An Empirical Study of Inter-Concept Similarities in Multimedia Ontologies

Markus Koskela*
Adaptive Informatics Research Centre
Helsinki University of Technology
P.O.Box 5400, FI-02015 TKK, Finland
markus.koskela@hut.fi

Alan F. Smeaton
Centre for Digital Video Processing and Adaptive Information Cluster, Dublin City University
Glasnevin, Dublin 9, Ireland
alan.smeaton@computing.dcu.ie

ABSTRACT

Generic concept detection has been a widely studied topic in recent research on multimedia analysis and retrieval, but the issue of how to exploit the structure of a multimedia ontology as well as different inter-concept relations, has not received similar attention. In this paper, we present results from our empirical analysis of different types of similarity among semantic concepts in two multimedia ontologies, LSCOM-Lite and CDVP-206. The results show promise that the proposed methods may be helpful in providing insight into the existing inter-concept relations within an ontology and selecting the most facilitating set of concepts and hierarchical relations. Such an analysis as this can be utilized in various tasks such as building more reliable concept detectors and designing large-scale ontologies.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.2.4 [**Database Management**]: Systems—*Multimedia databases*

General Terms

Experimentation, Algorithms, Human Factors

Keywords

Multimedia ontologies, semantic concept detection

1. INTRODUCTION & CONTEXT

Associating semantic information with visual data has attracted a lot of research attention recently in order to facilitate semantic indexing and concept-based retrieval of visual content. The predominant approach has been to employ various statistical modeling techniques for mid-level semantic concepts (events, objects, locations, people, etc.) based on low-level visual features. These models are then used to support high-level indexing and querying on visual data. Such approaches have shown promising results [1] as the semantic concept models can be trained off-line with considerably more relaxed requirements for computational efficiency and more positive and negative examples than would be available at query time.

Recently published large-scale multimedia ontologies as well as large manually annotated datasets have allowed an increase in multimedia concept lexicon sizes by orders of magnitude. The availability of ontologies like the Large Scale Concept Ontology for Multimedia (LSCOM) [11] and the MediaMill Challenge 101 [19] have been important developments in the research field. Both of these ontologies are accompanied by manual annotations on TREC Video Retrieval Evaluation (TRECVid) datasets. The TRECVid workshop [16] is an annual workshop series which encourages research in multimedia information retrieval by providing a large test collection, uniform scoring procedures, and a forum for comparing results for participating organisations. The TRECVid benchmarks have included a separate high-level concept detection task since 2002.

Semantic concepts do not exist in isolation but have different relationships between each other, including similarities in their semantic and visual (low-level) characteristics, co-occurrence statistics, and different hierarchical relations if a taxonomy has been defined for the concepts. Several techniques have recently been proposed to utilize these contextual inter-concept relations, but the question of how should they be utilized for various tasks still remains as an open research question. One approach is to concatenate the detection scores of either all available concept detectors or some "basis" concepts selected based on criteria such as coverage or reliability of detection. These individual concept detectors are gathered into a model vector space in which some statistical learning technique is then applied, as with any low-level feature space. In [17] and [5], the model vectors are used as input to a SVM to improve the performance of concept detection. In [18], the context vectors are incorporated into an authoring metaphor, which divides the indexing process into content, style, and context steps. A semiautomatic version of this approach was proposed in [6], in which a subset of detection scores is replaced with manually provided annotations. In [2], individual events are modeled as temporal processes within video shots using the temporal dynamics of the concurrent concepts.

*This work was performed while at the Centre for Digital Video Processing, Dublin City University, Ireland.

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.

An alternative approach to utilizing inter-concept relations is to analyze pairwise relations among all concepts in order to identify those that have a measurable effect (either positive or negative) to detecting or modeling a given concept. Various graphical models have been proposed for this purpose. In [14], models based on manually and automatically constructed Bayesian networks are used to capture the interaction between concepts. In [13], a factor graph framework for multimedia objects or "multijects" is used to model their interaction. Multiple graphical models for concept detection are studied in [22]. In [7], concept clustering and statistical G-tests are used for finding concept pairs that co-occur frequently. The use of frequent itemset mining, k-means clustering, and hidden Markov models to pattern mining of large-scale multimedia ontologies is studied in [21]. In [20], an existing ontology hierarchy is used to influence individual concept detectors. A boosting factor is used for top-down influence from parent concepts to children, and a confusion factor is defined for mutually exclusive concepts.

In the work reported here we analyze the inter-concept relations and levels of similarity between concepts in two large-scale multimedia ontologies and present preliminary results of an empirical study using these multimedia ontologies. While the definitive utility of any such analysis is derived from the application of such results to practical tasks of importance such as concept detection or annotation, such a study can provide additional insights into the ontologies as well as encourage the application of such methods provided that the results are plausible. We study here the overall characteristics and similarities among semantic concept models instead of processing the detection confidence scores of some individual data items. The analysis is based on image-level annotations, which has obvious limitations as most visual concepts are localized, i.e. they correspond to a distinct object or part of the scene. Unfortunately, producing localized annotations for large-scale multimedia ontologies is a challenging task.

