View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DCU Online Research Access Service

Measuring Concept Similarities in Multimedia Ontologies: Analysis and Evaluations

Markus Koskela, *Member, IEEE*, Alan F. Smeaton, *Member, IEEE*, and Jorma Laaksonen, *Senior Member, IEEE*

***Abstract*—The recent development of large-scale multimedia concept ontologies has provided a new momentum for research in the semantic analysis of multimedia repositories. Different methods for generic concept detection have been extensively studied, but the question of how to exploit the structure of a multimedia ontology and existing inter-concept relations has not received similar attention. In this paper, we present a clustering-based method for modeling semantic concepts on low-level feature spaces and study the evaluation of the quality of such models with entropy-based methods. We cover a variety of methods for assessing the similarity of different concepts in a multimedia ontology. We study three ontologies and apply the proposed techniques in experiments involving the visual and semantic similarities, manual annotation of video, and concept detection. The results show that modeling inter-concept relations can provide a promising resource for many different application areas in semantic multimedia processing.**

***Index Terms*—Clustering-based analysis, concept detection, inter-concept relations, multimedia ontology.**

I. Introduction

EXTRACTING semantic concepts from visual data has attracted a lot of research attention recently. The aim of the research has been to facilitate semantic indexing and concept-based retrieval of visual content. The leading principle has been to build semantic representations by extracting intermediate semantic levels from low-level features (see e.g., [1]–[3]). Statistical modeling of midlevel semantic concepts can be useful in supporting high-level indexing and querying on multimedia data, as such concept models can be trained off-line with considerably more positive and negative examples than what are available at query time.

In natural language processing, resources such as Cyc [4] and ConceptNet [5] can be used to extract commonsense assertions from a semantic network of concepts. In a similar fashion, the availability of recently published large-scale multimedia ontologies as well as large manually annotated datasets have enabled the semantic analysis of multimedia content as well as an increase in multimedia lexicon sizes by orders of magnitude. A major resource in this field is the Large Scale Concept Ontology for Multimedia (LSCOM) [6], [7], an expanded multimedia concept lexicon on the order of 1000 concepts, which also includes manual annotations for the concepts in the TREC Video Retrieval Evaluation (TRECVID) [8] 2005 dataset. These concepts relate to events, objects, locations, people, and program categories, and have been selected following a multistep process involving input solicitation, expert critiquing, comparison with related ontologies, and performance evaluation. Yet the design and construction of multimedia ontologies still remains an open research question as the definition of which semantic features are to be modeled tends to be fixed irrespective of the discriminative power of those semantic concepts. This means that the set of concepts in an ontology may be appealing from an ontological perspective, but may contain concepts which make little difference in their discriminative power, or there may be large ‘gaps’ in the resulting overall concept space.

The predominant approach to producing semantic concept models for multimedia data is to treat the problem as a generic learning problem, which makes it scalable to large numbers of concepts. Here, training data is used to learn independent models of different concepts over low-level feature distributions, and the set of concepts covered by such models generally forms part of a larger ontology. For building such models, one approach is to use discriminative approaches, such as support vector machines (SVMs), k-nearest neighbor classifiers, or decision trees, to classify between positive and negative examples of a certain concept. An alternative is to take a generative approach in which the probability density function of a semantic concept is estimated based on existing training data, e.g., with Gaussian mixture models or nonparametric density estimation.

We follow the generative approach and use global low-level features extracted from both video data and keyframe images for video shot representation. As the ground truth, we use manually annotated keyframes of various TRECVID collections, as they provide an unique and commonly used source of information for large-scale semantic concept modeling. The utilization of image-level annotations and global features in concept modeling has obvious limitations, as while some concepts correspond to the content of the image or video shot as a whole, most of them are localized, i.e., they correspond to a distinct object or part of the scene. As a result, we cannot distinguish between two concepts if they always co-occur with each other in the training data, and can expect certain unintuitive similarities as object-based concepts are mixed with commonly occurring backgrounds. Unfortunately, obtaining localized annotations for large-scale multimedia ontologies is an extremely challenging

Manuscript received September 29, 2006; revised April 5, 2007. This work was supported by the Irish Research Council for Science Engineering and Technology and by the Science Foundation Ireland under Grant 03/IN.3/I361. This work was supported by the Academy of Finland under Projects “Neural Methods in Information Retrieval Based on Automatic Content Analysis and Relevance Feedback” and “Finnish Centre of Excellence in Adaptive Informatics Research.” The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marcel Worring.

M. Koskela and J. Laaksonen are with the Adaptive Informatics Research Centre, Helsinki University of Technology, Helsinki FI-02015, Finland (e-mail: markus.koskela@hut.fi).

A. F. Smeaton is with the Centre for Digital Video Processing and Adaptive Information Cluster, Dublin City University, Dublin, Ireland.

Digital Object Identifier 10.1109/TMM.2007.900137

1520-9210/$25.00 © 2007 IEEE

task. Indeed, it is the aim of this paper to inspect the level of results that can be obtained with an image or shot level approach and whether the current methods benefit from such analysis.

In this paper, we study how multimedia concept models built over a general clustering method can be interpreted in terms of probability distributions and how the quality of such models can be assessed with entropy-based methods. The entropy of a certain feature vector's distribution is a measure of how uniformly the used feature distributes the concept over the set of clusters [9]. We make the assumption that a good model is such that the distribution is heavily concentrated on only a few clusters, resulting in a low value of entropy. This approach is readily scalable to large multimedia lexicons where each concept can be represented as a set of probability distributions over common clusterings based on different low-level feature spaces. In addition, and most interestingly for the work in this paper, the similarity of two distributions can be used to measure the overlap of the corresponding concepts. This enables us to produce a similarity matrix for all concepts in an ontology in order to study the inter-concept relations in the lexicon, and helps to determine the quality and coverage of the ontology covered by the set of concepts.

We propose a methodology for analyzing the low-level feature and semantic properties of three multimedia ontologies as flat concept lexicons, i.e., each of the concepts has been annotated separately and any taxonomical relations between the concepts are neglected, as the LSCOM ontology in its current form does not contain such a taxonomy. This kind of analysis can reveal existing inter-concept relations, but provides only a part of the picture as taxonomically related concepts are treated similarly as any two random concepts.

We adopt the term "visual" for all characteristics based on low-level feature spaces even though all such features might not be visual, e.g., for video analysis we also use audio features. The method we use for measuring the visual similarity among concepts was first presented in [10] and a similar method is applied here for pairwise co-occurrence properties of concepts. Furthermore, in the experiments section of the paper we examine the application of these techniques to other tasks, namely the analysis of large-scale multimedia ontologies, manual annotation of video by semantic concepts, and automatic detection of concepts.

The rest of the paper is organized as follows. Section II gives an overview of related work on semantic and concept-based indexing of multimedia content. Section III describes our clustering-based method for representing semantic concepts on low-level feature spaces and the evaluation of such representations with entropy-based methods. In Section IV, we discuss the estimation and analysis of different similarity relations between semantic concepts in an ontology. Section V presents a series of experiments in which the proposed methods are applied in different tasks. Finally, conclusions are given in Section VI.

II. Related Work

In this section, we provide a brief overview of related research in extracting semantic concepts from visual content. In general, we acknowledge three overlapping tasks in this general field, namely semantic categorization, annotation, and concept detection. Furthermore, we discuss two specific and open issues in current research that are directly relevant to the topic of this paper. The first issue is analyzing the usefulness and reliability of different concept models for multimedia content. Second, we briefly describe existing novel approaches to utilizing different inter-concept relations within an ontology.

In order to make image indexing by higher-level content possible, a fundamental requirement is to be able to capture the image's semantic content in such a way that corresponds to the human view of image semantics. Within broad domains, automatically extracted visual features often fail to do this adequately. In some cases, however, a certain level of semantic *categorization* with automatic methods is possible. For example, Szummer and Picard proposed a method to classify between indoor and outdoor images [11], and Vailaya *et al.* to classify between city images and landscape scenes [12].

An alternative to classification is the automatic *annotation* of images, where the input images are labeled with any of a set of available annotations if they fulfill the corresponding criteria. Unlike in classification, it is not assumed that the image collection can be divided to a set of classes, but rather that the images having a certain annotation constitute the representation of that semantic concept. A common approach to image annotation is to first obtain image regions by using a segmentation algorithm, partitioning the image area with a regular grid, or extracting interest points. A set of feature extraction methods is then employed for these regions and the extracted region-wise feature vectors over all images in the database are clustered to produce image blobs. Each image is described using a subset from the vocabulary of blobs, and the problem of image annotation can be viewed as a transformation from the blob representation to a keyword representation. A large number of methods have been proposed to achieve this, including the translation model of Barnard *et al.* [13], the cross-media relevance model of Jeon *et al.* [14], and models based on probabilistic latent semantic analysis [15], [16]. The main problem with the blob-based representation is the difficulty of obtaining both reliable results using weak segmentation algorithms and large-scale localized annotations to use as ground truth for large ontologies such as LSCOM.

