

Extracting Semantics from Multimedia Content: Challenges and Solutions

Lexing Xie, Rong Yan

Abstract Multimedia content accounts for over 60% of traffic in the current internet [74]. With many users willing to spend their leisure time watching videos on YouTube or browsing photos through Flickr, sifting through large multimedia collections for useful information, especially those outside of the open web, is still an open problem. The lack of effective indexes to describe the content of multimedia data is a main hurdle to multimedia search, and extracting semantics from multimedia content is the bottleneck for multimedia indexing. In this chapter, we present a review on extracting semantics from a large amount of multimedia data as a statistical learning problem. Our goal is to present the current challenges and solutions from a few different perspectives and cover a sample of related work. We start with a system overview with the five major components that extracts and uses semantic metadata: data annotation, multimedia ontology, feature representation, model learning and retrieval systems. We then present challenges for each of the five components along with their existing solutions: designing multimedia lexicons and using them for concept detection, handling multiple media sources and resolving correspondence across modalities, learning structured (generative) models to account for natural data dependency or model hidden topics, handling rare classes, leveraging unlabeled data, scaling to large amounts of training data, and finally leveraging media semantics in retrieval systems.

1 Introduction

Multimedia data are being captured, stored and shared at an unprecedented scale, yet the technology that helps people search, use, and express themselves with these

L. Xie and R. Yan are with IBM T J Watson Research Center, Hawthorne, NY, e-mail: {xlx, yanr}@us.ibm.com This material is based upon work funded in part by the U. S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government.

media is lagging behind. While no statistics is available about the total amount of multimedia content being produced, the following two statistics can provide us with an intuition about its scale : there are about 83 million digital still cameras sold in 2006 [37], and video already account for more than half of the internet traffic, with YouTube alone taking 10% [2, 74, 30]. A typical internet user actively gleans information from the web with several searches per day, yet their consumption of video content mostly remains passive and sequential, due to the inefficacy of indexing into video content with current practices. As an article on *Wired* overtly put: “Search engines cannot index video files as easily as text. That is tripping up the Web’s next great leap forward.” [3] The key to indexing *into* image and video files lies in the ability to describe and compare the media content in a way meaningful to humans, i.e. the grand challenge of closing the semantic gap [80] from the perceived light and sound to users’ interpretations.

One crucial step that directly addresses the semantic indexing challenge is to extract semantics from multimedia data. The advance in storage and computation power in recent years has made collecting and processing large amounts of image/video data possible – thus has shifted the solutions to semantic extraction from knowledge-drive to data-driven, similar to what has been practiced in speech recognition for several decades [72]. Algorithms and systems for data-driven semantics extraction are embodiments of statistical pattern recognition systems specialized in multimedia data. They learn a computational representation from a training data corpus labeled with one or more known semantic interpretations (such as face, human, outdoors). Statistical learning of multimedia semantics has significantly advanced performance and real-world practice in recent years, which made possible, for example, real-time face detectors [95].

This paper is intended to survey and discuss existing approaches on extracting multimedia semantics in a statistical learning framework. Our goal is to present the current challenges and solutions from a few different perspectives and cover a sample of related work. The scope of this chapter has two implications: (1) since the size of target semantics from media is usually very large (e.g., objects, scene, people, events, . . .), we put more emphasis on algorithms and systems designed generic semantics than those specialized in one or a few particular ones (e.g., faces); (2) we focus more on the new challenges for model design created by the scale of real-world multimedia data and the characteristics of learning tasks (such as rare classes, unlabeled data, structured input/output, etc.). Within this scope, the semantic extraction problem can be decomposed into several subproblems: the general processing steps of going from media data to features and then to semantic metadata; the semantic concept ontology, and how to leverage it for better detection; the challenge of dealing with *multi*-media, i.e. how to use a plurality of input types; dealing with real-world annotated training dataset: rare semantics, sparseness of labels in an abundance of unlabeled data, scaling to large datasets and large sets of semantics; accounting for the the natural dependencies in data with structured input and output, and using semantics in search and retrieval systems.

This said, learning to extract semantics from multimedia shall be of much broader interest than in the multimedia analysis community. Because (1) the abstract learn-

ing problems are very similar to those seen in many other domains: stream data mining, network measurement and diagnosis, bio-informatics, business processing mining, and so on; (2) multimedia semantics can in turn enable better user experiences and improve system design in closely related areas such as computer-human interaction, multimedia communication and transmission, multimedia authoring, etc. This work is also distinct from several other surveys on multimedia indexing [80, 84, 76, 110, 11] in that we present an in-depth discussion on semantic extraction, an important component in an entire indexing system, from an algorithmic perspective. For completeness, we briefly cover feature extraction and retrieval in Sections 2 and 7, leaving detailed discussion to the above-mentioned surveys.

The rest of this paper is organized as follows: Section 2 gives an overview to the entire workflow from multimedia content to media semantics; Section 3 discusses the design and use of a large multimedia lexicon; Section 4 studies strategies for multimodal fusion; Section 5 presents models for structured input, output, as well as hidden dimensions; Section 6 addresses three challenges in real-world training data; Section 7 contains examples for using multimedia semantics in search systems; Section 8 concludes the chapter with a brief discussion and outlook.

2 From Multimedia Content to Multimodal Semantics

Multimedia semantic extraction tried to answer the following question: does media clip x contain semantic concept c ? Many systems that answer this question consist of five broad conceptual components, as shown in Figure 1. The components include: the *image or video data* for training or classification, the definition of a *multimedia lexicon* containing the target semantics, the extraction of *content features*, the design and learning of computational *models* that map features to the target lexicon, as well as the *application* that will make use of the resulting semantic metadata being extracted.

Typical algorithmic components for semantic concepts detection include low-level feature extraction (box c) and feature-based model learning (box d). In this chapter we are mainly concerned with concept-independent designs of both components, i.e. generic feature extraction and learning paradigms that work for a wide-range of target semantics. Whenever warranted by need and performance, domain knowledge and constraints can be incorporated to build and improve specialized detectors, such as the events and highlight detectors for sports discussed in Chapter [107].

While not always considered part of the algorithm design for a semantic extraction system, the data domain and the target lexicon are essential design components that set the fundamental requirements for the system. These requirements include: what are the salient semantics in the domain, do they exist in the dataset being worked on, are they useful for content indexing and other applications, and are they detectable with the current design of algorithmic components. Answers to these questions apparently vary among the wide range of multimedia data domains such

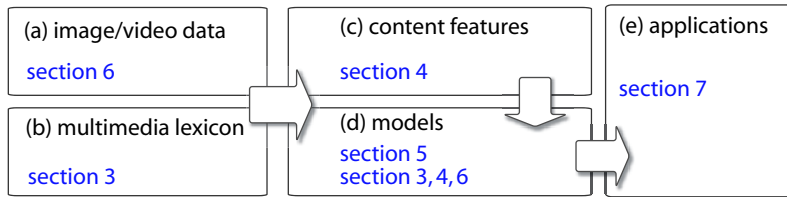


Fig. 1 Basic building blocks of a concept detection system. See section 2.

as web images, aerial photographs, consumer videos, broadcast content, instructional archive, or medical imagery. Domain knowledge plays an important role in coming up with good answers.

The rest of this section contains a brief overview of the two algorithm components: Section 2.1 reviews popular features in different content modalities, and Section 2.2 covers an effective baseline model using the support vector machine (SVM), adopted as the basic building block by numerous systems in the literature [36, 20, 82, 16]. Sections 3–7 present specific challenges originated from each of the five components alongside their current solutions, as annotated in Figure 1.

2.1 Features

Multimedia features are extracted from media sequences or collections, converting them into numerical or symbolic form. Good features shall be able to capture the perceptual saliency, distinguish content semantics, as well as being computationally and representationally economical. Here we briefly summarize commonly used features for completeness, and direct readers to respective surveys on image, video, speech and audio features for more details [33, 50, 80, 21].

2.1.1 Extracting features

Low-level features aim to capture the perceptual saliency of media signals. The procedures for computing them do not change with respect to the data collection, or the target semantics being detected. Mid-level features and detectors are computed using raw signal and/or low-level features. Their computation usually involve signal- or data-domain dependent decisions in order to cope with the change in the data domain and target semantics, sometimes training is needed. We now review low-level features by media modality and list a few examples of popular mid-level features.

- **Visual features.** Still images are usually described in three perceptual categories, i.e. color, texture, and shape [80]. While image sequences introduce one more dimension of perceptual saliency, i.e., motion. Color features are popular due to their ability to maintain strong cues to human perception with relatively less com-

putational overhead. The main concern in reliably extracting color information is to choose from a variety of color spaces and achieve perceptual resemblance and color constancy over different scene and imaging conditions. Local shapes capture conspicuous geometric properties in an image, this is among the most-studied image features, since psycho-visual studies have showed that the human visual system performs the equivalence of edge detection [39]. Local shapes are often computed over local gray-scale or color derivatives. Texture loosely describes an image aside from color and local shape, it typically reflects structure and randomness over a homogeneous part of an image. Filter families and statistical models such as Gabor filters and Markov analysis are popular choices for capturing texture. Motion provides information about short-term evolution in video. 2-D motion field can be estimated from image sequences by local appearance matching with global constraints, and motion can be represented in various forms of kinetic energy, such as magnitude histogram, optical flows and motion patterns in specific directions. Although color, shape, texture and motion can be described separately, there are features that provide integrated views such as correlogram [41] (color and texture) or wavelets (texture and local shape).

