Chapter XVI

Semantic Multimedia Information Analysis for Retrieval Applications

João Magalhães, Imperial College London, UK

Stefan Rüger, Imperial College London, UK

Abstract

Most of the research in multimedia retrieval applications has focused on retrieval by content or retrieval by example. Since the classical review by Smeulders, Worring, Santini, Gupta, and Jain (2000), a new interest has grown immensely in the multimedia information retrieval community: retrieval by semantics. This exciting new research area arises as a combination of multimedia understanding, information extraction, information retrieval, and digital libraries. This chapter presents a comprehensive review of analysis algorithms in order to extract semantic information from multimedia content. We discuss statistical approaches to analyze images and video content and conclude with a discussion regarding the described methods.

Introduction: Multimedia Analysis

The growing interest in managing multimedia collections effectively and efficiently has created new research interest that arises as a combination of multimedia understanding, information extraction, information retrieval, and digital libraries. This growing interest has resulted in the creation of a video retrieval track in TREC conference series in parallel with the text retrieval track (TRECVID, 2004).

Figure 1 illustrates a simplified multimedia information retrieval application composed by a multimedia database, analysis algorithms, a description database, and a user interface application. Analysis algorithms extract features from multimedia content and store them as descriptions of that content. A user then deploys these indexing descriptions in order to search the multimedia database. A semantic multimedia information retrieval application (Figure 1) differs eminently from traditional retrieval applications on the low-level analysis algorithms; its algorithms are responsible for extracting semantic information used to index multimedia content by its semantic. Multimedia content can be indexed in many ways, and each index can refer to different modalities and/or parts of the multimedia piece. Multimedia content is composed of the visual track, sound track, speech track, and text. All these modalities are arranged temporally to provide a meaningful way to transmit information and/or entertainment. The way video documents are temporally structured can be distinguished in two levels: semantic and syntactic structure (Figure 2).

At the syntactic level, the video is segmented into shots (visual or audio) that form a uniform segment (e.g., visually similar frames); representative key-frames are extracted from each shot, and scenes group neighboring similar shots into a single segment. The segmentation of video into its syntactic structure of video has been studied widely (Brunelli, Mich, & Modena, 1999; Wang, Liu, & Huang, 2000).

Figure 1. A typical multimedia information retrieval application

Figure 2. Syntactic and semantic structure of video

At the semantic level, annotations of the key-frames and shots with a set of labels indicate the presence of semantic entities, their relations, and attributes (agent, object, event, concept, state, place, and time (see Benitez et al., 2002, for details). Further analysis allows the discovery of logical sub-units (e.g., substory or subnarrative), logical units (e.g., a movie), and genres. A recent review of multimedia semantic indexing has been published by Snoek and Worring (2005).

The scope of this chapter is the family of semantic-multimedia analysis algorithms that automate the multimedia semantic annotation process. In the following sections, we will review papers on multimedia-semantic analysis: semantic annotation of key-frame images, shots, and scenes. The semantic analysis at the shot and scene level considers independently the audio and visual modalities and then the multi-modal semantic analysis. Due to the scope of this book, we will give more emphasis to the visual part than to the audio part of the multimedia analysis and will not cover the temporal analysis of logical substories, stories, and genres.

Key-Frame Semantic Annotation

Image analysis and understanding is one of the oldest fields in pattern recognition and artificial intelligence. A lot of research has been done since (Marr, 1983), culminating in the modern reference texts by Forsyth and Ponce (2003) and Hartley and Zisserman (2004). In the following sections we discuss different types of visual information analysis algorithms: single class models fit a simple probability density distribution to each label; translation models define a visual vocabulary and a method to translate from this vocabulary to keywords; hierarchical and network models explore the interdependence of image elements (regions or tiles) and model its structure; knowledge-based models improve the model's accuracy by including other sources of knowledge besides the training data (e.g., a linguistic database such as WordNet).

Single Class Models

A direct approach to the semantic analysis of multimedia is to learn a class-conditional probability distribution $p(w|x)$ of each single keyword w of the semantic vocabulary, given its training data x (see Figure 3). This distribution can be obtained by using Bayes' law

$$p(w|x) = \frac{p(x|w)p(w)}{p(x)}.$$

The data probability $p(x)$ and the keyword probability $p(w)$ can be computed straightforward, and the $p(x|w)$ can be computed with very different data density distribution models.

Several techniques to model the $p(x|w)$ with a simple density distribution have been proposed: Yavlinsky, Schofield, and Rüger (2005) used a nonparametric distribution, Carneiro and Vasconcelos (2005) a semi-parametric density estimation, Westerveld and de Vries (2003) a finite mixture of Gaussians, and Mori, Takahashi, and Oka (1999), Vailaya, Figueiredo, Jain, and Zhang (1999), and Vailaya, Figueiredo, Jain, and Zhang (2001) different flavors of vector quantization techniques.

Yavlinsky et al. (2005) modeled the probability density of images, given keywords as a nonparametric density smoothed by two kernels: a Gaussian kernel and an Earth Mover's Distance kernel. They used both global and 3×3 tile color and texture features. The best reported mean average precision (MAP) results with tiles achieved 28.6% MAP with the dataset of Duygulu, Barnard, de Freitas, and Forsyth (2002) and 9.2% with a Getty Images dataset.

Yavlinsky et al. (2005) showed that a simple nonparametric statistical distribution can perform as well or better than many more sophisticated techniques (e.g., translation models). However, the nonparametric density nature of their framework makes the task of running the

Figure 3. Inference of single class models

model on new data very complex. The model is the entire dataset meaning that the demands on CPU and memory increase with the training data.