A third widely studied and closely related technique for describing image content is semantic *concept detection*. In concept detection, after training models for the concepts to be used, the task is to detect those objects from a separate test set that are relevant for the given concept, in current research typically using nonlocalized annotations. Concept detection thus differs from automatic annotation and categorization and seems more suitable for generative models or density estimation, as the focus is less on learning exact discriminative boundaries between classes and more on identifying the regions of the input space likely to contain the principal portion of the relevant data items. Viewed this way, concept detection closely resembles query-by-example retrieval, with the fundamental difference that concept detection is typically performed off-line and with a greater number of training examples available. For generic concept detection, proposed methods include support vector machines [17], [18], Gaussian mixture models [19], and Bayesian learning using a constellation model [20].

When considering large-scale concept ontologies as we do in this paper, the usefulness of different concepts is an important question. It can be approached either from a task-oriented perspective or by directly analyzing the properties of the concept models, as is done in this paper. Using the former approach, Christel and Hauptmann study the benefits of incorporating concepts in a large number of potential multimedia queries [21]. Lin and Hauptmann then propose the use of mutual information between the relevance of a shot to a query and semantic concepts for determining concept utility [22]. The latter approach is taken in [23], where Kender and Naphade analyze a large database of concept annotations and use information gain to find the most reliable visual concepts. In a paper [24] resembling the approach proposed here, Yanai and Barnard analyze concept models using entropy of a blob-based representation to measure concept "visualness," i.e., concepts that can be reliably detected using low-level visual features.

Different semantic concepts do not exist in isolation, but form a part of a concept ontology. Concepts in an ontology can have different relationships between each other, including inter-concept semantic and visual similarities, statistical co-occurrence of two or more concepts, and different hierarchical relations. The problem of efficient utilization of these contextual inter-concept relations is currently a widely studied and open research issue, and several interesting approaches have recently been proposed. Naphade *et al.* propose a factor graph framework for inter-concept context analysis [25]. They use a probabilistic graphical network of multimedia objects or "multijects" to model their interaction. In [26], Wu *et al.* use an existing ontology hierarchy to influence individual concept detectors. A boosting factor is used for top-down influence from parent concepts to the children, and confusion factor is defined for mutually exclusive concepts. In [23], Kender and Naphade use the G-test for finding concept pairs that occur frequently together. Yan *et al.* use various graphical models to find relationships between concepts and study their effect in concept detection [27]. Snoek *et al.* propose an authoring metaphor to multimedia indexing [18]. They divide the indexing process to content, style, and context steps. In the last step, the individual concept detectors are combined into a semantic feature which is then used as input to a supervised learning stage. In [28], Xie and Chang study different data mining methods for static and temporal pattern mining of large-scale multimedia ontologies. They propose the use of frequent itemset mining, k-means clustering and hidden Markov models in new concept prediction.

III. Concept Representation and Analysis

In our representation of concepts we adopt a probabilistic model in which the probability density function of a semantic concept is modeled based on training data, instead of a binary classification approach where each database object is classified either as relevant or nonrelevant for the corresponding concept. Our aim here is thus not to perform classification or concept detection, but to analyze different properties of concept models directly via the properties of their density functions. In this paper, the properties we concentrate on include the robustness of the models and the similarity between pairs of concept models.

A. *Clustering-Based Concept Representations*

Clustering refers to partitioning data into k sets or clusters so that data items in a certain cluster are more similar to each other than to data items in other clusters. In the basic form (i.e., hard or crisp clustering), every data item belongs to exactly one cluster. Clustering is commonly used to *summarize* the observed data and to *generalize* it to unknown samples. We will concentrate here on the former purpose and view cluster analysis as similar to vector quantization, that is to provide a set of codewords to represent input data in a more compact way. Clustering-based probabilistic models are commonly used in image processing. Some examples include the use of clustering to obtain a codebook of blobs after an image segmentation step [13]–[16] and estimating a probability density in a high-dimensional feature space by first running a clustering algorithm and then using the cluster partitions to estimate the probability density using mixtures of Gaussians [29], [30].

Given a set of k cluster centroids, we can in theory calculate the *a priori* probability of each cluster being the best match for any vector $\mathbf{x}$ of the feature space. This is possible if the probability density function (pdf) $p(\mathbf{x})$ is known. If we denote the cluster by i and its surrounding Voronoi region by $\mathcal{V}_i$, we may calculate the cluster's *a priori* probability as

$$P_i = P(\mathbf{x} \in \mathcal{V}_i) = \int_{\mathcal{V}_i} p(\mathbf{x})\, \mathbf{dx}. \tag{1}$$

With discrete data, we replace the continuous pdf with a discrete probability histogram. Without danger of confusion, the probability can still be denoted as P_i

$$P_i = P(\mathbf{x} \in \mathcal{V}_i) = \frac{\#\{j \mid \mathbf{x}_j \in \mathcal{V}_i\}}{N} \tag{2}$$

where $\#\{\cdot\}$ stands for the cardinality of a set, and N is the size of the training data set whose members are $\mathbf{x}_j, j = 0, 1, \ldots, N-1$. The reader should note that the original probability density of the continuous feature space cannot be directly approximated with the discrete P_i, because the sizes of the histogram bins, i.e., the Voronoi regions, are not equal. It will suffice, however, that the one-directional mapping from the continuous distribution to the discrete one can be performed.

In what follows, we concentrate on the distributions of specific subsets of data. We may assume that the members of such a subset with N_m items fulfill a specific *ground truth* criterion by which each object can be classified as either a member or nonmember of the class. In this paper, considering the application domain, we denote these subsets as semantic *concepts* $\mathcal{C}$. Considering the mth concept $\mathcal{C}_m$, the corresponding probability histogram will be

$$\begin{aligned} P_i^m &= P(\mathbf{x} \in \mathcal{V}_i \mid \mathbf{x} \in \mathcal{C}_m) \\ &= \frac{\#\{j \mid \mathbf{x}_j \in \mathcal{V}_i, \mathbf{x}_j \in \mathcal{C}_m\}}{N_m} \end{aligned} \tag{3}$$

where N_m is the cardinality of the subset of vectors relevant for concept $\mathcal{C}_m$.

B. Latent Variable Models

The representation of visual data using global features is quite crude. For example, to accurately describe the content of an image, we should enumerate the objects contained in the image as well as their relationships and contexts. While some semantic concepts do correspond to the content of the image as a whole, some others are localized to a distinct object or a specific part of the background or scene. To facilitate localized image representation, a common approach is to use local appearance descriptors. There are different methods producing local descriptors, such as using automatic or manual segmentation, employing a regular grid, or extracting corner or interest points from the images.

The set of local descriptors can be quantized to produce "visual words" or *visterms*, which brings multimedia indexing back to text analysis, as the multimedia objects can now be regarded as documents consisting of a number of visterms. Analogously with text, a common method to represent multimedia documents is the *bag-of-visterms* model where relationships between visterms are ignored and only the observations of different visterms in documents are considered.

The bag-of-visterm representations can be used directly as concept models by concatenating the visterms of objects fulfilling the corresponding ground truth criteria. This approach is inadequate as the origins of the visterms are neglected. As an alternative, we may utilize probabilistic latent variable models such as *probabilistic latent semantic analysis* (PLSA) [31] which has been recently applied e.g., to image classification [15], [16]. In PLSA, a latent variable or aspect $z_k, k = 1, \ldots, n$ is associated with each observation. The joint probability of documents and visterms is then defined as the mixture

$$P(w_j, d_i) = P(d_i) \sum_k P(w_j \mid z_k) P(z_k \mid d_i) \tag{4}$$

where $P(w_j \mid z_k)$ is the class-conditional probability of the visterm w_j conditioned on the unobserved aspect z_k, and $P(z_k \mid d_i)$ denotes the probability distribution of the latent aspects given the document d_i. The PLSA model is fitted using the Expectation-Maximization (EM) algorithm [32]. After training the model, a new document d_q can be "folded-in" to the aspect space $P(z_k \mid d_q)$ by keeping the document independent probabilities $P(w_j \mid z_k)$ fixed and using EM. Using this approach, a concept $\mathcal{C}_m$ can be aggregated to a document $d_{\mathcal{C}_m}$ and the concept can be modeled as a distribution $P(z_k \mid d_{\mathcal{C}_m})$ over the latent aspects.