We study two ontologies, namely a subset of 39 concepts from the LSCOM ontology and an in-house ontology developed at the Centre for Digital Video Processing at Dublin City University, named CDVP-206 in the experiments. The reason for choosing the latter instead of other commonly used ontologies such as LSCOM or MediaMill 101 is that neither of these ontologies currently contains a hierarchical arrangement of its concepts, and we aim to study and exploit the hierarchical relations of concepts in the CDVP-206 ontology. The methodology for measuring concept similarity is the same as the one presented in [9], but in this paper we focus on the distinct and to a certain degree orthogonal types of inter-concept similarity in greater detail and we present an empirical study exploring these similarities.

The paper is organized as follows. Section 2 provides a brief overview of the used methods for analyzing and measuring different inter-concept similarity relations. In Section 3, the two ontologies used in the experimental part of the paper are described. A series of experiments with the proposed methods are presented in Section 4, and the paper is concluded in Section 5.

2. CONCEPT SIMILARITIES

In this section, we describe methods to measure the overlap among concepts. In many of the existing approaches to semantic concept modeling, the ontology, or the inter-concept relations in general, have not been usefully exploited. Each concept model is rather treated as a binary classifier and processed as if it were an independent entity. This approach may seem unsatisfactory, as there are obviously strong links between the occurrence patterns of related concepts, and these links could conceivably be exploited to improve the quality of the models. Moreover, some concepts are more inherently visual than others, and thereby easier to model with low-level features. One strategy could then be to use these visual concepts to improve the models for the other, more difficult, concepts.

We now consider four different similarity relations between semantic concepts.

2.1 Visual Similarity

When considering the automatic detection or classification of multiple concepts in an ontology, one question is the similarity among concepts based on low-level features. Despite the semantic gap and the resulting relatively low accuracy of independent content-based concept models, these models remain fundamental for many automatic processing tasks for multimedia data.

For measuring visual similarity among concepts, as our low-level representation we adopt a clustering-based probabilistic model in which we model the probability density function of the semantic concept over a set of k clusters common to the whole data set. For brevity, the treatment here is rather concise; see [9] for more details. The approach provides rather coarse concept models but has the benefit of being readily scalable to large multimedia lexicons as each concept is represented as a set of discrete distributions over clustered feature spaces. For simplicity, we adopt the term "visual" for all characteristics based on low-level features even though all such features might not be visual, e.g. for video analysis one might also include audio features.

2.1.1 Global Descriptors

A common approach to representing visual data in low-level feature spaces is to extract multiple global features from each media object. If a certain feature extraction method works favorably, semantically similar patterns will still be mapped in the feature space nearer to each other than semantically dissimilar ones. The dimensionality of the feature space can then be reduced e.g. with clustering, while preserving the similarity property. That is, after clustering we can still expect a non-uniform distribution of semantically related objects over the clustering provided that the low-level feature in question is able to capture enough of the semantic similarity between the visual objects. Given $i = 1, \ldots, k$ cluster centroids with $\mathcal{V}_i$ as the corresponding Voronoi region, a discrete probability histogram of a dataset can be written as

$$P_i = P(\mathbf{x} \in \mathcal{V}_i) = \frac{\#\{\, j \mid \mathbf{x}_j \in \mathcal{V}_i \,\}}{\#\{\, j \,\}} \quad . \tag{1}$$

where $\#\{\cdot\}$ stands for the cardinality of a set. With a specific subset of data c, fulfilling a certain ground truth criterion, the corresponding probability histogram will be

$$P_i^m = P(\mathbf{x} \in \mathcal{V}_i \mid \mathbf{x} \in c) = \frac{\#\{\, j \mid \mathbf{x}_j \in \mathcal{V}_i, \mathbf{x}_j \in c \,\}}{\#\{\, j \mid \mathbf{x}_j \in c \,\}} \quad . \tag{2}$$

In this paper, we refer to these subsets as semantic concepts.

2.1.2 Grid-Based Descriptors

The representation of visual data using global features is quite limited in many cases. While some semantic concepts do relate to the content of a video shot or image *as a whole*, some others are localized to a distinct object or a specific part of the background or scene. For region-based image representation, we use a localized appearance descriptor based on a regular grid dividing the image into a fixed number of regions. The set of grid-based descriptors is then quantized to produce visterms, after which the multimedia object can be regarded as a document consisting of a number of visterms. For modeling the documents in a lower-dimensional space, we use probabilistic latent semantic analysis (PLSA) [4], where a latent variable or aspect $z_k,\ k = 1, \ldots, n$ is associated with each observation. The joint probability of documents and visterms is defined as the mixture

$$P(w_j, d_i) = P(d_i) \sum_k P(w_j|z_k)P(z_k|d_i) \tag{3}$$

where $P(w_j|z_k)$ is the class-conditional probability of the visterm w_j conditioned on the unobserved aspect z_k and $P(z_k|d_i)$ denotes the probability distribution of the latent aspects given the document d_i. A new document d_q can be "folded-in" to the aspect space $P(z_k|d_q)$ by keeping the document-independent probabilities $P(w_j|z_k)$ fixed. A concept c_m can then be aggregated to a document d_{c_m} and modeled as a distribution $P(z_k|d_{c_m})$ over the latent aspects.

