumably because a smaller number of positive training examples will lead to greater variability among the different subsets of positive examples chosen in each fold to train the SVM classifier.

Some concepts consist of a few distinct, representative sounds that may be more successfully modeled by GMMs than by a single Gaussian. For instance, we have noticed that "beach" is mainly composed of two sound types, "wind" and "water" sounds; the AP for this concept is noticeably larger with the GMM than with 1G. This also suggests that performance could be improved by dividing some classes into more specific and hence more consistent subclasses (e.g., "animal" could be refined to "dog," "cat." etc).

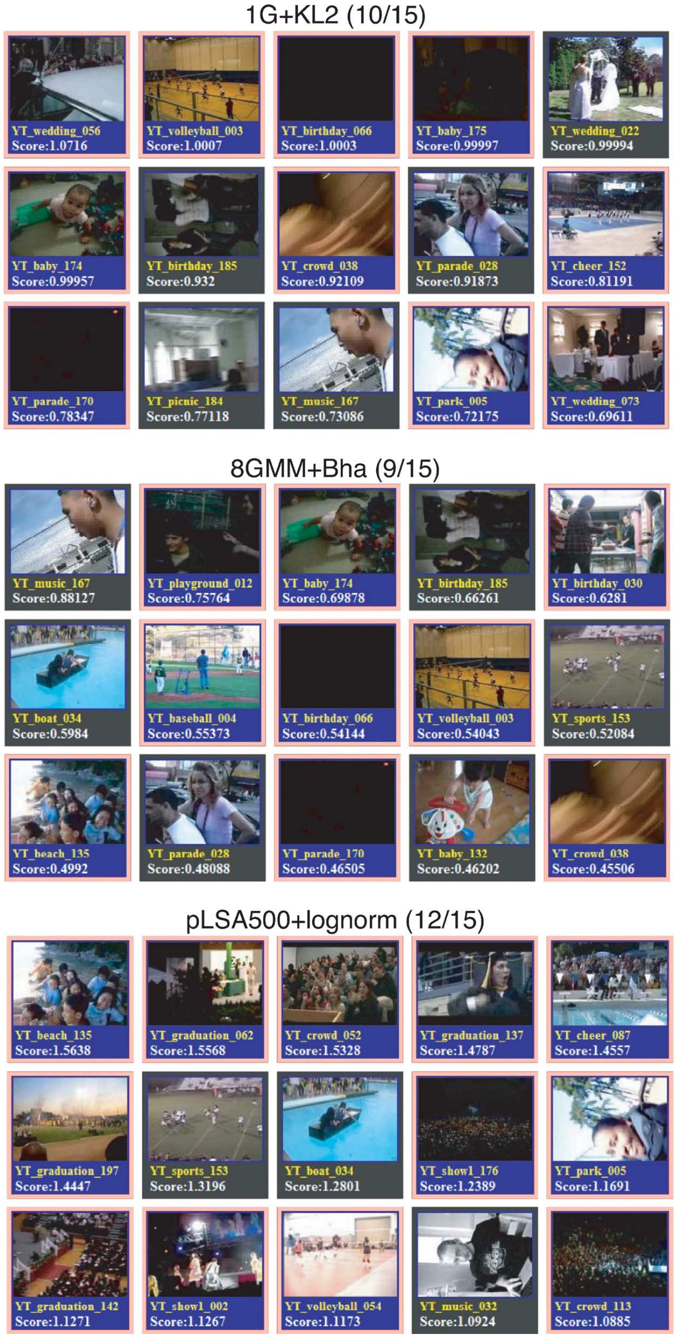
In addition, we have noticed that some concepts such as "cheer," "people," and "music" may be predominantly contained in other concepts such as "birthday," "sports," and "show." It is damaging to use such highly overlapped labels for SVM training with the 1G or GMM approaches because it is impossible to separate pure positive and negative segments at the scale of whole clips. The pLSA model is less sensitive to this problem, since it is able to represent the clip-level summary features directly as combinations of "topics," rather than trying to assign them to a single class. This may explain why its performance, averaged over all classes, appears superior to the other approaches.

Fig. 9 shows confusion matrices for each classification approach obtained by assigning each clip to the single class whose SVM gave the largest distance-to-margin, then looking at the distribution of labels assigned to all clips tagged with each specific class to obtain each row of the matrix. Because this approach does not allow the classifier to assign the multiple tags that each clip may bear, perfect performance is not possible and confusions may reflect label co-occurrence as well as imperfect classifiers. The 1G and GMM confusion patterns are more similar to each other than to the pLSA approach.

Fig. 10 gives example results for detecting the concept "cheer." Most "cheer" clips contain speech, music, and other background sounds that are more predominant than any cheering sound. On average, cheer sounds account for around 28% of the time within corresponding clips.