Westerveld and de Vries (2003) used a finite-mixture density distribution with a fixed number of components to model a subset of the DCT coefficients:

$$p(x|\theta) = \sum_{m=1}^k \alpha_m p(x|\mu_m, \sigma_m^2),$$

in which k is the number of components, θ represents the complete set of model parameters with mean μ_m , covariance σ_m^2 , and component prior α_m . The component priors have the constraints $\alpha_1, \dots, \alpha_k \geq 0$ and $\sum_{m=1}^k \alpha_m = 1$. Westerveld (2003) tested several scenarios to evaluate the effect (a) of the number of mixture components, (b) of using different numbers of DCT coefficients (luminance and chrominance), and (c) of adding the coordinates of the DCT coefficients to the feature vectors. The two first factors produced varying results, and optimal points were found experimentally. The third tested aspect, the presence of the coefficients position information, did not modify the results.

Marrying the two previous approaches, Carneiro and Vasconcelos (2005) deployed a hierarchy of semi-parametric mixtures to model $p(x|n)$ using a subset of the DCT coefficients as low-level features. Vasconcelos and Lippman (2000) had already examined the same framework in a content-based retrieval system.

The hierarchy of mixtures proposed by Vasconcelos and Lippman (1998) can model data at different levels of granularity with a finite mixture of Gaussians. At each hierarchical level l , the number of each mixture component k^l differs by one from adjacent levels. The hierarchy of mixtures is expressed as:

$$p(x|w_i) = \frac{1}{D} \sum_{m=1}^{k^l} \alpha_{i,m}^l p(x|\theta_{i,m}^l).$$

The level $l=1$ corresponds to the coarsest characterization. The more detailed hierarchy level consists of a nonparametric distribution with a kernel placed on top of each sample. The only restriction on the model is that if node m of level $l+1$ is a child of node n of level l , then they are both children of node p of level $l-1$. The EM algorithm computes the mixture parameters at level l , given the knowledge of the parameters at level $l+1$, forcing the previous restriction.

Carneiro and Vasconcelos (2005) report the best published retrieval MAP of 31% with the dataset of Duygulu et al. (2002). Even though we cannot dissociate this result from the pair of features and statistical model, the hierarchy of mixtures appears to be a very powerful density distribution technique.

Even though the approaches by Carneiro and Vasconcelos (2005) and Westerveld and de Vries (2003) are similar, the differences make it difficult to do a fair comparison. The DCT features are used in a different way, and the semi-parametric hierarchy of mixtures can model classes with very few training examples.

The relationship between finite-mixture density modeling and vector quantization is a well-studied subject (see Hastie, Tibshirani, & Friedman, 2001). One of the applications of vector

quantization to image retrieval and annotation was realized by Mori et al. (1999). Given the training data of a keyword, they divide the images into tiles and apply vector quantization to the image tiles in order to extract the codebook used to estimate the $p(x | w)$ density distribution. Later, they use a model of word co-occurrence on the image tiles in order to label the image. The words with the higher sum of probabilities across the different tiles are the ones assigned to that image.

Vailaya et al. (1999) and Vailaya et al. (2001) describe a Bayesian framework using a codebook to estimate the density distribution of each keyword. They show that the Minimum Description Length criterion selects the optimal size of the codebook extracted from the vector quantizer. The features are extracted from the global image, and there is no image tiling. The use of the MDL criterion makes this framework quite elegant and defines a statistical criterion to select every model parameter and without any user-defined parameters.

Translation Models

All of the previous approaches employ a direct model to estimate $p(x | w)$ with image global features and/or image tiles features. In contrast to this, the vector quantization (usually k -means) approach generates a codebook of image regions or image tiles (depending on the segmentation solution). The problem then is formulated as a translation problem between two representations of the same entity: English-Esperanto, word-blob codebook, or word-tile codebook.

Inspired by machine translation research, Duygulu et al. (2002) developed a method of annotating image regions with words. First, regions are created using a segmentation algorithm like normalized cuts (Shi & Malik, 2000). For each region, features are computed, and then blobs are generated by clustering the regional image features across an image collection. The problem then is formulated as learning the correspondence between the discrete vocabulary

Figure 4. Translation models

of blobs and the image keywords. The model consists of a mixture of correspondences for each word of each image in the collection:

$$p(w_j | I_n) = \sum_{i \in \{\text{blobs in } I_n\}} p(a_{nj} = i) p(w = w_{nj} | b = b_i),$$

in which $p(a_{nj} = i)$ expresses the probability of associating word j to blob i in image n , and $p(w = w_{nj} | b = b_i)$ is the probability of obtaining an instance of word w given an instance of blob b . These two probability distributions are estimated with the EM algorithm. The authors refined the lexicon by clustering indistinguishable words and ignoring the words with probabilities $p(w | b)$ below a given threshold.

The machine translation approach, the thorough experiments, and the dataset form strong points of this chapter (Duygulu et al., 2002). This dataset is nowadays a reference, and thorough experiments showed that (a) their method could predict numerous words with high accuracy, (b) increasing the probability threshold improved precision but reduced recall, and (c) the word clustering improved recall and precision.

