Automatic Multimedia Knowledge Discovery, Summarization and Evaluation

Ana B. Benitez *

Department of Electrical Engineering
Columbia University
1312 Mudd, #F6; 500 W. 120th St., MC 4712
New York, NY 10027, USA
Phone: 1-212-854-7473 / Fax: 1-212-932-9421
ana@ee.columbia.edu

Shih-Fu Chang

Department of Electrical Engineering
Columbia University
1312 Mudd; 500 W. 120th St., MC 4712
New York, NY 10027, USA
Phone: 1-212-854-6894 / Fax: 1-212-932-9421
sfchang@ee.columbia.edu

ABSTRACT

This paper presents novel methods for automatically discovering, summarizing and evaluating multimedia knowledge from annotated images in the form of images clusters, word senses and relationships among them, among others. These are essential for applications to intelligently, efficiently and coherently deal with multimedia. The proposed methods include automatic techniques (1) for constructing perceptual knowledge by clustering the images based on visual and text feature descriptors, and discovering similarity and statistical relationships between the clusters; (2) for constructing semantic knowledge by disambiguating the senses of words in the annotations using WordNet and the images clusters, and finding semantic relations between the senses in WordNet; (3) for reducing the size of multimedia knowledge by clustering similar concepts together; and (4) for evaluating the quality of multimedia knowledge using information and graph theory notions. Experiments show the potential of integrating the analysis of images and annotations for improving the performance of the image clustering and the word-sense disambiguation, the importance of good concept distance measures for knowledge summarization, and the usefulness of automatic measures for evaluating knowledge quality.

1. INTRODUCTION

This paper focuses on the discovery, summarization and evaluation of multimedia knowledge from annotated images such as image clusters, word senses and relationships among them. Existing techniques process each media independently or are domain specific so they do not generalize to arbitrary multimedia or knowledge. Knowledge is usually defined as facts about the world and represented as concepts and relationships among the concepts, i.e., semantic networks. Facts can refer to perceptions (color patterns) or semantics (meaning of word "car") in the world so two different kinds of knowledge can be distinguished, perceptual and semantic knowledge, respectively. Concepts are abstractions of objects, situations, events or patterns in the world (e.g., color pattern and car); relationships represent interactions among concepts (e.g., color cluster one visually Similar to color cluster two, and concept Sedan Specializes concept Car). In multimedia knowledge, concepts and relationships are exemplified by multimedia such as images and words.

The important proliferation of multimedia such as annotated images requires tools for discovering useful knowledge from multimedia to enable innovative and intelligent organization, filtering and retrieval of multimedia. Perceptual knowledge (e.g., image clusters and relationships between them) is essential for multimedia applications because it can be extracted automatically and is not inherently limited to a set of words as textual annotations. On the other hand, semantic knowledge (e.g., word senses and relationships between them) is the most powerful for multimedia applications because human communication often happens at this level. However, current approaches for extracting semantic knowledge are, at best, semi-automatic and very time consuming. Furthermore, it is often necessary to summarize multimedia knowledge in order to reduce the size of the knowledge. Hence, ways to quantify the consistency, completeness and conciseness of the multimedia knowledge, among others, are essential to evaluate and compare different techniques for knowledge discovery and summarization.

This paper presents new methods for discovering perceptual and semantic knowledge, and techniques for summarizing and evaluating multimedia knowledge from annotated images [3,4,5]. In contrast to prior work, the proposed knowledge discovery techniques integrate both the processing of images and

annotations. They also include novel ways for discovering perceptual relationships and semantic concepts. Furthermore, the proposed multimedia knowledge summarization and evaluation techniques are automatic and generic applicable to any multimedia knowledge that can be expressed as a set of concepts (e.g., image clusters and word senses), relationships among concepts (e.g., Visually Similar and Specializes relation), and examples of concepts (i.e., images and text examples of concepts). These methods are developed and used within the IMKA (Intelligent Multimedia Knowledge Application) system [6], which aims at extracting useful knowledge from multimedia and at implementing intelligent applications that use that knowledge. The IMKA system uses the MediaNet framework to represent multimedia knowledge [7], which is presented in section 3.

In the IMKA system, the process of extracting perceptual knowledge from a collection of annotated images [4] consists of discovering perceptual concepts by clustering the images based on visual and text feature descriptors. Perceptual relationships among the clusters are found based on descriptor similarity and statistical dependencies between the clusters. The semantic knowledge extraction process [5] discovers semantic concepts by disambiguating the sense of words in the annotations using WordNet [23] and the image clusters. Semantic relationships among the detected senses are found based on WordNet. The summarization of multimedia knowledge aims at reducing the size of the knowledge (in terms of number of concepts and relationships) by grouping similar concepts together. The IMKA system summarizes multimedia knowledge by calculating the distances between concepts using a novel concept distance measure and by grouping similar concepts together [3]. This paper also proposes automatic techniques for measuring the consistency, the completeness and the conciseness of multimedia knowledge based on information theory and graph notions such as entropy and graph density [3].

The paper is organized as follows. Section 2 reviews some prior work on multimedia knowledge discovery, summarization and evaluation. Section 3 defines and exemplifies the notion of multimedia knowledge by introducing the multimedia knowledge representation framework MediaNet. Sections 4 and 5 present the techniques for discovering perceptual and semantic knowledge from annotated image collections, respectively. Sections 6 and 7 describe the proposed methods for multimedia knowledge summarization and evaluation, respectively. Section 8 presents the experiment setup and results in

evaluating the proposed techniques. Finally, section 9 concludes with a summary and a discussion of future work.

2. PRIOR WORK

Relevant prior work on perceptual knowledge discovery includes visual thesauri construction [20,26], and joint visual-text clustering and retrieval of images [1,14]. The Texture Thesaurus [20] is a perceptual thesaurus limited to texture clusters of regions in satellite images constructed using neural network and vector quantization techniques. The Visual Thesaurus [26] adapts typical concepts and relationships from text thesauri to the visual domain. An approach for building the Visual Thesaurus involves grouping regions within and across images using visual descriptors. Then, relationships among groupings are learned through user interaction, so it is not a fully automatic system. Barnard et al. [1] clusters images by hierarchically modeling their distributions of words and visual feature descriptors; however, the clustering structures are limited to hierarchies. Grosky et al. [14] uses Latent Semantic Indexing and word weighting schemes to retrieve images using concatenated vectors of visual feature descriptors and category label bits. Limited experiments (50 images, 15 categories) have shown some performance improvement in image retrieval. In addition, Grosky et al. [14] does not try to discover relevant concepts or relationships from the images and their category labels.

Relevant work on semantic knowledge discovery includes word-sense disambiguation techniques for text documents [30,33]. Words in English may have more than one sense or meaning, for example "plant, industrial plant" and "plant, living organism" for the word “plant”. Word-sense disambiguation (WSD) is the process of finding the correct sense of a word within a document, which is a long-standing problem in Natural Language Processing. The reason for this is that although most English words have only one sense (80%), most words used in documents have more than one sense (80%) [30]. The two principles governing most word-sense disambiguation techniques are (1) that nearby words are semantically close or related and (2) that the sense of a word is often the same within a document [33]. In the literature, there are unsupervised [33] and supervised [30] approaches that often use WordNet as the electronic word-sense dictionary. WordNet [23] organizes English words into sets of synonyms (e.g., "rock, stone") and

connects them with semantic relations (e.g., Specializes, Contains and Entails). There are also image indexing approaches that disambiguate the senses of words in image annotations [1,28]. However, none of these approaches combine text and image features during the word-sense disambiguation.

Previous work on multimedia knowledge summarization has been limited to efforts in concept network reduction such as EZWordNet [22] and VISAR [9]. EZ.WordNet.1-2 [22] are coarser versions of WordNet generated by collapsing similar word senses and by dropping rare word senses. This process is governed by five rules manually designed by researchers for WordNet so they are not applicable to other kinds of knowledge such as perceptual knowledge. VISAR [9] is a hypertext system for the retrieval of textual captions. One of the functionalities of the VISAR system is the representation of the retrieved citations as networks of key concepts and relationships. Several reduction operators are used in this process (e.g., replace two concepts with a common ancestor) but the reduction operators are again manually defined and lacking generality.

Finally, prior work relevant on multimedia knowledge evaluation includes manual evaluation of semantic ontologies [13] and automatic but application-oriented evaluation of multimedia knowledge [2]. Typical criteria used by experts in the evaluation of semantic ontologies are consistency, completeness, conciseness, sensitiveness and expandability [13]. Barnard et al. [2] evaluates hierarchical image clusters using an automatic image and region annotation application. The performance of the image annotation is measured by comparing the words predicted by various models with words actually present in the data. The performance of the region annotation is measured both automatically based on the annotation performance and by manual inspection.