- **Audio features.** Audio signals can be characterized by a number of perceptual dimensions such as loudness, pitch, timber. Loudness can be captured by the signal energy, or energy in different frequency bands. Primitive pitch detection for monophonic tonal signals can be done with simple operations such as auto-correlation. Timber is typically captured by the amplitude envelop of spectrograms, i.e., the relative strength of different harmonics for tonal sounds. A number of simple features in the time or the STFT (Short Time Fourier Transform) domain has been effective in describing everyday sound types with one or more perceptual aspects. For instance, the zero-crossing rate in waveforms can both reflect pitch for monotonic sounds and reflect the voiced-ness of a speech segment; spectral centroid and entropy summarizes the timber and . More elaborate and robust features for modeling each of these aspects abound, such as robust pitch extractors [56, 26], LPC (linear prediction coefficients) [72], frequency-warped spectral envelops such as MFCC (mel-frequency cepstral coefficient) [33], as well as the dynamic aspects of timber such as sound onset and attack.
- **Text features.** Text information is often available alongside the image/video/audio content, features can be extracted from the transcripts obtained with automatic speech recognition (ASR) or closed caption (CC), optical character recognition (OCR) and production metadata; Techniques for extracting such features are similar to those in text retrieval, such as word counts in a bag-of-words representation. In addition to text-only features, speech signals have additional timing information upon which speaking rate and pause length can also be computed.
- **Metadata.** Metadata, sometimes called surface features, are additional information available to describe the structure or context of a media clip aside from the audio, visual, or textual part of the content itself. Examples include the name, time stamp, author, content source, the duration and location of video shots, and so forth. While not directly relevant to what are presented in the content, they can provide extra information on the content semantics. Useful metadata features are

usually coupled with suitable distance-metrics tailored to the nature of the specific data field, such as geographic proximity of GPS coordinates [59], semantic distances between locations [79].

Mid-level features capture perceptual intuitions as well as higher-level semantics derived from signal-level saliency. Examples of mid-level features and detectors include: tracked objects and segmented object parts [118], visual concepts pertaining objects, scenes and actions, such as people, airplane, greenery [90]; audio types, such as male/female speech, music, noise, mixture [77]; named entities extracted from text passages [24]. There are also mid-level features that are specific to a data domain, such as the crowd cheering detector, goal post detectors in sports videos [29].

Most approaches reviewed in this chapter uses low-level features, while mid-level features can be commonly found in domain-specific approaches for high-level events, such as in Chapter [107].

2.1.2 Feature aggregates

Feature aggregates are derived from content features and detectors, the purpose of the aggregate is to incorporate the inherent spatial-temporal structure in the media and align the features generated over different content units such as pixels, local regions, frames, short-time windows or text passages. The outcome of the aggregate is usually represented as numbers, vectors or sets, providing the data structure required by most statistical pattern recognition models while preserving the saliency of the target semantics. In practice, this aggregation is usually done with one or a combination of the following operations:

- Accumulative statistics such as histogram [87] and moments [89] provide simple yet effective means for aggregating features over space and time. They have the advantages of being insensitive to small local changes in the content, as well as being invariant to coordinate shift, signal scaling and other common transformations. While the associated disadvantage is in the loss of sequential or spatial information.
- The selection of possible feature detectors from candidate portions of the original signal aims to preserve perceptual saliency, provide better localization of important parts. Tracking and background subtraction can be viewed as one type of selection, as well as extracting salient parts and patches [55], silence detection in audio, or removing stop words.
- Features in an image or a image sequence can also be aggregated into sets. The sets can be unordered, e.g. bag of words, bag-of-features, or ordered in sequences or more general graph structures.

2.2 Learning semantics from features

The simplest semantic model can be a mapping function from features to the presence or absence of content semantics. One of the most common learning algorithms is support vector machines (SVMs) [47, 94], being preferred in the literature for its sound theoretical justifications and good generalization performances compared to other algorithms [13]. Built on the structural risk minimization principle, SVMs seek a decision surface that can separate the training data into two classes with the maximal margin between them. The decision function takes the form of a generalized linear combination of training samples:

$$y = \text{sign} \left(\sum_{i=1}^M y_i \alpha_i K(x, x_i) + b \right), \quad (1)$$

where x is the d -dimensional feature vector of a test example, $y \in \{-1, 1\}$ is the class label representing the absence/presence of the semantic concept, x_i is the feature vector of the i^{th} training example, M is the number of training examples, $K(x, x_i)$ is a kernel function representing the similarity measure between examples, the support vector weights $\alpha = \{\alpha_1, \dots, \alpha_M\}$ and offset b are the parameters of the model. The kernel function can take many different forms, such as the polynomial kernel $K(u, v) = (u \cdot v + 1)^p$, the Radial Basis Function (RBF) kernel $K(u, v) = \exp(-\gamma \|u - v\|^2)$ or kernels on structured data such as the string kernel. The RBF kernel is widely used due to its flexibility to model non-linear decision boundaries of arbitrary order and the perceived good performance on testing data. Note however, that the setting for the hyper-parameter γ in RBF kernels often exerts significant influence to the model performance, therefore it is usually chosen empirically with cross-validation [40].

Besides SVMs, there are a large variety of other models that have been investigated for multimedia semantics extraction, including Gaussian mixture models (GMM) [4, 98], hidden Markov models (HMM) [71], k Nearest Neighbor (kNN) [85], logistic regression [35], Adaboost [119] and so on. In Section 5 we will discuss models other than SVMs that cater to (1) structured input, especially temporal sequences (2) the natural but *hidden* topics in broadcast content collections.

Note that additional domain knowledge can be of great help to customize model design and improve performance for specific concepts such as faces, cars and sport events. Detailed methodologies of such design is outside the scope of this chapter, and we refer the interested readers to discussions in relevant literature [95, 75] and Chapter [107].

Before delving into examples and method variations in the rest of this chapter, we briefly define the evaluation measures used in this chapter and found in common benchmarks [91]. A decision function for a binary classification task, such as Equation 1, assigns a real-valued confidence score. We sort the test set at descending order of confidence scores, and evaluations of the scoring scheme concerns the number of correctly and incorrectly returned entries at any depth r . Denote as $rel(r)$

Concept	Avg Prec	Positive	Concept	Avg Prec	Positive
PERSON	0.8531	31161	ROAD	0.2481	2665
FACE	0.7752	17337	MICROPHONE	0.1947	2659
OUTDOOR	0.7114	15290	INTERVIEW	0.3019	2619
STUDIO	0.7541	4743	INTERVIEWSEQ	0.5237	2523
BUILDING	0.3048	4177	CAR	0.3151	2492
FEMALE	0.2632	3887	MEETING	0.1708	2262
WALKING	0.1635	3828	ANCHOR-STUDIO	0.8247	2392
URBAN	0.1127	3586	ARTIFICIAL-TEXT	0.6783	2373
LEADER	0.1822	3033	TREES	0.2522	2152
POLITICIANS	0.2782	2850	SPORTS	0.4481	1249
ASIAN-PEOPLE	0.4247	2776	MAPS	0.4816	610

Table 1 Example semantic detection results for 22 most frequent LSCOM concepts [61]. The SVM models are learned from the TRECVID-2005 development set with color moment features. For each concept, the column “positive” indicates the number of positive examples out of 55,932 keyframes, and the “Avg Prec” column is the average precision based on 2-fold cross validation.

the binary relevance indicator for the media clip at depth r , i.e. $rel(r) = 1$ iff. the clip contains the target semantic, 0 otherwise. For return list of size N , the precision $P(N)$, recall $R(N)$ and average precision $AP(N)$ are defined as follows:

$$P(N) = \frac{\sum_{r=1}^N rel(r)}{N} \quad (2)$$

$$R(N) = \frac{\sum_{r=1}^N rel(r)}{\sum_{r=1}^{\infty} rel(r)} \quad (3)$$

$$AP(N) = \frac{\sum_{r=1}^N P(r) \times rel(r)}{\sum_{r=1}^{\infty} rel(r)} \quad (4)$$

Average precision is a summary measure over a range of depths. It puts emphasis on returning more correct entries earlier, and has been shown to be more stable than other measures in common retrieval tasks [12].

As an example, Table 1 shows detection performances for twenty-two frequent concepts in multi-lingual broadcast news, and Figure 2 shows typical top-scored shots for four concepts. These detectors are trained on one visual feature (color moments) using SVMs. The performance measure shows that the basic strategy discussed in earlier parts of this section is indeed effective in extracting semantics even from visually diverse domains such as broadcast news.

3 The construction and use of multimedia ontology

Multimedia semantics do not exist in isolation. There are usually multiple concurrent semantics associated with any media clip, and the usage of semantics and their

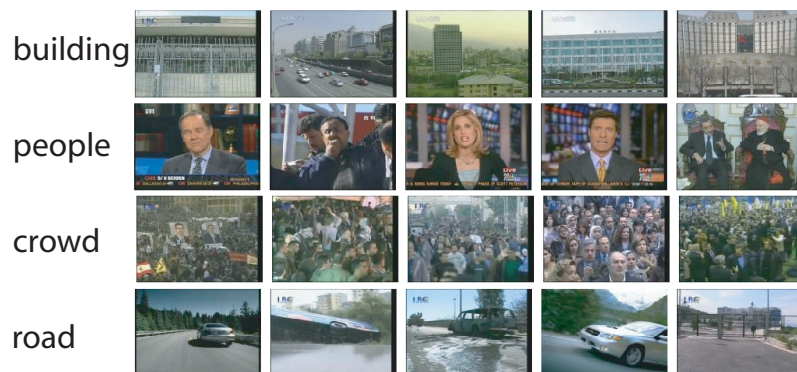


Fig. 2 Top detection result for concepts *building*, *people*, *crowd*, *road* on TRECVID 2005 broadcast news test set.

correlations often differ for collections of clips in different content domain. For example, the semantic annotations for a consumer photograph can be “*outdoors*, *mountain*, *vegetation*, *flower*”, and those for a broadcast news keyframe be “*studio*, *face*, *female anchor*, *computer or television screen*”. Between these two domains, the likelihood that we are seeing each tag is different, e.g., *female anchor* is rare in consumer photos, and we typically see less *flower* in news; which tags tend to occur together also changes, e.g., we can see more instances of *vehicle* together with *outdoors* and *mountain* in news.