Following a translation model, Jeon, Lavrenko, and Manmatha (2003), Lavrenko, Manmatha, and Jeon (2003), and Feng, Lavrenko, and Manmatha (2004) studied a model in which blob features $b_i^{(r)}$ of an image I are assumed to be conditionally independent of keywords w_i , that is:

$$\begin{aligned} p(w_i, b_i) &= \sum_{J \in D} P(J) P(w_i | J) P(b_i | J) \\ &= \sum_{J \in D} P(J) P(w_i | J) \prod_{r \in I} P(b_i^{(r)} | J) \end{aligned}$$

Note that $b_i^{(r)}$ and w_i are conditionally independent, given the image collection D and that $J \in D$ act as the hidden variables that generated the two distinct representations of the same process (words and features).

Jeon et al. (2003) recast the image annotation as a cross-lingual information retrieval problem, applying a cross-media relevance model based on a discrete codebook of regions. Lavrenko et al. (2003) continued their previous work (Jeon et al., 2003) and used continuous probability density functions $p(b_i^{(r)} | J)$ to describe the process of generating blob features and to avoid the loss of information related to the generation of the codebook. Extending their previous work, Feng et al. (2004) replaced blobs with tiles and modeled image keywords with a Bernoulli distribution. This last work reports their best results, a MAP of 30%, with a Corel dataset (Duygulu et al., 2002).

Latent semantic analysis is another technique of text analysis and indexing; it looks at patterns of word distributions (specifically, word co-occurrence) across a set of documents (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). A matrix M of word occurrences in documents is filled with each word frequency in each document. The singular value decomposition (SVD) of matrix M gives the transformation to a singular space in which projected documents can be compared efficiently.

Hierarchical Models

The aforementioned approaches assumed a minimal relation among the various elements of an image (blobs or tiles). This section and the following section will review methods that consider a hierarchical relation or an interdependence relation among the elements of an image (words and blobs or tiles).

Barnard and Forsyth (2001) studied a generative hierarchical aspect model, which was inspired by Hofmann and Puzicha's (1998) hierarchical clustering/aspect model. The data are assumed to be generated by a fixed hierarchy of nodes in which the leaves of the hierarchy correspond to soft clusters. Mathematically, the process for generating the set of observations O associated with an image I can be described by:

$$p(O|I) = \sum_c \left(p(c) \prod_{o \in O} \left(\sum_l p(o|l, c) p(l|c, I) \right) \right), \quad O = \{w_1, \dots, w_n, b_1, \dots, b_m\},$$

in which c indexes the clusters, o indexes words and blobs, and l indexes the levels of the hierarchy. The level and the cluster uniquely specify a node of the hierarchy. Hence, the probability of an observation $p(o|l, c)$ is conditionally independent given a node in the tree. In the case of words, $p(o|l, c)$ assumes a tabular form, and in the case of blobs, a single Gaussian models the regions' features. The model is estimated with the EM algorithm.

Blei and Jordan (2003) describe three hierarchical mixture models to annotate image data, culminating in the correspondence latent Dirichlet allocation model. It specifies the following joint distribution of regions, words, and latent variables (θ, z, y) :

$$p(r, w, \theta, z, y) = p(\theta | \alpha) \left(\prod_{n=1}^N p(z_n | \theta) p(r_n | z_n, \mu, \sigma) \right) \cdot \left(\prod_{m=1}^M p(y_m | N) p(w_m | y_m, z, \beta) \right).$$

This model assumes that a Dirichlet distribution θ (with α as its parameter) generates a mixture of latent factors: z and y . Image regions r_n are modeled with Gaussians with mean μ and covariance σ , in which words w_n follow a multinomial distribution with a β parameter.

This mixture of latent factors then is used to generate words (y variable) and regions (z variable). The EM algorithm estimates this model, and the inference of $p(w|r)$ is carried out by variational inference. The correspondence latent Dirichlet allocation model provides a clean probabilistic model for annotating images with multiple keywords. It combines the advantages of probabilistic clustering for dimensionality reduction with an explicit model of the conditional distribution from which image keywords are generated.

Li and Wang (2003) characterize the images with a hierarchical approach at multiple tiling granularities (i.e., each tile in each hierarchical level is subdivided into smaller sub-tiles). A color and texture feature vector represents each tile. The texture features represent the energy in high-frequency bands of wavelet transforms. They represent each keyword separately with two-dimensional, multi-resolution hidden Markov models. This method achieves a certain

degree of scale invariance due to the hierarchical tiling process and the two-dimensional multiresolution hidden Markov model.

Network Models

In semantic-multimedia analysis, concepts are interdependent; for example, if a house is detected in a scene, then the probability of existing windows and doors in the scene are boosted, and vice-versa. In other words, when inferring the probability of a set of interdependent random variables, their probabilities are modified iteratively until an optimal point is reached (to avoid instability, the loops must exist over a large set of random variables [Pearl, 1988]). Most of the papers discussed next model keywords as a set of interdependent random variables connected in a probabilistic network.

Various graphical models have been implemented in computer vision to model the appearance, spatial relations, and co-occurrence of local parts. Markov random fields and hidden Markov models are the most common generative models that learn the joint probability of the observed data (X) and the corresponding labels (Y). These models divide the image into tiles or regions (other approaches use contour directions, but these are outside the scope of our discussion). A probabilistic network then models this low-level division in which each node corresponds to one of these tiles or regions and its label. The relation among nodes depends on the selected neighboring method. Markov random fields can be expressed as:

$$P(x, y) = \frac{1}{Z} \cdot \prod_i \left(\phi_i(x_i, y_i) \cdot \prod_{j \in N_i} \varphi_{i,j}(y_i, y_j) \right),$$

in which i indexes the image's tiles, j indexes the neighbors of the current i tile, ϕ_i is the potential function of the current tile x_p and its possible labels y_p and $\varphi_{i,j}$ are the interaction functions between the current tile label and its neighbors. Figure 5 illustrates the Markov random field framework.