C. Entropy-Based Measure of Distributions

We will now turn to study the uniformity of the distributions of objects relevant to specific concepts over the clusters or latent aspects. A simple and commonly used measure for the randomness of a symbol distribution is its entropy. In our case, the cluster indexes for the vectors of the training set play the role of symbols. The entropy H of a distribution $P = (P_0, P_1, \ldots, P_{k-1})$ is

$$H(P) = -\sum_{i=0}^{k-1} P_i \log P_i \tag{5}$$

where k is the number symbols in the alphabet of the stochastic information source, in our case thus the number of clusters or latent aspects. P_i is the probability of cluster i being the correct one for an input vector, as defined before. Logarithm base of two is usually used and also assumed here.

If we assume that each of the k clusters is equally probable as the correct one for an input vector, we get the theoretical maximum for the entropy of a clustering $H_{\max^*} = \log k$. In the discrete case, the above definition for the maximum entropy to hold assumes that N is divisible by k. In general, this is not the case and $H_{\max^*}$ is biased toward smaller values. However, the produced error is insignificant with sufficient amount of data, i.e., if $N \gg k$. When studying the whole database, this can generally be assumed since the overall aim of clustering is to reduce computational requirements of the retrieval algorithm. With a concept $\mathcal{C}_m$ having only a few examples available, i.e., when N_m is relatively small, the difference may, however, be considerable so instead of $H_{\max^*}$ we calculate the empirical entropy maximum $H_{\max}$ for each concept by spreading its distribution over the k clusters as uniformly as possible with the given integer values of N_m and k, and using (5).

Instead of using entropy directly, often a more illustrative measure is perplexity $\mathrm{PPL} = 2^H$, commonly utilized in speech recognition. Perplexity can be considered as the weighted number of equal choices for a random variable; i.e., in this setting, the average number of equivalent clusters that have to be considered given the distribution. Thus, if entropy had a theoretical maximum value $H_{\max^*}$, the perplexity of a clustering would equal the total number of clusters, $\mathrm{PPL}_{\max^*} = k$. A suitable performance indicator for feature extraction and the associated clustering methods can be formed by the ratio of perplexity to its maximum value, denoted here as *normalized perplexity*

$$\overline{\mathrm{PPL}} = \frac{\mathrm{PPL}}{\mathrm{PPL}_{\max}} = \frac{2^H}{2^{H_{\max}}} \tag{6}$$

which is nonnegative and ≤ 1 in all cases. In general it can be assumed that when clustering distributes the input vectors roughly evenly over the clusters, the normalized perplexity of the whole data should thus be near unity. On the other hand, images with semantic similarity should be mapped to a small cluster subset, provided that the feature extraction and clustering methods have been favorable to that specific concept. In this case, $\overline{\mathrm{PPL}}$ should be clearly smaller than one. The normalized perplexity $\overline{\mathrm{PPL}}$ of a concept $\mathcal{C}_m$ can simply be calculated by replacing P_is in (5) by P_i^ms of (3).

A straightforward application of $\overline{\mathrm{PPL}}$ is to use it as a weight of the corresponding distribution in feature fusion. In general, different multimedia concepts are best represented using multiple features and combining their outputs based on their relative performances. A small value of $\overline{\mathrm{PPL}}$ corresponds to a well-concentrated distribution, so the relative weight of the corresponding feature should be increased. For example using softmax scaling on the inverse of $\overline{\mathrm{PPL}}$, the weight of the ith feature becomes

$$w_i = \frac{\exp(1/\overline{\mathrm{PPL}}_i)}{\sum_j \exp(1/\overline{\mathrm{PPL}}_j)}. \tag{7}$$

In the experiments of this paper, we will use clustering and $\overline{\mathrm{PPL}}$ to analyze different low-level features for their ability to produce nonuniform concept distributions and (7) for weighted feature combination.

IV. Inter-Concept Similarity

In order to analyze the overall utility of a concept ontology, we aim to measure the overlap among concepts based on a set of different characteristics. This enables us to produce a similarity matrix for the ontology in order to study the overall efficacy of the set of concepts as well as to analyze individual ones to find potential weaknesses, such as near-duplicate concepts as well as highly isolated ones. In many previous approaches to semantic concept modeling, the ontology, or inter-concept relations in general, have not usually been utilized, but each concept has rather been treated as a binary classifier and thus processed as if it were a separate entity. In the next section a set of quite diverse applications for the results of analysis of concept relations are proposed, in many cases combining two different similarity measures. In this section we now consider four different similarity relations between concepts.

A. Visual Similarity

Considering the multiple concepts in an ontology, an interesting question is the similarity between two concepts based on low-level features which can be automatically extracted. In Section III-C, we used entropy to measure concept distributions and so continuing with the information-theoretic approach, a natural measure of two concepts' similarity would be their *mutual information* as examined previously in [9]. Let us denote by P^m and P^n the probability distributions of concepts $\mathcal{C}_m$ and $\mathcal{C}_n$, over a set of either clusters or latent aspects. As entropy measures the randomness of a distribution, mutual information $I(P^m, P^n)$ can be used for studying the interplay between two distributions

$$I(P^m, P^n) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} P_{ij}^{mn} \log \frac{P_{ij}^{mn}}{P_i^m P_j^n} \tag{8}$$

where P^{mn} is the estimated joint probability of the two concepts.

However, when using mutual information in estimating inter-concept similarities, the sparsity of the data quickly becomes a problem. In order to obtain an accurate enough model of a concept, the value of k has to be relatively large, resulting in a sparse joint probability matrix P^{mn} unless we have a lot of training data. Therefore, we take a different approach and use a bin-to-bin histogram distance measure in estimating the concept similarities. A number of such measures are available, including intersection, Euclidean distance, χ^2-statistic, and Kullback–Leibler divergence. Based on earlier experiments [10], [33], we decided to use Jeffrey divergence [34]

$$d_{\mathrm{JD}}(P^m, P^n) = \sum_{i=0}^{k-1} \left(P_i^m \log \frac{P_i^m}{\hat{P}_i} + P_i^n \log \frac{P_i^n}{\hat{P}_i} \right) \tag{9}$$

where $\hat{P} = (P^m + P^n)/2$ is the mean distribution, as it is symmetric and numerically stable with empirical distributions and usually gives rather consistent results.

B. Co-Occurrence

A complementary view of concept similarity can be provided by considering co-occurrence statistics or *collocations* of pairs of concepts. In computational linguistics, a collocation is defined as a sequence of words or terms which co-occur more often than would be expected by chance. A similar analysis can also be used in multimedia ontologies, since the presence or absence of certain concepts may be a valuable cue in predicting the presence of other concepts in a multimedia object. A number of methods for analyzing co-occurrence patterns have been proposed in recent research, including the G-test [23], frequent itemsets [28] and shot clustering [28]. We examine concept occurrence data as a binary variable over the items in the training data. Thus, here we denote P^m as a vector of length equalling the size of the training set with $P_i^m = 1$ if the ith item is relevant for concept $\mathcal{C}_m$ and $P_i^m = 0$ otherwise. Again, there exist different suitable distance measures, including the Hamming distance, Dice measure, point-wise mutual information, and the Cosine measure

$$d_{\cos}(P^m, P^n) = \frac{\sum_{i=0}^{k-1} P_i^m P_i^n}{\sqrt{\sum_{i=0}^{k-1} \left(P_i^m\right)^2} \sqrt{\sum_{i=0}^{k-1} \left(P_i^n\right)^2}}. \tag{10}$$

C. Semantic Similarity

The third type of concept similarity we discuss is the similarity between concepts based on their semantic meanings. By nature, these properties are rather different from the two similarities discussed above as we cannot use a ground truth set of annotated multimedia objects to deduce semantics of concepts. Instead, there are two basic ways to quantify semantic similarity: either using a lexical resource such as WordNet [35] or by gathering similarity assessments from human subjects (see e.g., [36]). We take the latter approach in this paper and gather subjective assessments of different concepts' semantic similarity from a group of test subjects. Gathering such assessments for a small number of concepts is straightforward, but with large-scale ontologies this becomes infeasible due to the quadratic increase of numbers of concept pairs compared to the size of the ontology. As a result, we limit our study of semantic similarities to comparisons within the other concept relations. In the experiments in Section V, we will present two studies of semantic similarity compared with visual similarity.

D. Hierarchical Structure

The final similarity relation considered in this paper is based on the tree-structured construction or taxonomy used in a multimedia ontology. The most common relation in multimedia ontologies is the subsumption or *is-a* relation. A valid ontology should contain such a hierarchy of concepts, otherwise it is simply just a multimedia lexicon. Therefore, it is unfortunate that, at its current stage of development, the LSCOM ontology is not organized in a hierarchical manner, so our work here is limited. The hierarchical structure of concepts can be directly used to improve the detection of individual concepts, and an example of such a method is given in [26].