Given these observations, defining suitable ontologies for multimedia, as well as using them to help semantic extraction has become important tasks in order for multimedia analysis to be useful and realistic.

3.1 Making a visual lexicon

The process of developing a good multimedia lexicon involve several steps: (1) defining a list of semantic concepts from prior knowledge; (2) ground these concepts on a database by finding examples; (3) build detectors for the the lexicon. These steps are sometimes iterated in order to obtain good results.

As a example, the TRECVID benchmark [90] started with ten semantic concepts in 2002: *Outdoors*, *indoors*, *face*, *people*, *cityscape*, *landscape*, *text overlay*, *speech*, *instrumental sound*, and *monologue*. This list only covered a subset of the important semantics in video, so in TRECVID-2003 the list was enlarged to 831 semantic concepts on a 65-hour development video collection, 17 of which were selected for benchmarking the detection performance. These annotation were collected in a common annotation forum, and the annotation tool, VideoAnnex [53], allowed each user to add/edit the ontology independently. TRECVID 2005 and after adopted a fixed lexicon for concept annotation, partly to address the lack of convergence in

user-assigned free text labels. This effort has led to a Large-Scale Concept Ontology for Multimedia (LSCOM) [61], and an interim result of this has resulted in 39 high-level features (concepts) definitions and annotations dubbed LSCOM-lite [60].

The LSCOM-lite concepts went through the three step life cycle mentioned above. (1) The broadcast news domain were first divided into seven major categories using domain knowledge, these categories are: program category, setting/scene/site, people, object, activity, event, and graphics. Several representative concepts are then selected from each category, where the selection involved mapping them to real-world query logs and the semantic knowledgebase WordNet, as well as validating with past TRECVID queries [1]. (2) A collaborative annotation effort is then carried out among participants in the TRECVID 2005 benchmark, with human subjects judging the presence or absence of each concept in the key frame, producing annotations for the 39 concepts on the entire TRECVID 2005 development set (over 60,000 keyframes from more than 80 hours of multi-lingual broadcast news). (3) Ten of the LSCOM-Lite concepts were evaluated in the TRECVID 2005 high-level feature detection task, twenty of them were evaluated at TRECVID 2006, and another twenty were evaluated at TRECVID 2007 on a different content domain (documentary).

The full LSCOM effort has developed an expanded multimedia concept lexicon well-beyond the previous efforts. Concepts related to events, objects, locations, people, and programs have been selected following a multi-step process involving input solicitation, expert critiquing, comparison with related ontologies, and performance evaluation. Participants include representatives from intelligence community, ontology specialists, and researchers in multimedia analysis. In addition, each concept has been qualitatively assessed according to the following three criteria:

- Utility, or a high practical relevance in supporting genuine use cases and queries;
- Observability, or a high frequency of occurrence within video data sets from the target domain;
- Feasibility, or a high likelihood of automated extraction considering a five-year technology horizon.

An annotation process was completed in late 2005 by student annotators at Columbia University and Carnegie Mellon University. The first version of the LSCOM annotations consist of keyframe-based labels for 449 visual concepts, out of the 834 initial selected concepts, on the TRECVID 2005 development set [1]. Here are sample concept definitions in LSCOM, note their emphasis on visual salience, and their wide coverage in many multimedia domains.

- *Waterscape-Waterfront*: Shots depicting a waterscape or waterfront
- *Mountain*: Shots depicting a mountain or mountain range with the slopes visible.
- *Sports*: Active sports scenes included jogging/running and players performing sport; excluded: fans at sporting events (including benched players); sports in music; video sports equipment; celebrating after/before sporting event.
- *People-Marching*: Shots showing one or more people marching

3.2 *Multimedia ontology and semantic extraction*

A multimedia ontology *de-isolates* semantics in two different ways: (1) putting concepts in context with each other with pre-defined semantic relationships such as those found in WordNet and Cyc; (2) linking concepts with their joint presence in multimedia datasets. The extraction of multimedia semantics can in turn use the related semantic interpretations along with the co-occurrence patterns in image/video collections to improve the detection of each semantic concept. For example, when the concept “bus” is observed in a video, we know that its hypernym “vehicle” is also valid, concept “wheel” is likely to be visible since wheels are parts of a bus; the concept “road” have high likelihood to appear, while the concept “office” is less likely to co-occur.

These two types of concept relationships are commonplace in multimedia collections, and they can be useful in two complementary ways, i.e., using multi-concept relationship to improve the concept detection accuracy, and using correlated context from data to construct, refine, or discover semantic relationships in a video collection. The rest of this section will briefly review several approaches on using and constructing semantic knowledge.

Pattern recognition techniques have been used to automatically exploit multi-concept relationships. For example, Naphade et al. [62] explicitly modeled the linkages between various semantic concepts via a Bayesian network, where the semantic ontology were encoded in the network topology, and data correlations were captured in the model parameters. Snoek et. al. [82] used a multi-concept “context link” layer for the same purpose in the MediaMill concept detection architecture. This link aims to filter raw concept detector outputs by either learning a meta-classifier or with ontological common sense rules. Hauptmann et al. [35] constructed an additional logistic regression classifier atop uni-concept detection results, to capture the inter-concept causations and fuse the multi-concept predictions. Amir et al. [4] concatenated concept prediction scores into a long vector called model vectors and used a support vector machine as the meta-classifier. Wu et. al. [100] proposed an ontology-based multi-classification algorithm, attempting to model the possible influence relations between concepts based on a predefined ontology hierarchy. Yan et al. [116] described several approaches for mining the relationship between video concepts with several probabilistic graphical model representations. We have experimented the effect of a large lexicon on concept detection performance, as shown in Figure 3. This experiment uses naive-Bayes classifiers to model the relationship between target concept ground-truth and concept detection scores. We can see that concept detection performance can be improved for more than 10% using statistical models on cross-concept relationship, the improvement saturates around 200 concepts, similar to what was observed by Snoek et. al. in video retrieval [81].

There are several recent studies on discovering and refining semantic relationship from data, esp. on the broadcast news ontology [1] where a complete labeling of several concepts is available. Kender [51] analyzed the dependencies and redundancies in the LSCOM ontology, confirmed several intuitive ontological relationships and suggested a few revisions based on the concept co-occurrence in the data. Xie

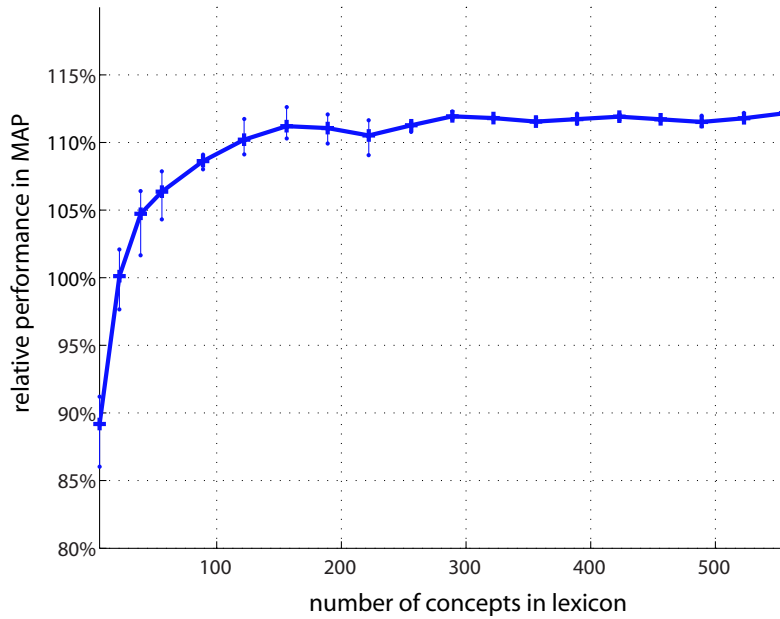


Fig. 3 The effect of a large concept ontology on concept detection performance on TRECVID 2007 test data. X-axis: number of concepts in the ontology, randomly selected from a total of 556 concepts; Y-axis: relative mean-average precision over 20 concepts, with respect to that of visual-only single-concept detectors. See [17] for details.

and Chang [101] found that co-occurrence and temporal precedence are effective in correlating concepts, and the discovered concept tuples either confirm generic ontological rules or reflect data domain characteristics. The problem of reliably mining large-scale relations from data remains a challenging one, and the progress of which can be facilitated with creating large annotated datasets in multiple domains, as well as more research efforts into the mining and analysis methodologies.