The Markov condition implies that a given node only depends on its neighboring nodes. This condition constitutes a drawback for these models, because only local relationships are incorporated into the model. This makes it highly unsuitable for capturing long-range relations or global characteristics.

Figure 5. Two types of random fields

In order to circumvent this limitation, Kumar and Herbert (2003a) propose a multi-scale random field (MSRF) as a prior model on the class labels on the image sites. This model implements a probabilistic network that can be approximated by a 2D hierarchical structure such as a 2D-tree. A multiscale feature vector captures the local dependencies in the data. The distribution of the multiscale feature vectors is modeled as a mixture of Gaussians. The features were selected specifically to detect human-made structures, which are the only types of objects that are detected.

Kumar and Herbert's (2003) second approach to this problem is based on discriminative random fields, an approach inspired on conditional random fields (CRF). CRFs, defined by Lafferty, McCallum, and Pereira (2001), are graphical models, initially for text information extraction, that are meant for visual information analysis in this approach. More generally, a CRF is a sequence-modeling framework based on the conditional probability of the entire sequence of labels (Y), given the all image (X). CRFs have the following mathematical form:

$$P(y|x) = \frac{1}{Z} \cdot \prod_i \left(\phi_i(y_i, x) \cdot \prod_{j \in N_i} \phi_{i,j}(y_i, y_j; x) \right),$$

in which i indexes the image's tiles, j indexes the neighbors of the current i tile, ϕ_i is the association potential between the current tile and the image label, and $\phi_{i,j}$ is the interaction potential between the current tile and its neighbors (note that it is also dependent on the image label). Figure 5 illustrates the conditional random field framework. The authors showed that this last approach outperformed their initial proposal of a multiscale random field as well as the more traditional MRF solution in the task of detecting human-made structures.

He, Zemel, and Carreira-Perpiñán (2004) combine the use of a conditional random field and data at multiple scales. Their multiscale conditional random field (mCRF) is a product of individual models, each model providing labeling information from various aspects of the image: a classifier that looks at local image statistics; regional label features that look at local label patterns; and global label features that look at large, coarse label patterns. The mCRF is shown to detect several types of concepts (i.e., sky, water, snow, vegetation, ground, hippopotamus, and bear) with classification rates better than a traditional Markov random field.

Quattoni, Collins, and Darrell (2004) extend the CRF framework to incorporate hidden variables and combine class-conditional CRFs into a unified framework for part-based object recognition. The features are extracted from special regions that are obtained with the scale-invariant feature transform or SIFT (Lowe, 1999). The SIFT detector finds points in locations at scales in which there is a significant amount of variation. Once a point of interest is found, the region around it is extracted at the appropriate scale. The features from this region then are computed and plugged into the CRF framework. The advantage of this method is that it needs fewer regions by eliminating redundant regions and selecting the ones with more energy on high-frequency bands.

One should note that all these approaches require a ground truth at the level of the image's tiles/regions as is common in computer vision. This is not what is found traditionally in multimedia information retrieval datasets in which the ground truth exists rather at a global level.

Knowledge-Based Models

The previous methods have only visual features as training data to create the statistical models in the form of a probabilistic network. Most of the time, these training data are limited, and the model's accuracy can be improved by other sources of knowledge. Prior knowledge can be added to a model either by a human expert who states the relations between concept variables (nodes in a probabilistic network) or by an external knowledge base in order to infer the concept relations (e.g., with a linguistic database such as WordNet) (Figure 6).

Tansley (2000) introduces a multimedia thesaurus in which media content is associated with appropriate concepts in a semantic layer composed by a network of concepts and their relations. The process of building the semantic layer uses Latent Semantic Indexing to connect images to their corresponding concepts, and a measure of each correspondence (image concept) is taken from this process. After that, unlabeled images (test images) are annotated by comparing them with the training images using a k -nearest-neighbor classifier. Since the concepts' interdependences are represented in the semantic layer, the concepts' probability computed by the classifier are modified by the others concepts.

Other authors have explored not only the statistical interdependence of context and objects but also have used other knowledge that is not present in multimedia data, which humans use to understand (or predict) new data. Srikanth, Varner, Bowden, and Moldovan (2005) incorporated linguistic knowledge from WordNet (Miller, 1995) in order to deduce a hierarchy of terms from the annotations. They generate a visual vocabulary based on the semantics of the annotation words and their hierarchical organization in the WordNet ontology.

Benitez and Chang (2002) and Benitez (2005) took this idea further and suggested a media ontology (MediaNet) to help to discover, summarize, and measure knowledge from annotated images in the form of image clusters, word senses, and relationships among them. MediaNet, a Bayesian network-based multimedia knowledge representation framework, is composed by a network of concepts, their relations, and media exemplifying concepts and relationships. The MediaNet integrates classifiers in order to discover statistical rela-

Figure 6. Knowledge-based models

tionships among concepts. WordNet is used to process image annotations by stripping out unnecessary information. The summarization process implements a series of strategies to improve the images' description qualities, for example using WordNet and image clusters to disambiguate annotation terms (images in the same clusters tend to have similar textual descriptions). Benitez (2005) also proposes a set of measures to evaluate the knowledge consistency, completeness, and conciseness.

Tansley (2000) used a network at the concept level, and Benitez (2005) used the MediaNet network to capture the relations at both concept and feature levels. In addition, Benitez (2005) utilized WordNet, which captures human knowledge that is not entirely present in multimedia data.