2.1.3 Similarity Measurement

For both the global and grid-based descriptors, any bin-to-bin histogram distance measure can be used in estimating the visual distance $d_{\text{vis}}(c_m, c_n)$ of concepts c_m and c_n. In this paper, we use Jeffrey divergence

$$d_{\text{vis}} = d_{\text{JD}}(P^m, P^n) = \sum_{i=0}^{k-1} \left(P_i^m \log \frac{P_i^m}{\hat{P}_i} + P_i^n \log \frac{P_i^n}{\hat{P}_i} \right) , \tag{4}$$

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

2.2 Co-occurrence

A complementary view of concept similarity can be obtained by considering the co-occurrence statistics of pairs of concepts. The concepts in an ontology are interrelated and certain concept pairs co-occur more often or more rarely in the same or neighboring multimedia objects than would be expected by chance. For example, the concepts *car* and *road* will in all likelihood appear frequently together whereas *indoor* and *outdoor* are almost always mutually exclusive. The knowledge of the presence or absence of certain concepts may therefore be a valuable cue in predicting the presence of other concepts. This can be exploited in many kinds of applications, such as concept detection or annotation.

The presence of a semantic concept in a multimedia object is usually assumed as a binary variable, making it straightforward to analyze co-occurrence patterns with standard data mining techniques. A number of such methods have been proposed in recent research, including the G-test [7], frequent itemsets [21] and shot clustering [7, 21]. In a similar manner, we examine concept occurrence as a binary variable over the data items in the training set. We denote C^m as a vector of length equalling the size of the training set, with $C_i^m = 1$ if the ith item is relevant for concept c_m and $C_i^m = 0$ otherwise. To measure co-occurrence distance $d_{\text{co}}(c_m, c_n)$, we use the Cosine measure

$$d_{\text{co}} = d_{\cos}(C^m, C^n) = \frac{\sum_{i=0}^{k-1} C_i^m C_i^n}{\sqrt{\sum_{i=0}^{k-1} (C_i^m)^2} \sqrt{\sum_{i=0}^{k-1} (C_i^n)^2}} . \tag{5}$$

2.3 Semantic Similarity

The third type of concept similarity we consider is similarity among concepts based on their semantic meaning. By nature, these properties differ from the two similarities discussed above and one cannot use a ground truth set of annotated multimedia objects to deduce semantics of concepts. Semantics are also highly subjective, so any group of annotators will likely produce conflicting results on the semantic similarity between concepts.

For our experiments, we gather subjective assessments of different concepts' semantic similarity from a group of human subjects. The details of the data gathering procedure are given in Section 4.1.1. Gathering such assessments for a small number of concepts is straightforward but with large-scale ontologies this becomes infeasible due to the quadratic increase in the number of concept pairs compared to the size of the ontology. As a result, the semantic similarities based on human assessments have to be restricted to some subset of concept pairs. Here, we limit our study of semantic similarities to comparisons with visual similarity relations.

2.4 Hierarchical Structure

The fourth similarity relation considered in this paper is based on an hierarchical construction or taxonomy of concepts. A taxonomy provides a natural way of organizing and processing large ontologies. Concepts near each other in such a hierarchy, e.g. sibling concepts or concept pairs with a parent-child relation, will have close association to each other and should be treated differently than random pairs of concepts. The concept hierarchy can also be used directly to construct a tree distance, in which the distance of two concepts is a function of the number of edges between them. The most common relation in multimedia taxonomies is the subsumption or *is-a* relation (e.g. *dog* is-a *animal* is-a *object*). The whole hierarchy of the CDVP-206 ontology is based on the subsumption relation.

The concept taxonomy can also highlight the fundamental differences between concepts of different types. Generally, top-level concepts are common and thus have lots of training data, but are often too generic for accurate modeling. For example, concepts such as *person*, *outdoor* and *sky* can often appear in more than half of all multimedia objects. On the other hand, rare leaf concepts may be distinctive but suffer from a lack of positive examples. In practice, the rarest concepts have to be excluded when building concept models as they cannot be reliably modeled. As a result, the mid-frequency concepts often provide the most fruitful portion of concepts for automatic analysis and modeling.

An hierarchical structure among concepts can be directly utilized for different purposes. One straightforward top-down approach to concept detection is to consider the concept hierarchy as a decision tree where each detector differentiates among its immediate descendants. This approach is suited for sibling concepts that are mutually exclusive, which is a common property with multimedia ontologies but not

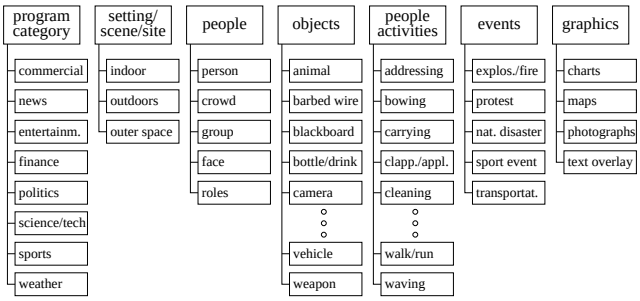

Figure 1: The two topmost levels of the CDVP-206 ontology.

universal. For example, the positive result set of an *animal* detector probably should not be further divided into distinct species unless it is known that animals of only one species may exist in any given multimedia object.