TABLE I
THE SET OF 280 CONCEPTS FROM THE LSCOM ONTOLOGY USED IN THE EXPERIMENTS

Concept	Abbr	Size	$\overline{PPL}$	Feat
actor	act	0.177	0.73	eh
address or speech	add	0.023	0.56	eh
adobehouses	ado	0.005	0.35	cs
adult	adu	0.658	0.90	dc
agricultural people	agr	0.003	0.32	cs
aircraft cabin	ai0	0.001	0.49	cl
airplane flying	ai1	0.002	0.19	eh
airplane	ai2	0.007	0.36	eh
airport or airfield	ai3	0.002	0.46	eh
airport	ai4	0.005	0.35	cl
animal pens and cages	an0	0.002	0.29	cs
animal	an1	0.348	0.81	eh
antenna	ant	0.002	0.45	cl
apartment complex	ap0	0.004	0.37	cs
apartments	ap1	0.001	0.37	cs
armed person	ar0	0.026	0.54	cl
armored vehicles	ar1	0.006	0.32	eh
asian people	asi	0.062	0.62	cs
athlete	ath	0.024	0.44	cl
attached body parts	att	0.015	0.64	sc
baby	bab	0.001	0.35	cs
backpack	ba0	0.004	0.32	cs
backpackers	ba1	0.004	0.32	cs
baseball	ba2	0.001	0.16	cs
basketball	ba3	0.005	0.22	cs
beach	be0	0.003	0.35	eh
beards	be1	0.008	0.43	cs
bicycles	bic	0.001	0.50	sc
birds	bir	0.002	0.20	cs
blank frame	bla	0.005	0.02	eh
boat ship	boa	0.006	0.30	eh
body parts	bod	0.023	0.70	sc
boy	boy	0.008	0.50	eh
bride	br0	0.005	0.10	cs
bridges	br1	0.002	0.27	cl
building	bui	0.098	0.75	cs
bus	bu0	0.003	0.43	cs
business people	bu1	0.003	0.29	sc
cables	cab	0.006	0.42	cl
camera	cam	0.007	0.39	cs
canal	can	0.001	0.47	sc
capital	cap	0.002	0.26	cs
car crash	ca0	0.001	0.56	cs
car	ca1	0.055	0.65	eh
caucasians	cau	0.128	0.68	cs
celebration or party	ce0	0.008	0.31	cs
celebrity entertainment	ce1	0.041	0.56	eh
cell phones	ce2	0.004	0.24	sc
charts	cha	0.008	0.39	cs
cheering	ch0	0.009	0.37	eh
cheerleader	ch1	0.001	0.06	cs
child	chi	0.010	0.60	cs
cityscape	cit	0.004	0.34	cl
civilian person	civ	0.607	0.90	dc
classroom	cla	0.002	0.46	cs
clearing	cle	0.001	0.44	eh
clouds	clo	0.014	0.39	cl
commentator or studio expert	co0	0.015	0.16	cs
commercial advertisement	co1	0.313	0.81	cs
computer or television screens	co2	0.038	0.39	eh
computers	co3	0.020	0.31	eh
conference room	co4	0.009	0.38	cs
congressman	co5	0.005	0.26	cs
corporate leader	cor	0.017	0.56	eh
court	co6	0.003	0.28	cs
courthouse	co7	0.001	0.36	cs
crowd	cro	0.133	0.66	eh
cul-de-sac	cul	0.001	0.39	cs
dancing	dan	0.017	0.25	cs
dark-skinned people	dar	0.009	0.37	cs
daytime outdoor	day	0.310	0.81	cs
demonstration or protest	dem	0.009	0.40	cs
desert	des	0.009	0.35	eh
dining room	din	0.001	0.34	sc
dirt gravel road	dir	0.006	0.38	cs
dogs	dog	0.002	0.24	sc
dresses of women	dr0	0.021	0.60	eh
dresses	dr1	0.006	0.18	eh
driver	dri	0.004	0.25	cs
earthquake	ear	0.001	0.37	cs
election campaign address	el0	0.005	0.27	cs
election campaign convention	el1	0.002	0.37	cs
election campaign greeting	el2	0.003	0.30	cs
election campaign	el3	0.009	0.32	cs
emergency vehicles	eme	0.002	0.49	eh
entertainment	ent	0.201	0.74	cs
exiting car	exi	0.004	0.30	eh
exploding ordinance	ex0	0.009	0.32	eh
explosion fire	ex1	0.012	0.45	eh
eyewitness	eye	0.001	0.47	cs
face	fa0	0.399	0.81	eh
factory worker	fa1	0.002	0.40	ht
female anchor	fe0	0.021	0.14	cs
female newsject	fe1	0.006	0.48	cs
female person	fe2	0.219	0.81	eh
female reporter	fe3	0.010	0.28	cs
fields	fie	0.004	0.22	eh
fighter combat	fig	0.003	0.33	sc
finance busines	fin	0.012	0.56	ht
first lady	fir	0.004	0.31	cs
flags	fla	0.035	0.61	cs
flood	fl0	0.001	0.30	cs
flowers	fl1	0.012	0.31	sc
flying objects	fly	0.002	0.23	eh
food	fo0	0.018	0.37	cs
football	fo1	0.002	0.42	cl
forest	for	0.007	0.25	cs
free standing structures	fre	0.007	0.47	cs
funeral	fun	0.012	0.40	cs
furniture	fur	0.095	0.70	cs
george bush	geo	0.008	0.39	cs
girl	gir	0.004	0.36	cs
glass	gl0	0.005	0.30	cs
glasses	gl1	0.020	0.48	cs
golf course	go0	0.002	0.19	cs
golf player	go1	0.001	0.26	cs
golf	go2	0.002	0.26	cs
government leader	gov	0.066	0.69	eh
grandstands bleachers	gr0	0.025	0.42	eh
grassland	gr1	0.005	0.22	cl
greeting	gre	0.006	0.48	cs
ground combat	gr2	0.014	0.44	cl
ground vehicles	gr3	0.052	0.59	eh
group	gr4	0.175	0.76	eh
guard	gua	0.003	0.40	cl
guest	gue	0.009	0.15	cs
hand	ha0	0.009	0.58	cs
handshaking	ha1	0.002	0.51	eh
harbors	har	0.004	0.28	cs
head and shoulder	he0	0.168	0.70	eh
head of state	he1	0.013	0.48	cs
helicopters	hel	0.001	0.23	eh
high security facility	hi0	0.002	0.55	eh
highway	hi1	0.005	0.29	eh
hill	hil	0.005	0.26	eh
hospital	ho0	0.004	0.26	cs
host	ho1	0.010	0.15	cs
house of worship	ho2	0.003	0.42	cl
house	ho3	0.003	0.38	cs
hu jintao	huj	0.003	0.27	cs
individual	in0	0.258	0.74	eh
indoor sports venue	in1	0.006	0.27	eh
industrial setting	in2	0.004	0.37	cs
infants	inf	0.002	0.43	cs
insurgents	ins	0.002	0.44	cl
interview on location	in3	0.069	0.65	eh
interview sequences	in4	0.051	0.32	cs
john kerry	joh	0.002	0.48	cs
kitchen	kit	0.002	0.26	sc
laboratory	lab	0.001	0.48	cs
lakes	lak	0.004	0.20	sc
landlines	la0	0.001	0.30	cs
landscape	la1	0.015	0.37	eh
lawn	law	0.012	0.32	cl
logos full screen	log	0.026	0.30	cs
machine guns	mac	0.025	0.49	cl
male anchor	ma0	0.029	0.14	cs
male newsject	ma1	0.032	0.66	cs
male person	ma2	0.443	0.88	dc
male reporter	ma3	0.024	0.29	cs
maps	map	0.015	0.25	cs
medical personnel	med	0.003	0.41	eh
meeting	mee	0.057	0.66	cs
microphones	mic	0.052	0.71	eh
military base	mi0	0.002	0.40	sc
military personnel	mi1	0.045	0.60	cs
moonlight	moo	0.003	0.23	cs
mosques	mos	0.001	0.58	cs
motorcycle	mot	0.002	0.55	cs
mountain	mou	0.010	0.34	eh
muddy scenes	mud	0.002	0.26	cs
mug	mug	0.002	0.17	sc
muslims	mus	0.007	0.42	eh
natural disasters	nat	0.006	0.38	cs
network logo	net	0.018	0.26	cs
news studio	ne0	0.105	0.32	cs
newspapers	ne1	0.003	0.13	cs
nighttime	nig	0.028	0.48	dc
non-uniformed fighters	no0	0.006	0.47	cs
non-us national flags	no1	0.016	0.49	cs
oceans	oce	0.004	0.12	cs
office building	of0	0.016	0.46	cs
office	of1	0.029	0.64	cs
officers	of2	0.002	0.47	cs
old people	old	0.003	0.29	cs
outdoor	out	0.339	0.84	cs
overlaid text	ove	0.355	0.81	eh
parade	pa0	0.008	0.35	eh
parking lot	pa1	0.001	0.53	sc
pavilions	pav	0.001	0.43	cs
pedestrian zone	ped	0.002	0.43	sc
people crying	pe0	0.002	0.40	cs
people marching	pe1	0.021	0.47	eh
person	per	0.666	0.90	dc
pickup truck	pic	0.003	0.43	cs
pipes	pip	0.002	0.45	cs
police private security personnel	po0	0.009	0.49	cs
police	po1	0.002	0.39	cs
politics	po2	0.072	0.68	cs
powerplants	pow	0.060	0.62	cs
press conference	pre	0.038	0.65	cs
prisoner	pri	0.002	0.53	cs
processing plant	pr0	0.001	0.55	cs
protesters	pr1	0.006	0.28	cs
rainy	rai	0.002	0.26	cs
religious figures	rel	0.010	0.46	sc
reporters	rep	0.010	0.39	cs
residential buildings	res	0.013	0.53	cs
rifles	rif	0.022	0.50	cl
riot	rio	0.003	0.38	cs
river bank	ri0	0.002	0.34	cs
river	ri1	0.006	0.30	cs
road overpass	ro0	0.001	0.34	cs
road	ro1	0.055	0.65	eh
rocky ground	roc	0.013	0.40	eh
room	roo	0.015	0.27	cs
rpg	rpg	0.001	0.54	cs
ruins	rui	0.006	0.30	cs
running	ru0	0.011	0.39	cl
runway	ru1	0.006	0.41	eh
scene text	sce	0.108	0.79	cs
school	sch	0.002	0.49	sc
science technology	sci	0.003	0.45	cs
security checkpoint	sec	0.002	0.34	eh
ship	shi	0.001	0.48	ht
shooting	sho	0.004	0.31	cl
sidewalks	sid	0.025	0.62	cs
singing	si0	0.012	0.22	cs
single family homes	si1	0.006	0.36	cs
single person female	si2	0.080	0.64	eh
single person male	si3	0.152	0.70	eh
single person	si4	0.233	0.72	eh
sitting	sit	0.164	0.72	cs
sky	sky	0.115	0.66	cl
smoke	smo	0.013	0.42	ht
snow	sno	0.006	0.23	cs
soccer	soc	0.009	0.09	cl
soldiers	sol	0.029	0.55	cs
speaker at podium	sp0	0.020	0.51	eh
speaking to camera	sp1	0.081	0.48	cs
sports	spo	0.046	0.47	cl
stadium	st0	0.004	0.26	eh
standing	st1	0.211	0.88	dc
steeple	ste	0.002	0.42	cs
still image	sti	0.013	0.54	cs
stock market	st0	0.002	0.27	cs
store	st1	0.009	0.35	cs
street battle	st2	0.014	0.42	cs
streets	st3	0.007	0.37	eh
studio with anchorperson	stu	0.053	0.15	cs
suburban	sub	0.011	0.54	eh
suits	sui	0.111	0.68	eh
sunglasses	su0	0.002	0.49	cs
sunny	su1	0.033	0.60	cs
swimmer	swi	0.002	0.29	sc
talking	tal	0.165	0.76	eh
tanks	tan	0.004	0.29	cl
telephones	tel	0.006	0.31	sc
tennis	ten	0.003	0.12	cl
text labeling people	te0	0.023	0.46	cs
text on artificial background	te1	0.051	0.31	cs
ties	tie	0.171	0.70	cs
tower	tow	0.005	0.32	cl
trees	tre	0.057	0.65	cs
tropical settings	tro	0.001	0.49	cs
truck	tru	0.010	0.49	eh
urban park	ur0	0.002	0.33	cs
urban scenes	ur1	0.067	0.73	cs
us flags	usf	0.008	0.33	cs
vegetation	veg	0.084	0.68	cs
vehicle	veh	0.068	0.65	eh
walking running	wa0	0.086	0.78	cs
walking	wa1	0.033	0.68	cs
waterscape waterfront	wa2	0.017	0.39	eh
waterways	wa3	0.010	0.27	eh
weapons	we0	0.043	0.54	cl
weather	we1	0.009	0.25	cs
windows	wi0	0.097	0.81	dc
windy	wi1	0.019	0.58	cs
yasser arafat	yas	0.005	0.44	cs