4 Multimodality: information fusion and cross-modal association

To detect semantics from multimedia streams, it is almost always beneficial to combine detection outputs from multiple modalities that provide complementary information (Figure 4(b)). For example, the presence of “Bill Clinton” (without performing face recognition), usually involves one or more persons in the image, and the word “Clinton” in the spoken content; and the visual concept “clear sky” can be identified by both its color (blue with gradient) and texture (very smooth). Note however, more does not easily lead to better. The challenges of multi-modal fusion mainly lies in the broad diversity among the modalities, which can be summarized into the following three aspects: (1) Representation difference, e.g. bags-of-words

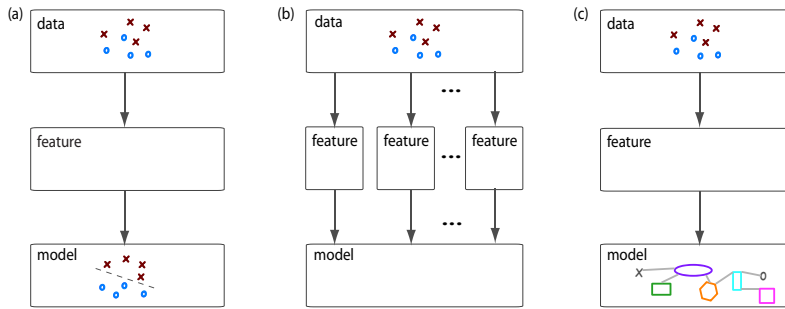


Fig. 4 Semantic concept modeling problems brought up by multi-modality and structured input/output. (a) Basic concept learning architecture, as described in Section 2.2. (b) Concept learning from multimodal cues, described in Section 4. (c) Concept Learning with structured models, described in Section 5.

for text or filter responses off image patches for texture; (2) Distribution diversity, e.g., word distributions are typically multinomial, while color and texture features are mostly modeled with one or more Gaussians (3) Domain dependency of the underlying modalities, e.g., the color variations over one news program is much larger than those in a typical surveillance video. For multimodal fusion, efforts has been devoted to answering three important questions: *when* to combine, *what* to combine, and *how* to combine. While a lot of progress has been made in the recent years, the definite answer is still open, and there is very likely more than a few good answers.

The rest of the section contains two parts – we first review several general learning approaches for multi-modal fusion and discuss their strengths and weaknesses, we will then cover models for a cross-modal association: a special case of multi-modal learning widely seen in real-world image collections and their surrounded text annotations (e.g. web images, or Flickr).

4.1 Multimodal Fusion

Multi-modal fusion approaches can be categorized into two families, i.e., early fusion and late fusion, with a dichotomy on *when* to combine. The early fusion methods merge multi-modal features into a longer feature vector before it is used as the input of classifiers. In contrast, the late fusion methods directly fuse detection outputs after multiple uni-modal classifiers are generated. Neither of the fusion methods are perfect [83]. Early fusion can implicitly model the correlations between different feature components by combining them into a long feature. However, early fusion be caught in trouble if the feature constitution of different modalities is too heterogeneous with skewed length distribution and numerical scales. This is less of a problem for late fusion, since the features from each modality will not interact with each other before the final fusion stage. Moreover, late fusion allows the system to adopt various detection techniques according to specific feature types. Also, it usu-

ally requires less computation power compared with the early fusion counterpart. Therefore, late fusion techniques appear to be more popular and more extensively studied than early fusion techniques in the literature.

The confidence scores and features generated from different modalities/models usually need to be normalized before fusion. Typical normalization schemes include rank normalization [115], range normalization, logistic normalization [69], and Gaussian normalization [4]. The final detection results are then produced by merging the normalized confidences. For *what* to combine in late fusion, this merge step can operate on one or more types of inputs: 1) combine multiple detection models, 2) combine the detection models of the same class with different underlying features, or 3) combine the models with the same underlying features but different parameter configurations. To address *how* to combine, approaches range from pre-define combination rules based on domain knowledge, to a large body of machine learning methods aiming for further performance improvement. For example, Amir et al. [4] studied min, max and unweighted linear combination function for multi-modality and multi-model fusion. Among the machine learning approaches, simple models such as linear combinations and logistic regressions [35] has been explored, super-kernel fusion [99], as an example extension of simple discriminative models, constructs a hierarchy of kernel machines to model the non-linear decision boundaries. Yang et al. [117] specifically consider the problem of detecting news subjects in news video archives by linearly combining the multi-modal information in videos, including transcripts, video structure and visual features. The weights are learned from SVMs. Snoek et al. [83] compare the early fusion and late fusion methods with SVMs as the base classifiers and meta-level classifiers for fusing text and images. Their experiments on 184 hours of broadcast video and 20 semantic concepts show that late fusion on average has slightly better performance than early fusion for most concepts, but if the early fusion is better for a concept, the improvement will be more significant than later fusion.

4.2 Cross-modal association and image annotation

Viewing semantic concepts as binary detection on low-level multi-modal features is not the only way for multimedia semantics extraction. An emerging direction for (image/video) concept detection is to jointly model the associations between annotated concept words and image features. This scenario has very wide appeal since the image+text data model fits many real-world image collections: professional stock photo catalogues, personal pictures, images on the web with surrounding HTML text, or images on media-rich social sites such as Flickr and Facebook. These approaches typically assume that words and image features are generated by one set of hidden information sources, i.e. the hidden semantics. Hence, image features and concepts are no longer marginally independent to each other. Once image features are given, associated words can be inferred by the information flow passed through the hidden layer. Actually, most of these approaches have been designed

under a slightly different name called “image annotation”, which aims to match the associating keywords to their corresponding images, and automatically predict new keywords for a given image.

A number of learning algorithms have been applied in the task of automatic image annotation, such as machine translation models [5], relevance language models [44], graphical models [6] and graph random-walk methods [67]. Barnard et al. [5] interpreted regions in images and the words in annotation as aligned bi-text and used machine translation models to learn their joint probabilities (with and without word orderings) in order to uncover their statistical correspondence. Blei et al. [6] developed a gaussian-multinomial latent Dirichlet allocation(GM-LDA) model and a correspondent latent Dirichlet allocation(GM-LDA) model that simultaneously capture the information from image regions and associated text keywords via a directed graphical model. Jeon et al. [44] used the framework of cross-lingual retrieval to formulate the image/video annotation. They proposed an annotation model called cross-media relevance model(CMRM) which directly computed the probability of annotations given the image. It was shown to outperform the translation models in the image/video annotation task. By representing the terms and image features in a unified graph, Pan et al. [67] proposed a random walk with restart(RWR) approach to capture the correlation between words and images. Jin et al. [45] proposed a coherent language model for image annotation that can model the word-to-word relationship in the annotation process. This approach allows the annotation length to be automatically determined and the annotated number of examples to be reduced by using active learning technique. Iyengar et al. [42] described a joint text/image modeling approach for video retrieval that allows the full interaction between multi-modalities to result in a considerable performance improvement in TRECVID datasets.

5 Structured Models

Semantic modeling as discussed in Section 2 and 3 treats each data instance as an independent unit, and models there learn a direct mapping function from the input features to the target class. In many real-world scenarios, however, the problems calls for structured models (Figure 4(c)). And the resulting model structure accounts for either the natural data dependencies, such as those in temporal data streams, e.g. a foul in soccer usually leads to a throw-in in the next few shots; or inherent structure in the data collection, e.g. arrow charts and maps are recurrent visual themes in news programs and they can mean “financial news”, “weather report”, “war coverage” or “natural disaster”.

5.1 Models for semantics in temporal sequences

Graphical models are natural choices as stochastic representations for temporal evolutions in streams, or intuitive dependencies in data. Hidden Markov Models (HMMs) [73] is one of the most popular graphical models due to its simple structure and available efficient inference algorithms. HMMs are used by Schlenzig, Hunter and Jain [78] to recognize four types of gestures from continuous recordings, and by Starner, Weaver and Pentland [86] to recognize American Sign language from wearable computers. In produced video streams features reflect both the content and the production conventions, such as segmenting stories in news programs [18], or detecting plays in sports broadcast [102]. Flavors of Dynamic Bayesian Network (DBN) are extensions of HMM used to encode more complex dependencies. Brand, Oliver and others [9, 66] develop coupled HMM (CHMM) to account for multiple interacting streams for multi-object multi-agent action recognition. Zhang et al. [57] analyze multi-camera/microphone meeting captures for group interaction events such as discussion, monologue, presentation + note-taking. Two-layer HMM is used to infer individual action and group action in cascade, each state in the HMMs are assigned domain-specific meanings and the parameters are learned from data.

The ability of graphical models to capture the data dependency can be used in conjunction with discriminant models, such as kernel-based classifiers, to improve detection performance on known semantics. For instance, for the problem of distinguishing shots that do or do not contain a generic event (e.g., *airplane landing*, *riot*), Ebadollahi et al. [31] use SVM on representations generated from HMM likelihoods and parameters from input feature streams; Xu and Chang [106] use bag-of-features representation of temporal streams, compute similarity metric among shots using earth mover’s distance (EMD) or pyramid match kernel, and then resort to SVM for the final decision; Xie et al. [104] use multiple kernel learning in the final decision stage to optimally combine multiple input streams.

5.2 Uncover the hidden semantic layer with unsupervised learning

Most semantic extraction algorithms learn a direct mapping from annotated data to a pre-defined list of semantic concepts. While simple and effective, these approaches does not make full use of the inherent data distributions and latent structures.