Summary

The described algorithms vary in many different aspects such as in their low-level features, segmentation methods, feature representation, modeling complexity, or required data. While some concepts require a lot of data to estimate its model (e.g., a car), others are very simple and require just a few examples (e.g., sky). So, we advocate that different approaches should be used for different concept complexities.

Single-class models assume that concepts are independent and that each concept has its own model. These are the simplest models that can be used and the ones with better accuracy (e.g., Yavlinsky et al., 2005).

Translation models, hierarchical models, and network models capture a certain degree of the concept's interdependence (co-occurrence) from the information present in the training data. The difference between the models is linked to the degree of interdependence that can be represented by the model. In practice, when interdependencies information is incorporated in the model, it also inserts noise in the form of false interdependencies, which causes a decrease in performance. So, the theoretical advantage of these models is in practice reduced by this effect.

All these models rely exclusively on visual low-level features in order to capture complex human concepts and to correctly predict new unlabeled data. Most of the time, the training data are limited, and the model's accuracy can be improved by other sources of knowledge. Srikanth et al. (2005) and Benitez (2005) are two of the few proposals that exploit prior knowledge that is external to the training data in order to capture the interdependent (co-occurrence) nature of concepts.

At this time, knowledge-based models seem to be the most promising semantic analysis algorithms for information retrieval. Text information retrieval already has shown great improvement over exclusively statistical models when external linguistic knowledge was used (Harabagiu et al., 2000). Multimedia retrieval will go through a similar progress but at a slower pace, because there is no multimedia ontology that offers the same knowledge base as WordNet offers to linguistic text processing.

Shot and Scene Semantic Annotation

Shot and scene semantic analysis introduces the time dimension to the problem at hand. The time dimension adds temporal frames, resulting in more information to help the analysis. To take advantage of the sequential nature of the data, the natural choices of algorithms are based on hierarchical models or network models. The section is organized by modality, and within each modality, we don't detail the algorithms by technique due to space constraints. This way, we shed some light on multimodality shot and scene semantic analysis and keep the chapter's emphasis on visual information analysis.

Audio Analysis

Audio analysis becomes a very important part of the multimodal analysis task when processing TV news, movies, sport videos, and so forth. Various types of audio can populate the sound track of a multimedia document, the most common types being speech, music, and silence. Lu, Zhang, and Jiang (2002) propose methods to segment audio and to classify each segment as speech, music, silence, and environment sound. A k -nearest neighbor model is used at the frame level followed by vector quantization to discriminate between speech and nonspeech. A set of threshold-based rules is used in order to discriminate among silence, music, and environment sound. The authors also describe a speaker change detection algorithm based on Gaussian-mixture models (GMM); this algorithm continuously compares the model of the present speaker's speech with a model that is created dynamically from the current audio frame. After a speaker change has been detected, the new GMM replaces the current speaker's GMM.

In most TV programs and sport videos, sound events do not overlap, but in narratives (movies and soap operas), these events frequently occur simultaneously. To address this problem, Akutsu, Hamada, and Tonomura (1998) present an audio-based approach to video indexing by detecting speech and music independently, even when they occur simultaneously. Their framework is based on a set of heuristics over features histograms and corresponding thresholds. With a similar goal, Naphade and Huang (2000) define a generic statistical framework based on hidden Markov models (Rabiner, 1989) in order to classify audio segments into speech, silence, music, and miscellaneous and their co-occurrences. By creating an HMM for each class and every combination of classes, the authors achieved a generic framework that is capable of modeling various audio events with high accuracy.

Another important audio analysis task is the classification of the musical genre of a particular audio segment. This can capture the type of emotion that the director wants to communicate (e.g., stress, anxiety, happiness). Tzanetakis and Cook (2002) describe their work on categorizing music as rock, dance, pop, metal, classical, blues, country, hip-hop, reggae, or jazz (jazz and classical music had more subcategories). In addition to the traditional audio features, they also use special features to capture rhythmic characteristics and apply simple statistical models such as GMM and KNN to model each class' feature histogram. Interestingly, the best reported classification precision (61%) is in the same range as human performance for genre classification (70%).

All these approaches work as a single class model of individual classes/keywords. Note that the hidden Markov model is, in fact, a probabilistic network for modeling a single temporal event that corresponds to a given concept/keyword. So, even though it is a network model, it is used as a single class model.

Visual Analysis

Many of the visual video analysis methods are based on heuristics that are deduced empirically. Statistical methods are more common when considering multimodal analysis. Most of the following papers explore the temporal evolution of features to semantically analyze video content (e.g., shot classification, logical units, etc.). Video visual analysis algorithms are of two types: (a) heuristics-based, in which a set of threshold rules decides the content class, and (b) statistical algorithms that are similar to the ones described in Section 2.

Heuristic methods rely on deterministic rules that were defined in some empirical way. These methods monitor histograms, and events are detected if the histogram triggers a given rule (usually a threshold). They are particularly adequate for sport videos because broadcast TV follows a set of video production rules that result in well-defined semantic structures that ease the analysis of the sports videos. Several papers have been published on sports video analysis, such as football, basketball and tennis, in order to detect semantic events and to semantically classify each shot (Li & Sezan, 2003; Luo & Huang, 2003; Tan, Saur, Kulkarni, & Ramadge, 2000).