V. EXPERIMENTS

In this section, we present a set of experiments in which the methods proposed in Sections III and IV are applied. As the first experiment, we study different low-level features' abilities in producing well-concentrated concept-wise distributions using the normalized perplexity measure. Next, we examine the visual and semantic similarities between concepts in respective experiments. We then apply the concept co-occurrence statistics in manual annotation, and finally study the usefulness of inter-concept similarity assessments in concept detection. Due to the space limitations, the performed experiments are described concisely, and are mostly intended as example applications of the methods proposed in this paper, rather than exhaustive experiments in their respective application areas. First of all, we begin the section by describing the settings for the three ontologies used in the experiments.

A. Experiment Settings

In the following experiments, we study three different multimedia ontologies: LSCOM, LSCOM-Lite, and CDVP-206. Each of these ontologies is accompanied by manual annotations on TRECVID datasets. The concepts in the ontologies have been annotated one by one on the shot level based on extracted keyframes, typically resulting in multiple annotations for each shot. In particular, this implies that the annotations are not localized within the keyframe. The TREC Video Retrieval Evaluation (TRECVID) workshop [8] is an annual workshop series aimed to encourage research in multimedia information retrieval by providing a large test collection, uniform scoring procedures, and a forum for comparing results for participating organizations. We have extracted various multimodal low-level features from each of these datasets. Due to the lack of localized annotations and the difficulty of the segmentation problem, we use global features for the LSCOM-Lite and LSCOM ontologies, and features extracted using a fixed 5 × 5 grid for the CDVP-206 ontology. The details for each ontology are given below.

1) LSCOM-Lite: LSCOM-Lite [37] is a subset of the ontology developed in the ARDA/NRRC workshop on LSCOM (see below). It contains 39 semantic concepts listed in Table II. A joint effort was organized to annotate the TRECVID 2005

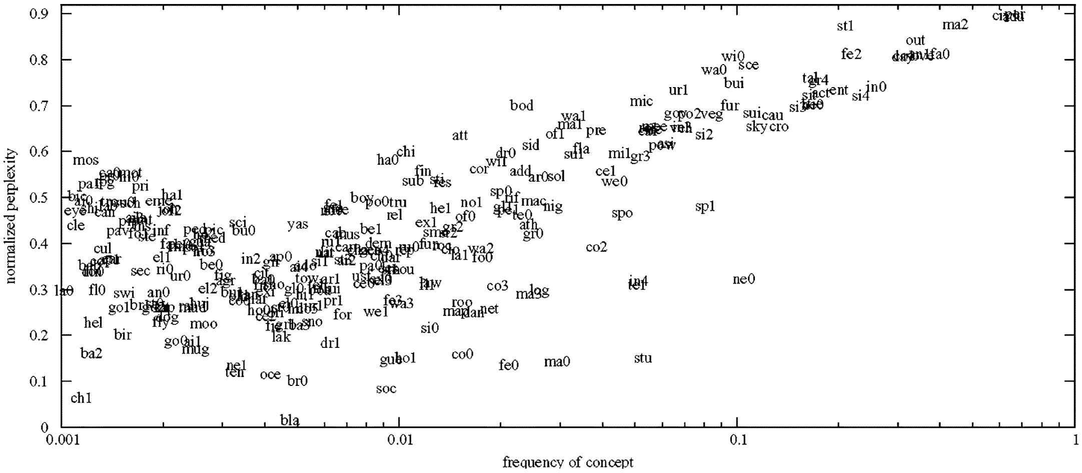


Fig. 1. Minimum values of normalized perplexity over relative sizes of LSCOM concepts based on the training data. The three-letter abbreviations for the concepts are given in Table I. Note that the x axis is logarithmic.

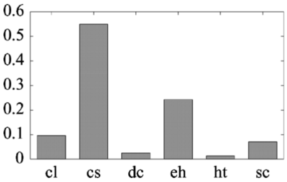


Fig. 2. Proportions of concepts for which each feature yields the minimum value of $\overline{\mathrm{PPL}}$.