The idea of latent structures and hidden *topics* was first explored in in text retrieval. There, each document d in collection \mathcal{D} consists of words w in vocabulary \mathcal{W} . The text data in a collection are summarized into a feature matrix $M^{D \times W}$ containing the word counts for each document, i.e., the bag-of-words representation, with $D = |\mathcal{D}|$ and $W = |\mathcal{W}|$. The algorithms then finds K latent topics $Z^{K \times W}$ to best represent M . Latent semantic indexing (LSI) [27] considers each document $M_d = [m_{d1}, \dots, m_{dW}]$ as a linear combination of latent *topics* $M_d = \sum_k w_{dk} Z_k$, where $Z_k = [z_{d1}, \dots, z_{dW}]$ being the k^{th} topic vector denoted by the relative strength of

each feature (word), and $W_d = [w_{d1}, \dots, w_{dK}]^T$ being the mixing weights for document d . The hidden topics and weights are then uncovered with singular value decomposition:

$$M = USZ^T \approx M_K = U_K S_K Z_K^T$$

. This provides a rank- K approximation to matrix M with minimum least-square error, with the rows of Z_K represent the K topics, and the d^{th} row of $U_K S_K$ being topic weights W_d for each document. There are a few probabilistic extensions to LSI. Probabilistic latent semantic indexing (pLSI) [38] expresses the joint probability of a word w and a document d as

$$p(d, w) = p(d) \sum_z p(w|z)p(z|d),$$

with z the unobserved topic variable, and $p(z|d)$ taking the role of the top-mixing weights. The mixture of unigrams model [7] is a special case of pLSI where each document is associated with only one topic. The latent Dirichlet allocation (LDA) model [7] offers even more flexibility by modeling the top-mixing weights as random variables observing *prior* distributions.

Topic models has been extended to handle multimedia data. Gemert [93] applied LSI to capture the joint latent semantic space of text and images. Blei and Jordan [6] have extended the mixture of unigrams and the LDA model into a Gaussian-multinomial mixture (GM-Mix) and a Gaussian-multinomial LDA (GM-LDA) to model captioned images. Hierarchical HMM model (HHMM) [103] is another variant of directed graphical models that captures latent topic structures with temporal dependencies. Undirected graphical model has also been explored. The *dual-wing harmonium* (DWH) model for multimedia data [105] can be viewed as an undirected counterpart of the two-layer directed aspect models such as LDA, with the topic mixing as document-specific and feature-specific combination of aspects rather than via a cumulative effect of single topic draws. Inference on DWH is fast due to the conditional independence of the hidden units, although the offline learning process could take longer due to an intractable normalization factor.

We illustrate the effectiveness of learning the latent semantic topics in Fig 5. Each topic is described by the top 10 words and the top 5 key images with highest conditional probabilities on the latent topic. Intuitively, the first three topics correspond to scenes of Weather News, Basketball and Airplane, respectively, the formation of which is based on the evidence from both words and images. The fourth topic (a CNN anchor person) is very consistent in the visuals, and diverse in the words, it is likely to be primarily determined by image similarities. The last topic is mainly driven by word similarities – its interpretation is not obvious at the first sight, due to its apparent visual diversity in weather and sports reports. However scanning the top words tells us that mentions of places (york, jersey), numbers (six), and certain verbs (stopping, losing) are indeed common across these two topic themes.

The uncovered hidden topics and multimodal concepts can be used to help semantic extraction in three ways: (1) presented as topics in themselves [103, 6], (2)

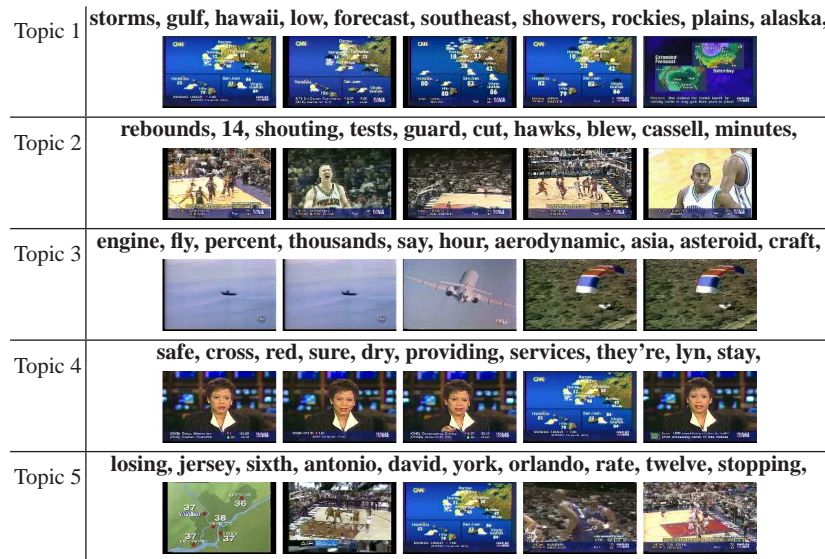


Fig. 5 Illustration of five latent topics learned from broadcast news videos with DWH [105]. The top 10 words and the top 5 images associated with each topic are shown.

used as intermediate representations for supervised classification [105], (3) used to initiate the labeling and user feedback process in supervised learning.

6 Training data

Although multimedia semantics extraction can be formulated as a straightforward supervised learning problem, not every supervised learning algorithm is directly applicable in this scenario. The difficulty partly stems from several distinct properties of training data in multimedia domains, such as unbalanced training distribution, limited positive labels and a large number of general examples. In this section we provide a brief overview and discussion on three directions to address these issues, which includes methods for predicting rare classes (Figure 6(a)), leveraging unlabeled data (Figure 6(b)) and scaling to large datasets (Figure 6(c)).

6.1 Predicting rare classes

Standard machine learning algorithms often assume that positive/negative data has a balance distribution, however, multimedia collections usually contain only a small fraction of positive examples for each concept. For example, only less than 8% of all shots are labeled as cityscape and less than 3% labeled as landscape in the

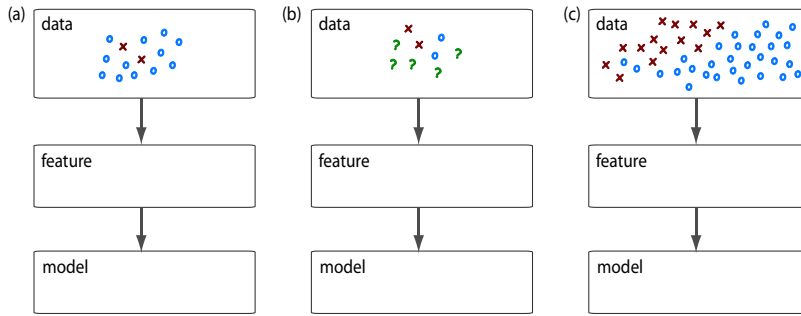


Fig. 6 Semantic concept modeling challenges brought up by data characteristics. (a) Learning a rare class, as described in Section 6.1. (b) Learning with unlabeled data, as described in Section 6.2. (c) Scaling to large amounts of training data, as described in Section 6.3.

TRECVID’02 development data. This is because the positive examples of a semantic concept is typically a coherent subset of images (e.g. cityscape, landscape and sunrise), but the negative class is less well-defined as “everything else” in the collection. Unfortunately, many learning algorithms will get into trouble when dealing with imbalanced datasets [70]. For instance, when the class distribution is too skewed, SVMs will generate a trivial model by predicting everything to the majority class. Japkowicz [43] shows that the data imbalance issue can significantly degrade prediction performance, especially when training data are non-linearly separable. Therefore, it is of crucial importance for us to address the rare data problem in the context of semantic extraction.

To date, there have been a few attempts to address the rare class problems in several applications, such as fraud detection [19], network intrusion, text categorization and web mining [49]. Two of the most popular solutions are named “over-sampling” which replicates positive data, and “under-sampling” which throws away part of negative data. They were designed to balance the data distribution and thus mitigate the data skewness problem in the training collection [97]. Although it is still an open question if artificially varying the training distribution can improve prediction performance with theoretical guarantee, Foster [97] provided some insights and qualitative analysis of the effectiveness on why tuning training distribution can be beneficial. To demonstrate, we apply over-sampling to the TRECVID’02 data using SVMs, altering the positive data distribution from 10% - 60%. Figure 7 shows the detection performance for “cityscape” with respect to precision, recall and F1-measure. We observe that SVMs always predict test examples as negative and thus yields zero precision/recall until the size of rare class examples is roughly comparable to the size of negative class examples. This observation again suggests that balancing training distribution is useful to improve the detection performance.

However, both under-sampling and over-sampling bear known drawbacks. Under-sampling is likely to eliminate some of the potentially useful examples and such loss of information may hurt the classification performance. Over-sampling, on the other hand, significantly increases the number of training data and thus consumes more

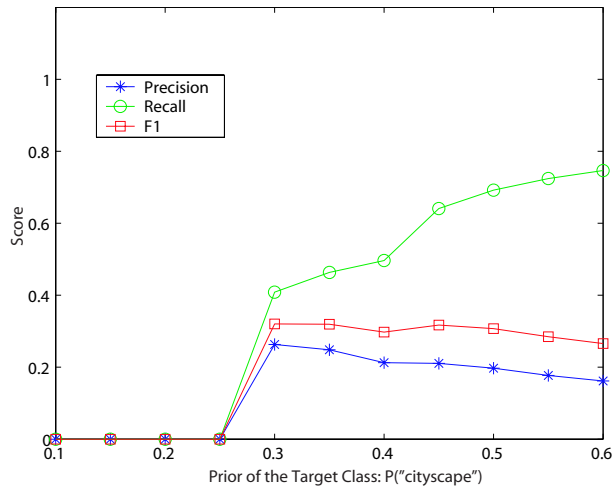


Fig. 7 The effect of modifying training distributions. Performance of the *Cityscape* concept classifier on TRECVID-2002 collection.

time in the learning process. This problem is critical to SVMs, since the training time complexity for SVMs is empirically close to quadratic of the number of support vectors, and cubic in the worse case [46]¹. In addition, overfitting is more likely to occur with replicated minor examples [97].