On the other hand, one may also employ a bottom-up approach where objects detected which are relevant to a child concept with a high confidence are also deemed relevant for the parent concept. This may be beneficial especially if the child concept can be detected more reliably. Both the top-down and bottom-up approaches can also be utilized in semi-automatic annotation.

3. MULTIMEDIA ONTOLOGIES

In the empirical study presented in this paper, we use two ontologies developed for news video material.

3.1 LSCOM-Lite

LSCOM-Lite [12] is an interim subset of the LSCOM [11] ontology, developed in the ARDA/NRRC workshop on Large Scale Ontology for Multimedia (LSCOM). LSCOM-Lite contains the following 39 semantic concepts:

> *airplane, animal, boat/ship, building, bus, car, charts, computer/tv screen, corporate leader, court, crowd, desert, entertainment, explosion/fire, face, flag us, government leader, maps, meeting, military, mountain, natural disaster, office, outdoor, people marching, person, police/security, prisoner, road, sky, snow, sports, studio, truck, urban, vegetation, walking/running, waterscape/waterfront, weather.*

Video shots from the entire TRECVid 2005 [16] development collection of about 80 hours of TV news recorded in November 2004, were manually annotated for the LSCOM-Lite concepts in a joint effort by TRECVid participants. As global low-level features for this dataset 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 the clustering method for these features, we use the Self-Organizing Map [8] with $k = 256$ (16×16 map units), and the different clusterings are fused using weighted linear combination.

3.2 CDVP-206

The CDVP-206 ontology is based on an hierarchical multimedia taxonomy of 213 concepts developed in the Centre for Digital Video Processing at Dublin City University [3]. Figure 1 shows the two topmost levels of the CDVP-206 ontology hierarchy.

A subset of 6 656 shots from the TRECVid 2004 dataset was annotated using this ontology [3], after which 206 concepts in the ontology had been assigned to at least one shot. The *is-a* hierarchy of the ontology was utilized in complementing the annotations of parent concepts with their children's annotations. With this dataset we take a keyframe-based approach and use three grid-based image features (color histogram, Gabor texture and Canny edge detection) extracted over a regular 5×5 grid. For each of these features, we use k-means clustering with $k = 256$. By concatenating these feature-wise clusters, we obtain a bag-of-visterms representation of 75 ($5\times5\times3$) visterms out of a vocabulary of 768 for each keyframe. To obtain the final representation for the keyframes, we perform PLSA with 50 latent aspects.

4. EMPIRICAL STUDY

4.1 Inter-Concept Similarity

We begin our ontology analysis by examining semantic and visual similarities and distinctness of the concepts in both ontologies. For purposes of illustration, we classify concepts as either similar, neutral, or distinct, in relation to other concepts in the ontology using heuristically set thresholds. In actual application of the similarity values, it would presumably be preferable to use the proper values instead.

In these experiments, we consider concepts c_m and c_n to be visually similar if the visual distance (Eq. (4)) between them is smaller than a threshold, i.e. if $d_{\text{vis}}(c_m, c_n) < T^s_{\text{vis}}$. Similarly, a concept c_m is considered visually distinct if

$$\underset{c_i \in \mathcal{O} \setminus c_m}{\operatorname{argmin}} \, d_{\text{vis}}(c_m, c_i) > T^d_{\text{vis}} \tag{6}$$

where $\mathcal{O}$ is the full ontology, i.e. a set containing all concepts.

4.1.1 Assessments of Semantic Similarity

We ran two separate experiments in which we measured how the visual inter-concept similarity correlated with human observations of concept similarity. In the first experiment, we presented sets of six concepts from the LSCOM-Lite ontology to users in random order. These sets always contained a seed concept and its five visually most similar concepts. Users were then asked to nominate the odd one out, namely the concept which was conceptually most distant to the others. This process was repeated for each of the 39 concepts by 30 different users. In the second experiment, the procedure was repeated for each concept in the CDVP-206 ontology with the exception that the sets of six concepts were now composed of the seed concept, its four visually most similar concepts, and a randomly selected concept, again in random order. For each concept as the seed, we gathered results from 30 users.

Table 1 gives an overview of the results of the semantic similarity experiments. The upper and lower rows show the normalized mean and the median of the number of times the corresponding concept was chosen as the odd one out. The seed column shows the proportion of cases where the seed concept was selected, and the other concepts are listed in decreasing order of visual similarity. In the case of CDVP-206, the rightmost value corresponds to the random concept. We observe a non-uniform distribution of selections for the non-seed concepts as was to be expected. On aver-

Table 1: An overview of semantic similarity assessments.

LSCOM-Lite	seed	other concepts				
normalized mean	0.186	0.076	0.068	0.121	0.217	0.331
median	3	1	1	1	4	7
CDVP-206	seed	other concepts				
normalized mean	0.131	0.071	0.092	0.106	0.130	0.470
median	1	1	1	1	1	14

Table 2: Semantically and visually distinct concepts in LSCOM-Lite with $T^d_{\text{vis}} = 0.3$ and $T^d_{\text{sem}} = 10$.

visually distinct:	**semantically distinct:**
animal, bus, charts, court, flag us, maps, prisoner, snow, weather	airplane, animal, boat/ship, court, entertainment, flag us, natural disaster, vegetation

visually and semantically distinct:
animal, court, flag us

age, visually more similar concepts are selected more rarely than less similar ones. The random concept introduced in the CDVP-206 experiment is chosen in almost half of the answers as the odd one out. The median rows help to illustrate the asymmetric form of the distributions; for the majority of the concepts, the visually most similar concepts are chosen only very rarely.