[8] development set for the LSCOM-Lite concepts. The dataset consists of about 80 h or 43 907 shots of TV news recorded in November 2004. For our experiments we use two video features (MPEG-7 Motion Activity and temporal color moments), three MPEG-7 image descriptors calculated from the main shot keyframe (Color Layout, Edge Histogram, and Homogeneous Texture), and one audio feature (mel-scaled cepstral coefficient) as shot-level, low-level features. We use the Self-Organizing Map (SOM) [38] as the clustering method with $k = 256$ (16×16 map units) for all these features.

2) LSCOM: The Large Scale Concept Ontology for Multimedia (LSCOM) [6], [7] is an expanded multimedia concept lexicon in development, aimed to contain on the order of 1000 concepts. The current version 1.0 has 856 concepts defined, of which 449 have been used to annotate the TRECVID 2005 development set after a collaborative annotation process was completed in late 2005. Of these 449 concepts, 430 have been used to annotate at least one shot. In the work reported here, we study a set of 280 concepts, selected on the basis that the proportion of relevant shots in the training data was required to exceed 0.001. These concepts are shown in Table I. In the annotation effort for LSCOM the dataset was annotated based on the keyframes of sub-shots. In TRECVID, a time constraint is posed on shot lengths, with adjacent brief sub-shots being concatenated together until they fulfill the time constraint. Therefore, the end result is a total of 61 901 annotated sub-shots. We take a keyframe-level approach and use the sub-shot annotations with the used features extracted from the associated keyframes. We use six MPEG-7 image descriptors, viz. Color Layout (cl), Color Structure (cs), Dominant Color (dc), Edge Histogram (eh), Homogeneous Texture (ht), and Scalable Color (sc). As with LSCOM-lite, we train 16×16-sized SOMs for all these features separately.

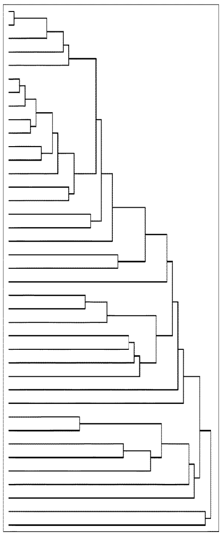
3) CDVP-206: A hierarchical multimedia ontology of 213 concepts was developed in the Centre for Digital Video Processing at Dublin City University [39]. A subset of 6656 shots from the TRECVID 2004 dataset was annotated using this ontology, after which 206 concepts in the ontology had at least one relevant shot. The *is-a* hierarchy of the ontology was utilized in complementing the annotations of parent concepts with their children's annotations. We used three local image features (color histogram, Gabor texture and Canny edge detection) extracted over a regular 5×5 grid. For each of these features we used k-means clustering with $k = 256$. We then concatenate the feature-wise clusters and thus have a bag-of-visterms representation of 75 ($5 \times 5 \times 3$) visterms out of a vocabulary of 768 for each keyframe. To obtain the final representation for the keyframes, we perform probabilistic latent semantic analysis (cf. Section III-B) with 50 latent aspects.

B. Normalized Perplexity

In the experiments in this section, we use the LSCOM ontology. The minimum values of normalized perplexity $\overline{\mathrm{PPL}}$ (6), i.e., the most nonuniform concept distribution, among the six image features are listed in Table I. Fig. 1 shows the minimum $\overline{\mathrm{PPL}}$ values plotted over the proportion of relevant shots in the training data. The three-letter abbreviations used for the concepts are given in Table I. The first observation we can make on Fig. 1 is that the proportion of relevant shots has a clear effect on $\overline{\mathrm{PPL}}$ values. Frequent concepts tend to have high values, which is understandable as common concepts are bound to be heterogeneous, i.e., there is large intra-class variance. To a certain level this is also affected by the fact that a large number of relevant shots will inevitably inhabit a large number of clusters. This effect is not dominant, however, and we have obtained rather similar $\overline{\mathrm{PPL}}$ values by using equal-size concepts obtained by sampling the ground truth. The second observation we make is that rare concepts have a larger variance in minimum $\overline{\mathrm{PPL}}$ values. The reason for the artifact of higher $\overline{\mathrm{PPL}}$ values for very rare

TABLE II
RESULTS OF EXPERIMENTS WITH THE LSCOM-LITE ONTOLOGY

Concept	Five visually most similar concepts	Base AP	Aux. AP
face (2)	person(0), government leader (0), outdoor (10), building (4), walking/running (14)	-	-
person (0)	face (0), outdoor (6), government leader (3), walking/running (14), building (7)	-	-
government leader (1)	face (2), person (1), meeting (3), outdoor (20), building (3)	-	-
corporate leader (0)	face (1), person (0), government leader (0), meeting (2), outdoor (27)	0.002	0.001
meeting (5)	government leader (3), face (2), person (0), building (13), crowd (7)	0.086	**0.126**
outdoor (2)	urban (1), building (1), road (0), walking/running (23), car (3)	-	-
urban (0)	outdoor (0), building (2), road (0), car (3), walking/running (25)	-	-
building (1)	urban (4), outdoor (5), road (0), car (2), person (18)	-	-
car (0)	road (2), outdoor (0), urban (1), building (1), walking/running (26)	0.077	**0.092**
road (0)	urban (0), car (3), outdoor (1), building (4), walking/running (22)	-	-
crowd (0)	walking/running (5), outdoor (10), urban (9), people marching (1), person (5)	-	-
walking/running (10)	outdoor (3), urban (6), crowd (1), road (4), person (6)	-	-
vegetation (14)	outdoor (0), building (1), walking/running (12), urban (1), road (2)	-	-
military (8)	outdoor (0), urban (3), building (1), walking/running (15), road (3)	0.049	**0.056**
sky (8)	outdoor (2), building (0), urban (0), road (1), military (19)	-	-
entertainment (15)	person (0), face (5), outdoor (1), walking/running (5), urban (4)	-	-
sports (5)	walking/running (3), outdoor (0), car (13), vegetation (9), person (0)	0.329	**0.345**
office (3)	person (1), face (0), outdoor (6), entertainment (20), building (0)	0.044	0.010
people marching (8)	crowd (1), walking/running (3), outdoor (2), urban (11), military (5)	0.006	**0.015**
police/security (11)	crowd (0), walking/running (4), urban (5), outdoor (4), road (6)	0.007	**0.014**
natural disaster (10)	building (5), urban (2), outdoor (1), road (0), military (12)	-	-
mountain (0)	sky (2), waterscape/waterfront (3), outdoor (4), road (0), car (21)	0.018	**0.022**
waterscape/waterfront (0)	sky (1), outdoor (5), mountain (2), boat/ship (9), building (13)	0.025	**0.031**
boat/ship (20)	waterscape/waterfront (0), sky (0), mountain (2), road (4), outdoor (4)	-	-
desert (3)	sky (0), mountain (1), explosion/fire (24), waterscape/waterfront (1), outdoor (1)	0.015	**0.021**
explosion/fire (5)	sky (13), outdoor (1), urban (5), military (4), building (2)	0.020	0.020
airplane (11)	sky (1), waterscape/waterfront (3), outdoor (6), road (1), building (8)	0.005	0.003
truck (0)	road (1), car (0), urban (7), outdoor (12), building (10)	0.015	**0.019**
animal (21)	waterscape/waterfront (2), outdoor (0), sky (1), car (6), road (0)	0.002	0.001
snow (2)	mountain (0), sky (0), waterscape/waterfront (0), airplane (16), animal (12)	-	-
computer/tv screen (3)	studio (4), face (3), person (1), meeting (8), building (11)	0.114	**0.135**
studio (6)	computer/tv screen (1), face (2), person (0), meeting (12), maps (9)	-	-
maps (2)	weather (10), studio (5), face (11), person (1), charts (1)	0.076	0.076
weather (5)	maps (0), person (1), face (4), charts (1), outdoor (19)	0.385	**0.418**
charts (2)	weather (12), person (2), face (5), maps (2), computer/tv screen (7)	0.071	**0.116**
flag us (15)	government leader (0), face (0), person (0), crowd (0), walking/running (15)	0.037	**0.074**
bus (0)	road (1), car (0), urban (1), building (9), outdoor (19)	-	-
court (20)	meeting (2), government leader (0), person (0), face (7), corporate leader (1)	-	-
prisoner (0)	military (6), government leader (0), person (0), face (4), walking/running (20)	-	-



concepts is due to normalization using the actual entropy maximum. The size of the clustering, $k = 256$, equals roughly a value of 0.04 in the x axis of Fig. 1, and concepts having fewer relevant shots than the number of clusters begin to approach the maximum value if none of the features is working particularly well. A smaller value of k could therefore be used to study rarer concepts. In the midfrequency range, concepts with the lowest values of $\overline{\mathrm{PPL}}$ are quite what is to be expected, including concepts like *blank frame*, *soccer*, *tennis*, *oceans*, and various news studio related concepts.