As an alternative to modifying skewed data distribution, ensemble-based approaches have been proposed in recent studies, of which the basic idea is to combine multiple individual classifiers on balanced data distributions. In [19], a multi-classifier meta-learning approach has been devised to deal with skewed class distributions. Joshi et al. [49] provided insights into the cases when AdaBoost, a strong ensemble-based learning algorithm, can achieve better precision and recall in the context of rare classes. It was found that the performance of AdaBoost for rare class is critically dependent on the learning abilities of the base classifiers. To bring the strengths of under-sampling and over-sampling together, Yan et al. [111] proposed an ensemble approach that first partitions negative data into small groups, constructs multiple classifiers using positive data as well as each group of negative data, and finally merges them via a top-level meta-classifier. Various classifier combination strategies are investigated including majority voting, sum rule, neural network and hierarchical SVMs. Experimental results show that this approach can achieve higher and more stable performance than over/under-sampling strategies in the TRECVID datasets.

Beyond augmenting learning algorithms, we can also consider modifying the training data sets. For example, it is possible to perturb the original positive examples (by adding white noises or information from other modalities) and create a

¹ Although linear-time algorithm has been derived for linear SVMs with an alternative formulation [48], no speedup of general SVMs is known.

larger set of synthetic positive examples so as to balance the data distribution. In this scenario, however, how to produce semantically correct but visually distinctive examples will become a major problem to address.

6.2 Leveraging Unlabeled Data

Successful concept detection outputs usually rely on a large annotated training corpus that contains a sufficient number of labeled image/video samples. In practice, however, the number of labeled video samples is usually few for most semantic concepts, since manual annotation is such a labor-intensive process. For instance, annotating 1 hour of broadcast news video with a lexicon of 100 semantic concepts can take anywhere between 8 to 15 hours [53]. This problem is further worsened given a large number of infrequently-appearing semantic concepts in video collections.

As a remedy for the label sparseness, a variety of semi-supervised learning algorithms have been developed in an attempt to leverage additional unlabeled data in the training collection. Moreover, multiple modalities in video streams further prompt us to consider multi-view learning strategies which can explicitly split the feature space into multiple subsets, or views. Combining semi-supervised learning and multi-view setting offers powerful tools to learn with unlabeled data and these approaches are generally called “multi-view semi-supervised learning”. Co-training [8] is one of the most well-known multi-view semi-supervised learning algorithms. It starts with two initial classifiers learned from separate views. Both classifiers are then incrementally updated in every iteration using an augmented labeled set, which includes additional unlabeled samples with the highest classification confidence in each view. Co-EM [65] can be viewed as a probabilistic version of co-training, which requires each classifier to provide class probability estimation for all unlabeled data. Collins and Singer [25] introduced the CoBoost algorithm which attempts to minimize the disagreement on the unlabeled data between classifiers of different views. This class of co-training type algorithms has been successfully applied to a variety of domains, including natural language processing [68], web page classification [8], information extraction [25] and visual detection [52].

Although identified as a potential application domain by the original co-training authors [8], applying co-training as-is yields poor performance in video concept detection. After examining the real-world video data, we realized that the failure of co-training in this domain can be partially attributed to the violation of its underlying assumptions which requires that each view be sufficient for learning the target concepts. For example, when color histograms are used to learn the video concept “airplane” of two video frames that have the same color histogram (e.g. white/gray on blue background), one can contain an airplane but the other may contain an eagle. Therefore, the view from low-level color features alone will not be sufficient to learn the underlying concepts. Empirically, Yan et. al. [113] found that co-training tends to produce lower average precision with more unlabeled data introduced with noisy

labels. In the domain of natural language processing, Pierce et al. [68] also observed the similar degradation of the co-training algorithm if the labeled data introduced by the other view is not accurate enough.

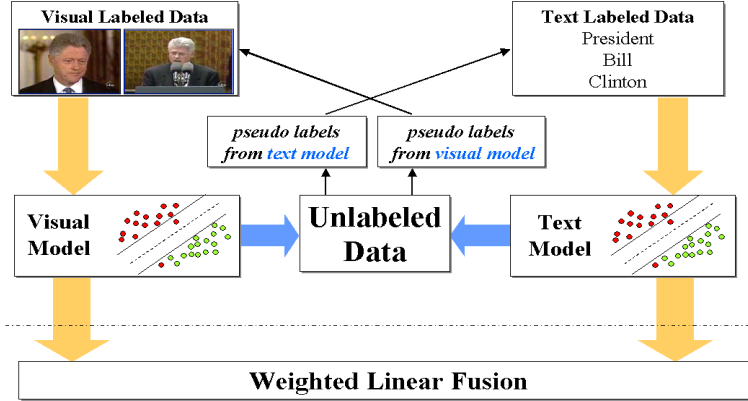


Fig. 8 Illustration of co-training in multi-modal learning.

Better semi-supervised learning algorithms should be able to guarantee that unlabeled data will at worst result in no significant performance degradation and at best improve performance over the use of the labeled data sets alone. Yan et al. [113] proposed a more effective algorithm called semi-supervised cross feature learning (SCFL) for concept detection. Unlike co-training which updates each classifier by incorporating the selected unlabeled data to augment the labeled set, SCFL learns separate classifiers from selected unlabeled data and combines them with the classifiers learned from noise-free labeled data. One advantage of the proposed approach is that it can theoretically prevent its performance from being significantly degraded even when the assumption of view sufficiency fails.

If further manual annotation is possible, we can enhance the semi-supervised learning by iteratively inquiring a human annotator to review and provide the correct labels for some selected unlabeled data. This type of problem is called “active learning” [22] or “selective sampling” [15] in the literature. An active learner begins with a pool of unlabeled data, selects a set of unlabeled examples to be manually labeled as positive or negative and learn from the newly obtained knowledge repetitively. Typically, the unlabeled examples can be selected by means of either minimizing the learner’s expected error [15] or maximizing the information gain or version space reduction [92]. The effectiveness of active learning for reducing annotation cost in semantic concept detection has been demonstrated by a number of investigations [22, 63, 109, 92]. Note that active learning and co-training can be combined, e.g., corrected co-training [68] and co-testing [58], which require users to annotate the selected unlabeled data from the co-training algorithm. Applying corrected co-training [112] to semantic concept detection shows a considerable performance improvement over initial classification results.

6.3 Scalability

Real-world multimedia collections easily contain hundred thousands or even millions of data. For example, photo sharing site Flickr.com has 3500~4000 new photos every minute, translating to 5 million per day, and 2 billion per year. Moreover, the target lexicon is scaled from a few concepts to several hundred concepts, just as the research and development has progressed in the past few years [34, 1]. The computational requirements for concept detection are increasing significantly under the dual growth of data \times concepts. However, most of the existing algorithms do not scale well to such a high computational demand. For example, current support vector machine(SVM) implementations have a learning time of $O(mn^2)$ and a prediction time of $O(mn)$ on a non-linear kernel, with m the feature dimensions and n the dataset size. Therefore, the computational resources needed to learn millions of data will be prohibitive even after negative data are down sampled. A simultaneous focus on learning and classification efficiency is needed to perform detection over the large lexicon of concepts, it should be at a speed at least an order of magnitude faster than the current processing without compromising the detection accuracy.

To speed up machine learning process without performance degradation, one approach is to exploit the information redundancy in the learning space. There are a large body of previous work on reducing the computational complexity of SVMs, such as [28, 14, 54]. The attempt is to either reduce the number of training samples offered to the learner, sample the large number of support vectors that are generated by the learner, or create new learning functions to approximate the current prediction function without losing the generalization ability or accuracy. Along another line, researchers also proposed several efficient ensemble learning algorithms based on random feature selection and data bootstrapping. Breiman has developed *random forest* [10], which aggregates an ensemble of unpruned classification/regression trees using both bootstrapped training examples and random feature selection, outperforming a single tree classifier in experiments. Ensemble learning approaches are not limited to tree classifiers. For instance, asymmetric bagging and random subspace classifiers [88] were used in an image retrieval task, with a strategy similar to that of random forest.

To further reduce the information redundancy across multiple labels, Yan et al. [114] proposed a boosting-type learning algorithm called model-shared subspace boosting (MSSBoost). It can automatically find, share and combine a number of random subspace models across multiple labels. This algorithm is able to reduce the information redundancy in the label space by jointly optimizing the loss functions over all the labels. Meanwhile, this approach enjoys the advantage of being built on small base models, learned on a small number of bootstrap data samples and a randomly selected feature subspace. The experimental results on a synthetic dataset and two real-world multimedia collections have demonstrated that MSSBoost can outperform the non-ensemble baseline classifiers with a significant speedup on both the learning and prediction process. It can also use a smaller number of base models to achieve the same classification performance as its non-model-shared counterpart.

7 Retrieval with semantic concepts

By introducing semantic concepts as intermediate layer in multimedia retrieval, a new retrieval approach called concept-based retrieval has recently emerged. It utilizes a set of semantic concepts to describe visual content in multimedia collections, and maps the user queries to identify the relevant/irrelevant concepts for combination. Since semantic concepts can serve as a bridge between query semantics and content semantics, concept-based retrieval is able to capture the information needs in a more effective way and thus improve the retrieval performance. In the following discussion, we briefly describe the use and utility of semantic concepts in assisting multimedia retrieval.

7.1 *The utility of semantic concepts*

Semantic concepts can be categorized into two types. One type consists of general concepts with frequent appearances and sufficient training examples to represent their characteristics. These concepts can often be learned with a reasonable prediction accuracy. For instance, in broadcast news collection, *anchor person*, *outdoors*, *cars and roads* belong to this type of concepts. In contrast, the other type of concepts consists of more specific concepts with less frequent occurrence. Thus, the number of their training examples is usually insufficient and less representative. In some sense, the detection of rare concepts is similar to a retrieval problem (with few training examples) rather than a classification problem. *Prisoner*, *physical violence* are two examples of this type of semantic concepts.