For the purposes of the following experiments, we define a concept c_m as semantically distinct if $n^*_m > T^d_{\text{sem}}$ where n^*_m is the number of times c_m was selected as the odd concept out; the asterisk is used to mark c_m as a seed concept. In a similar manner, we consider c_n to be semantically similar to c_m if $\min(n^*_m, n_n) < T^s_{\text{sem}}$. Due to the nature of the experiments, the defined semantic similarity relationship between concepts is necessarily asymmetric.

4.1.2 Similar and Distinct Concepts

For LSCOM-Lite, by using threshold $T^s_{\text{vis}} = 0.1$ we can observe a rather large clique of visually similar concepts consisting of *face*, *person*, *government leader*, *outdoor*, *urban*, *building*, *car*, *road*, *crowd*, and *walking/running* in addition to a smaller clique consisting of *vegetation*, *outdoor*, and *sky*. Then, by taking into account the semantic similarity with $T^s_{\text{sem}} = 3$, we obtain two separate cliques of concepts that can be regarded as both semantically and visually similar. These cliques are formed by the following concepts: *face – person – government leader* and *outdoor – urban – building – car – road*. The threshold values naturally have a direct effect on the results; the number of similar concepts can be increased by raising the thresholds.

The most distinct concepts of LSCOM-Lite, according to our analysis, are listed in Table 2. With the thresholds set as $T^d_{\text{vis}} = 0.3$ and $T^d_{\text{sem}} = 10$, the sets of visually and semantically distinct concepts contain 9 and 8 concepts, respectively. With the said thresholds, there are three concepts, viz. *animal*, *flag us*, and *court*, that satisfy both criteria.

With the larger CDVP-206 ontology, the semantically most similar concepts are quite possibly not among the four visually most similar concepts, so with the experimental settings we used we cannot determine the overall semantic distinctness of concepts. Furthermore, due to the larger size and the existing concept taxonomy, there is a large number of both

Table 3: Some visually distinct but semantically similar concept pairs in CDVP-206 ($T^d_{\text{vis}} = 0.4$ and $T^s_{\text{sem}} = 3$).

airplane – airplane landing	entering – government leader
airplane – pilot	entering – standing
airplane landing – sky	ice skating – playing
bank setting – chair	ice skating – sport event
baseball – tennis	ice skating – sports
bicycle – bird	pilot – road
bridge – tractor	pilot – sky
dancing – tennis	prisoner – transportation event
department store setting – supermarket setting	prisoner – truck
	street light – tent
dome – tent	supermarket setting – table
driving – pilot	tractor – tree

visually and semantically similar concepts, many of which have a direct relation with each other in the concept hierarchy. For these reasons, we consider it to be more fruitful to explore interesting anomalies existing in the CDVP-206 ontology. We focus here on concept pairs for which the two distance measures have contrary values, i.e. concepts that are visually similar but have a high semantic distance, or vice versa. Unfortunately, due to the nature of the semantic similarity experiments we carried out, the list of visually distinct but semantically similar concept pairs is partial as we only gathered data on the semantic distance of each concept to four other concepts. We are thus limited to studying the concepts which are semantically similar but visually distinct from their visually nearest concepts. Table 3 shows a list of such concepts with threshold values $T^d_{\text{vis}} = 0.4$ and $T^s_{\text{sem}} = 3$.

In the opposite case, that is with concepts scoring high values for semantic distinctness but being visually similar to one or more concepts, we observe that the resulting concepts are to a large extent either generic, upper-level concepts in the taxonomy (Fig. 1), or object-related concepts for which the visual similarity is strongly affected by the background due to the use of global features. For example, with $T^s_{\text{vis}} = 0.1$ and $T^d_{\text{sem}} = 10$ the list of semantically distinct concepts consists of two program categories (*commercial* and *science/technology*), two other generic concepts (*indoor* and *roles*), in addition to *animal*, *bottle/drink*, *keyboard*, and *transportation setting*.

4.2 Co-occurrence

We now turn our attention to the co-occurrence patterns among multiple concepts. We measure the co-occurrence distance between two concepts using Eq. (5).

Figures 2 and 3 illustrate the relationship between co-occurrence distance and visual similarity among all concept pairs in the LSCOM-Lite and CDVP-206 ontologies, respectively. The correlation between the two properties is evident: concept pairs that co-occur frequently together are also visually similar. This is however largely due to the global natures of both the low-level features and the ground-truth annotations, as the positive training sets for co-occurring concepts are highly overlapping. Table 4 shows the ten pairs of concepts that have the smallest co-occurrence distances in both ontologies. The results are hardly surprising as the concept pairs are also semantically very similar, even redundant, or, in the case of CDVP-206, often have a direct parent-child relationship in the concept hierarchy.