Tan et al. (2000) introduced a model for estimating camera movements (pan, tilt, and zoom) from the motion vectors of compressed video. The authors further showed how camera motion histograms could be used to discriminate various basketball shots. Prior to this, the video is segmented into shots based on the evolution of the intensity histogram across different frames. Shots are detected if the histogram exceeds a predefined threshold; then, they are discriminated based on (a) the accumulated histogram of camera motion direction (fast breaks and full-court advances), (b) the slope of this histogram (fast breaks or full-court advances), (c) sequence of camera movements (shots at the basket), and (d) persistence of camera motion (close-ups).

Other heuristic methods deploy color histograms, shot duration, and shot sequences to automatically analyze various types of sports such as football (Ekin, Tekalp, & Mehrotra, 2003) and American football (Li & Sezan, 2003).

The statistical approaches reviewed previously can be applied to the visual analysis of video content with the advantage that shapes obtained by segmentation are more accurate due to the time dimension. Also, analyzing several key-frames of the same shot and then combining the results facilitate the identification of semantic entities in a given shot.

Luo and Hwang's (2003) statistical framework tracks objects within a given shot with a dynamic Bayesian network and classifies that shot from a coarse-grain to a fine-grain level. At the coarse-grain level, a key-frame is extracted from a shot every 0.5 seconds. From these key-frames, motion and global features are extracted, and their temporal evolution is modeled with a hierarchical hidden Markov model (HHMM). Individual HHMMs (a single-class model approach) capture a given semantic shot category. At the fine-grain level

analysis, Luo and Hwang (2003) employ object recognition and tracking techniques. After the coarse-grain level analysis, segmentation is performed on the shots to extract visual objects. Then, invariant points are detected in each shape to track the object movement. These points are fed to a dynamic Bayesian network to model detailed events occurring within the shot (e.g., human body movements in a golf game).

Souvannavong, Merialdo, and Huet (2003) used latent semantic analysis to analyze video content. Recall that latent semantic analysis algorithm builds a matrix M of word occurrences in documents, and then the SVD of this matrix is computed to obtain a singular space. The problem with multimedia content is that there is no text corpus (a vocabulary). A vector quantization technique (k -means) returns a codebook of blobs, the vocabulary of blobs from the shots' key-frames. In the singular feature space, a k -nearest-neighbor ($k=20$) and a Gaussian mixture model technique are used to classify new videos. The comparison of the two techniques shows that GMM performs better when there is enough data to correctly estimate the 10 components. The k -nn algorithm has the disadvantages of every nonparametric method—the model is the training data, and for the TRECVID dataset (75,000 key-frames), training can take considerable time.

Multimodal Analysis

In the previous analysis, the audio and visual modalities were considered independently in order to detect semantic entities. These semantic entities are represented in various modalities, capturing different aspects of that same reality. Those modalities contain co-occurring patterns that are synchronized in a given way because they represent the same reality. Thus, synchronization and the strategy to combine the multimodal patterns is the key issue in multimodal analysis. The approaches described in this section explore the multimodality statistics of semantic entities (e.g., pattern synchronization).

Sports video analysis can be greatly improved with multimodal features; for example, the level of excitement expressed by the crowd noise can be a strong indicator of certain events (foul, goal, goal miss, etc). Leonardi, Migliotari, and Prandini (2004) take this into account when designing a multimodal algorithm to detect goals in football videos. A set of visual features from each shot is fed to a Markov chain in order to evaluate their temporal evolution from one shot to the next. The Markov chain has two states that correspond to the goal state and to the nongoal state. The visual analysis returns the positive pair shots, and the shot audio loudness is the criterion to rank the pair shots. Thus, the two modalities never are combined but are used sequentially. Results show that audio and visual modalities together improve the average precision when compared only to the audio case (Leonardi et al., 2004).

In TV news videos, text is the fundamental modality with the most important information. Westerveld, et al. (2003) build on their previous work described previously to analyze the visual part and to add text provided by an Automatic Speech Recognition (ASR) system. The authors further propose a visual dynamic model to capture the visual temporal characteristics. This model is based on the Gaussian mixture model estimated from the DCT blocks of the frames around each key-frame in the range of 0.5 seconds. In this way, the most significant moving regions are represented by this model with an evident applicability

to object tracking. The text retrieval model evaluates a given $Shot_i$ for the queried keywords $\mathcal{Q} = q_1, q_2, q_3, \dots$:

$$RSV(Shot_i) = \frac{1}{|\mathcal{Q}|} \sum_{k=1}^{|\mathcal{Q}|} \log(\lambda_{Shot} p(q_k | Shot_i) + \lambda_{Scene} p(q_k | Scene_i) + \lambda_{Coll} p(q_k)).$$

This measure evaluates the probability that one or more queried keywords appear in the evaluated shot, $p(q_k | Shot_i)$, or in the scene, $p(q_k | Scene_i)$, under the prior $p(q_k)$. The λ variables correspond to the probabilities of corresponding weights. This function, inspired by language models, creates the scene-shot structure of video content. The visual model and the text model are combined under the assumption that they are independent; thus, the probabilities are simply multiplied. The results with both modalities are reported to be better than using just one.

Naphade and Huang (2001) characterize single-modal concepts (e.g., indoor/outdoor, forest, sky, water) and multimodal concepts (e.g., explosions, rocket launches) with Bayesian networks. The visual part is segmented into shots (Naphade et al., 1998), and from each key-frame, a set of low-level features is extracted (color, texture, blobs, and motion). These features then are used to estimate a Gaussian mixture model of multimedia concepts at region level and then at frame level. The audio part is analyzed with the authors' algorithm described previously (Naphade & Huang, 2000). The outputs of these classifiers are then combined in a Bayesian network in order to improve concept detection. Their experiments show that the Bayesian network improves the detection performance over individual classifiers. IBM's research by Adams et al. (2003) extend the work of Naphade and Huang (2001) by including text from Automatic Speech Recognition as a third modality and by using Support Vector Machines to combine the classifiers' outputs. The comparison of these two combination strategies showed that SVMs (audio, visual, and text) and Bayesian networks (audio and visual) perform equally well. However, since in the latter case, speech information was ignored, one might expect that Bayesian networks can, in fact, perform better. More details about IBM's research work can be found in Naphade and Smith (2003), Natsev, Naphade, and Smith (2003), and Tseng, Lin, Naphade, Natsev, and Smith (2003).