Furthermore, Table I shows for each concept the feature that resulted in the minimum value of $\overline{\mathrm{PPL}}$, and Fig. 2 lists the percentages of the concepts for which each feature yields the minimum value. The feature abbreviations are the ones given in Section V-A. It can be observed that each of the six features yields the lowest $\overline{\mathrm{PPL}}$ value for some concepts, highlighting the need for using diverse features for modeling multimedia concepts. Still, for most of the concepts either Color Structure or Edge Histogram gives the minimum value, with shares of 55% and 24% of the concepts, respectively.

C. Visual Similarity

In the second set of experiments, we examine visual similarities among the 39 LSCOM-Lite concepts. We use a linear combination of the multimodal features described in Section V-A, using softmax scaling (7). The distance between concepts in the six clusterings—each corresponding to a different low-level visual feature—are measured by using the Jeffrey divergence of (9). The visual distance between concepts $\mathcal{C}_m$ and $\mathcal{C}_n$ is thus

$$d_{\mathrm{vis}}(\mathcal{C}_m, \mathcal{C}_n) = \sum_{i=1}^{6} \frac{1}{2} \left(w_i^m + w_i^n\right) d_{\mathrm{JD}}(P^{m,i}, P^{n,i}) \qquad (11)$$

where $P^{m,i}$ denotes the probability distribution for concept $\mathcal{C}_m$ in the ith clustering. A full matrix of inter-concept similarities is difficult to illustrate, even for the relatively small LSCOM-Lite ontology. Therefore, the visual similarities are shown in two alternative ways in Table II, namely listing the five most similar concepts for each concept, and as a visual similarity dendrogram, constructed using weighted pair-group average linkage. An examination of this table shows that the visual similarities do seem to group semantically related concepts. In the next experiment we study how well this grouping appears to work for endusers.

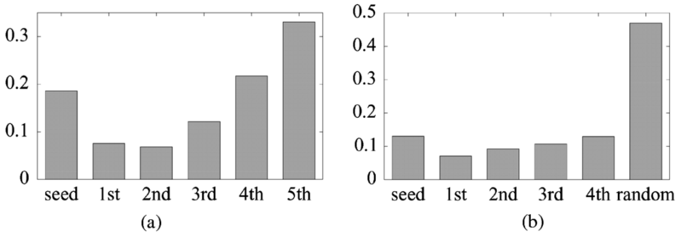


Fig. 3. Proportion of odd-concept-out selections in the user experiments for (a) the 39 concepts of LSCOM-Lite and (b) the CDVP-206 ontology. (a) LSCOM-Lite and (b) CDVP-206.

D. Semantic Similarity Assessments

We ran a set of experiments in which we measured how our visual inter-concept similarity, described above, correlated with human observations of concept similarities. From the set of 39 concepts in LSCOM-Lite, we presented users in random order with six concepts: a seed concept and its five visually most similar concepts (Table II). The users were then asked to nominate the odd one out, namely the one which was conceptually most distant from the others. We repeated this process for each of the 39 concepts. We then performed the same procedure for each concept in the CDVP-206 ontology except that the set of six concepts was composed of a seed concept, its four most similar ones plus another randomly selected concept. Each selection of an odd-one-out for each of the sets of 39 or 206 concepts was performed independently by 30 different users.

The results of the user choices of the most dissimilar concept from the set of six are shown in Fig. 3 and the actual concept-wise results of the user experiments for LSCOM-Lite in Table II (in parentheses after the concepts). The null hypothesis is that we would obtain a uniform distribution among the options. This is however clearly disproven since in the case of LSCOM-Lite the most frequently chosen options as odd-one-out are the 4th and 5th most similar concepts and in the case of the CDVP-206 ontology by far the most frequently chosen option is the randomly selected concept. In both cases the seed concept is also selected rather often, in 18.6% or 13.1% of the cases, respectively. This might suggests that there are some concepts in the ontologies that are semantically distinct, at least from the visually most similar concepts. On the other hand, there is a tendency towards frequent concepts in the lists of visually most similar concepts, and the common concepts are less likely to be selected as odd concepts. With LSCOM-Lite, for example, the ten most frequent concepts in the ontology appear among the five visually most similar concepts more than twice the number of times expected by uniform distribution.

E. Assisted Annotation

Manual annotation of visual media using an ontology of any reasonable size requires the annotator to be fairly familiar with the organization and structure of the ontology in order to achieve inter-annotator consistency. However, when the annotator is not familiar with the ontology, as is the case in the growing amount of nonprofessional annotation activities taking place on the web for example, then there is a challenge to offer automatic assistance in the process. In this experiment to test our concept-concept similarity, ten users, not familiar with the LSCOM ontology, each annotated 40 video shots using different functionalities of an annotation tool. The users were either working within a defined time limit of one minute, or with unlimited time to complete the task, but always under instruction to be as exhaustive in choosing annotation concepts as they could. The annotation tool functionality included i) text search through the concept names, ii) browsing through themed groupings of concepts, and iii) recommendation of concepts to use for a shot based on concepts already assigned. Concept recommendation was based on co-occurrence similarities between concepts chosen up to that point and other concepts in the ontology.

We measured the average number of annotations per shot, the annotation rate (annotations per minute) and the average number of annotations in the limited time of 1 min. These are shown in Table III. The results clearly show that adding the recommendation of concepts, based on concept-concept similarity, increases the number of annotations per shot and the rate of annotation. With the P-value of 0.05, the increases are statistically significant compared to using both the search only and search with themes functionalities. In the limited time experiment, the corresponding differences were not statistically significant. Instead, it was observed that the inclusion of the themed concept groups resulted in statistically significant loss of efficiency with limited time. In all likelihood this was due to the unfamiliarity of the ontology among the test subjects.

TABLE III
RESULTS FROM THE ANNOTATION EXPERIMENT

	Search Only	Search + Themes	Search + Recmd.
Avg annotations per shot	6.9	7.2	11.3
Annotation rate	4.2	3.6	6.1
Avg annotations in fixed 1 min.	6.3	5.2	7.7

F. Concept Detection

In the final set of experiments to assess the usefulness of our concept-concept similarity, we study the utilization of visual and co-occurrence inter-concept relations for individual concept detection. A conventional approach to building an automatic concept detector is to train some machine learning or other detector with positive and negative examples of that concept, independently of other concepts. In these experiments we add both positive and negative *auxiliary* concepts from the LSCOM ontology to the LSCOM-Lite concept detectors of the TRECVID 2006 high-level feature detection task. In practice, there are few if any concepts that co-occur frequently, but are visually very different in the setting studied here. This is because of the lack of localization information in the annotations and the use of global features. Still, if such concepts exists, they can be considered potentially helpful for building concept detectors as they may reveal such shots relevant to a concept that would be otherwise easily neglected. The opposite holds for concepts useful as negative auxiliaries: a visually similar but seldom co-occurring concept is likely to produce false positives. Using these criteria, we pick out five positive and negative candidate concepts and check one by one whether their inclusion improves detection results by using cross-validation with the development set. Typically this resulted in 1–4 additional concepts, the majority of which were negative. As an example, based on the analysis, *military* was augmented with *foxhole* as a positive and with *news studio* and *windows* as negative concepts, and *charts* had a total of four negative concepts assigned: *logos full screen*, *commercial advertisement*, *overlaid text*, and *person*.

In the actual detection of individual concepts, we have used considerably larger (256 × 256 units) SOM-based feature indexes in the PicSOM retrieval system [40]. For details on this, see [41]. The result of the experiment was that by including these additional concepts, the mean inferred average precision (AP) increased from 0.069 to 0.080 over the 20 concepts that were analyzed in TRECVID 2006. The concept-wise AP values for these concepts are shown in the rightmost two columns of Table II. The results in the Aux. AP column appear in bold where the auxiliary concept approach leads to improved detection.