We have argued that pLSA is successful because it can represent soundtracks as mixtures of "topics" that may correspond to varying kinds of sounds within the overall soundtrack duration. To give greater insight, Fig. 11 shows the weights associated with each class for each of the anonymous topics for a 100 topic model based on 1024 component GMM occupancy histograms. While many pLSA topics are strongly identified with a single concept, many others make significant contributions to several classes, such as concepts 26 to 28 that occur in both "beach" and "sunset," or topics 96 and 97 that contribute to "park" and "picnic." The conjecture is that these topics correspond to the GMM states that cover the common sounds that occur in these classes; however, this needs to be confirmed by a closer examination of the time frames corresponding to the GMM states associated with these topics.

The pLSA approach gives consistently the best results. For instance, pLSA achieves a higher Average Precision than the next best approach (1G) for 18 out of the 25 categories; this is statistically significant under a binomial model. However, the margin of improvement is relatively small and might not be im-



1G+KL2 (10/15)

8GMM+Bha (9/15)

pLSA500+lognorm (12/15)

YT_wedding_056 Score:1.0716
YT_volleyball_003 Score:1.0007
YT_birthday_066 Score:1.0003
YT_baby_175 Score:0.99997
YT_wedding_022 Score:0.99994
YT_baby_174 Score:0.99957
YT_birthday_185 Score:0.932
YT_crowd_038 Score:0.92109
YT_parade_028 Score:0.91873
YT_cheer_152 Score:0.81191
YT_parade_170 Score:0.78347
YT_picnic_184 Score:0.77118
YT_music_167 Score:0.73086
YT_park_005 Score:0.72175
YT_wedding_073 Score:0.69611
YT_music_167 Score:0.88127
YT_playground_012 Score:0.75764
YT_baby_174 Score:0.69878
YT_birthday_185 Score:0.66261
YT_birthday_030 Score:0.6281
YT_boat_034 Score:0.5984
YT_baseball_004 Score:0.55373
YT_birthday_066 Score:0.54144
YT_volleyball_003 Score:0.54043
YT_sports_153 Score:0.52084
YT_beach_135 Score:0.4992
YT_parade_028 Score:0.48088
YT_parade_170 Score:0.46505
YT_baby_132 Score:0.46202
YT_crowd_038 Score:0.45506
YT_beach_135 Score:1.5638
YT_graduation_062 Score:1.5568
YT_crowd_052 Score:1.5328
YT_graduation_137 Score:1.4787
YT_cheer_087 Score:1.4557
YT_graduation_197 Score:1.4447
YT_sports_153 Score:1.3196
YT_boat_034 Score:1.2801
YT_show1_176 Score:1.2389
YT_park_005 Score:1.1691
YT_graduation_142 Score:1.1271
YT_show1_002 Score:1.1267
YT_volleyball_054 Score:1.1173
YT_music_032 Score:1.0924
YT_crowd_113 Score:1.0885

Fig. 10. Examples of retrieval results for the "cheer" concept. Shown are the top 15 results for each of the best-performing detection systems, 1G+KL2, 8GMM+Bha, and pLSA500+lognorm. Results that are correct according to manual labeling have pale borders. The proportion of correct results is shown in the heading for each pane.

portant in some applications. The baseline single-Gaussian, or likelihood-based GMM systems perform relatively well in comparison and are much simpler to construct and to evaluate. Thus, depending on the nature of the database and the value of the highest possible precision, these may be valid approaches. However, this pattern could change with larger training databases and needs to be reevaluated.

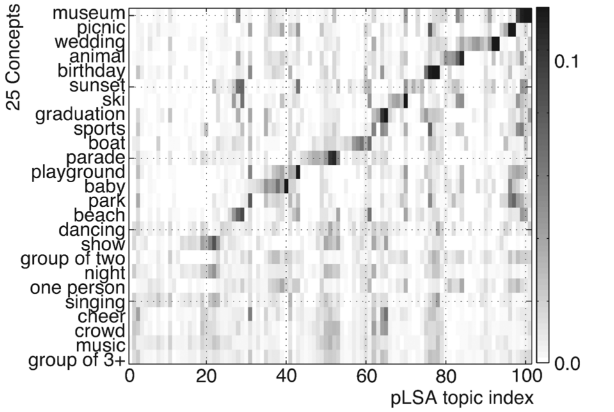


Fig. 11. Example pLSA topic weights (i.e., $p(z|c)$) across all concepts for a 100-topic model. Topic columns are sorted according to the concept for which they have the largest weight.