Figures 2 and 3 also show that among concepts having a

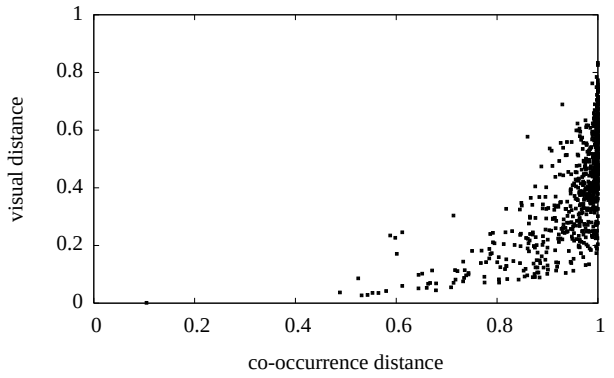

Figure 2: Visual vs. co-occurrence distances of all concept pairs in the LSCOM-Lite ontology.

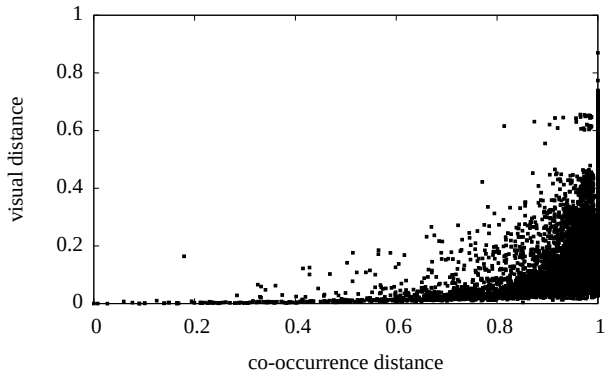

Figure 3: Visual vs. co-occurrence distances of all concept pairs in the CDVP-206 ontology.

high co-occurrence distance, the visual similarities are notably dispersed. There are pairs of concepts for which both distances are high, but also a considerable number of concept pairs with high visual similarity. This demonstrates a well-known major problem for concept detection as it shows that the low-level features we used are unable to make a distinction between many pairs of concepts even when they tend not to occur simultaneously. The problem seems to be less severe for LSCOM-Lite which is understandable as many of the 39 concepts included are quite distinct.

The majority of the concept pairs have a high co-occurrence distance between each other. This highlights the potential utility of discovering and exploiting mutually exclusive or almost exclusive concepts pairs in an ontology. Provided that the presence of certain concepts is known in given multimedia objects, e.g. by reliable detection or user annotations, it becomes possible to exclude the presence of a number of other concepts, with high accuracy.

4.2.1 Auxiliary Concepts for Concept Detection

As a potential application for the kind of concept-concept analysis we have presented here, we explore the utilization of visual and co-occurrence relations in concept detection. Based on the analysis, we add both positive and negative auxiliary concepts to aid individual concept detectors. Such auxiliary concepts were utilized in our TRECVid 2006 experiments [15], where they increased the mean (inferred) average precision of our detectors by 16%.

Table 4: Ten pairs of concepts having the smallest co-occurrence distances in both ontologies.

LSCOM-Lite	CDVP-206
face – person	adult – person
car – road	people activities – people
outdoor – sky	american flag – flag
outdoor – urban	graphics – text overlay
building – urban	gun – weapon
road – urban	city street – cityscape
building – outdoor	adult – people activities
outdoor – road	people activities – person
face – studio	protest – protesting
boat/ship – waterscape	adult – male

As discussed earlier and clearly seen in Figures 2 and 3, there are few if any concepts that co-occur together frequently but are visually very different in our setting. Still, when concept pairs exist that show such a tendency, it is conceivable that this relation may be helpful when building concept detectors for either one of the concepts. The second concept might highlight such properties that are not clearly manifested in the primary model for the concept in question and thus reveal objects that, while being relevant, would potentially otherwise be missed.

The opposite holds for concepts potentially useful as negative auxiliaries: objects relevant for a visually similar but seldom co-occurring concept are likely to be false positives in our current detector. In fact, this property is likely to be more useful than the positive auxiliary concepts in improving individual detectors, for two reasons. Firstly, there are a lot more potential pairs of concepts having the said property, as can be seen from Figures 2 and 3 and secondly, using positive auxiliaries requires a delicate tuning of weight parameters for the concepts. The positive training data for the concept itself is the most important source of information for the concept detector, and the additional concepts are responsible for fine-tuning the detector. The negative auxiliaries, on the other hand, offer an additional source of information about potential false positives and are thus less sensitive to the values of the weight parameters.

To select auxiliary concepts based on the criteria described above, we apply a distance measure d from the concept pairs to the "optimal" relations for positive and negative auxiliaries in the two-dimensional space spanned by the visual and co-occurrence distances. The optimal points are defined as the maximal visual distance ($d_{\text{vis}} = 1$) and zero co-occurrence distance ($d_{\text{co}} = 0$) for the positive auxiliaries, and vice versa for the negative auxiliaries. We denote these distances to the positive and negative optimal points as d^+ and d^-, respectively. In order to emphasize mid-range concept pairs, i.e. pairs that have non-extremal values on both dimensions d_{vis} and d_{co}, we chose here to use Minkowski distance of order ∞ instead of Euclidean distance. Thus, for concept pair c_m and c_n, we get

$$d^+(c_m, c_n) = \max\{d_{\text{co}}(c_m, c_n), |d_{\text{vis}}(c_m, c_n) - 1|\} \quad (7)$$

$$d^-(c_m, c_n) = \max\{|d_{\text{co}}(c_m, c_n) - 1|, d_{\text{vis}}(c_m, c_n)\} \; . \quad (8)$$

When building a detector for a specific concept c_m, the concepts in $c_i \in \mathcal{O} \setminus c_m$ that have the lowest values of $d^+(c_m, c_i)$ and $d^-(c_m, c_i)$ are determined and can then be used as positive and negative auxiliary concepts, respectively.