The distinctions between these two concept types consequently suggest different utilities in the retrieval task [23]. For instance, the common semantic concepts often have universal predictive powers over a large number of queries, and their association with query topics can probably be learned from a large training collection. On the other hand, the usefulness of rare semantic concepts is limited to merely a small number of domains. Therefore, they are more appropriate to be applied in domain-specific queries.

7.2 *Use of semantic concepts for automatic retrieval*

To illustrate how semantic concepts can be used in multimedia retrieval, we discuss four most common types of concept-based retrieval methods. The simplest approach is to match each concept name with query terms. If a concept is found to be relevant, its detection outputs can be used to refine the retrieval results. For example, the concept “building” will be helpful for retrieving the query of “finding the scenes containing buildings in New York City”. This method is intuitive to understand and simple to implement. However, it is unrealistic to expect a general user to explicitly

indicate all related concepts in a query description. For example, the concept of “outdoor” could be useful for the query of “finding people on the beach”, but it does not show up in the query directly.

To extend the power of simple query matching, we can follow the idea of global query analysis in text retrieval, which attempts to enrich query description from external knowledge sources, such as WordNet [32]. These approaches have shown promising retrieval results [96, 64] by leveraging extra concepts. However, these approaches are also likely to bring in noisy concepts, and thus lead to unexpected deterioration of search results. Moreover, even when the subset of relevant concepts are perfectly identified, it remains a challenge to derive a good strategy to combine semantic concepts with other text/image retrieval results.

As an alternative, we can leverage semantic concepts by learning the combination strategies from training collections, e.g., learning query-independent combination models [4] and query-class dependent combination models [115]. These approaches can automatically determine concept weights and handle hidden semantic concepts. However, since these learning approaches can only capture the general patterns that distinguish relevant and irrelevant training documents, their power is usually limited by the number of available training data.

Finally, we can also consider local analysis approaches that adaptively leverage semantic concepts on a per query basis. The essence of local, or re-ranking strategies is to utilize initial retrieved documents to select expanded discriminative query concepts to improve the retrieval performance. For example, we proposed a retrieval approach called probabilistic local context analysis (pLCA) [108], which can automatically leverage useful high-level semantic concepts based on initial retrieval output. However, the success of these approaches usually relies on reasonably accurate initial search results. If initial retrieval performance is unsatisfactory, it is possible for local analysis approaches to degrade the retrieval results.

A more complete survey for multimedia retrieval can be found at [110]. To summarize, all four types of approaches have proved to be successful in utilizing high-level semantic concepts for video retrieval, despite their own limitations. Moreover, these methods are not mutually exclusive, a composite strategy usually produces better results than any single approach. How to automatically determine the best strategy or strategies to incorporate high-level concepts into video retrieval is an interesting direction for future exploration.

8 Discussions and summary

In this chapter we presented the general approaches and active research directions to semantic extraction from multimedia. We discussed the five main components in semantic modeling, followed by a selection of challenges and solutions in real-world media processing tasks for each component: the design of a multimedia lexicon and the use of it to help concept detection; handling multiple sources of input and a special case of resolving correspondence between images and text annotations; us-

ing structured (generative) models to account of natural data dependency or model hidden topics; handling rare classes, leveraging unlabeled data, and scale to large amounts of training data; finally the use of media semantics in automatic and interactive retrieval systems.

At the end of this review, we would like to present our views on a few challenges ahead: (1) Scale concept detection with high accuracy to massive amounts of training data and a large number of concepts. Currently some concepts have notably higher performance than others, for instance, *people, face, outdoors, nature* typically have very accurate top results, while *court, desert, glacier* are yet to improved due to their diverse appearance lack of sufficient labeled examples., In order to scale to thousands of concepts, the algorithms and the computational architecture also need to evolve and keep up the pace, this may mean both new paradigms for semantic learning and efficient parallel computing structures. (2) Generalize semantic extraction to many data domains. Currently the tight coupling of training and testing data makes lengthy cycles for learning and deploying semantic models. Clever algorithms are called for in deciding which concepts will generalize well, and how to quickly adapt to domain characteristics. (3) Effective use of unstructured media and structured metadata. Media semantics do not exist in isolation, neither do the people who capture and consume them. Successful use of structured metadata, such as time, location, author or social relationships should mitigate semantic diversity and alleviate the problem of insufficient training data.

References

1. LSCOM lexicon definitions and annotations: Dto challenge workshop on large scale concept ontology for multimedia. <http://www.ee.columbia.edu/dvmm/lscom/>.
2. YouTube comprises 10% of all internet traffic. <http://www.webpronews.com/topnews/2007/06/19/youtube-comprises-10-of-all-internet-traffic>.
3. Looking high and low: Who will be the google of video search? *Wired*, june 2007. <http://www.wired.com/techbiz/media/magazine/15-07/st.videominig>.
4. A. Amir, W. Hsu, G. Iyengar, C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang. IBM research TRECVID-2003 video retrieval system. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, Nov 2003.
5. K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2002.
6. D. Blei and M. Jordan. Modeling annotated data. In *Proc. of the 26th ACM Intl. Conf. on SIGIR*, 2003.
7. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993 – 1022, 2003.
8. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. of the Workshop on Computational Learning Theory*, 1998.
9. M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 994, Washington, DC, USA, 1997. IEEE Computer Society.
10. L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

11. R. Brunelli, O. Mich, and C. M. Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78–112, June 1999.
12. C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, New York, NY, USA, 2000. ACM.
13. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):955–974, 1998.
14. C. J. C. Burges and B. Schölkopf. Improving the accuracy and speed of support vector machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 375. The MIT Press, 1997.
15. C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *Proc. 17th International Conference on Machine Learning(ICML00)*, pages 111–118, 2000.
16. M. Campbell, S. Ebadollahi, M. Naphade, A. P. Natsev, J. R. Smith, J. Tesic, L. Xie, K. Scheinberg, J. Seidl, A. Haubold, and D. Joshi. IBM research trecvid-2006 video retrieval system. In *NIST TRECVID Workshop*, Gaithersburg, MD, November 2006.
17. M. Campbell, A. Haubold, M. Liu, A. P. Natsev, J. R. Smith, J. Tesic, L. Xie, R. Yan, and J. Yang. IBM research trecvid-2007 video retrieval system. In *NIST TRECVID Workshop*, Gaithersburg, MD, November 2006.
18. L. Chaisorn, T.-S. Chua, and C.-H. Lee. A multi-modal approach to story segmentation for news video. *World Wide Web*, 6(2):187–208, 2003.
19. P. Chan and S. Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proc. Fourth Intl. Conf. Knowledge Discovery and Data Mining*, pages 164–168, 1998.
20. S.-F. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D. Zhang. Columbia university TRECVID-2005 video search and high-level feature extraction. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2005.
21. S. F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):688–695, June 2001.
22. M.-Y. Chen, M. Christel, A. Hauptmann, and H. Wactlar. Putting active learning into multimedia applications - dynamic definition and refinement of concept classifiers. In *Proceedings of ACM Intl. Conf. on Multimedia*, Singapore, November 2005.
23. M. Christel and A. G. Hauptmann. The use and utility of high-level semantic features. In *Proc. of Intl. Conf. on Image and Video Retrieval (CIVR)*, Singapore, 2005.
24. M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 189–196, 1999.
25. M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proc. of EMNLP*, 1999.
26. A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002.
27. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 1990.
28. T. Downs, K. E. Gates, and A. Masters. Exact simplification of support vector solutions. *J. Mach. Learn. Res.*, 2:293–297, 2002.
29. L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu. A mid-level representation framework for semantic sports video analysis. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 33–44, New York, NY, USA, 2003. ACM Press.
30. J. Duffy. Video drives net traffic. *PC World*, august 2007. <http://www.pcworld.com/article/id,136069-pg,1/article.html>.
31. S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith. Visual event detection using multi-dimensional concept dynamics. In *Interational Conference on Multimedia and Expo (ICME)*, Toronto, Canada, July 2006.