3. MEDIANET

MediaNet is a unified knowledge representation framework that uses multimedia for representing semantic and perceptual facts about the world (e.g., a texture pattern Visually Similar to another, and concept Sedan Specializes concept Car). The main components of MediaNet include concepts, relations among concepts, and media exemplifying concepts and relationships (see example in Figure 1). MediaNet extends and differs from related work such as the Multimedia Thesaurus [32] in two ways: (1) in

combining perceptual and semantic concepts in the same network, and (2) in supporting perceptual and semantic relationships that can be exemplified by media.

In MediaNet, concepts can represent either semantically meaningful objects (e.g., car) or perceptual patterns in the world (e.g., texture pattern). Concepts are defined and/or exemplified by multimedia such as images, video, audio, text, and audio-visual feature descriptors. MediaNet models the traditional semantic relations such as Specializes and Contains but also adds additional functionality by modeling perceptual relations based on feature descriptor similarity and constraints (e.g., condition on the distance of color histograms). Feature descriptors can also be associated to multimedia (e.g., color histogram for images and tf*idf for textual annotations). An example of multimedia knowledge represented using MediaNet is shown in Figure 1. This example illustrates the concepts Human and Hominid represented by text, an image region, audio and a shape feature descriptor. The two concepts are related by a semantic relationship (i.e., Specializes) and a perceptual relationship (i.e., Similar Shape); the latter is exemplified by a condition on a shape descriptor similarity.

4. PERCEPTUAL KNOWLEDGE DISCOVERY

Our proposed approach for extracting perceptual knowledge from an collection of annotated images consists of three steps, as shown in Figure 2. First, visual feature descriptors and text feature descriptors are extracted from the images and textual annotations. Then, perceptual concepts are formed by clustering the images based on their visual and text feature descriptors. Finally, perceptual relationships are discovered based on descriptor similarity and statistical dependencies between the clusters. This section discusses each step in detail. Examples of images and annotations are shown in Figure 3.(a).

The methods for image and text processing, and for discovering perceptual concepts described below are not new. Instead, our focus is on the discovery of relations between the extracted perceptual concepts. We also extensively test and report on the effectiveness of different clustering methods and feature combinations for perceptual concept extraction in section 8.

4.1 Basic Image and Text Processing

During this step, visual and text feature descriptors are extracted from the images and the textual annotations, respectively. Each media is processed independently.

The images are first segmented into regions with homogeneous visual features. Image segmentation consists of grouping visually similar pixels in an image into regions. There are many proposed methods for segmenting images [29]. The IMKA system uses Columbia University's automatic region segmentation method, which fuses color and edge pixel information [34]. This method has been proven to provide excellent segmentation results. After segmenting the images, descriptors are extracted from the images and the regions for representing visual features such as color, texture and shape. The IMKA system uses color histogram, Tamura texture, and edge direction histogram globally for images [19,29]; and mean LUV color, aspect ratio, number of pixels, and position locally for segmented regions [34]. This feature descriptor set covers the important visual features; moreover, each descriptor has been independently shown to be effective in retrieving visually similar images or videos in visual databases [19,29,34].

In the basic text processing, the words in the annotations are tagged with their part-of-speech information (e.g., noun and verb). WordNet is used to stem words down to their base form (e.g., "burned" is reduced to "burn") and to correct some POS tagging errors (e.g., "dear" in Figure 4 can not be a verb based on WordNet). Then, stopwords, (i.e., frequent words with little information such as "be"), non-content words (i.e., words that are not nouns, verbs, adjectives or adverbs such as "besides"), and infrequent words are discarded because of their small relevance. The remaining words for each image are represented as a feature vector using word-weighting schemes [12], which assign weights to words reflecting their discriminating power in a collection. The IMKA system uses tf*idf, term frequency weighted by inverse document frequency; and log tf*entropy, logarithmic term frequency weighted by Shannon entropy of the terms over the documents. The latter has been proven to outperform the former in Information Retrieval [12].

4.2 Perceptual Concept Extraction

The second step is to find perceptual concepts by clustering the images based on visual and text feature descriptors. Each cluster is considered a perceptual concept. Clustering is the process of discovering natural patterns in data by grouping similar data items [16]. The IMKA system uses a diverse set of clustering algorithms: the k-means algorithm, the Ward algorithm, the k-Nearest Neighbors algorithm (KNN), the Self-Organizing Map algorithm (SOM), and the Linear Vector Quantization algorithm (LVQ). The rationale for selecting each algorithm follows. The k-means algorithm is one of the simplest and fastest clustering algorithms. The Ward algorithm has been shown to outperform other hierarchical clustering methods. The k-nearest neighbor does not require Euclidean metric and can avoid the globular biases of other clustering algorithms. SOM and LVQ are neural network clustering algorithms that are capable of learning feature descriptor weights. Besides, LVQ allows treating annotations as labels driving the clustering; this is one way to integrate text and images in the clustering.

The IMKA system can cluster images based on any visual and text feature descriptor. Regarding local visual feature descriptors, the system can cluster the images based on the feature descriptors of all the regions, the largest regions, or the center region. The system can also generate clusters based on any combination of text and visual feature descriptors by concatenating the corresponding feature vectors. The mean and the variance of the bins of concatenated feature vectors are normalized to zero and one, respectively. Concatenation and normalization is the second way of integrating visual and text descriptors in the clustering. The dimension of feature vectors can be reduced using Latent Semantic Indexing (LSI) [10], which has the effect of uncorrelating feature vector bins.

4.3 Perceptual Relationship Extraction

SOM/LVQ and Ward clustering algorithms already provide Similar and Specializes relationships among the clusters, respectively. Additional relations between clusters are discovered by analyzing the descriptor similarity and the statistical dependencies between the clusters. Each cluster is said to have Similar relationships with its k nearest cluster neighbors. The distance between two clusters is calculated as the distance between the centroids. The number of neighbors could be set from 2 to 4 because that is the

cluster neighbor range for SOM and LVQ clusters. In the IMKA system, we propose new methods for extracting Equivalent, Specializes, Co-occurs, and Overlaps relationships between clusters based on cluster statistics as summarized in Table 1. For example, if two clusters use the same feature descriptors and their conditional probabilities are one or very close to one, they are considered equivalent. Such statistical relationships will prove to be useful in summarizing multimedia knowledge.

5. SEMANTIC KNOWLEDGE DISCOVERY

The proposed approach for extracting semantic knowledge from annotated images, which have already been clustered as described in section 4, consists of three steps, as shown in Figure 4. First, words are tagged with their part-of-speech information, and annotations chunked into word phrases. Then, semantic concepts are extracted by disambiguating the senses of the words with a novel methods that uses both WordNet and the image clusters. Finally, semantic relations and additional concepts relating the detected senses are found in WordNet. This section discusses each step in detail.

5.1 Basic Text Processing

During this step, the words are stemmed and tagged with their part-of-speech and stopwords and non-content words are discarded, as described in section 4.1. Then, the textual annotations are chunked into noun and verb phrases (e.g., the sentence "I love New York" has two noun phrases "I" and "New York", and one verb phrase "love") [21]. In addition, single words are grouped into compound words (e.g., "New York" in Figure 4 is one compound word with one meaning). For the recognition of compound words, the IMKA system detects noun and verb phrases containing only nouns or verbs, respectively. Then, different combinations of the words, starting from the ones with more words and preserving word ordering, are searched in WordNet. If a word search is successful, the words are removed from the following word combinations until all the combinations have been searched. As an example, the noun phrase "New York" in Figure 4 will cause the following word searches: "New York ", "New" and "York"; the first search is successful so no additional searches are executed.

5.2 Semantic Concept Extraction

The second step in the semantic knowledge extraction process is to disambiguate the senses of the remaining words using WordNet and the images clusters. Each detected sense is a semantic concept.

The intuition behind the proposed approach is that the images that belong to the same image cluster are likely to share some common semantics although very general (e.g., images of animals and flowers in vegetation have similar global color and share semantics such as "nature" and "vegetation"). The proposed technique also follows the two principles for word-sense disambiguation: consistent sense for a word and semantically relatedness of nearby words in the annotations of the clusters. The word-sense disambiguation procedure consists of two basic steps (see Figure 4). First, the senses of words annotating the images in a cluster are ranked based on all the annotations of the cluster. An image can belong to several clusters; the second step is to add the ranks of the senses for the same word and image for the different clusters to which the image belongs. The more relevant the concept, the higher the rank. Therefore, the detected sense for a word is the highest ranked sense.

