Multimedia Knowledge Integration, Summarization and Evaluation

Ana B. Benitez
Department of Electrical Engineering
Columbia University
New York, NY 10027, USA
ana@ee.columbia.edu

Shih-Fu Chang
Department of Electrical Engineering
Columbia University
New York, NY 10027, USA
sfchang@ee.columbia.edu

ABSTRACT

This paper presents new methods for automatically integrating, summarizing and evaluating multimedia knowledge. These are essential for multimedia applications to efficiently and coherently deal with multimedia knowledge at different abstraction levels such as perceptual and semantic knowledge (e.g., image clusters and word senses, respectively). The proposed methods include automatic techniques (1) for interrelating the concepts in the multimedia knowledge using probabilistic Bayesian learning, (2) for reducing the size of multimedia knowledge by clustering the concepts and collapsing the relationships among the clusters, and (3) for evaluating the quality of multimedia knowledge using notions from information and graph theory. Experiments show the potential of knowledge integration techniques for improving the knowledge quality, the importance of good concept distance measures for clustering and summarizing knowledge, and the usefulness of automatic measures for comparing the effects of different processing techniques on multimedia knowledge.

KEYWORDS

Multimedia knowledge, knowledge integration, knowledge summarization, knowledge evaluation, concept distance, concept clustering, Bayesian networks

1. INTRODUCTION

This paper focuses on the integration, summarization and evaluation of multimedia knowledge representing perceptual or semantic information about the world depicted by, or related to an annotated image collection. Existing techniques are domain specific and do not generalize to arbitrary multimedia knowledge. Knowledge is usually defined as facts about the world and is often represented as concepts and relationships among the concepts, i.e., semantic networks. Concepts are abstractions of objects, situations, events or perceptual patterns in the world (e.g., a color pattern and concept Car); relationships represent interactions among concepts (e.g., color pattern one visually similar to color pattern two, and "sedan" specialization of "car").

Automatic knowledge integration, summarization and evaluation are essential for multimedia applications because multimedia applications often deal with multimedia knowledge at different abstraction levels such as perceptual and semantic knowledge (e.g., image clusters and word senses, respectively), which are usually extracted using different techniques. This diverse multimedia knowledge needs to be integrated to be uses in a coherent and meaningful way by applications. Furthermore, it is often necessary to reduce the multimedia knowledge in order to keep the most representative and useful multimedia knowledge, before or after the knowledge integration. Hence, ways to quantify the consistency, completeness and conciseness of the multimedia knowledge are essential to evaluate and compare any of these knowledge integration and summarization techniques.

Related work on multimedia knowledge integration includes generic pattern classification techniques. In particular, Bayesian Networks (BNs) allow the discovery of the statistical structure of a domain but they are not optimized for multimedia. There is a lot of work in the literature on building and fine-tuning classifiers for recognition of objects and scenes in images [17,20,22], among other multimedia; however, these are usually constrained to a specific domain and trained on skewed data sets. Prior work on multimedia knowledge summarization has been limited to efforts in network and concept reduction such as EZWordNet [14] and VISAR [7]. EZ.WordNet.1-2 are coarser versions of the English dictionary WordNet generated by collapsing similar word senses and by dropping rare word senses [14]. This process is governed by five rules manually designed by researchers for WordNet so they are not applicable to other knowledge bases or other kinds of knowledge such as perceptual knowledge. WordNet organizes English words into sets of synonyms (e.g., "rock, stone") and connects them with semantic relations (e.g., generalization) [15]. VISAR is a hypertext system for the retrieval of textual captions [7]. One of the functionalities of the VISAR system is the representation of the retrieved citations as a network of key concepts and relationships. Several reduction operators are used in this process (e.g., replace two concepts for a common ancestor) but the reduction operators are again manually defined and

lacking generality. Furthermore, the methodology followed by some of the reduction operators is not clearly specified. Prior work relevant to multimedia knowledge evaluation includes manual evaluation of semantic ontologies [9] and automatic but application-oriented evaluation of multimedia knowledge [1].

This paper presents new methods for integrating, summarizing and evaluating multimedia knowledge. In contrast to prior work, our techniques are automatic and generic applying 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., feature descriptor similarity, and generalization and aggregation relations), and instances of concepts (i.e., images and/or text representing the concepts). These methods are developed and used within the IMKA (Intelligent Multimedia Knowledge Application) system [4], which aims at extracting useful knowledge from multimedia and implementing intelligent applications that use that knowledge. The IMKA system uses the MediaNet framework to represent multimedia knowledge [5], which is presented in the next section.

In the IMKA system, the integration of multimedia knowledge consists of discovering new relationships between the concepts in the knowledge. The proposed approach for multimedia knowledge integration is based on building meta-classifiers for the concepts and learning statistical dependencies among them using a Bayesian network. 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, by grouping similar concepts into super-concepts, and by collapsing the relationships among super-concepts. Knowledge summarization could either precede or proceed knowledge integration; in fact, multimedia knowledge can be integrated and summarized in multiple stages and in different order. 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. Experiments show the potential of knowledge integration techniques for improving the knowledge quality, the importance of good concept distance measures for clustering and summarizing knowledge, and the usefulness of automatic measures for comparing the effects of different processing techniques on multimedia knowledge.