32. C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
33. B. Gold and N. Morgan. *Speech and audio signal processing*. Wiley New York, 2000.
34. G. Griffin, A. Holub, and P. Perona. Caltech-256 Object Category Dataset. Technical report, Caltech, 2007.
35. A. Hauptmann, M.-Y. Chen, M. Christel, C. Huang, W.-H. Lin, T. Ng, N. Papernick, A. Velivelli, J. Yang, R. Yan, H. Yang, and H. D. Wactlar. Confounded Expectations: Informedia at TRECVID 2004. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2004.
36. A. G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang. CMU Informedia's TRECVID 2005 Skirmishes. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2005.
37. A. Hesseldahl. Micron's megapixel movement. *BusinessWeek*, 2006.
38. T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd Intl. ACM SIGIR conference*, pages 50–57, Berkeley, California, United States, 1999.
39. B. Horn. *Robot Vision*. McGraw-Hill College, 1986.
40. C. Hsu, C. Chang, C. Lin, et al. A practical guide to support vector classification. *National Taiwan University, Tech. Rep., July*, 2003.
41. J. Huang, S. Ravi Kumar, M. Mitra, W. Zhu, and R. Zabih. Spatial Color Indexing and Applications. *International Journal of Computer Vision*, 35(3):245–268, 1999.
42. G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S. P. Khudanpur, D. Klakow, M. R. Krause, R. Manmatha, H. J. Nock, D. Petkova, B. Pytlik, and P. Virga. Joint visual-text modeling for automatic retrieval of multimedia documents. In *Proceedings of ACM Intl. Conf. on Multimedia*, November 2005.
43. N. Japkowicz. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI Workshop on Learning from Imbalanced Data Sets. Tech Rep. WS-00-05, Menlo Park, CA: AAAI Press*, 2000.
44. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual ACM SIGIR conference on informaion retrieval*, pages 119–126, Toronto, Canada, 2003.
45. R. Jin, J. Y. Chai, and S. Luo. Automatic image annotation via coherent language model and active learning. In *Proceedings of ACM Intl. Conf. on Multimedia*, November 2004.
46. T. Joachims. Making large-scale svm learning practical. In *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf and C. Burges and A. Smola (ed.), Springer, 1995.
47. T. Joachims. Making large-scale support vector machine learning practical. In A. S. B. Schölkopf, C. Burges, editor, *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA, 1998.
48. T. Joachims. Training linear svms in linear time. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226, New York, NY, USA, 2006. ACM.
49. M. Joshi, R. Agarwal, and V. Kumar. Predicting rare classes: Can boosting make any weak learner strong? In *the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, July 2002*.
50. D. Jurafsky and J. Martin. *Speech and language processing*. Prentice Hall, Upper Saddle River, NJ, 2000.
51. J. Kender. A large scale concept ontology for news stories: Empirical methods, analysis, and improvements. In *IEEE International Conference on Multimedia and Expo (ICME)*, Beijing, China, 2007.
52. A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using cotraining. In *Proc. of the Intl. Conf. on Computer Vision*, 2003.
53. C. Lin, B. Tseng, and J. Smith. VideoAnnEx: IBM MPEG-7 annotation tool for multimedia indexing and concept learning. In *IEEE International Conference on Multimedia and Expo*, Baltimore, MD, 2003.
54. K. Lin and C. Lin. A study on reduced support vector machines. *IEEE Transactions on Neural Networks*, 14(6), 2003.

55. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.
56. J. Markel. The SIFT algorithm for fundamental frequency estimation. *Audio and Electroacoustics, IEEE Transactions on*, 20(5):367–377, 1972.
57. I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3):305–317, 2005.
58. I. Muslea, S. Minton, and C. A. Knoblock. Active semi-supervised learning = robust multi-view learning. In *Proc. of Intl. Conf. on Machine Learning*, 2002.
59. M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in geo-referenced digital photo collections. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 196–203, New York, NY, USA, 2004. ACM Press.
60. M. Naphade, L. Kennedy, J. Kender, S. Chang, J. Smith, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for TRECVID 2005. Technical report, IBM Research, 2005.
61. M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.
62. M. R. Naphade, T. Kristjansson, B. Frey, and T. Huang. Probabilistic multimedia objects (multijets): A novel approach to video indexing and retrieval in multimedia systems. In *Proc. of IEEE International Conference on Image Processing (ICIP)*, pages 536–540, 1998.
63. M. R. Naphade and J. R. Smith. Active learning for simultaneous annotation of multiple binary semantic concepts. In *Proceedings of IEEE International Conference On Multimedia and Expo (ICME)*, pages 77–80, Taipei, Taiwan, 2004.
64. S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *Proceedings of the Conference on Image and Video Retrieval (CIVR)*, pages 370–379, Singapore, 2006.
65. K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proc. of CIKM*, pages 86–93, 2000.
66. N. M. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.
67. J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Gcap: Graph-based automatic image captioning. In *Proc. of the 4th International Workshop on Multimedia Data and Document Engineering (MDDE 04), in conjunction with Computer Vision Pattern Recognition Conference (CVPR 04)*, 2004.
68. D. Pierce and C. Cardie. Limitations of co-training for natural language learning from large datasets. In *Proc. of EMNLP*, 2001.
69. J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 1999.
70. F. Provost. Machine learning from imbalanced data sets 101/1. In *AAAI Workshop on Learning from Imbalanced Data Sets. Tech Rep. WS-00-05, Menlo Park, CA: AAAI Press*, 2000.
71. B. Pytlík, A. Ghoshal, D. Karakos, and S. Khudanpur. Trecvid 2005 experiment at Johns Hopkins University: Using hidden Markov models for video retrieval. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2005.
72. L. Rabiner and B. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1993.
73. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, feb 1989.
74. W. Roush. Tr10: Peering into video's future. *Technology Review*, march 2007. <http://www.technologyreview.com/Infotech/18284/?a=f>.
75. H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):23–38, 1998.

76. Y. Rui, T. Huang, and S. Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, 1999.
77. E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *ICASSP '97: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)-Volume 2*, page 1331, Washington, DC, USA, 1997. IEEE Computer Society.
78. J. Schlenzig, E. Hunter, and R. Jain. Recursive identification of gesture inputs using hidden Markov models. In *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, pages 187–194. IEEE Computer Society Press, 1994.
79. B. Shevade, H. Sundaram, and L. Xie. Modeling personal and social network context for event annotation in images. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 2007.
80. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval: the end of the early years. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 12:1349 – 1380, 2000.
81. C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Trans. Multimedia*, 2007.
82. C. Snoek, M. Worring, J. Geusebroek, D. Koelma, and F. Seinstra. The MediaMill TRECVID 2004 semantic video search engine. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2004.
83. C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of ACM Intl. Conf. on Multimedia*, pages 399–402, Singapore, November 2005.
84. C. G. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.
85. M. Srikanth, M. Bowden, and D. Moldovan. LCC at trecvid 2005. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2005.
86. T. Starner, A. Pentland, and J. Weaver. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1371–1375, 1998.
87. M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.
88. D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(7):1088–1099, 2006.
89. C. Teh and R. Chin. On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496–513, 1988.
90. The National Institute of Standards and Technology (NIST). TREC video retrieval evaluation, 2001–2007. <http://www-nlpir.nist.gov/projects/trecvid/>.
91. The National Institute of Standards and Technology (NIST). Common evaluation measures, 2002. <http://trec.nist.gov/pubs/trec11/appendices/MEASURES.pdf>.
92. S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of ACM Intl. Conf. on Multimedia*, pages 107–118, 2001.
93. J. van Gemert. Retrieving images as text, 2003. Master Thesis, University of Amsterdam.
94. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
95. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proc. CVPR*, 1:511–518, 2001.
96. T. Volkmer and A. Natsev. Exploring automatic query refinement for text-based video retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 765–768, Toronto, ON, 2006.
97. G. Weiss and F. Provost. The effect of class distribution on classifier learning. Technical report, Department of Computer Science, Rutgers University, 2001.
98. T. Westerveld. *Using generative probabilistic models for multimedia retrieval*. PhD thesis, CWI, Centre for Mathematics and Computer Science, 2004.

99. Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 572–579, New York, NY, USA, 2004.
100. Y. Wu, B. L. Tseng, and J. R. Smith. Ontology-based multi-classification learning for video concept detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, 2004.
101. L. Xie and S.-F. Chang. Pattern mining in visual concept streams. In *International Conference on Multimedia and Expo (ICME)*, Toronto, Canada, July 2006.
102. L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden Markov models. In *Proc. International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Orlando, FL, 2002.
103. L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. *Unsupervised Mining of Statistical Temporal Structures in Video*, chapter 10. Kluwer Academic Publishers, 2003.
104. L. Xie, D. Xu, S. Ebadollahi, K. Scheinberg, S.-F. Chang, and J. R. Smith. Pattern mining in visual concept streams. In *Proc. 40th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Oct 2006.
105. E. P. Xing, R. Yan, and A. G. Hauptmann. Mining associated text and images using dual-wing harmoniums. In *Uncertainty in Artificial Intelligence (UAI)'05*, 2005.
106. D. Xu and S.-F. Chang. Visual event recognition in news video using kernel methods with multi-level temporal alignment. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, June 2007.
107. C. Xu et. al. *Sports Video Analysis: from Semantics to Tactics*. Springer, 2008.
108. R. Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2006.
109. R. Yan and A. G. Hauptmann. Multi-class active learning for video semantic feature extraction. In *Proceedings of IEEE International Conference On Multimedia and Expo (ICME)*, pages 69–72, Taipei, Taiwan, 2004.
110. R. Yan and A. G. Hauptmann. A review of text and image retrieval approaches for broadcast news video. *Inf. Retr.*, 10(4-5):445–484, 2007.
111. R. Yan, Y. Liu, R. Jin, and A. Hauptmann. On predicting rare class with SVM ensemble in scene classification. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, 2003.
112. R. Yan and M. R. Naphade. Co-training non-robust classifiers for video semantic concept detection. In *Proc. of IEEE Intl. Conf. on Image Processing (ICIP)*, 2005.
113. R. Yan and M. R. Naphade. Semi-supervised cross feature learning for semantic concept detection in video. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, San Diego, US, 2005.
114. R. Yan, J. Tesic, and J. R. Smith. Model-shared subspace boosting for multi-label classification. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 834–843, New York, NY, USA, 2007. ACM.
115. R. Yan, J. Yang, and A. G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 548–555, New York, NY, USA, 2004.
116. R. Yan, M. Yu Chen, and A. G. Hauptmann. Mining relationship between video concepts using probabilistic graphical model. In *IEEE International Conference on Multimedia and Expo (ICME)*, Toronto, Canada, 2006.
117. J. Yang, M. Y. Chen, and A. G. Hauptmann. Finding person X: Correlating names with visual appearances. In *Proc. of the Intl. Conf. on Image and Video Retrieval (CIVR)*, pages 270–278, Dublin, Ireland, 2004.
118. A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.
119. Y. Zhai, X. Chao, Y. Zhang, O. Javed, A. Yilmaza, F. Rafi, S. Ali, O. Alatas, S. Khan, and M. Shah. University of Central Florida at TRECVID 2004. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD, 2004.