Table 5: Ten most positive and negative concept pairs in LSCOM-Lite.

positive pairs	negative pairs
maps – weather	military – urban
person – studio	urban – vegetation
face – studio	building – military
boat/ship – waterscape	building – crowd
computer/tv screen – face	car – walking/running
computer/tv screen – person	car – vegetation
car – truck	vegetation – walking/running
mountain – waterscape	road – vegetation
outdoor – people marching	crowd – road
person – sports	car – face

Table 6: Potential auxiliary concepts for selected concepts in the CDVP-206 ontology.

concept	positive	negative
building	car crash, dome, bridge, steps and staircases, explosion/fire	car, indoor, office setting, juvenile/child/teenager, table
golf	sports, sport event, events, vegetation, greenery	tree, land, water body, outdoors, natural disaster
indoor	hockey/ice hockey, charts, sports, playing, airport setting	outdoors, building, cityscape/urban setting, car, vehicle
tree	car crash, golf, weather, commercial, snow	car, governm. leader, indoor, carrying, politician
weapon	missile, soldier, missile launch, store setting, military personnel	outdoors, building, standing, setting/scene/site, people

Table 5 lists the ten concept pairs that have the overall smallest values of d^+ and d^- in the LSCOM-Lite ontology, and Table 6 shows five positive and negative auxiliaries for a number of selected concepts in the CDVP-206 ontology. In both tables, we observe concepts that tend to co-occur in the positive columns. The concepts are, however, not semantically identical or even remarkably similar but rather they tend to express specific instances of the presence of the related concept. The negative columns, on the other hand, show concepts that would be likely to produce false positives for each other. It should be noted that not all concepts listed in Tables 5 and 6 are likely to function as beneficial auxiliaries, and the final positive and negative auxiliary concepts should be determined e.g. with cross-validation.

4.3 Concept Hierarchy

In the third part of this study, we take into account the concept taxonomy existing for the CDVP-206 ontology. For many applications, such an hierarchical ordering of concepts is crucial and multimedia ontologies such be designed accordingly. A concept taxonomy brings a structured representation to a large-scale ontology and enables a multitude of types of top-down and bottom-up processing. Especially the similarity relations between parent and child concepts as well as between sibling concepts are of particular interest.

In this preliminary study, we present the analysis of two sets of sibling concepts in the CDVP-206 ontology, namely the children of the concept *indoor*:

> *laboratory setting, meeting/board room, court, house setting, press conference, bank setting, factory setting, departm. store setting, hospital setting, store setting, church setting, studio setting, restaurant setting, school setting, supermarket setting, night club setting airport setting, transport. setting, office setting,*

and the concept *outdoors*:

> *cityscape/urban setting, land, building, sky, bridge, rural setting, statue/monument, beach, steps and staircases, cloud, vegetation, waterfall, mountain, water body, desert, snow, road.*

These were selected as they provide a realistic setting that could be obtained top-down as the results of an indoor-outdoor detector or, alternatively, of manual annotation. A second reason for using these is that in the CDVP-206 ontology they represent both the mutually exlusive and non-exclusive sets of sibling concepts. The children of *indoor* in CDVP-206 are almost completely mutually exclusive whereas some concept pairs among the children of *outdoors* co-occur rather frequently. As a result, the *indoor* concepts can be processed with a multi-class classifier whose confidence scores can be updated e.g. using a confusion factor [10, 20]. The children of *outdoors*, however, require a different approach, such as the one taken in Section 4.2.

First of all, we examine the visual distances between the sibling concepts. With thresholds $T^s_{\text{vis}} = 0.1$ and $T^d_{\text{vis}} = 0.2$ for both sets, we observe both distinct and similar concepts in both cases. Among the children of *indoor*, the list of distinct concepts contains *laboratory setting*, *bank setting*, *church setting*, *supermarket setting*, and *night club setting*. We also observe a large clique of visually similar concepts consisting of *meeting/board room*, *court*, *house setting*, *press conference*, *factory setting*, and *office setting*, as well as some additional similar pairs (*transp. setting – house setting* and *hospital setting – office setting*). The presence of such a large clique of similar concepts suggests a problem for further automatic classification of indoor-type concepts.

Table 7: Five most positive and negative concept pairs among the children of *outdoors* in CDVP-206.

positive pairs	negative pairs
rural setting – desert	water body – road
beach – water body	land – building
building – bridge	rural setting – road
mountain – snow	land – road
vegetation – snow	land – water body

For the *outdoors* concepts, the visually distinct concepts are *bridge*, *statue/monument*, *beach*, *waterfall*, and *desert*. There are also two cliques of similar concepts, both of which contain the common concepts *building*, *sky*, *vegetation*, *water body*, and *road*. In addition, the first clique contains the concept *cityscape/urban setting* and the second concepts *land* and *rural setting*. There are also two additional similar pairs (*sky – cloud* and *sky – mountain*). As mentioned earlier, in this case the visual similarity provides only one viewpoint on the concept relations as the *outdoors* concepts do also co-occur in the same objects, with the concept pair having the smallest co-occurrence distance being *vegetation – sky*. Therefore, it is justified to analyze these concepts as in Section 4.2 where the co-occurrence distance was taken into account using Eqs. (7) and (8). Accordingly, Table 7 lists the five concept pairs that have the overall smallest values of d^+ and d^- among the children concepts of *outdoors*.