The approach by Snoek and Worring (2005) is unique in the way synchronization and time relations between various patterns are modeled explicitly. They propose a multimedia semantic analysis framework based on Allen's (1983) temporal interval relations. Allen showed that in order to maintain temporal knowledge about any two events, only a small set of relations is needed to represent their temporal relations. These relations, now applied to audio and visual patterns, are the following: precedes, meets, overlaps, starts, during, finishes, equals, and no relation. The framework can include context and synchronization of heterogeneous information sources involved in multimodal analysis. Initially, the optimal pattern configuration of temporal relations of a given event is learned from training data by a standard statistical method (maximum entropy, decision trees, and SVMs). New data are classified with the learned model. The authors evaluate the event detection on a soccer video (goal, penalty, yellow card, red card and substitution) and TV news (reporting anchor, monologue, split-view and weather report). The differences among the various classifiers (maximum entropy, decision trees, and SVMs) appear to be not statistically significant.

Summary

When considering video content, a new, very important dimension is added: time. Time adds a lot of redundancy that can be explored effectively in order to achieve a better segmentation and semantic analysis. The most interesting approaches consider time either implicitly (Westerveld et al., 2003) or explicitly (Snoek & Worring, 2005).

Few papers show a deeper level of multimodal combination than Snoek and Worring (2005) and Naphade and Huang (2001). The first explicitly explores the multimodal co-occurrence of patterns resulting from the same event with temporal relations. The latter integrates multimodal patterns in a Bayesian network to explore pattern co-occurrences and concept interdependence.

Natural language processing experts have not yet applied all the techniques from text to the video's extracted speech. Most approaches to extract information from text and combine this with the information extracted from audio and video are all very simple, such as a simple product between the probabilities of various modalities' classifiers.

Conclusion

This chapter reviewed semantic-multimedia analysis algorithms with special emphasis on visual content. Multimedia datasets are important research tools that provide a means for researchers to evaluate various information extraction strategies. The two parts are not separate, because algorithm performances are intrinsically related to the dataset on which they are evaluated.

Major developments in semantic-multimedia analysis algorithms will probably be related to knowledge-based models and multimodal fusion algorithms. Future applications might boost knowledge-based model research by enforcing a limited application domain (i.e., a constrained knowledge base). Examples of such applications are football game summaries and mobile photo albums.

Multimodal analysis algorithms already have proven to be crucial in semantic multimedia analysis. Large developments are expected in this young research area due to the several problems that wait to be fully explored and to the TRECVID conference series that is pushing forward this research area through a standard evaluation and a rich multimedia dataset.

We believe that semantic-multimedia information analysis for retrieval applications has delivered its first promises and that many novel contributions will be done over the next years. To better understand the field, the conceptual organization by different statistical methods presented here allows readers to easily put into context novel approaches to be published in the future.

References

- Adams, W. H. et al. (2003). Semantic indexing of multimedia content using visual, audio and text cues. *EURASIP Journal on Applied Signal Processing*, 2, 170–185.
- Akutsu, M., Hamada, A., & Tonomura, Y. (1998). Video handling with music and speech detection. *IEEE Multimedia*, 5(3), 17–25.
- Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11), 832–843.
- Barnard, K., & Forsyth, D. A. (2001). Learning the semantics of words and pictures. In *Proceedings of the International Conference on Computer Vision*, Vancouver, Canada.
- Benitez, A. (2005). *Multimedia knowledge: Discovery, classification, browsing, and retrieval* [doctoral thesis]. New York: Columbia University.
- Benitez, A. B., & Chang, S. F. (2002). Multimedia knowledge integration, summarization and evaluation. In *Proceedings of the International Workshop on Multimedia Data Mining in conjunction with the International Conference on Knowledge Discovery & Data Mining*, Alberta, Canada.
- Benitez, A. B. et al. (2002). Semantics of multimedia in MPEG-7. In *Proceedings of the IEEE International Conference on Image Processing*, Rochester, NY.
- Blei, D., & Jordan, M. (2003). Modeling annotated data. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada.
- Brunelli, R., Mich, O., & Modena, C. M. (1999). A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2), 78–112.
- Carneiro, G., & Vasconcelos, N. (2005). Formulating semantic image annotation as a supervised learning problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Duygulu, P., Barnard, K., de Freitas, N., & Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*, Copenhagen, Denmark.
- Ekin, A., Tekalp, A. M., & Mehrotra, R. (2003). Automatic video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7), 796–807.
- Feng, S. L., Lavrenko, V., & Manmatha, R. (2004). Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Cambridge, UK.
- Forsyth, D., & Ponce, J. (2003). *Computer vision: A modern approach*. Prentice Hall.
- Harabagiu, S., et al. (2000). Falcon: Boosting knowledge for answer engines. In *Proceedings of the Text Retrieval Conference*, Gaithersburg, MD.