The IMKA system ranks the different senses of a word for an image in a cluster by matching the definitions of the possible senses listed by WordNet to the annotations of all the images in the cluster using word weighting schemes, i.e., tf*idf and log tf*entropy. In this process, the definition of each sense is considered to be a document in a collection; and the query keywords, the annotations of all the images in the cluster. The definition of a sense (e.g., sense "rock, stone" in Figure 4) is constructed by concatenating the synonym set (e.g., "rock, stone"), the meaning (e.g., "mineral aggregate making up the Earth's crust") and the usage examples of the sense (e.g., "the work was brown") together with the definition of directly or indirectly related senses (e.g., sense "lava", which Specializes sense "rock, stone") provided by WordNet. Different weights are assigned to the synonym set, the meaning, and the usage examples of a sense, and the definitions of related senses. As an example, higher weights should be assigned to the synonym set (e.g., 1.0 for "rock" and "stone") compared to the usage examples (e.g., 0.8 for "rock" and "brown"), and to the definition of a sense compared to the definition of related senses (e.g., 1.0 for definition of sense "rock, stone", 0.8 for definition of sense "lava").

5.3 Semantic Relationship Extraction

The third step is to discover semantic relationships among the semantic concepts. During this process, the IMKA system finds the paths connecting every pair of detected senses in WordNet, either directly or through intermediate senses. All the semantic relationships and the intermediate senses on these paths are added to the extracted semantic knowledge. Therefore, the constructed knowledge will not be restricted to the detected senses. For example, in Figure 4, senses "mountain" and "rock, stone" are connected through the concept "object", their common ancestor, and Specializes relationships between them. Table 2 lists the semantic relations in WordNet together with definitions and examples.

6. MULTIMEDIA KNOWLEDGE SUMMARIZATION

Our proposed approach for summarizing multimedia knowledge consists of reducing the number of concepts and relationships of the knowledge in three steps, as shown in Figure 5. First, the distances among the concepts in the multimedia knowledge are calculated. Then, similar concepts are clustered together based on the concept distances. Finally, the knowledge summary is generated based on the concept clusters. This section discusses each step in detail. In a preliminary stage, the least frequent concepts can be discarded and weights assigned to concepts for personalized knowledge summarization.

6.1 Concept Distances

The first step in summarizing multimedia knowledge is to calculate the distances among concepts using a novel technique based on concept statistics and knowledge topology.

There are many proposed methods for calculating concept distance or similarity among concepts in semantic concept networks such as WordNet [18,31]. Some methods rely uniquely on hierarchical specialization/generalization relationships among concepts [18] whereas others take into account all the semantic relations [31]. There are methods that use exclusively the concept network topology [31] while others combine both concept network topology and concept statistics [18]. Recent work [8] evaluated five concept distance measures using WordNet in a real-word spelling error correction system in which [18] was found to outperform the rest. The concept distance measure presented in [18] only considers the

specialization/generalization hierarchy among concepts. The distance of a relationship in the concept hierarchy is the information content, as defined in information theory, of the child concept given the parent concept, i.e., of encountering an instance of the child concept given an instance of the parent concept. We propose a new concept distance measure that generalizes the measure proposed in [18] to an arbitrary concept network with different relations among concepts.

The IMKA system uses a novel concept distance measure based on concept statistics and network topology that is not limited to specialization/generalization relationships. Instead, it supports relations such as Contains, Entails and Overlaps. First, the distance of each relationship in the concept network is calculated based on concept statistics. The distance between any two concepts is then obtained as the distance of the shortest distance path (which may consist of several hops) between the two concepts in the network. The distance of a relationship r connecting concepts c and c' is calculated as follows:

$$dist(c, c', r) = p(c)\, IC(c', r \mid c) + p(c')\, IC(c, r \mid c') = -p(c)\, \log(p(c', r \mid c)) - p(c')\, \log(p(c, r \mid c')) \quad (1)$$

where IC(x) is the information content of x, p(c) is the probability of encountering an instance of concept c, and p(c,r|c') is the probability of encountering an instance of concept c through relationship r given an instance of concept c'. We assume binary relationships, i.e., relationships that only have two vertices, a source and a target. The intuition behind Equation (1) is the following: the distance of a relationship between two concepts increases with the concept probabilities but decreases with the conditional probabilities for that relationship. The proposed concept distance satisfies the properties of a distance function. The concept distance proposed in [18] corresponds to the first information content term in Equation (1), IC(c',r|c), when concept c is the parent node of concept c' in the specialization/ generalization concept hierarchy.

There are different approaches toward calculating statistics of concepts such as WordNet's senses in a text corpus. The approach [27], often used in conjunction with measure [18], obtains the frequency of a concept by adding the occurrences of specialized concepts to the strict occurrence of the concept. In a related way, the IMKA system first finds strict concept frequencies for each concept, freq'(c), by

summing up the number of times the concept is instantiated in each image. As an example, the concept House would have a frequency of two for an image whose annotation contains the sense "house" twice. The inferred concept frequencies are then propagated in the concept network, e.g., an instance of concept Dog is also an instance of concept Animal. Considering a relationship r between concepts c and c', a different fraction of the frequency of concept c will be included in the frequency of concept c' based on relationship r, and vice versa. In formulistic terms, the inferred frequency of concept c is calculated using the following system:

$$\text{freq}(c) = \text{freq}_o(c) + \sum_{r \in \text{relationsWithSrc}(c)} w_{t \to s}(r)\ \text{freq}(\text{tgt}(r)) + \sum_{r \in \text{relationsWithTgt}(c)} w_{s \to t}(r)\ \text{freq}(\text{src}(r)) \quad (2)$$

where relationsWithSrc(c) and relationsWithTgt(c) are the sets of relationships that have c as source and target, respectively; src(r) and tgt(r) are the source and target nodes of relationship r; $w_{s \to t}(r)$ and $w_{t \to s}(r)$ are the relation propagation weights for relationship r from source to target and from target to source, respectively; concepts(K) is the set of concepts in multimedia knowledge K; and $\text{freq}_o(c)$ is proportional to freq'(c). This system specified by Equation (2) represents the relationship of the final inferred concept frequencies, which is solved using a simple, iterative method similar to the one that the Google search engine uses to calculate PageRanks for rating web pages [25]. The relations in the multimedia knowledge affect the inferred concept frequencies and, therefore, the concept distances through the relation propagation weights, $w_{s \to t}(r)$ and $w_{t \to s}(r)$, which are learned or specified by experts (see Table 3 for examples). Finally, the probability of concept c and the conditional probability of concept c through relation r given concept c' is calculated as follows:

$$p(c) = \frac{\text{freq}(c)}{\sum_{c \in \text{concepts}(K)} \text{freq}'(c)} \quad (3)$$

$$p(c, r \mid c') = \frac{w_{s \to t}(r)\ \text{freq}(\text{src}(r)) + w_{t \to s}(r)\ \text{freq}(\text{tgt}(r)) - w_{s \to t}(r)\ w_{t \to s}(r)\ \text{freq}(\text{tgt}(r) \cap \text{src}(r))}{\text{freq}(c')}. \quad (4)$$

6.2 Concept Clustering

The second step in the multimedia knowledge summarization process is to cluster the concepts based on the distances among them. The concepts are clustered into as many clusters as the desired number of concepts in the multimedia knowledge summary.

The IMKA system uses a modified KNN clustering algorithm to group concepts into a given number of clusters. The KNN clustering algorithm was selected because of the continuity and the non-globular shape of the resulting clusters. Moreover, the KNN clustering algorithm does not require a specific distance function. Whereas the KNN clustering algorithm merges the clusters of two data items with at least k_t shared neighbors within k neighbors [17], the modified KNN clustering algorithm merges the clusters of the two data items with the largest number of shared neighbors until a given number of clusters is reached. The input to the clustering algorithm is the k nearest concepts of each concept and the desired number of clusters. Different weighting schemes of shared neighbors [17] are supported in the IMKA system as well as the reduction of the number of shared neighbors based on data item weights. If a data item is more important (i.e., higher weighted), then, the data item will have fewer shared neighbors and be clustered with fewer other data items; it will tend to remain alone in a cluster. A centroid for each cluster is obtained as the data item in the cluster with maximum accumulated weighted shared neighbors to the rest of the data items in the cluster.

6.3 Knowledge Reduction

The knowledge summary is generated using the concept clusters. Each concept cluster becomes a super-concept, i.e., group of concepts, in the knowledge summary inheriting the text and image examples of the cluster members. If all the concepts in a concept cluster are semantic concepts, then the type of the super-concept is set to semantic; otherwise, it is set to perceptual. The relationships among super-concepts are the relationships between concepts in the corresponding concept clusters. For visualizing summarized knowledge, among others, decisions should be made for labeling and simplifying the super-concepts and the their relationships in the knowledge summary. For example, we can use the text examples corresponding to the centroid or the most probable concept in the concept cluster as labels of the entire

super-concept. Regarding the relationships between two super-concepts, they can be represented using only the highest distance relationship, among other strategies. In the sub-network of the knowledge summary of 32 super-concepts shown in Figure 7, the concept labels were manually chosen for illustration purposes. In addition, all the relationships of the same type between any two super-concepts were represented with only one arc in the figure.

7. MULTIMEDIA KNOWLEDGE EVALUATION

In this section, we propose automatic ways for measuring the consistency, completeness and conciseness of multimedia knowledge. These are three of the five criteria identified in [13] in the expert evaluation and assessment of semantic ontologies. The other two criteria, expandability and sensitiveness (i.e., how new definitions or changes in existing definitions affect the ontology properties, respectively), are not considered because they usually depend on the ontology management. The goal of the proposed measures was the automatic, application-independent techniques for evaluating the goodness of multimedia knowledge.

7.1 Consistency

Consistency refers to whether it is possible to obtain contradictory conclusions from valid input definitions. In terms of concept distances, the consistency of multimedia knowledge can be calculated based on the differences on the distances between two concepts through different paths. The larger the distance spread among concepts, the more inconsistent or contradictory the different paths connecting the concepts.

We propose to measure the inconsistency of multimedia knowledge by calculating the spread of the total distances of the k shortest distance paths between every pair of concepts with respect to the shortest distance path, as follows:

$$\mathrm{ICST(K)} = \log\left(\frac{\sum_{c,c' \in \mathrm{concepts(K)}} \sum_{i=1}^{i=k} (d(c,c',i) - d(c,c',1))^2}{|\mathrm{concepts(K)}|^2 \; k} + 1\right) \tag{5}$$

where |concepts(K)| is the number of concepts in multimedia knowledge K, k is the number of shortest distance paths considered between any two concepts, and d(c,c',i) is the distance between concepts c and c' through path i. The k shortest distance paths are ordered from shortest to longest distance starting at i=1 up to i=k. The lower ICST(K), the more consistent the multimedia knowledge.

7.2 Completeness

Completeness refers to the completeness of both the ontology and the definitions in the ontology. The only way for measuring the completeness of multimedia knowledge would be the direct comparison with the target knowledge, which is not available. Instead, we can measure the information (randomness), the graph completeness, and the category correlation of the multimedia knowledge. The more informative (random), the more complete the graph, and the higher correlation with category labels; the more complete the knowledge.

We propose to measure the randomness of information of the multimedia knowledge by calculating the concept entropy, as follows:

$$\mathrm{CPT\_H(K)} = -\sum_{c \in \mathrm{concepts(K)}} p(c) \log(p(c)) \tag{6}$$

where p(c) is the probability of concept c obtained as described in section 6.1. We measure the graph completeness of the multimedia knowledge by adapting the formula of graph density (i.e., ratio of the number of relations in the graph by the maximum number of possible relationships) to weighted relationships, as follows:

$$\mathrm{CPT\_D(K)} = \frac{\sum_{r \in \mathrm{relations(K)}} [1 - d(r)/dmax]}{|\mathrm{concepts(K)}|(|\mathrm{concepts(K)}| - 1)} \tag{7}$$

where relations(K) is the set of relationships in multimedia knowledge K, d(r) is the distance of relationship r, and dmax is the maximum distance for a relationship. The higher CPT_H(K) and CPT_D(K), the more complete the multimedia knowledge.

If category labels are available for the images, we propose an entropy-based criterion to evaluate the completeness or correlation of the concepts with the category labels. If L {l1, … lm} and C = concepts(K) = {c1, … cn} are the sets of category labels and concepts, respectively, the correlation of multimedia knowledge K with respect to category labels L can be calculated as the harmonic mean of one minus the mean entropies of the categories within each concept (INH(C)), and of each category over the concepts (INH(L)), normalized from 0 to 1, as follows:

$$\mathrm{CPT\_CH(K, L)} = \frac{2\ \mathrm{INH(C)}\ \mathrm{INH(L)}}{\mathrm{INH(C)} + \mathrm{INH(L)}} \tag{8}$$

where

$$\mathrm{INH(C)} = 1 - \sum_{j=1}^{n} p(c_j) \sum_{i=1}^{m} p(l_i \mid c_j) \log(p(l_i \mid c_j)) \Big/ \log(m) \qquad \mathrm{INH(L)} = 1 - \sum_{i=1}^{m} p(l_i) \sum_{j=1}^{n} p(c_j \mid l_i) \log(p(c_j \mid l_i)) \Big/ \log(n) \tag{9}$$

The harmonic mean ensures that CPT_CH(K,L) is close to one only if both INH(C) and INH(L) are close to one. The closer PCT_CH(K,L) to one, the better the knowledge K fits the labels L. PCT_CH(K,L) is equal to one if and only if the number of the concepts and the categories are the same and the images in each concept are exactly the same as the images in a category.

7.3 Conciseness

Conciseness refers to whether all the information in the ontology is precise, necessary and useful. In terms of concepts distances, the redundancy of some multimedia knowledge can be calculated as the number of null eigen values in the concept distance matrix. The larger the number of null eigen values, the more redundant (i.e., less concise) the multimedia knowledge.

Our proposed way to evaluate the conciseness of multimedia knowledge is by comparing the number of

concepts and the rank of the concept distance matrix, as follows:

$$\mathrm{ICCS(K)} = \frac{|\mathrm{concepts(K)}| - \mathrm{rank(M)}}{|\mathrm{concepts(K)}|} \tag{10}$$

where M is the concept distance matrix, and rank(M) is the rank of the matrix M. The elements of the matrix M are the pair-wise distances between every pair of concepts. The lower ICCS(K), the more concise the multimedia knowledge.

8. EXPERIMENTS

Semantic and perceptual multimedia knowledge was extracted from a collection of 3,624 annotated images as described in sections 4 and 5, respectively. CPT_CH(K,L) and the word-sense disambiguation accuracy were used to evaluate the resulting image clusters and word senses, respectively. Multimedia knowledge extracted from a smaller collection of 271 images was summarized as described in section 6. ICST(K), CPT_H(K), CPT_D(K), and ICCS(K) were used to compare the multimedia knowledge at different steps in the proposed approaches with respect to several baseline approaches.

8.1 Experiment Setup

The test set was a diverse collection of 3,624 *nature* and *news* images from the Berkeley's CalPhotos collection (http://elib.cs.berkeley.edu/photos/) and the ClariNet news newsgroups (http://www.clari.net/), respectively. The images in CalPhotos were already labeled as *plants* (857), *animals* (818), *landscapes* (660) or *people* (371). The *news* images from ClariNet were categorized into *struggle* (236), *politics* (257), *disaster* (174), *crime* (84) and *other* (67) by researchers at Columbia University. The *nature* and *news* images had annotations in the form of keywords and sentences, respectively (see Figure 3.(a)).

During the perceptual knowledge extraction process, the images were scaled down to a maximum height and width of 100 pixels and segmented to at most 16 regions. Words that appeared less than 5 times in the images annotations were discarded for the extraction of the text feature descriptors, whose dimensionality was further reduced to 500 and 125 using LSI. Clustering was done using different algorithms -k-means,

SOM, LVQ and KNN-, different feature descriptors -color histogram, Tamura texture, edge direction histogram, the descriptors of the 16 regions, the largest region's descriptors, the center region's descriptors, the tf*idf descriptor, and the log tf*entropy descriptor; and different number of clusters - ranging from 9 to 529. The SOM and LVQ maps were made square. The labels used for LVQ clustering algorithms were the category labels listed above. During the semantic knowledge extraction process, in addition, the sense definitions were generated assigning different weights to the synonym set with respect to the meaning and usage examples of a sense, and to the definitions of directly and indirectly related senses. Lof tf * entropy was used to match sense definitions and cluster annotations using the cosine metric.

The perceptual and semantic knowledge constructed for 271 randomly chosen *nature* images by clustering the images based on color histogram, log tf * entropy, and an integrated color histogram / log tf * entropy feature vector into 16 clusters for each descriptor was summarized. The nouns in the *what* annotations of the images were the only ones used in the semantic knowledge extraction. The initial multimedia knowledge had 790 semantic concepts, 48 perceptual concepts, 842 Specializes relations, 414 Contains relations, and 9 Association relations. Summaries of different sizes were generated from the initial knowledge using the propagation relation weights shown in Table 3, among others.

8.2 Experiment Results

In this section, we present and discuss the experimental results performed for evaluating the proposed techniques for perceptual and semantic knowledge discovery, and for multimedia knowledge summarization.

8.2.1 Perceptual Knowledge Discovery

The criterion used to evaluate the image clusters generated during the perceptual knowledge extraction process was CPT_CH(K,L) using the primary categories {*nature, news*} and the secondary categories {*plant, animal, landscape, people, struggle, politics, disaster, crime, other*} as the labels L. Figure 6.(a) and Figure 6.(b) show the results obtained for the best clustering algorithm for each feature descriptor for

both category sets. The results for the local (region) visual feature descriptors were excluded from the figure because they were the worst due, likely, to the use of the Euclidean metric in building the clusters instead of specialized metrics [34]. Figure 6 also displays results in concatenating the 125-bin log tf*entropy descriptor and each visual feature descriptor with bin normalization but no LSI. The results with normalization and LSI were very similar that together with the small number of null eigen values for concatenated visual-text feature vectors (e.g., 3 bins for 166-bin color histogram + 125-bin log tf * entropy) shows the *high independence of visual and text feature descriptors*. The figure also includes results for randomly generated clusters for baseline comparison.

As can be seen in Figure 6, both *text and visual feature descriptors enable the discovery of useful knowledge* because their results are well above random behavior. As expected text feature descriptors are more powerful than visual feature descriptors and the log tf*entropy descriptor outperforms the tf*idf descriptor. Some concatenated visual-text feature descriptors slightly outperform the individual text feature descriptor for the primary categories but not for the secondary categories probably because the latter categories are less visually separable. This indicates that *both kinds of descriptors should be integrated* in the knowledge extraction process in providing different kinds of useful knowledge. Please, note that 500-bin log tf * entropy descriptor was not integrated with the visual feature descriptors although it provides the best overall results. Although not shown in the results, the trend of INH(C) and INH(L) is monotonically increasing and decreasing, respectively.

8.2.2 Semantic Knowledge Discovery

The criterion to evaluate the word-sense disambiguation process was the word-sense disambiguation accuracy, in other words, the percentage of words correctly disambiguated. The first author of this paper generated the ground truth for the annotations of 10% of randomly selected images in the collection; no training was needed. Table 4 shows the accuracy results for best image clusters (BI), worst image clusters (WI), cluster-per-image (TT, equivalent to WSD using only text), selecting most frequent senses (MF), and selecting random senses (RD). The results for the last three approaches are provided for baseline comparison. The accuracy results are separated for the *nature* and the *news* images, and for nouns, verbs,

adjectives, adverbs and all the content words.

As shown in Table 4, for both image sets, best image clusters consistently outperforms cluster-per-image and random senses. For *nature images, best image clusters and, often, worst image clusters provide better results* than most frequent senses. The results for the *news images* are quite different: *most frequent sense outperforms* even best image clusters except for adjectives and adverbs. Several factors can explain the result differences between *nature* and *news* images: (1) WordNet has a more comprehensive coverage of nature concepts because several animal and plant thesauri where used in its construction; (2) the textual annotations of *news* images are well-formed phrases so there are more words that can potentially confuse the word-sense disambiguation process; (3) *news* images are more diverse and, therefore, their clusters may not be as "meaningful"; and (4) the gap between concepts and the visual features for *news* images is larger. The proposed approach for word-sense disambiguation outperforms most frequent sense for images with short annotations (e.g., one word “plant”) that are clustered with more-extensively annotated and semantically related images (e.g., image annotated with “plant, flower, buttercup”); an example is shown in Figure 3.(b).

Although not shown in Table 4, the best word-sense disambiguation results were obtained for 9 to 25 clusters, which reinforces the fact that visual clusters are useful for word-sense disambiguation. The use of different visual feature descriptors or clustering algorithms had no obvious impact in the results. For generating the sense definitions, a reduction factor of 0.8 in the weights of the meaning and usage examples with respect to the synonym set, and of the definition of each relationship between the original sense and the related sense provided good results. The only exception being that log tf * entropy in some instances concatenated with visual feature descriptors (i.e., color histogram) consistently provided the best clusters for word-sense disambiguation.

8.2.3 Multimedia Knowledge Summarization

Table 5 show the values for ICST(K), CPT_H(K), CPT_D(K) and ICCS(K) obtained in the experiments evaluating the proposed techniques for summarization of multimedia knowledge, respectively. The first two rows of Table 5 show the results for the multimedia knowledge extracted from the image collection

using the proposed concept distance (dist, see Equation (1)) and the semantic distance [18] ($dist_{Jiang}$), and for a random version of this multimedia knowledge. The random multimedia knowledge was generated by randomly rearranging the images and concepts in the knowledge network. Table 5 also shows the results in summarizing the extracted multimedia knowledge into knowledge summaries of 2, 4, 8, 16, 32, 64, 128, 256 and 512 super-concepts using the proposed concept distance and the semantic distance [18]. Table 6 lists the most frequent words in the annotations, concepts before and after word-sense disambiguation, and concepts in the summary of 32 super-concepts. Figure 7 shows part of the knowledge summary of 32 super-concepts, the sub-network that includes the most frequent concepts.

As expected, Table 5 shows that the random multimedia knowledge has much higher entropy and lower concept redundancy, but also considerable larger distance spread. The graph density for the random multimedia knowledge remains constant because the number of relations does not change. In addition, summarizing multimedia knowledge increases the graph density and the distance spread, and decreases the concept entropy and the concept redundancy. The sharp entropy increase (and distance spread decrease) indicates the considerable difference between from the summaries of 16 and 32 super-concepts. In fact, the summary of 16 super-concepts is composed of very common and very rare super-concepts. On the contrary, the summary of 32 super-concepts has a more uniform distribution of concept occurrences. We can then conclude that the proposed *knowledge evaluation measures are useful* for distinguishing between extracted and randomized knowledge and estimating the quality of knowledge. For example, these knowledge evaluation measures can be used to decide the number of super-concepts in which to summarize some multimedia knowledge.

The use of different *concept distances seems to have a considerable impact* in the quality of the resulting summaries in spite of the large number of Specializes relations in the extracted multimedia knowledge. The results for summaries of just a few super-concepts are almost identical for either of the concept distance measures. However, the proposed concept distance results in considerably higher entropy and lower distance spread especially for summaries of 32 super-concepts and more. Both distance measures present similar trends in terms of random and summarized multimedia knowledge because of the large number of Specializes relationships in the multimedia knowledge in these specific experiments.

9. CONTRIBUTIONS AND CONCLUSIONS

This paper proposes novel techniques for automatically discovering, summarizing and evaluating multimedia knowledge including techniques for discovering perceptual and semantic knowledge for annotated image collections. In particular, it has proposed (1) new techniques for discovering perceptual relations among perceptual concepts; (2) a novel technique for disambiguating the words in image annotations that uses not only the annotations and WordNet but also the image features; (3) a new technique for calculating distances among concepts used by a modified KNN algorithm to cluster concepts for generating multimedia knowledge summaries; and (4) automatic ways of measuring the consistency, completeness and conciseness of multimedia knowledge.

Experiments have shown that both visual and text feature descriptors are uncorrelated but useful in extracting perceptual knowledge from annotated images; therefore, the integration of both kinds of descriptors has potential to improve performance compared to individual descriptors. The evaluation of the proposed word-sense disambiguation approach has shown that using perceptual knowledge in the form of image clusters can improve performance compared to most frequent senses and text word-sense disambiguation (above 5% for *nature* images). Additional experiments have shown the importance of good concept distance measures, as the proposed one, for clustering and summarizing knowledge, and the usefulness of the proposed automatic measures for measuring the quality of multimedia knowledge.

Future work aims at improving the efficiency of the implementation of these techniques in terms of processing time, memory usage and scalability by developing heuristic approximations of some proposed techniques (e.g., incrementally learning the Bayesian network). Applications that use the constructed multimedia knowledge for automatic image classification, browsing, retrieval and annotation will also be implemented and evaluated. Future work will also consist of proposing a complexity-constraint framework for personalizing the quality of the multimedia knowledge for specific user applications.

ACKNOWLEDGMENTS

This research is partly supported by a Kodak fellowship awarded to the first author of the paper.

REFERENCES

1. Barnard, K., P. Duygulu, and D. Forsyth, "Clustering Art", *CVPR-2001*, Hawaii, USA, Dec. 9-14, 2001.
2. Barnard, K., P. Duygulu, and D. Forsyth, N. de Freitas, D. Blei, and M.I. Jordan, "Matching Words and Pictures", *JMLR, Special Issue on Text and Images*, Vol. 3, pp. 1107-1135, 2003.
3. Benitez, A.B., and S.-F. Chang, "Multimedia Knowledge interrelation, Summarization and Evaluation", *MDM/KDD-2002*, Edmonton, Alberta, Canada, July 23-26, 2002.
4. Benitez, A.B., and S.-F. Chang, "Perceptual Knowledge Construction From Annotated Image Collections", *ICME-2002*, Lausanne, Switzerland, Aug 26-29, 2002.
5. Benitez, A.B., and S.-F. Chang, "Semantic Knowledge Construction From Annotated Image Collections", *ICME-2002*, Lausanne, Switzerland, Aug 26-29, 2002.
6. Benitez, A.B., S.-F. Chang, and J.R. Smith, "IMKA: A Multimedia Organization System Combining Perceptual and Semantic Knowledge", *ACM MM-2001*, Ottawa, CA, 2001.
7. Benitez, A.B., J.R. Smith, and S.-F. Chang, "MediaNet: A Multimedia Information Network for Knowledge Representation", *IS&T/SPIE-2000*, Vol. 4210, Boston, MA, Nov 6-8, 2000.
8. Budanitsky, A., and G. Hirst, "Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures", *NAACL-2001*, Pittsburgh, PA, June 2001.
9. Clitherow, P., D. Riecken, and M. Muller, "VISAR: A System for Inference and Navigation in Hypertext", *ACM Conference on Hypertext*, Pittsburgh, PA USA, Nov. 5-8, 1989.
10. Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Indexing", *JASIS*, Vol. 41, No. 6, pp. 391-407, 1990.
11. Duda, R.O., P.E. Hart, and D.G. Stork, "*Pattern Classification*", John Wiley & Sons, Second Edition, United States of America, 2001.
12. Dumais, S.T., "Improving the retrieval of information from external sources", *Behavior Research Methods, Instruments and Computers*, Vol. 23, No. 2, pp. 229-236, 1991.
13. Gomez-Perez, A., "Evaluation of Taxonomic Knowledge in Ontologies and Knowledge Bases", *KAW-1999*, Alberta, Canada, Oct. 16-21, 1999.
14. Grosky, W.I., and R. Zhao: "Negotiating the Semantic Gap: From Feature Maps to Semantic Landscapes", *SOFSEM-2001*.
15. Hastings, W.K., "Monte Carlo Sampling Methods Using Markov Chains and their Applications", Biometrika, Vol. 57, No. 1, pp. 97-109, 1970.
16. Jain, A.K., M.N. Murty, and P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, pp.264-323, Sept. 1999.
17. Jarvis, R.A., and E.A. Patrick, "Clustering Using a Similarity Measure Based on Shared Near Neighbors", *IEEE Transaction on Computers*, Vol. c-22, No. 11, Nov. 1973.
18. Jiang, J.J., and D.W. Conrath, "Semantic Similarity based on Corpus Statistics and Lexical Taxonomy", *International Conference on Research in Computational Linguistics*, Taiwan, 1997.
19. Kumar, R., and S.-F. Chang, "Image Retrieval with Sketches and Coherent Images", *ICME-2000*, New York, Aug. 2000.
20. Ma, W.Y., and B.S. Manjunath, "A Texture Thesaurus for Browsing Large Aerial Photographs", *JASIS*, Vol. 49, No. 7, pp. 633-648, May 1998.
21. McKelvie, D., C. Brew and H. Thompson, "Using SGML as a basis for data-intensive NLP", *Applied Natural Language Processing*, Washington, USA, April 1997.
22. Mihalcea, R., and D. Moldovan, "Automatic Generation of a Coarse Grained WordNet", *NAACL-2001*, Pittsburgh, PA, June 2001.
23. Miller, G.A., "WordNet: A Lexical Database for English", *Comm. of the ACM*, Vol. 38, No. 11, pp. 39-41, Nov. 1995.

24. Murphy, K., "The Bayes Net Toolbox for Matlab", *Computing Science and Statistics*, Vol. 33, 2001.
25. Page, L., S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bring Order to the Web", *Stanford Digital Library Working Paper*; available at http://dbpubs.stanford.edu:8090/pub/1999-66, Sept 2002.
26. Picard, R.W., "Toward a Visual Thesaurus", *Workshops in Computing (MIRO-1995)*, Glasgow, UK, Sep. 1995.
27. Richardson, R., and A.F. Smeaton, "Using WordNet in a Knowledge-Based Approach to Information Retrieval", Working paper, CA-0395, School of Computer Applications, Dublin City University, Ireland, 1995.
28. Rowe, N.C., "Precise and Efficient Retrieval of Captioned Images: The MARIE Project", *Library Trends*, Fall 1999.
29. Rui, Y., T. Huang, and S.-F. Chang, "Image Retrieval: Current Techniques, Open Issues, and Promising Directions", *Journal of Visual Communication and Image Representation*, 1999.
30. Stetina, J., S. Kurohashi, M. Nagao, "General Word Sense method based on a full sentential context", *COLING-ACL Workshop*, Montreal, CA, July 1998.
31. Sussna, M., "Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network", *CIKM-1993*, pp. 67-74, 1993.
32. Tansley, R., "The Multimedia Thesaurus: Adding A Semantic Layer to Multimedia Information", Ph.D. Thesis, Computer Science, University of Southampton, Southampton UK, August 2000.
33. Yarowsky, D., "Unsupervised Word-sense Disambiguation Rivaling Supervised Methods", *Association of Computational Linguistics*, 1995.
34. Zhong, D., and S.-F. Chang, "Video Object Model and Segmentation for Content-Based Video Indexing", *IEEE Int. Symposium on Circuits and Systems (ISCAS-1997)*, Hong Kong, 1997.

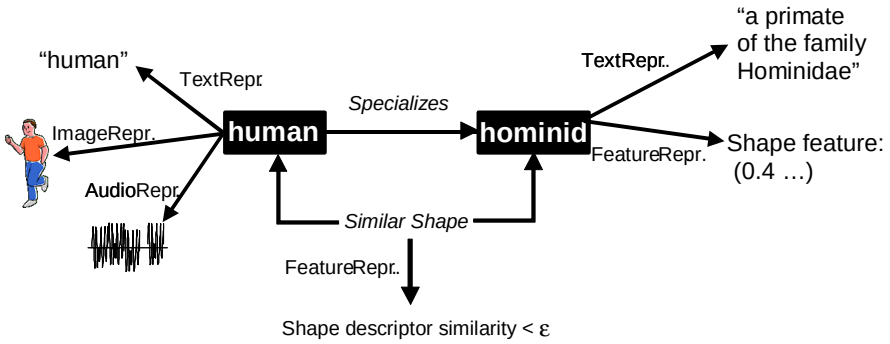


Figure 1: Example of multimedia knowledge that illustrates the concepts Human and Hominid (boxes) represented by text, an image region, audio and a shape feature descriptor. The two concepts are related by a semantic relationship, (i.e., Specializes) and a perceptual relationship (i.e., Similar Shape); the latter relationship is represented by a condition on a shape descriptor similarity.

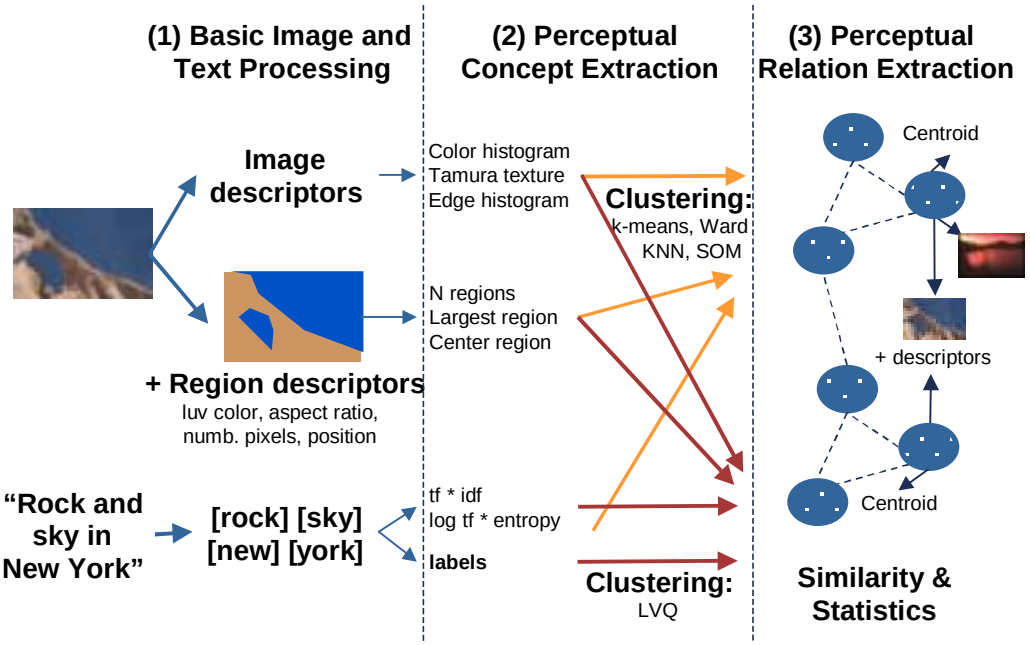
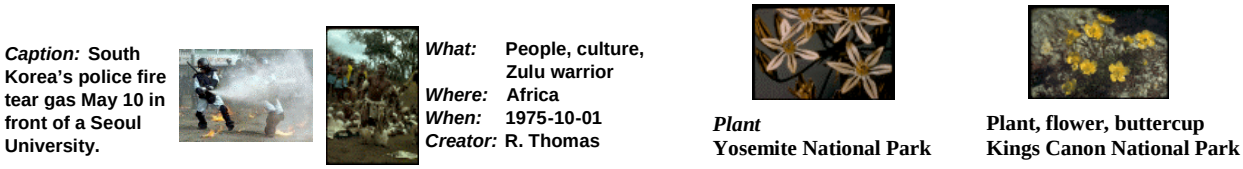


Figure 2: Perceptual knowledge extraction process: first, visual and text feature descriptors are extracted from the images and the textual annotations, respectively; then, perceptual concepts (dotted ellipses) are obtained by clustering the images based on the feature descriptors; and, finally, perceptual relationships (dash lines) among clusters are discovered based on cluster similarity and conditional probabilities.



(a) (b)

Figure 3: (a) Example of a *news* image (left) and a *nature* image (right) with their textual annotations. (b) Example where proposed method correctly disambiguates the sense of word *Plant* but most frequent sense fails.

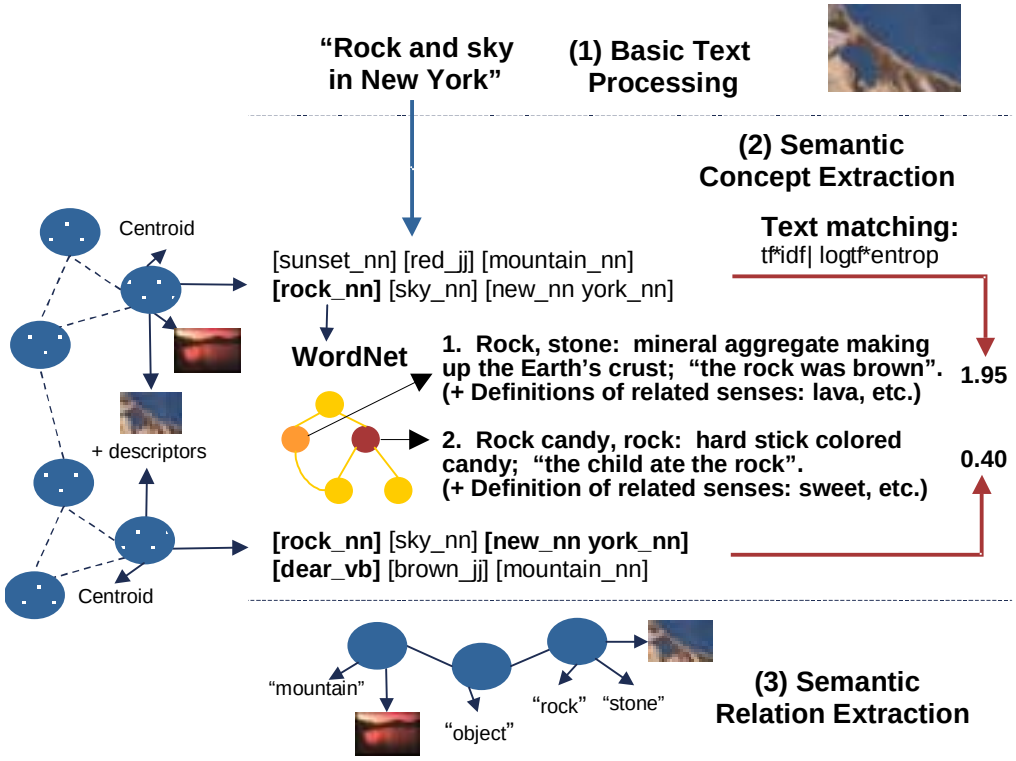


Figure 4: Semantic knowledge extraction process: first, the textual annotations are tagged with their part-of-speech ("_nn", "_vb", "_jj" and "_rb" for nouns, verbs, adjective and adverbs, respectively) and chunked into word phrases; then, semantic concepts (plain ellipses) are extracted by disambiguating the senses of the words using WordNet and the image clusters; and, finally, semantic relationships (plain lines) among senses are found in WordNet.

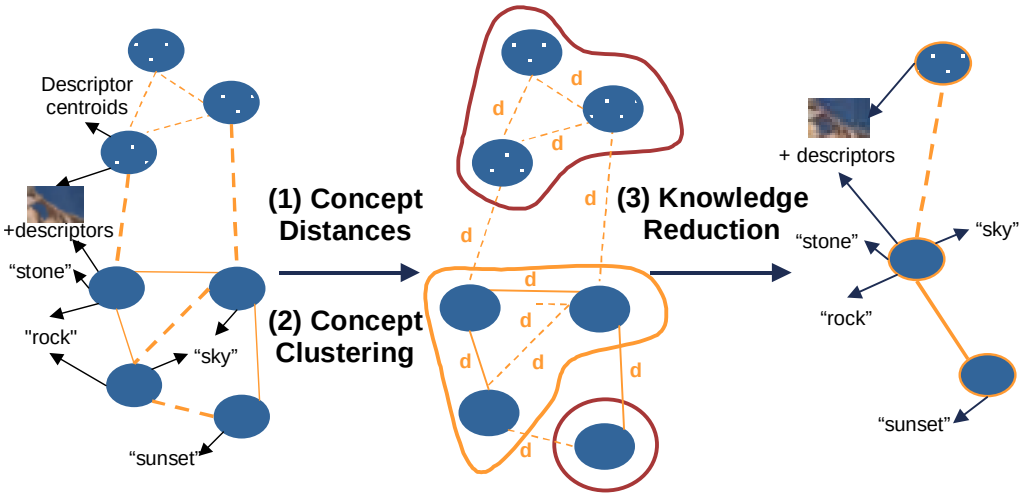


Figure 5. Multimedia knowledge summarization process: first, distances among concepts are calculated; then, similar concepts are clustered together; and, finally, the summary is constructed based on the concept clusters.

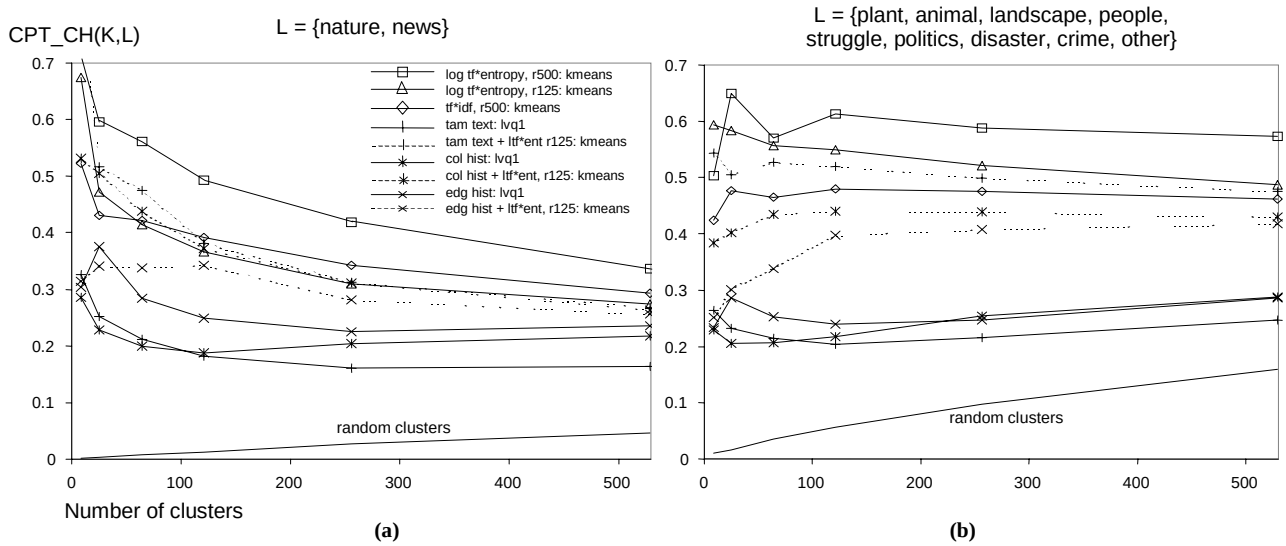


Figure 6: Entropy (CPT_CH) results (y axis) per number of clusters (x axis) for (a) the primary categories and (b) the secondary categories. "col hist" is color histogram, "tam text" is Tamura texture, "edg hist" is edge direction histogram, "ltf*ent" is log tf*entropy, "lvq1" is LVQ with primary categories, and "r125/500" is LSI for a reduced dimensionality of 125/500 bins.