VI. CONCLUSION

In this paper, we presented our work on analyzing large-scale concept ontologies using a clustering-based framework. The shape of a semantic concept's distribution mapped on a set of clusters depends on factors like the distribution of the original data in the very-high-dimensional pattern space, the feature extraction technique, the overall shape of the dataset, and the distribution of the studied concept. The mapping of semantically similar patterns is highly nonrandom provided that the used feature is able to capture enough of the patterns' high-level similarity. We proposed the use of an entropy-based measure to quantify this property and the application of the measure to finding concepts that are relatively more "visual", i.e., easier to model with low-level visual features, as well as automatic weighting of multiple low-level feature spaces. Furthermore, we used the similarities of different concept distributions to measure the strength of the relationships between the concepts. In particular, the similarities in visual content and co-occurrence patterns of concepts can be used to analyze different inter-concept relations within an ontology. However, due to the use of nonlocalized concept annotations and global features, the distinction between these properties becomes somewhat blurred, as the visual representations of region-based concepts are influenced by the corresponding backgrounds and contexts. Still, we were able to obtain meaningful results in the experiments. It is our view that there are a lot of potential applications for this kind of analysis of multimedia data, and in this paper we aimed to illustrate the potential in a variety of experiments including assisted video annotation and automatic concept detection. Further work on similar applications is planned.

REFERENCES

[1] C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in visual information retrieval," *IEEE Multimedia*, vol. 6, no. 3, pp. 38–53, 1999.

[2] M. R. Naphade and T. S. Huang, "Extracting semantics from audiovisual content: The final frontier in multimedia retrieval," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 793–810, Jul. 2002.

[3] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proc. ACM Multimedia (MM'04)*, New York, Oct. 2004, pp. 660–667.

[4] D. B. Lenat, "CYC: A large-scale investment in knowledge infrastructure," *Commun. ACM*, vol. 38, no. 11, pp. 33–38, Nov. 1995.

[5] H. Liu and P. Singh, "ConceptNet—A practical commonsense reasoning tool-kit," *BT Technol. J.*, vol. 22, no. 4, pp. 211–225, Oct. 2004.

[6] LSCOM Lexicon Definitions and Annotations Version 1.0 Columbia University, Tech. Rep. #217-2006-3, Mar. 2006, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia.

[7] M. Naphade, J. R. Smith, J. Tešić, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[8] A. F. Smeaton, "Large scale evaluations of multimedia information retrieval: The TRECVid experience," in *Proc. 4th Int. Conf. Image and Video Retrieval*, Singapore, Jul. 2005, pp. 11–17.

[9] J. Laaksonen, M. Koskela, and E. Oja, "Class distributions on SOM surfaces for feature extraction and object retrieval," *Neural Networks*, vol. 17, no. 8–9, pp. 1121–1133, Oct.–Nov. 2004.

[10] M. Koskela and A. F. Smeaton, "Clustering-based analysis of semantic concept models for video shots," in *Proc. Int. Conf. Multimedia Expo (ICME 2006)*, Toronto, ON, Canada, Jul. 2006.

[11] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *Proc. IEEE Int. Workshop on Content-Based Access of Image and Video Database*, Bombay, India, Jan. 1998, pp. 42–51.

[12] A. Vailaya, A. Jain, and H. J. Zhang, "On image classification: City images versus landscapes," *Pattern Recognit.*, vol. 31, no. 12, pp. 1921–1935, Dec. 1998.

[13] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, Feb. 2003.

[14] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. 26th ACM SIGIR Conf. Information Retrieval*, Toronto, ON, Canada, Jul.–Aug. 2003, pp. 119–126.

[15] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections," in *Proc. Int. Conf. Computer Vision (ICCV 2005)*, Beijing, China, 2005, pp. 370–377.

[16] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *Proc. IEEE Int. Conf. Computer Vision (ICCV 2005)*, Beijing, China, Oct. 2005, vol. 1, pp. 883–890.

[17] M. R. Naphade and J. R. Smith, "Learning visual models of semantic concepts," in *Proc. Int. Conf. Image Processing (ICIP 2003)*, Barcelona, Spain, Sep. 2003, vol. 2, pp. 531–534.

[18] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The Semantic Pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 10, pp. 1678–1689, Oct. 2006.

[19] M. R. Naphade, S. Basu, J. R. Smith, C.-Y. Lin, and B. Tseng, "Modeling semantic concepts to support query by keywords in video," in *Proc. IEEE Int. Conf. Image Processing (ICIP 2002)*, Rochester, NY, Sep. 2002, vol. 1, pp. 145–148.

[20] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Workshop on Generative-Model Based Vision*, Washington, DC, Jun. 2004.

[21] M. G. Christel and A. G. Hauptmann, "The use and utility of high-level semantic features in video retrieval," in *Proc. 4th Int. Conf. on Image and Video Retrieval*, Singapore, Jul. 2005, pp. 134–144.

[22] W.-H. Lin and A. Hauptmann, "Which thousand words are worth a picture? Experiments on video retrieval using a thousand concepts," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME 2006)*, Toronto, Canada, Jul. 2006.

[23] J. R. Kender and M. R. Naphade, "Visual concepts for news story tracking: Analyzing and exploiting the NIST TRECVID video annotation experiment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, USA, Jun. 2005.

[24] K. Yanai and K. Barnard, "Image region entropy: A measure of "visualness" of web images associated with one concept," in *Proc. 13th ACM Int. Conf. on Multimedia*, Singapore, Nov. 2005.

[25] M. R. Naphade, I. Kozintsev, and T. Huang, "A factor graph framework for semantic video indexing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 1, pp. 40–52, Jan. 2002.

[26] Y. Wu, B. L. Tseng, and J. R. Smith, "Ontology-based multi-classification learning for video concept detection," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME 2004)*, Taipei, Taiwan, R.O.C., Jun. 2004, pp. 1003–1006.

[27] R. Yan, M.-Y. Chen, and A. Hauptmann, "Mining relationship between video concepts using probabilistic graphical models," in *Proc. IEEE Int. Conf. Multimedia Expo*, Toronto, ON, Canada, Jul. 2006.

[28] L. Xie and S.-F. Chang, "Pattern mining in visual concept streams," in *Proc. IEEE Int. Conf. Multimedia & Expo (ICME 2006)*, Toronto, ON, Canada, Jul. 2006.

[29] K. Popat and R. W. Picard, "Cluster-based probability model and its application to image and texture processing," *IEEE Trans. Image Processing*, vol. 6, no. 2, pp. 268–284, Feb. 1997.

[30] T. D. Rikert, M. J. Jones, and P. Viola, "A cluster-based statistical model for object detection," in *Proc. Int. Conf. Computer Vision*, Corfu, Greece, Sep. 1999, vol. 2, pp. 1046–1053.

[31] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn. J.*, vol. 42, no. 1, pp. 177–196, 2001.

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. B-39, no. 1, pp. 1–38, 1977.

[33] M. Koskela and J. Laaksonen, "Semantic annotation of image groups with self-organizing maps," in *Proc. 4th Int. Conf. Image and Video Retrieval (CIVR 2005)*, Singapore, Jul. 2005, pp. 518–527.

[34] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1997.

[35] A. Budanitsky and G. Hirst, "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures," in *Proc. NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, Jun. 2001.

[36] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965.

[37] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann, A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005 IBM, 2005, Tech. Rep..

[38] T. Kohonen, *Self-Organizing Maps*, 3rd ed. New York: Springer-Verlag, 2001, vol. 30.

[39] G. Gaughan, "Novelty Detection in Video Retrieval: Finding New News in TV News Stories," Ph.D. dissertation, Dublin City University, Dublin, Ireland, 2006.

[40] J. Laaksonen, M. Koskela, and E. Oja, "PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 841–853, Jul. 2002.

[41] M. Koskela and J. Laaksonen, "Semantic concept detection from news videos with self-organizing maps," in *Proc. 3rd IFIP Conf. Artificial Intelligence Applications and Innovations*, Athens, Greece, Jun. 2006.

Markus Koskela (M'06) received the M.Sc. degree in engineering physics and mathematics in 1999 and the Dr.Sci. degree in technology in computer science in 2003, both from Helsinki University of Technology, Helsinki, Finland.

He is presently a Postdoctoral Researcher at the Adaptive Informatics Research Centre, Helsinki University of Technology. He was a Postdoctoral Fellow with Dublin City University, Dublin, Ireland, from 2005 to 2007. His research interests are in image and video indexing, relevance feedback, multimedia understanding, and content-based information retrieval.

Alan F. Smeaton (M'05), photograph and biography unavailable at the time of publication.

Jorma Laaksonen (SM'02) received the Dr.Sci. in technology degree in 1997 from Helsinki University of Technology, Helsinki, Finland.

He is presently Academy Research Fellow of Academy of Finland at the Laboratory of Computer and Information Science. He is an author of several journal and conference papers on pattern recognition, statistical classification, and neural networks. His research interests are in content-based information retrieval and recognition of handwriting.

Dr. Laaksonen is a founding member of the SOM and LVQ Programming Teams and the PicSOM Development Group, and a member of the International Association of Pattern Recognition (IAPR) Technical Committee 3: Neural Networks and Machine Learning.