- Hartley, R., & Zisserman, A. (2004). *Multiple view geometry in computer vision* (2nd ed.). Cambridge University Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. Springer.
- He, X., Zemel, R. S., & Carreira-Perpiñán, M. Á. (2004). Multiscale conditional random fields for image labeling. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Cambridge, UK.
- Hofmann, T., & Puzicha, J. (1998). *Statistical models for co-occurrence data* (No. 1635 A. I. Memo). Massachusetts Institute of Technology.
- Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada.
- Kumar, S., & Herbert, M. (2003a). Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proceedings of the IEEE International Conference on Computer Vision*, Nice, France.
- Kumar, S., & Herbert, M. (2003b). Man-made structure detection in natural images using causal multiscale random field. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Madison, WI.
- Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, San Francisco.
- Lavrenko, V., Manmatha, R., & Jeon, J. (2003). A model for learning the semantics of pictures. In *Proceedings of the Neural Information Processing System Conference*, Vancouver, Canada.
- Leonardi, R., Migliotari, P., & Prandini, M. (2004). Semantic indexing of soccer audio-visual sequences: A multimodal approach based on controlled Markov chains. *IEEE Transactions on Circuits Systems and Video Technology*, 14(5), 634–643.
- Li, B., & Sezan, I. (2003). Semantic sports video analysis: Approaches and new applications. In *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain.
- Li, J., & Wang, J. Z. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9), 1075–1088.
- Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, Kerkyra, Corfu, Greece.
- Lu, L., Zhang, H.-J., & Jiang, H. (2002). Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7), 293–302.
- Luo, Y., & Hwang, J. N. (2003). Video sequence modeling by dynamic Bayesian networks: A systematic approach from coarse-to-fine grains. In *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain.
- Marr, D. (1983). *Vision*. San Francisco: W.H. Freeman.
- Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41.

- Mori, Y., Takahashi, H., & Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, Orlando, FL.
- Naphade, M., et al. (1998). A high performance shot boundary detection algorithm using multiple cues. In *Proceedings of the IEEE International Conference on Image Processing*, Chicago.
- Naphade, M., & Smith, J. (2003). Learning visual models of semantic concepts. In *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain.
- Naphade, M. R., & Huang, T. S. (2000). *Stochastic modeling of soundtrack for efficient segmentation and indexing of video*. In *Proceedings of the Conference on SPIE, Storage and Retrieval for Media Databases*, San Jose, CA.
- Naphade, M. R., & Huang, T. S. (2001). A probabilistic framework for semantic video indexing filtering and retrieval. *IEEE Transactions on Multimedia*, 3(1), 141–151.
- Natsev, A., Naphade, M., & Smith, J. (2003). Exploring semantic dependencies for scalable concept detection. In *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Los Angeles: Morgan Kaufmann Publishers.
- Quattoni, A., Collins, M., & Darrell, T. (2004). Conditional random fields for object recognition. In *Proceedings of the Neural Information Processing Systems Conference*, Vancouver, Canada.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2), 257–286.
- Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905.
- Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), 1349–1380.
- Snoek, C. G. M., & Worring, M. (2005). Multimedia event based video indexing using time intervals. *IEEE Transactions on Multimedia*, 7(4), 638–647.
- Snoek, C. G. M., & Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1), 5–35.
- Souvannavong, F., Merialdo, B., & Huet, B. (2003). Latent semantic indexing for video content modeling and analysis. In *Proceedings of the TREC Video Retrieval Evaluation Workshop*, Gaithersburg, MD.
- Srikanth, M., Varner, J., Bowden, M., & Moldovan, D. (2005). Exploiting ontologies for automatic image annotation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil.
- Tan, Y-P., Saur, D. D., Kulkarni, S. R., & Ramadge, P. J. (2000). Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1), 133–146.

- Tansley, R. (2000). *The multimedia thesaurus: Adding a semantic layer to multimedia information* [doctoral thesis]. University of Southampton, UK.
- TRECVID. (2004). *TREC video retrieval evaluation*. Retrieved November 2005, from <http://www-nlpir.nist.gov/projects/trecvid/>
- Tseng, B. L., Lin, C.-Y., Naphade, M., Natsev, A., & Smith, J. (2003). Normalised classifier fusion for semantic visual concept detection. In *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain.
- Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 293–302.
- Vailaya, A., Figueiredo, M., Jain, A., & Zhang, H. (1999). A Bayesian framework for semantic classification of outdoor vacation images. In *Proceedings of the SPIE: Storage and Retrieval for Image and Video Databases VII*, San Jose, CA.
- Vailaya, A., Figueiredo, M., Jain, A. K., & Zhang, H. J. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1), 117–130.
- Vasconcelos, N., & Lippman, A. (1998). A Bayesian framework for semantic content characterization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA.
- Vasconcelos, N., & Lippman, A. (2000). A probabilistic architecture for content-based image retrieval. In *Proceedings of the IEEE Computer Vision and Pattern Recognition, Hilton Head, SC*.
- Wang, Y., Liu, Z., & Huang, J.-C. (2000). Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing*, 17(6), 12–36.
- Westerveld, T., & de Vries, A. P. (2003). Experimental result analysis for a generative probabilistic image retrieval model. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada.
- Westerveld, T., de Vries, A. P., Ianeva, T., Boldareva, L., & Hiemstra, D. (2003). Combining information sources for video retrieval. In *Proceedings of the TREC Video Retrieval Evaluation Workshop*, Gaithersburg, MD.
- Yavlinsky, A., Schofield, E., & Rüger, S. (2005). Automated image annotation using global features and robust nonparametric density estimation. In *Proceedings of the International Conference on Image and Video Retrieval*, Singapore.