VI. Conclusion

In this paper, we have described several variants of a system for classifying consumer videos into a number of semantic concept classes, based on features derived from their soundtracks. Specifically, we have experimented with various techniques for summarizing low-level MFCC frames into fixed-size clip-level summary features, including single Gaussian models, Gaussian mixture models, and probabilistic latent semantic analysis of the Gaussian component histogram. We constructed SVM classifiers for each concept using the Kullback–Leibler, Bhattacharyya, and Mahalanobis distances. In spite of doubts over whether soundtrack features could be effective for identifying content classes with no obvious acoustic correlates such as "picnic" and "museum," we show that our classifiers are able to achieve APs far above chance, and in many cases at a level useful in real retrieval tasks.

Acknowledgment

Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Youtube—Broadcast Yourself, 2006. [Online]. Available: http://www.youtube.com

[2] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 1996, pp. 993–996.

[3] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, pp. 1331–1334.

[4] G. Williams and D. P. W. Ellis, "Speech/music discrimination based on posterior probability features," in *Proc. Eurospeech*, Budapest, Sep. 1999.

[5] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 441–457, Jul. 2001.

[6] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Commun.*, vol. 40, pp. 351–363, 2003.

[7] K. Lee and D. P. W. Ellis, "Detecting music in ambient audio by long-window autocorrelation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 9–12.

[8] E. Wold, T. Blum, and J. Wheaton, "Content-based classification, search and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[9] J. Foote, "Content-based retrieval of music and audio," in *Proc. SPIE*, 1997, vol. 3229, pp. 138–147.

[10] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 209–215, Jan. 2003.

[11] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *Proc. 1st ACM Workshop Continuous Archival and Retrieval of Personal Experiences*, New York, Oct. 2004 [Online]. Available: http://www.ee.columbia.edu/~dpwe/pubs/carpe04-minimpact.pdf

[12] R. G. Malkin and A. Waibel, "Classifying user environment for mobile applications using linear autoencoding of ambient audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Pittsburgh, PA, Mar. 2005, pp. 509–512.

[13] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–22, 2006.

[14] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Content analysis for acoustic environment classification in mobile robots," in *Proc. AAAI Fall Symp., Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Syst.*, 2006, pp. 16–21.

[15] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[16] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. Loui, and J. Luo, "Kodak consumer video benchmark data set: Concept definition and annotation," in *Proc. MIR Workshop, ACM Multimedia*, Germany, Sep. 2007.

[17] M. I. Mandel and D. P. W. Ellis, "A web-based game for collecting music metadata," *J. New Music Res.* vol. 37, no. 2, pp. 151–165, 2008 [Online]. Available: http://www.ee.columbia.edu/~dpwe/pubs/MandelE08-majorminer.pdf

[18] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. Loui, and J. Luo, "Large-scale multimodal semantic concept detection for consumer video," in *Proc. MIR Workshop, ACM Multimedia*, Germany, Sep. 2007.

[19] J. Shawe-Taylor and N. Cristianini*, Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[20] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *J. Mach. Learn. Res., JMLR, Special Topic on Learning Theory.*, pp. 819–844, 2004.

[21] D. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Orlando, FL, 2002, pp. 4072–4075.

[22] M. I. Mandel and D. P. W. Ellis, "Song-level features and support vector machines for music classification," in *Proc. Int. Conf. Music Inf. Retrieval ISMIR*, London, U.K., Sep. 2005, pp. 594–599 [Online]. Available: http://www.ee.columbia.edu/~dpwe/pubs/ismir05-svm.pdf

[23] J. Hershey and P. Olsen, "Variational bhattacharyya divergence for hidden Markov models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 4557–4560.

[24] F. Monay and D. Gatica-Perez, "Plsa-based image auto-annotation: Constraining the latent space," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, New York, Oct. 2004.

[25] R. Lienhart and M. Slaney, "pLSA on large scale image database," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Honolulu, HI, Apr. 2007, pp. 1217–1220.

[26] T. Hoffmann, "Probabilistic latent semantic indexing," in *Proc. 1999 Int. Conf. Res. Develop. Inf. Retrieval (SIGIR'99)*, Berkeley, CA, Aug. 1999.

[27] J. Arenas-Garcia, A. Meng, K. Petersen, T. Lehn-Schioler, L. Hansen, and J. Larsen, "Unveiling music structure via plsa similarity fusion," in *Proc. IEEE Workshop Mach. Learn. Signal Process.*, Thessaloniki, Aug. 2007, pp. 419–424.

Keansub Lee received the B.S. degree in electronics engineering from Kyung-Hee University, Seoul, Korea, in 1996, the M.S. degree in electrical engineering from Korea University, Seoul, in 2001, and the Ph.D. degree in electrical engineering from Columbia University, New York, in 2009.

He is currently a Postdoctoral Researcher at Prof. D. Ellis' Laboratory for Recognition and Organization of Speech and Audio (LabROSA). His research interests lie at the analyzing and indexing environmental sounds and soundtracks of consumer videos.

Daniel P. W. Ellis received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He is an Associate Professor in the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound. He worked at MIT, where he was a Research Assistant in the Media Lab, and he spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA.