

Large-Scale Multimodal Semantic Concept Detection for Consumer Video

Shih-Fu Chang¹, Dan Ellis¹, Wei Jiang¹, Keansub Lee¹, Akira Yanagawa¹, Alexander C. Loui², Jiebo Luo²

Columbia University, New York, NY
{sfchang, dpwe, wjiang, kslee,
akira}@ee.columbia.edu

Eastman Kodak Company
Rochester, NY
{Alexander.loui, Jiebo.luo}@kodak.com

ABSTRACT

In this paper we present a systematic study of automatic classification of consumer videos into a large set of diverse semantic concept classes, which have been carefully selected based on user studies and extensively annotated over 1300+ videos from real users. Our goals are to assess the state of the art of multimedia analytics (including both audio and visual analysis) in consumer video classification and to discover new research opportunities. We investigated several statistical approaches built upon global/local visual features, audio features, and audio-visual combinations. Three multi-modal fusion frameworks (ensemble, context fusion, and joint boosting) are also evaluated. Experiment results show that visual and audio models perform best for different sets of concepts. Both provide significant contributions to multimodal fusion, via expansion of the classifier pool for context fusion and the feature bases for feature sharing. The fused multimodal models are shown to significantly reduce the detection errors (compared to single modality models), resulting in a promising accuracy of 83% over diverse concepts. To the best of our knowledge, this is the first work on systematic investigation of multimodal classification using a large-scale ontology and realistic video corpus.

Categories and Subject Descriptors

Information Search and Retrieval; Multimedia Databases; Video Analysis

General Terms

Algorithms, Management, Performance

Keywords

Video classification, semantic classification, consumer video indexing, multimedia ontology

1. INTRODUCTION

With the explosive growth of user generated content, there has been tremendous interest in developing next-generation technologies for organizing and indexing multimedia content including photos, videos, and music. One of the major efforts in recent years involves automatic semantic classification of media content into a large number of predefined concepts that are both relevant to practical needs and amenable to automatic detection. The outcomes of such classification processes are high-level semantic descriptors, analogous to textual terms describing document content, and can be very useful for developing powerful retrieval or filtering systems for consumer media.

Large-scale semantic classification systems require several critical components. First, a large ontology is needed to define the list of important concepts and the relations among the concepts. Such ontologies may be constructed from the results of formal user studies or data mining of user interaction with online systems. Second, a large corpus consisting of realistic data are needed for training and testing automatic classifiers. An annotation process is also needed to obtain the concept labels of the defined concepts over the corpus. Third, signal processing and machine learning tools are needed to develop robust classifiers (also called models or concept detectors) that can be used to detect presence of each concept in any test data.

Recently, developments of such large-scale semantic classification systems have been reported for generic classes (e.g., car, airplane, flower) [17] and multimedia concepts in news videos [15]. In the consumer media domain, only limited efforts have been conducted to categorize consumer photos or videos into a small number of classes. In a companion paper [10], we have described a systematic effort to establish the first large-scale ontology and benchmark data set for consumer video classification. It consists of over 100 relevant and potentially detectable concepts, and annotation of 25 selected concepts over a set of 1338 consumer videos. The availability of such large ontology and rigorously annotated benchmark data set brings about a unique opportunity for evaluating state-of-the-art machine learning tools and multimedia analytics in automatic semantic classification.

In this paper, we present several novel statistical models and multimodal fusion frameworks for automatic audio-visual content classification. On the visual side, we investigate different approaches using both global and local features and ensemble fusion with multiple parameter sets. On the audio side, we develop techniques based on simple Gaussian models as well as advanced statistical methods such as probabilistic latent semantic analysis. One of our main goals is to understand the individual contributions of audio and visual models and find the optimal fusion strategies. To this end, we have developed and evaluated several fusion frameworks, ranging from simple weighted averaging, multimodal context fusion by boosted conditional random field, to multi-class joint boosting.

Through extensive experiments, we have demonstrated promising detection accuracy of the proposed classification methods, and more valuably, important insights about the contributions of individual algorithms and modalities in detecting a diverse set of semantic concepts. The multimodal multi-concept classification system is shown to reduce the detection errors by as much as 15%

(in terms of equal error rate) compared to alternatives using single modalities only. Audio models, though not as effective as the visual counterpart in terms of average performance, play an indispensable role – several concepts exclusively rely on the audio models and audio models provide significant contributions to the performance gains in model fusion.

We briefly review the ontology and semantic concepts for consumer videos in Sec. 2. Visual and audio models are described in Sec. 3 and 4 respectively. We present three multimodal fusion frameworks in Sec. 5. Extensive experiments for performance evaluation and discussion of results are included in Sec. 6.

2. SELECTION OF THE SEMANTIC CONCEPTS

Our research focuses on semantic concept detection over a collection of consumer videos, and an ontology of concepts derived from user studies, both originated at the Eastman Kodak company [10]. The videos were shot by about 100+ participants in a year-long user study, using the video mode of current-generation consumer digital cameras, which can capture videos of arbitrary duration at TV-quality resolution and frame rate. The full ontology of over 100 concepts was developed to cover real consumer needs as revealed by the studies. For our experiments, we further pared these down to 25 concepts that were simultaneously useful to users, practical both in terms of the anticipated viability of automatic detection and of annotator labeling, and sufficiently represented in the video collection. The concepts fall into several broad categories including activities (e.g. skiing, dancing), occasions (e.g. birthday, graduation), locations (e.g. beach, park), or particular objects in the scene (e.g. baby, boat, groups of three or more people). Most concepts were intrinsically visual, although some concepts, such as music and cheering, were primarily acoustic.

The Kodak video collection comprised over 1300 videos with an average length of 30 s. We had annotators label each video with each of the concepts; for most concepts, this was done on the basis of keyframes taken every 10 s, although some concepts (particularly the acoustic ones) relied on watching and hearing the full video. This resulted in labels for 5166 keyframes.

We also experimented with gathering additional data from the video sharing site YouTube. Using each of our concept terms as a query, we downloaded several hundred videos for each concept. We then manually filtered these results to discard videos that were not consistent with the consumer video genre (e.g. edited or broadcast content), resulting in 1874 videos with an average duration of 145 s. The YouTube videos were then manually relabeled with the 25 concepts, but only at the level of entire videos instead of keyframes.

More details on the video collections and labels are provided in a companion paper [10].

3. VISUAL-BASED DETECTORS

We first define some terminology. Let C_1, \dots, C_M denote M semantic concepts we want to detect, and let \mathcal{D} denote the set of training data $\{(\mathbf{I}, \mathbf{y}_I)\}$. Each \mathbf{I} is an image and the

corresponding $\mathbf{y}_I = \{y_I^1, \dots, y_I^M\}$ is the vector of concept labels, where $y_I^i = +1$ or -1 denotes, respectively, the presence or absence of concept C_i in image \mathbf{I} .

3.1 Global Visual Features & Baseline Models

The visual baseline model uses three attributes of color images: texture, color and edge. Specifically, three types of global visual features are extracted: *Gabor texture (GBR)*, *Grid Color Moment (GCM)*, and *Edge Direction Histogram (EDH)*. These features have been shown effective and efficient in detecting generic concepts in several previous works [2], [3], [15]. The GBR feature is used to estimate the image properties related to structures and smoothness; GCM approximates the color distribution over different spatial areas; and EDH is used to capture the salient geometric cues like lines. A detailed description of these features can be found in [16].

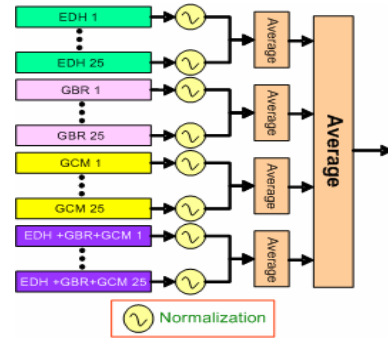


Figure 1: The workflow of the visual baseline detector.

Based on these global visual features, two types of support vector machine (SVM) classifiers are learned for detecting each concept: (1) one SVM classifier is trained over each of the three features individually; and (2) these features are concatenated into one feature vector over which a SVM classifier is trained. Then the detection scores from all different SVM classifiers are averaged to generate the baseline visual-based concept detector.

The SVMs are implemented using LIBSVM (Version 2.81) [1] with the RBF kernel. For learning each SVM classifier, we need to determine the parameter setting for both the RBF kernel (γ) and the SVM model (C) [1]. Here we employ a *multi-parameter set model* instead of cross-validation so that we can reduce the degradation of performance in the case that the distribution of the validation set is different from the distribution of the test set. Instead of choosing the best parameter set from cross-validation, we average the scores from the SVM models with 25 different sets of parameters C and γ :

$$C = \{2^0, 2^2, 2^4, 2^6, 2^8\}, \gamma = \{2^{k-4}, 2^{k-2}, 2^k, 2^{k+2}, 2^{k+4}\},$$

where $k = \text{ROUND}(\log_2(1/|D_f|))$ and $|D_f|$ is the dimensionality of the feature vector based on which the SVM classifier is built ($\gamma = 2^k$ is the recommend parameter in [1]). The multi-parameter set approach is applied to each of the three features mentioned above, as well as the aggregate feature, as shown in

Fig. 1. Note the scores (i.e., distances to the SVM decision boundary) generated by each SVM are normalized before averaging. Various normalization strategies are described in Sec. 5.1.

3.2 Visual Models Using Local Features

Complementary to the global visual features, local descriptors such as *SIFT* features [11] have been shown very useful for detecting specific objects. Recently, an effective *bag-of-features* (*BOF*) representation [4] has been proposed for image classification. In BOF images are represented by a *visual vocabulary* constructed by clustering the original SIFT descriptors into a set of *visual tokens*. BOF provides a uniform middle-level representation through which the original orderless SIFT descriptors of an image can be mapped to a feature vector, and based on this feature vector the learning-based algorithms, such as the SVM classifier, can be applied for concept detection. Lately, using the BOF representation, the *Spatial Pyramid Matching* (*SPM*) approach [9] and the *Vocabulary-Spatial Pyramid Matching* (*VSPM*) approach [7] have been developed to fuse information from multiple resolutions in the spatial domain and multiple visual vocabularies of different granularities. Promising performance has been obtained for detecting generic concepts like bike and person. In this work, we experimented with the VSPM approach [7] to investigate the power of the local SIFT features in detecting diverse concepts in the consumer domain.

3.2.1 Local SIFT Descriptor

The 128-dimensional SIFT feature proposed in [11] has been proven effective in detecting objects, because it is designed to be invariant to relatively small spatial shift of region positions, which often occurs in real images. Computing the SIFT descriptor over the affine covariant regions results in local description vectors which are invariant to affine transformations of the image. In this work, instead of computing SIFT features over the detected interest points as in the traditional feature extraction algorithms [11], we extract SIFT features for every image patch with 16x16 pixels over a grid with spacing of 8 pixels as in [9]. This dense sampling method has been shown more effective in detecting generic concepts [9] than the traditional method using selected interest points only.

3.2.2 Vocabulary-Spatial Pyramid Match Kernel

For each concept C_i , the SIFT features from all the positive training images for this concept are first aggregated together, and through hierarchical clustering these SIFT features are clustered into $L+1$ sets of clusters $\mathcal{V}_i^0, \dots, \mathcal{V}_i^L$ with level 0 being the coarsest and level L the finest. \mathcal{V}_i^l represents a visual vocabulary comprised of n_l visual tokens $\mathcal{V}_i^l = \{v_{i,1}^l, \dots, v_{i,n_l}^l\}$. The visual vocabularies are expected to include the most informative visual descriptors that are characteristic of images sharing the same concept.

Given the visual vocabulary at each level \mathcal{V}_i^l , the local features of an image are mapped to tokens in the vocabulary and counts of tokens are computed to form a token histogram

$H_i^l(\mathbf{I}) = [h_{i,1}^l(\mathbf{I}), \dots, h_{i,n_l}^l(\mathbf{I})]$. In the *Spatial Pyramid Match Kernel* (*SPMK*) method, each image is further decomposed into 4^s blocks in a hierarchical way ($s = 0, \dots, S$), with a separate token histogram $H_{i,k}^{l,s}(\mathbf{I})$ associated with each spatial block.

To compute matches between two images \mathbf{I}_p and \mathbf{I}_q , histogram intersection is used.

$$\mathcal{M}_i^{l,s}(\mathbf{I}_p, \mathbf{I}_q) = \sum_{k=1}^{4^s} \sum_{j=1}^{n_l} \min\{h_{i,k,j}^{l,s}(\mathbf{I}_p), h_{i,k,j}^{l,s}(\mathbf{I}_q)\}.$$

The final vocabulary-spatial pyramid match kernel defined by vocabulary \mathcal{V}_i^l is given by weighted sum of matches at different spatial levels:

$$\mathcal{K}_i^l(\mathbf{I}_p, \mathbf{I}_q) = \frac{\mathcal{M}_i^{l,0}(\mathbf{I}_p, \mathbf{I}_q)}{2^s} + \sum_{s=1}^S \frac{\mathcal{M}_i^{l,s}(\mathbf{I}_p, \mathbf{I}_q)}{2^{s-s+1}}.$$

The above measure is used to construct a kernel matrix, whose elements represent similarities (or distances) between all pairs of training images (including both positive and negative samples) for concept C_i . Images coming from C_i are likely to share common

visual tokens in \mathcal{V}_i^l and thus have high matching scores in the kernel matrix. The process of constructing VSPM kernels for multi-level vocabularies is illustrated in Fig. 2. The VSPM kernels provide important complementary visual cues to the global visual features and are utilized in two ways for concept detection: (1) For each individual concept C_i , the VSPM kernels $\mathcal{K}_i^0, \dots, \mathcal{K}_i^L$ are combined with weights into an ensemble kernel:

$$\mathcal{K}_i^{ensemble} = \sum_{l=0}^L w_l \mathcal{K}_i^l,$$

where weights w_l can be heuristically determined in a way similar to [6] or optimized through experimental validation. Then the ensemble kernel is directly used for learning a one-vs.-all SVM classifier for detection of concept C_i ; (2) VSPM kernels from different concepts are shared among different concept detectors through a joint boosting framework which will be described in detail in Section 5.3.

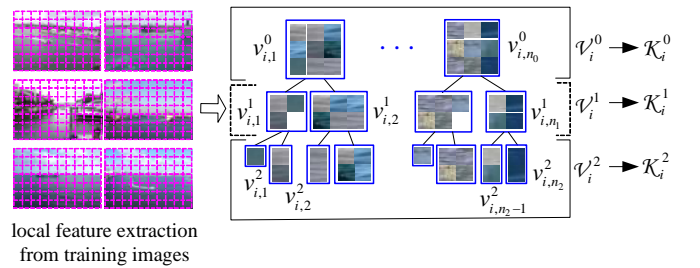


Figure 2: Illustration of the kernel construction process used in the Vocabulary-Spatial Pyramid Match (VSPM) model.

4. AUDIO-BASED DETECTOR

The soundtracks of each video are described and classified by two techniques, single Gaussian modeling, and probabilistic latent

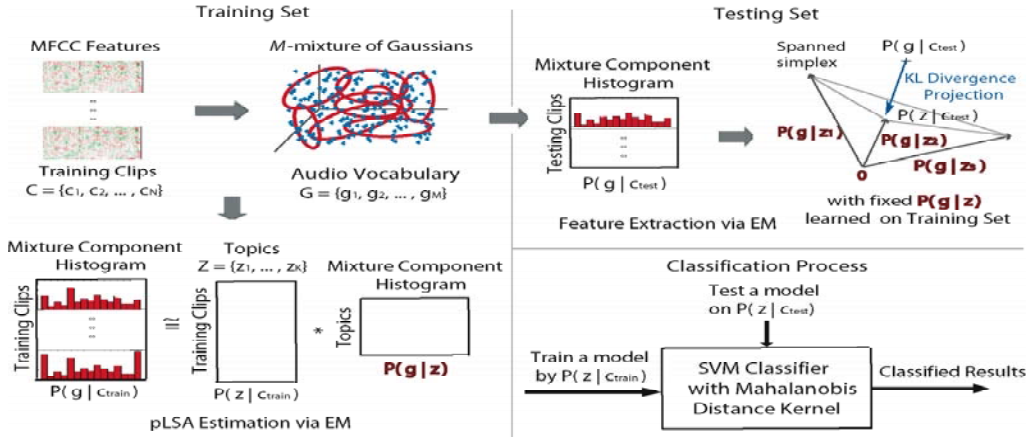


Figure 3: Illustration of the calculation of audio features as the pLSA weights describing the histogram of GMM component utilizations. Top left shows the formation of the global GMM; bottom left shows the formation of the topic profiles, $p(g|z)$; top right shows the analysis of each clip into topic weights by matching each histogram to a combination of topic profiles, and bottom left shows the final classification by SVM.

semantic analysis (pLSA) [18] of Gaussian mixture model (GMM) component occupancy histograms, both described below. All systems start with the same basic representation of the audio, as 25 Mel-frequency Cepstral Coefficients (MFCCs) extracted from frequencies up to 7 kHz over 25 ms frames every 10 ms. Since each video has a different duration, it will result in a different number of feature vectors; these are collapsed into a single clip-level feature vector by the two techniques described below. Finally, these fixed-size summary features are compared to one another, and this matrix of distances (comparing positive examples with a similar number of randomly-chosen negative examples) is used to train a SVM classifier for each concept. The distance-to-boundary values from the SVM are taken to indicate the strength of relevance of the video to the concept, either for direct ranking or to feed into the fusion model.

4.1 Single Gaussian Modeling

After the initial MFCC analysis, each soundtrack is represented as a set of $d = 25$ dimensional feature vectors, where the total number depends on the length of the original video. (In some experiments we augmented this with 25 dimensions of ‘delta MFCCs’ giving the local time-derivative of each component, which slightly improved results.) To describe the entire dataset in a single feature vector, we ignore the time dimension and treat the set as samples from a distribution in the MFCC feature space, which we fit with a single 25-dimensional Gaussian by measuring the mean and (full) covariance matrix of the data. This approach is based on common practice in speaker recognition and music genre identification, where the distribution of cepstral features, ignoring time, is found to be a good basis for classification.

To calculate the distance between two distributions, as required for the gram-matrix input (kernel matrix as defined in Sec. 3.2) to the SVM, we have tried two approaches. One is to use the Kullback-Leibler (KL) divergence between the two Gaussians, namely, if video clip i has a set of MFCC features denoted X_i , described by mean vector μ_i and covariance matrix Σ_i , then the KL distance between videos i and j is:

$$D_{KL}(X_i, X_j)^2 = (\mu_i - \mu_j)^T (\Sigma_i^{-1} + \Sigma_j^{-1}) (\mu_i - \mu_j) + \text{tr}(\Sigma_i^{-1} \Sigma_j + \Sigma_j^{-1} \Sigma_i) - 2d$$

The second approach simply treats the d -dimensional mean vector μ_i concatenated with the $d(d+1)/2$ unique values of the covariance matrices Σ_i as a point in a new $(25+325)$ dimensional feature space, normalizes each dimension by its standard deviation across the entire training set, then builds a gram matrix from the Euclidean distance between these normalized feature statistic vectors.

4.2 Probabilistic Latent Semantic Analysis

The Gaussian modeling assumes that different activities are associated with different sounds whose average spectral shape, as calculated by the cepstral feature statistics, will be sufficient to discriminate categories. However, a more realistic assumption is that each soundtrack will consist of many different sounds that may occur in different proportions even for the same category, leading to variation 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 particular classes of video.

To build a model better able to capture this idea, we first trained a large Gaussian mixture model, comprising $M = 256$ Gaussian components, on a subset of MFCC frames chosen randomly from the entire training set. (The number of mixtures was optimized in pilot experiments.) These 256 mixtures are considered as anonymous sound classes from which each individual soundtrack is assembled – the analogues of words in document modeling. Then, we classify every MFCC frame in a given soundtrack to one of the mixture components, and describe the overall soundtrack with a histogram of how often each of the 256 Gaussians was chosen when quantizing the original representation. Note that this representation also ignores temporal structure, but it

is able to distinguish between nearby points in cepstral space, depending on how densely that part of feature space is represented in the entire database, and thus how many Gaussian components it received in the original model. The idea of using histograms of acoustic tokens to represent the entire soundtrack is also similar to that in using visual token histograms for image representation (Sec. 3.2).

We could use this histogram directly, but to remove redundant structure and to give a more compact description, we go on to explain the histogram with probabilistic Latent Semantic Analysis (pLSA) [18]. This approach, originally developed to generalize the distributions of individual words in documents on different topics, 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. During training, the topic definitions are driven to a local optimum by using the EM algorithm. Specifically, the histogram representation gives the probability $p(g|c)$ that a particular component, g , will be used in clip c as the sum of the distribution of components for topic z , $p(g|z)$, weighted by the specific contributions of each topic to clip c , $p(z|c)$, i.e.

$$p(g|c) = \sum_z p(g|z)p(z|c)$$

The topic profiles $p(g|z)$ (which are shared between all clips), and the per-clip topic weights $p(z|c)$, are optimized by EM. 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, and found that around 160 topics was the best number for our task. Representing a test item similarly involves finding the best set of weights to match the observed histogram as a combination of the topic profiles; we match in the sense of minimizing the KL distance, which requires an iterative solution. Finally, each clip is represented by its vector of topic weights, and the SVM’s gram matrix (referred to as kernel \mathcal{K}_{audio} in Section 5.3) is calculated as the Mahalanobis (i.e. covariance-normalized Euclidean) distance in that 160-dimensional space. The process of pLSA feature extraction is illustrated in Fig. 3.

5. FUSION OF AUDIO-VISUAL FEATURES AND MODELS

Semantic concepts are usually defined by both visual and audio characteristics. For example, “dancing” is usually accompanied with background “music”. It can be expected that by combining the audio and visual features and corresponding models, better performance can be obtained than using any single modality. In the section, we develop three fusion strategies for combining audio and visual features and models.

5.1 Ensemble Fusion

One intuitive strategy to fuse the audio-based and visual-based detection results is *ensemble fusion*, which typically combines independent detection scores by weighted sum along with some normalization procedures to adjust the raw scores before fusion.

For normalization, we utilize z-score Eqn.(1), sigmoid Eqn.(2), and sigmoid after normalization with z-score (sigmoid2) Eqn.(3).

$$f(x) = (x - \mu) / \sigma \quad (1)$$

$$f(x) = 1 / [1 + \exp(-x)] \quad (2)$$

$$f(x) = 1 / [1 + \exp(-v)], \quad v = (x - \mu) / \sigma \quad (3)$$

where x is the raw score, μ and σ are mean and standard deviation respectively.

Such ensemble fusion method has been applied to combining the SVM models using different parameters and features (as illustrated in Fig. 1). Here, we extend the fusion process to include audio models, using optimal weights that are determined by maximizing the performance of the fused model over a separate validation data set. The cross-modal fusion architecture is shown in Fig. 4.

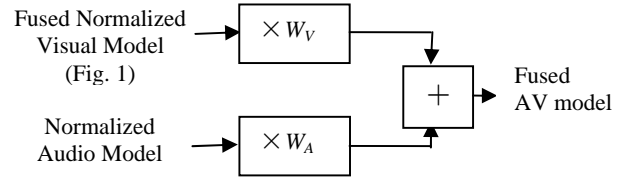


Figure 4: Ensemble fusion of audio and visual models.

5.2 Audio-Visual BCRF (AVBCRF)

In all of the approaches mentioned above, each concept is detected independently from each other in the one-vs.-all manner. However, semantic concepts do not occur in isolation -- knowing the information about certain concepts (e.g. “person”) of an image is expected to help detection of other concepts (e.g. “wedding”). Based on this idea, in the following two subsections, we propose to use context-based concept detection methods for multimodal fusion by taking into account the inter-conceptual relationships. Specifically, two algorithms are developed under two different fusion frameworks: (1) an *Audio-Visual Boosted Conditional Random Field (AVBCRF)* method where a two-stage *Context-Based Concept Fusion (CBCF)* framework is utilized; (2) an *Audio-Visual Joint Boosting (AVJB)* algorithm where both audio-based and visual-based kernels are combined to train multi-class concept detectors jointly. The former can be categorized as late fusion since it combines prediction results from models that have been trained separately. On the contrary, the latter is considered as an early fusion approach as it utilizes kernels derived from individual concepts in order to learn joint models for detecting multiple concepts simultaneously. In addition, on the visual side, CBCF fuses baseline models using global features, while AVJB further explores the potential benefits of local visual features. We will introduce AVBCRF in this subsection, and the AVJB algorithm will be described in the next subsection.

The *Boosted Conditional Random Field (BCRF)* algorithm is proposed in [8] as an efficient context-based fusion method for improving concept detection performance. Specifically, the relationships between different concepts are modeled by a *Conditional Random Field (CRF)*, where each node represents a concept and the edges between nodes represent the pairwise

relationships between concepts. This BCRF algorithm has a two-layer framework (as shown in Fig. 5). In the first layer, independent visual-based concept detectors are applied to get a set of initial posterior probabilities of concept labels on a given image. Then in the second layer the detection results of each individual concept are updated through a context-based model by considering the detection confidence of the other concepts. Here we extend BCRF to include models using both visual and audio modalities.

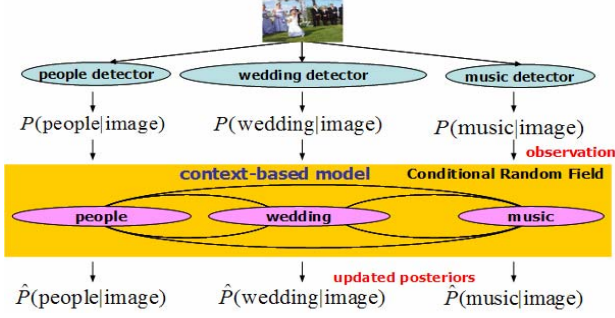


Figure 5: The context-based concept fusion framework based on Boosted Conditional Random Field.

For each image \mathbf{I} , the input observations are the initial posterior probabilities $\mathbf{h}_1 = [\mathbf{h}_{vis,1}, \mathbf{h}_{aio,1}]$, including the visual-based independent detection results $\mathbf{h}_{vis,1} = \{h_{vis,1}^1, \dots, h_{vis,1}^M\}$ as well as the audio-based independent detection results $\mathbf{h}_{aio,1} = \{h_{aio,1}^1, \dots, h_{aio,1}^M\}$. Then these inputs are fed into the CRF to get the improved posterior probabilities $P(\mathbf{y}_1 | \mathbf{I})$ through inference based on the inter-conceptual relationships. After inference the belief b_1^i on each node C_i is used to approximate the posterior probability: $P(y_1^i = \pm 1 | \mathbf{I}) \approx b_1^i(\pm 1)$. The aim of CRF modeling is to minimize the total loss J for all concepts over all the training data (D):

$$J = -\prod_{\mathbf{I} \in D} \prod_{i=1}^M b_1^i(+1)^{(1+y_1^i)/2} b_1^i(-1)^{(1-y_1^i)/2}. \quad (4)$$

Eqn.(4) is an intuitive function: the minimizer of J favors those posteriors closest to training labels.

To avoid the difficulty of designing potential functions in CRF, the Boosted CRF framework developed in [14] is incorporated and generalized to optimize the logarithm of Eqn.(4):

$$\arg \min_{b_1^i} \{\log J\} = \arg \min_{F_1^i, G_1^i} \left\{ \sum_{\mathbf{I} \in D} \sum_{i=1}^M \log e^{-y_1^i (F_1^i + G_1^i)/2} \right\} \quad (5)$$

in an iterative boosting process by finding the optimal F_1^i and G_1^i , where F_1^i and G_1^i are additive models:

$$F_1^i(T) = \sum_{t=1}^T f_1^i(t), \quad G_1^i(T) = \sum_{t=1}^T g_1^i(t).$$

$f_1^i(t)$ is a discriminant function (e.g. SVM or logistic) with input \mathbf{h}_1 as the feature, and $g_1^i(t)$ is a discriminant function (e.g. SVM in our algorithm) with the current belief $b_1^i(t)$ as the feature in iteration t . Both $f_1^i(t)$ and $g_1^i(t)$ can be considered weak classifiers learned by the standard boosting procedure, but over

different features. The contributions from other concept scores to detection of a specific concept are explored in each iteration since the whole set of concept detection scores are used as input to the classifiers in each iteration. More details about the formula derivation can be found in [8], [14].

5.3 Audio-Visual Joint Boosting (AVJB)

In this section, we will introduce a systematic early fusion framework to combine the audio-based and visual-based features/kernels for training multi-class concept detectors. Instead of training independent detectors based on visual features and audio features separately, the visual features/kernels and audio features/kernels can be used together to learn concept detectors at the first place. To this end, we adopt the joint boosting and kernel sharing framework developed in [7] which utilizes a two-stage framework: (1) the kernel construction stage; and (2) the kernel selection and sharing stage. In the first stage, concept-specific features/kernels such as the VSPM kernels described in Sec. 3.2.2, are constructed to capture the most representative characteristics of the visual content for each concept individually. Note local visual features (e.g., SIFT-based visual tokens) are used here. Then in the second stage, these kernels are shared by different concepts through a joint boosting algorithm which can automatically select the optimal kernels from the kernel pool to learn a multi-class concept detector jointly. This two-stage framework can be directly generalized to incorporate audio-based kernels. That is, in the first stage, based on acoustic analysis various features/kernels can be constructed (such as the audio vocabulary and kernel described in Sec. 4.2), and these kernels can be added into the rich kernel pool together with all the visual-based kernels, and in the second stage the optimal subset of kernels are selected and shared through the joint boosting learning algorithm.

The process of joint boosting is illustrated in Fig. 6. By sharing good kernels among different concept detectors, individual concepts can be enhanced by incorporating the descriptive power from other concepts. Also by sharing the common detectors among concepts, required kernels and training samples for detecting individual concepts will be reduced [7], [13].

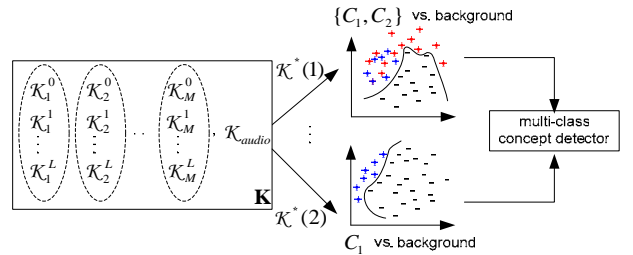


Figure 6: Illustration of kernel and classifier sharing using joint boosting. A kernel pool \mathbf{K} is shared by different detectors. First, using kernel $\mathcal{K}^*(1)$ a binary classifier is used to separate C_1 and C_2 from the background. Then using $\mathcal{K}^*(2)$ a binary classifier further picks out C_1 .

In Section 3.2.2 we obtained $L+1$ concept-specific VSPM kernels $\mathcal{K}_i^0, \dots, \mathcal{K}_i^L$ for each concept C_i corresponding to the multi-resolution visual vocabularies $\mathcal{V}_i^0, \dots, \mathcal{V}_i^L$. In addition, in Section

4.2 we have the audio-based kernel \mathcal{K}_{audio} . Then the joint boosting framework from [7] can be directly adopted here for sharing visual and audio based kernels for concept detection. Specifically, during each iteration t , we select the optimal kernel $\mathcal{K}^*(t)$ and the optimal subset of concepts $\mathcal{S}^*(t)$ to share the optimal kernel. Then a binary classifier is trained using kernel $\mathcal{K}^*(t)$ which tries to separate concepts in subset $\mathcal{S}^*(t)$ from the background (for the other concepts not in $\mathcal{S}^*(t)$, a prediction $k_c^i(t)$ is given based on the prior). After that, we calculate the training error of this binary classifier and re-weight the training samples similar to the Real AdaBoost algorithm. Finally all weak classifiers from all iterations are fused together to generate the multi-class concept detector.

6. EXPERIMENTS

In this section, we evaluate the performance of features, models, and fusion methods described earlier. We conduct extensive experiments using the Kodak benchmark video set described in Section 1. Among the 25 concepts annotated over the video set, we use 21 visual-dominated concepts to evaluate the performance of visual methods and impact of incorporating additional methods based on audio features. Audio-based methods are also evaluated by using three additional audio-dominated concepts (singing, music, and cheer). In the discussion following each experiment, we highlight main findings and important insights in *italic text*.

6.1 Experimental Setup & Performance Metrics

Each concept detection algorithm is evaluated in five runs and the average performances over all runs are reported. The data sets in the runs are generated as follows: the entire data set \mathcal{D} is randomly split to 5 subsets $\mathcal{D}_1, \dots, \mathcal{D}_5$. By rotating these 5 subsets, we generate the training set, validation set, and test set for each run. That is, for run 1, training set = $\{\mathcal{D}_1, \mathcal{D}_2\}$, validation set = \mathcal{D}_3 , test set = $\{\mathcal{D}_4, \mathcal{D}_5\}$. Then we switch one subset for run 2, where training set = $\{\mathcal{D}_2, \mathcal{D}_3\}$, validation set = \mathcal{D}_4 , test set = $\{\mathcal{D}_5, \mathcal{D}_1\}$. Similarly, we can keep switching to generate the data sets for run 3, run 4, and run 5. For each run, all algorithms are trained over the training set and evaluated over the test set, except for the AVBCRF algorithm in which the validation set is used to learn the joint boosting model that fuses individual detectors learned using the training set separately.

The average precision (AP) and mean average precision (MAP) are used as performance metrics. AP is related to multi-point average precision value of a precision-recall curve. AP is an official performance metric used by TRECVID [12]. To calculate AP for concept C_i we first rank the test data according to the classification posteriors of concept C_i . Then from top to bottom, the precision after each positive sample is calculated. These precision values are averaged over the total number of positive samples for C_i . AP favors highly ranked positive samples and combines precision and recall values in a balanced way. MAP is the average of per-concept APs across all concepts. To help readers compare performance, in some cases, we also report the detection accuracy based on Equal Error Rate (EER).

6.2 Performance Comparison and Discussions

6.2.1 Baseline Approaches

Visual Baseline

First, we evaluate the visual baseline detector with multiple parameter sets described in Sec. 3.1. For score normalization, we used sigmoid which was shown to outperform other options. Fig. 7 shows the performance when different numbers of SVMs with distinct parameter settings are fused. “Top(n)” denotes the fused model that computes average of detection scores from n detectors that achieve top performance over the validation set. The objective here is to study the effect of varying the number of models during ensemble fusion. Intuitively, the more models used in fusion the more stable the fused performance will be when testing over unseen data set. Such conjecture has been confirmed in our experiments – Top25 gives the best MAP performance as well as good APs over different concepts. On the other hand, APs of Top1 are not stable across different concepts and the MAP is the worse among all compared methods. This indicates that in our data sets the distribution of the validation set is quite different from that of the test set, and the conventional method optimizing a single set of parameters by cross-validation suffers from over fitting. In comparison, the multi-parameter set model can get relatively stable performance in such case. Based on this observation, in the following experiments, the “Top25” results are used and referred to as the visual-based baseline detection results. Fig. 7 also shows the AP of random guess, which is proportional to the number of positive samples of each concept.

From the above results, we found that *in general frequent concepts enjoy higher detection accuracy. However, other factors such as concept definition specificity and content consistency are also important*. For example, concepts like “sunset”, “parade”, “sports”, “beach” and “boat”, though infrequent (# of positive samples < 100), can be detected with high accuracy. On the other hand, some frequent concepts like “group of 3” and “one person” have much lower accuracy. This confirms that *careful choices and definitions of concepts play a critical role in developing robust semantic classification systems*.

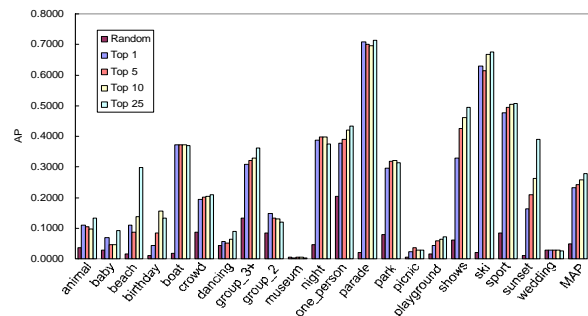


Figure 7: Performance of visual baseline detectors fusing varying numbers of models with different parameter sets

Audio Baseline

Fig. 8 shows the results of the three different audio-based approaches (single Gaussians with either KL or Mahalanobis distance measure, or the pLSA modeling of GMM component histograms). We see that all three approaches perform roughly

the same, with different models doing best for individual concepts. There is also a wide variation in performance depending on the concept, which is to be expected since different labels will be more or less evident in the soundtrack. However, *the main determinant of performance of audio-based 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.* For example, although the infrequent classes “wedding”, “museum”, and “parade” have APs similar to more common classes “cheer” and “one person”, their variation is much larger among the 5-fold cross-validation. Such a relationship between the frequency and the performance variance was also found in the visual detectors. Though not shown in Fig. 7 (due to space limit in the graph), the infrequent concepts (“boat”, “parade”, and “ski”) have accuracy similar to common concepts (“one person”, “shows”, and “sports”), but much larger performance variance among cross validation. Since different approaches have similar performances, in the following experiments, the single Gaussian with KL distance measure is used as the audio-based baseline detector.

Since most of the selected concepts are dominated by the visual cues, the results show *the visual-based models as expected achieve higher accuracy than the audio models for most concepts.* However, audio models also provide significant benefits. For example, concepts like “music”, “singing”, and “cheer” can be detected by audio models only due to the nature of the concepts. Even for some visually dominated concepts (like “museum” and “animal”), audio methods were found to be more reliable than visual counterparts. The soundtracks of video clips from these concepts provide rather consistent audio features for classification. This also suggests these two concepts may need to be refined to be more specific so that the corresponding visual content may be more consistent (e.g., “animal” refined to “dog” and “cat” etc).

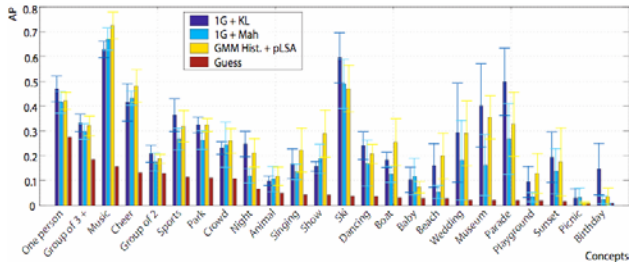


Figure 8: Performance of audio-based classifiers on Kodak data using MFCC+delta-MFCC base features. Labels are sorted by prior probability (guessing). Error bars indicate standard deviation over 5-fold cross-validation testing.

6.2.2 Audio-Visual Fusion Approaches

Ensemble Fusion

We evaluate different normalization strategies used in ensemble fusion described in Section 5.1. Specifically, we compare normalization methods based on z-score, sigmoid, or sigmoid 2 (i.e., z-score followed by sigmoid). Additionally, we test two different score fusion methods – uniform average and weighted average.

We found uniform averaging between audio and visual baseline models does not perform as well as visual models alone. This is

reasonable as most of the selected concepts have stronger cues from visual appearances than audio attributes; thus equal weighting is not expected to be the best option. This is indeed confirmed in results shown in Fig. 9, which compares weighted audio-visual combination with different normalization strategies. Among different score normalization strategies, the z-score method performs best, outperforming the visual-only model by 4% in MAP. The improvement is especially significant for several concepts, “dance”, “parade” and “show”, with 6% - 24% gains in terms of AP. Note the optimal weights for combining audio and visual models are determined through validation, and thus vary across different concepts. For most concepts, the visual models dominate, with the visual weight ranging from 0.6 to 1.

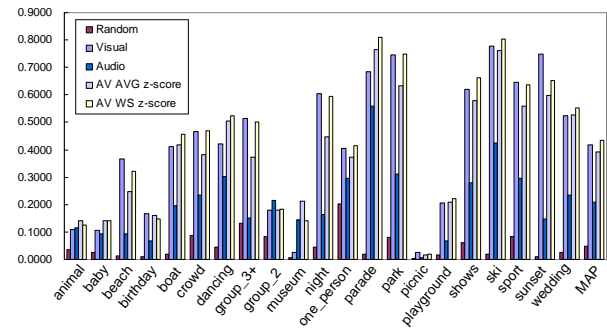


Figure 9: Comparison of weighted fusion of audio and visual models with different score normalization processes.

The above results show that *with simple weighted averaging schemes, audio and visual models can be combined to improve the concept detection accuracy.* However, additional care is needed to determine the appropriate weights and score normalization strategies.

Audio-Visual Boosted CRF & Audio-Visual Joint Boosting

Fig. 10 shows the per-concept AP of different audio-visual fusion algorithms, where “AVBCRF + baseline” corresponds to the method that computes average of the posteriors from AVBCRF and the visual baseline, and “AVJB + baseline” corresponds to the method that computes average of the posteriors from AVJB and the visual baseline. “ALL” corresponds to the method that we average the posteriors from AVBCRF, AVJB, and the visual baseline model. From our previous experiences [3], combining the advanced algorithms (e.g. AVBCRF and AVJB) with the visual baseline usually gives better performance than using these advanced algorithms alone. For comparison, the best performing ensemble fusion method (weighted combination of audio and visual based detection scores with z-score normalization) is also shown in the figure.

By combining visual baseline detectors and audio baseline detectors through context fusion, the AVBCRF algorithm improves the performance by more than 10% when it is fused with the visual baseline. The improvements over many concepts are significant, e.g. 40% over “animal”, 51% over “baby”, 228% over “museum”, 35% over “dancing”, and 21% over “parade”. These results confirm the power of incorporating inter-concept relations into the context fusion model. Our experiments also show that *context fusion among visual models only does not provide performance gain on the average. Only when the audio*

models are incorporated into the context fusion, clear performance gain is achieved. This is interesting and important – the audio models provide non-trivial complementary benefits in addition to the visual models.

Compared to straightforward weighted averaging over audio and visual models for each concept, the AVBCRF context fusion method shows more consistent improvement over the diverse set of concepts. Most importantly, it avoids the problem of large performance degradation by weighted average model over a few concepts (“sunset” and “museum”), when models from one modality are significantly worse than the others. In other words, by fusing multimodal models over a large pool of concepts, the stability of the detectors can be greatly improved.

Fig. 11 gives an example of the top 20 detected video clips for the “parade” concept (ranked based on the detection scores in descending order) using both AVBCRF and visual based baseline. Many irrelevant videos (marked by red rectangular) are included in the top result when using only visual based baselines. This is because most of these irrelevant videos contains crowd in the outdoor scene and the visual appearances are similar to those of “parade” images. By using AVBCRF, such irrelevant videos are removed largely because of the help from the audio models. Parade scenes are usually accompanied with noisy sound from the crowd and loud music associated with the parade. The visual appearances plus audio together can distinguish “parade” videos more effectively than only using a single type of features.

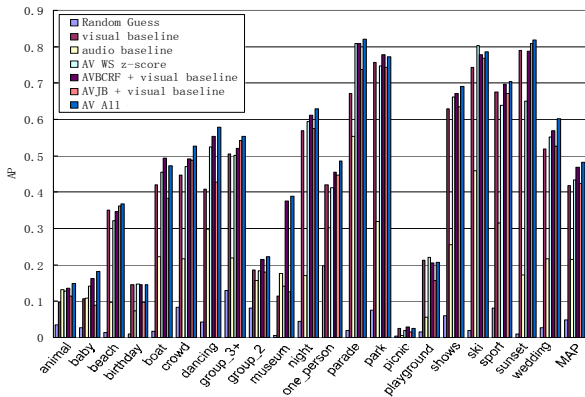


Figure 10: comparison of different audio-visual fusion algorithms.

AVJB does not result in improved performance when it is applied alone or combined with the visual baseline. This indicates that the use of local features and feature sharing in AVJB is not as effective as the exploration of inter-concept context modeling in AVBCRF. However, AVJB does provide complementary benefits – by combining AVJB with AVBCRF and visual baseline, we achieved further improvements over many concepts, e.g. 10% over “animal”, 12% over “baby”, 7% over “beach”, 7% over “crowd”, 7% over “one person”, etc. It is interesting to see that most concepts benefiting from feature sharing (AVJB) overlap with concepts benefiting from context fusion (AVBCRF). More research is needed to gain deeper understanding of the mechanism underlying this phenomenon, and develop techniques that may automatically discover such concepts.

Analysis of the results from the AVJB models also allows us to

investigate the relative contributions of features extracted from images of individual concepts, and how they are shared across classifiers of multiple concepts. Fig. 12 shows the frequency of individual kernels used by the AVJB algorithm in simultaneously detecting 21 concepts through 200 iterations. Only 25 out of the total 64 kernels (3 visual-based kernels for each concept and 1 audio kernel for all concepts) are selected by the feature selection /sharing procedures. It’s surprising to see that single audio kernel turns out to be the most frequently used kernel, more than any other kernels constructed from visual features (described in Sec. 3.2.2). This again confirms the importance of multimodal fusion – despite the lower accuracy achieved by the audio models (compared to their visual counterparts), the underlying audio features play an important role in developing multimodal fusion models.



Figure 11: Top 20 video clips from the “parade” concept. The irrelevant videos are marked by red rectangles. Video clips are ranked based on the detection scores in descending order.

The feature selection and sharing processes used in AVJB are useful in pruning the feature pool in order to make the models more compact. Kernels learned from “birthday”, “museum”, and “picnic” are discarded because of their relatively poor quality. Images from these concepts have highly diverse visual content and thus the learned visual vocabularies and associated kernels can not capture meaningful characteristics of these concepts.

To allow comparison with other classification systems, we also measure the detection accuracy using a common metric, Equal Error Rate (EER). EER values of the visual model, audio model, the final fused model (“AV ALL” shown in Fig. 10) are shown in Fig. 13. It can be seen that the proposed fusion framework is effective, reducing the overall error rates from 0.2 (using visual models alone) to 0.17 – a 15% improvement. It is also encouraging to see that with sound approaches of audio-visual content analytics and machine learning, a satisfactory accuracy of

83% can be achieved in detecting the diverse set of semantic concepts over consumer videos.

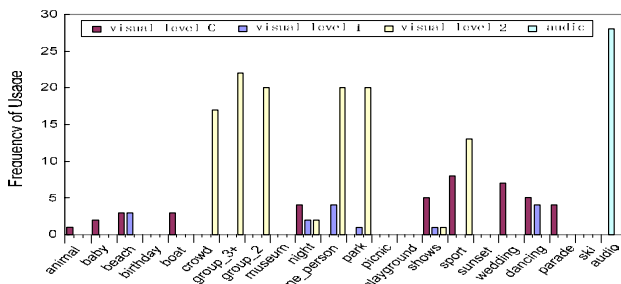


Figure 12: Frequency of kernels used by the AVJB algorithm throughout 200 iterations.

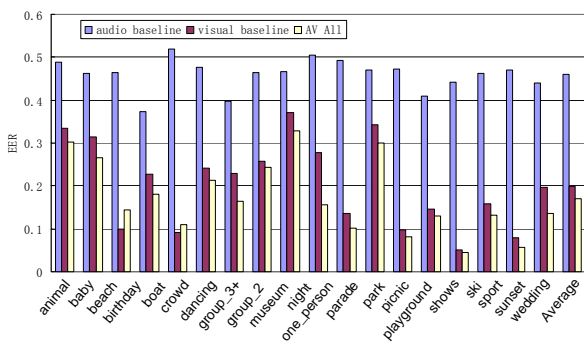


Figure 13: EER comparison of different algorithms.

7. CONCLUSIONS

We develop new methods and assess the state of the art in automatic classification of consumer videos into a large set of semantic concepts. Experiments of 24 diverse concepts over 1300+ videos from real users reveal several important findings – specificity of concept definitions and numbers of training samples play important roles in determining the detector performance; both audio and visual features contribute significantly to the robust detection performance; inter-concept context fusion is more effective than the use of complex local features; and most importantly a satisfactory detection accuracy as high as 83% over diverse semantic concepts is demonstrated. The results confirm the feasibility of semantic classification of consumer videos and suggest novel ideas for further improvements. One important area is to incorporate other contextual information such as user profile and social relations. Another direction is to explore advanced frameworks that model the synchronization and the temporal evolution among audio and visual features of temporal events.

8. ACKNOWLEDGEMENT

This project has been supported in part by a grant from Eastman Kodak. Wei Jiang is also a Kodak Graduate Research Fellow.

9. REFERENCES

[1] C.C. Chang and C.J. Lin. LIBSVM: a Library for Support Vector Machines. 2001,

<http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

[2] S.F. Chang, et al. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. In *NIST TRECVID workshop*, Gaithersburg, MD, 2005.

[3] A. Amir, et al. IBM Research TRECVID-2004 Video Retrieval System. In *NIST TRECVID 2004 Workshop*, Gaithersburg, MD, 2004..

[4] R.Fergus, P. Perona, A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *IEEE Proc. CVPR*, 2003, pp. 264-271.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Dept. Statistics, Stanford University Technical Report, 1998.

[6] K. Grauman and T. Darrel. Approximate correspondences in high dimensions. *Advances in NIPS*. 2006.

[7] W. Jiang, S.F. Chang, and A.C. Loui. Kernel sharing with joint boosting for multi-class concept detection. In *CVPR Workshop on Semantic Learning Applications in Multimedia*, Minneapolis, MN, 2007.

[8] W. Jiang, S.F. Chang, and A.C. Loui. Context-based concept fusion with boosted conditional random fields. In *IEEE Proc. ICASSP*. vol.1, 2007, pp. 949-952.

[9] S. Lazebnic, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, vol. 2, 2006, pp. 2169-2178.

[10] A.C. Loui, et al. Kodak Consumer Video Benchmark Data Set: Concept Definition & Annotation. ACM Multimedia Information Retrieval Workshop, Sept. 2007.

[11] D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, 1999, pp.1150-1157.

[12] NIST. TREC Video Retrieval Evaluation (TRECVID). 2001 -- 2006, <http://www-nlpir.nist.gov/projects/trecvid/>

[13] A. Torralba, K. Murphy, and W. Freeman. Sharing features: effective boosting procedure for multi-class object detection. In *Proc. CVPR*, vol. 2, 2004, pp. 762-769.

[14] A. Torralba, K. Murphy, and W. Freeman. Contextual models for object detection using boosted random fields. *Advances in NIPS*, 2004.

[15] A. Yanagawa, et al. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Columbia University ADVENT Tech. Report # 222-2006-8, March 2007, <http://www.ee.columbia.edu/dvmm/columbia374> .

[16] A. Yanagawa, W. Hsu, and S.-F. Chang. Brief Descriptions of Visual Features for Baseline TRECVID Concept Detectors. Columbia University ADVENT Tech. Report #219-2006-5, July 2006.

[17] Caltech 101 data sets, http://www.vision.caltech.edu/Image_Datasets/Caltech101

[18] T. Hoffmann. Probabilistic latent semantic indexing. In *Proc. SIGIR*, 1999.