5. CONCLUSIONS

In this paper, we presented an approach to multimedia ontology analysis which we feel has potential in building concept detectors for large-scale ontologies, as well as in ontology design and diagnostics. Multimedia ontologies are more than lists of independent concepts, and the inter-concept relationships provide important cues that should be employed in semantic multimedia analysis.

In this study, we used non-localized or image-level annotations for concepts. This inevitably leads to overlapping positive sets in the training data and thus makes it problematic to distinguish between co-occurrence and visual similarity properties. We were still able to obtain meaningful results in our experiments, but using localized annotations and region-based indexing could lead to more accurate concept models. The used method for measuring semantic similarity using test subjects does not scale well to large ontologies, and further studies should consider the use of semantic concept networks instead. The concept taxonomy provides a natural structure to large ontologies, and especially the parent–child and sibling relations should be utilized comprehensively.

A yet another viewpoint for ontology analysis is to consider the reliability of the concept models, as concepts that can be more accurately modeled should generally be favored. This was ignored in this study and can provide a natural direction for further analysis.

6. ACKNOWLEDGMENTS

This work was supported by The Irish Research Council for Science, Engineering and Technology, and by Science Foundation Ireland, under grant number 03/IN.3/I361.

7. REFERENCES

[1] M. G. Christel and A. G. Hauptmann. The use and utility of high-level semantic features in video retrieval. In *Proceedings of 4th International Conference on Image and Video Retrieval (CIVR 2005)*, pages 134–144, Singapore, July 2005.

[2] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith. Visual event detection using multi-dimensional concept dynamics. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME 2006)*, Toronto, Canada, July 2006.

[3] G. Gaughan. *Novelty Detection in Video Retrieval: Finding New News in TV News Stories*. PhD thesis, Dublin City University, 2006.

[4] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 42(1):177–196, 2001.

[5] G. Iyengar, H. J. Nock, and C. Neti. Discriminative model fusion for semantic concept detection and annotation in video. In *Proceedings of ACM Multimedia*, Berkeley, CA, November 2003.

[6] W. Jiang, S.-F. Chang, and A. C. Loui. Active context-based concept fusion with partial user labels. In *Proceedings of IEEE International Conference on Image Processing (ICIP 06)*, Atlanta, GA, USA, 2006.

[7] J. R. Kender and M. R. Naphade. Visual concepts for news story tracking: Analyzing and exploiting the NIST TRECVID video annotation experiment. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, June 2005.

[8] T. Kohonen. *Self-Organizing Maps*, 3rd edition. 2001.

[9] M. Koskela, A. F. Smeaton, and J. Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. *IEEE Transactions on Multimedia*, 2007. To appear.

[10] B. Li, K. Goh, and E. Y. Chang. Confidence-based dynamic ensemble for image annotation and semantics discovery. In *Proceedings of ACM Multimedia*, pages 195–206, Berkeley, CA, November 2003.

[11] 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*, 13(3):86–91, July-September 2006.

[12] 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. TR, IBM, 2005.

[13] M. R. Naphade, I. Kozintsev, and T. Huang. A factor graph framework for semantic video indexing. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(1):40–52, January 2002.

[14] S. Paek and S.-F. Chang. Experiments in constructing belief networks for image classification systems. In *Proceedings of International Conf. on Image Processing* , Vancouver, Canada, September 2000.

[15] M. Sjöberg, H. Muurinen, J. Laaksonen, and M. Koskela. PicSOM experiments in TRECVID 2006. In *Proceedings of the TRECVID 2006 Workshop*, Gaithersburg, MD, USA, November 2006.

[16] A. F. Smeaton. Large scale evaluations of multimedia information retrieval: The TRECVid experience. In *Proc. 4th International Conference on Image and Video Retrieval*, pages 11–17, Singapore, July 2005.

[17] J. R. Smith, M. Naphade, and A. P. Natsev. Multimedia semantic indexing using model vectors. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2003)*, volume 2, pages 445–448, Baltimore, MD, USA, July 2003.

[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 Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1678–1689, October 2006.

[19] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of ACM Multimedia 2006*, Santa Barbara, CA, October 2006.

[20] Y. Wu, B. L. Tseng, and J. R. Smith. Ontology-based multi-classification learning for video concept detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2004)*, pages 1003–1006, Taipei, Taiwan, June 2004.

[21] L. Xie and S.-F. Chang. Pattern mining in visual concept streams. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME 2006)*, Toronto, Canada, July 2006.

[22] R. Yan, M.-Y. Chen, and A. Hauptmann. Mining relationship between video concepts using probabilistic graphical models. In *Proc. Int'l Conf. on Multimedia & Expo*, Toronto, Canada, July 2006.