	FD (c1) = FD(c2)	**FD (c1) ≠ FD(c2)**
$p(c1|c2), p(c2|c1) > \alpha$	c1 equivalent to c2; and vice versa	c1 co-occurs with c2; and vice versa
$0 < p(c1|c2) < \beta; p(c2|c1) > \alpha$	c1 specializes c2; c2 generalizes c1	c2 co-occurs with c1
$0 < p(c1|c2), p(c2|c1) < \alpha$	c1 overlaps c2; and vice versa	--

Table 1: Statistical relationship discovery rules where FD(c) is the feature descriptors used to generate cluster c, $p(c1|c2)$ is the probability of an image to belong to cluster c1 if it belongs to cluster c2, α is a positive real number smaller but close to one, and β is a positive real number smaller than α.

Relationship	**Definition**	**Example**
Synonymy	Equivalent	rock ↔ stone
Antonymy	Opposite	white ↔ black
Hypernymy Hyponymy	Generalizes Specializes	animal → dog rose → flower
Meronymy Holonymy	Is contained in Contains	ship → fleet martini → gin
Troponymy	Manner of	whisper → speak
Entailment	Causes or requires	divorce → marry

Table 2: Some relations in WordNet with definitions and examples.

Relation	Source to Target	Target to Source
Equivalent	1.0	1.0
Specializes	1.0	0.0
Contains	0.5	0.0
Causes	0.5	0.0

Table 3: Propagation weights for several semantic and perceptual relations from source to target and vice versa. These weights are used to calculate the concept frequencies (see section 6.2).

	Nature Images						News Images					
	%	**BI**	**WI**	**TT**	**MF**	**RD**	**%**	**BI**	**WI**	**TT**	**MF**	**RD**
Nouns	93.0	91.75	87.94	82.92	85.92	74.44	60.8	66.06	57.07	57.88	68.59	45.86
Verbs	1.08	66.67	40.74	59.26	44.44	44.44	25.0	48.16	36.61	39.07	58.48	24.08
Adjectives	5.72	58.04	41.43	40.85	55.71	44.29	12.3	71.00	56.50	54.68	72.00	46.77
Adverbs	0.20	100.0	33.33	37.50	100.0	75.00	1.90	80.65	50.00	62.50	74.19	45.16
All words	*100*	*89.20*	*85.29*	*84.72*	*83.80*	*72.42*	*100*	*59.95*	*52.46*	*52.33*	*66.58*	*40.52*

Table 4: Word-sense disambiguation accuracy (in percentages) for best image clusters (BI), worst image clusters (WI), image-per-cluster (TT), most frequent senses (MF), and random senses (RD). The results are provided separately for nature and news images, and for nouns, verbs, adjectives, adverbs and all words. Column % indicates word percentages.

Knowledge	ICST		CPT_H		CPT_D		ICCS	
	dist	$dist_{Jiang}$	dist	$dist_{Jiang}$	dist	$dist_{Jiang}$	dist	$dist_{Jiang}$
Extracted	*0.002*	*0.156*	*24.583*	*14.559*	*0.002*	*0.001*	*0.288*	*0.287*
Randomized	7.027	9.000	49.561	37.759	0.002	0.001	0.045	0.043
Summary 2	0.000	0.000	0.074	0.074	0.500	0.500	0.000	0.000
Summary 4	3.824	3.824	0.086	0.086	0.250	0.250	0.000	0.000
Summary 8	4.713	4.713	0.105	0.105	0.125	0.125	0.000	0.000
Summary 16	*4.312*	*4.312*	*0.618*	*0.618*	*0.071*	*0.071*	*0.000*	*0.000*
Summary 32	*0.187*	*0.699*	*5.212*	*4.300*	*0.229*	*0.220*	*0.125*	*0.000*
Summary 64	0.079	0.836	7.619	5.533	0.089	0.068	0.000	0.000
Summary 128	0.008	0.556	10.681	6.884	0.030	0.018	0.008	0.008
Summary 256	0.020	0.523	14.503	8.269	0.010	0.005	0.004	0.000
Summary 512	0.003	0.379	18.083	9.727	0.004	0.002	0.008	0.008

Table 5: Inconsistency (distance spread, ICST), completeness (entropy, CPT_H, and graph density, CPT_D) and inconciseness (concept distance redundancy, ICCS) results for extracted knowledge, randomized knowledge, and summaries of 2, 4, 8, 16, 32, 64, 128, and 512 super-concepts using the proposed concept distance, *dist*, and the distance [18], $dist_{Jiang}$.

Words		Summary of 32 super-concepts	
Plant	5.78	Entity, living entity	100.0
Animal	5,51	Plant life, vascular plant	33.92
Flower	4.91	Biological group	32.60
Landscape	4.44	Family (biology)	29.88
Habitat	4.43	Woody plant	29.88
People	2.49	Genus (biology)	29.77
Bird	1.88	Bush, flower, herb, grass	29.77
Culture	1.88	Landscape, ocean, land	29.45
Chordata	1.21	Animal, vertebrate	22.07
Mammal	1.14	Amphibian, fish, frog	17.69

Table 6: Most frequent words in annotations and concepts in the summary of 32 super-concepts. Occurrence probabilities are given in percentages.

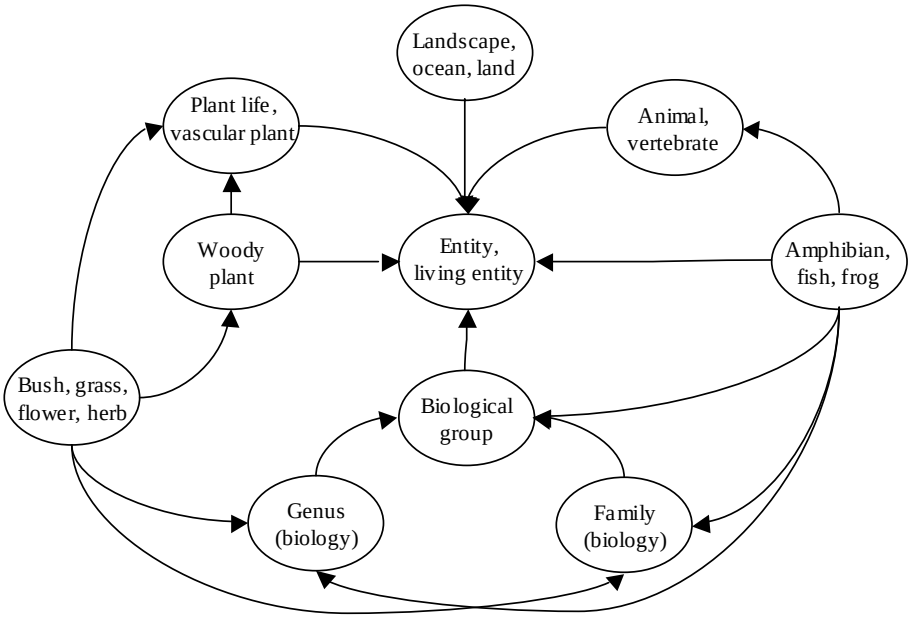


Figure 7: Sub-network of the knowledge summary of 32 super-concepts. The relationships are Specializes relationships. The arrows indicate the direction of the relationship from source to target. The concept labels were manually chosen for illustration purposes. In addition, all the relationships of the same type between any two super-concepts were represented with only one arc in the figure. In general, the super-concepts may include both perceptual and semantic concepts and there may be multiple relationships, both uni- and bi-directional, of different types between super-concepts, which are not shown in the graph.