The paper is organized as follows. Section 2 defines and exemplifies multimedia knowledge by presenting the multimedia knowledge representation framework MediaNet. Sections 3, 4 and 5 describe the proposed methods for multimedia knowledge integration, summarization and evaluation, respectively. Section 6 presents the experiment setup and results in evaluating the proposed techniques. Finally, section 7 concludes with a summary and a discussion of future work.

2. MEDIANET

MediaNet is a unified knowledge representation framework that uses multimedia information for representing semantic and perceptual information about the world. The main components of MediaNet include concepts, relations among concepts, and media representing concepts and relationships. Examples of media are images, text and feature descriptors such as color histogram. MediaNet extends and differs from related work such as the Multimedia Thesaurus [21] 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 represented by media.

Concepts can represent either semantically meaningful objects (e.g., car) or perceptual patterns in the world (e.g., texture pattern). MediaNet models the traditional semantic relations such as generalization and aggregation but adds additional functionality by modeling perceptual relations based on feature descriptor similarity and constraints (e.g., condition on the distance of the color histograms). For example, perceptual knowledge for an image collection could be image clusters constructed based on visual and text feature descriptor similarity, and feature descriptor similarity and statistical relationships among the clusters [2]. Semantic knowledge for an annotated image collection could be the senses of the words in the textual annotations and semantic relationships among them as given by the electronic dictionary WordNet; the sense of each word could be disambiguated by matching the textual annotations of all the images in a cluster with the definitions of each possible sense [3]. In MediaNet, both concepts and relationships are defined and/or exemplified by multimedia information such as images, video, audio, graphics, text, and audio-visual feature descriptors. Feature descriptors can also be associated to the multimedia content (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. Weights and probabilities can be assigned to the concepts, relationships, and media representations in MediaNet to capture positive and negative examples of concepts and user feedback, in other words, the process of extracting semantics from percepts (i.e., automatic text annotation using visual feature descriptors). MPEG-7 is an international standard for the description of multimedia that has the potential to revolutionize current multimedia representation and applications [16]. Multimedia knowledge expressed using

the MediaNet framework can be encoded using MPEG-7 description tools, in particular, using the tools for describing semantics and models of multimedia [5].

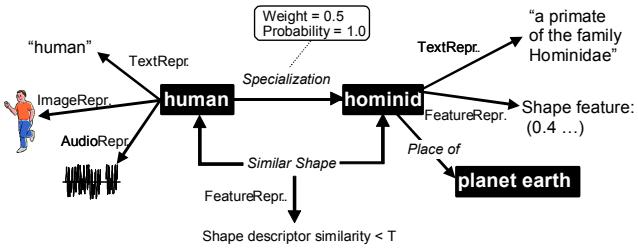

Figure 1: Example of multimedia knowledge.

3. MULTIMEDIA KNOWLEDGE INTEGRATION

The process of integrating multimedia knowledge consists of discovering relationships among concepts in multimedia knowledge to enable applications to make a coherent and meaningful use of diverse multimedia knowledge. As described in the previous section, the input multimedia knowledge is a set of concepts and relationships among concepts where both concepts and relationships can be either semantic or perceptual, and represented by different media such as images and text. Feature descriptors can also be associated with the images and the textual annotations.

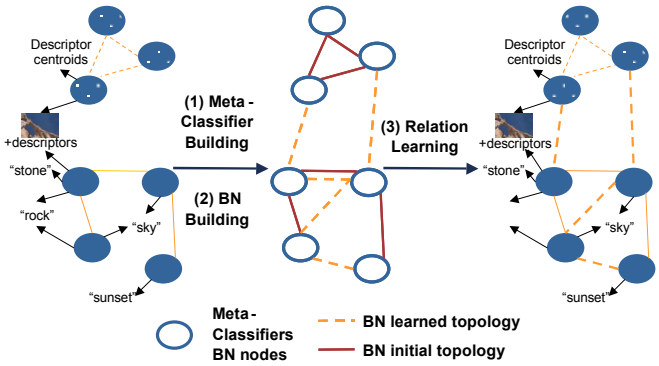

Figure 2. Multimedia knowledge integration process.

The proposed approach for multimedia knowledge integration consists of three steps, as shown in Figure 2: (1) building meta-classifiers for the concepts, (2) building a Bayesian Network (BN) whose nodes are the trained meta-classifiers and whose initial topology is the one of the known multimedia knowledge; and (3) adding the learned statistical relationships from the Bayesian network to the multimedia knowledge. This section describes each step. In Figure 2, dotted ellipses and dash lines represent perceptual concepts and relationships, respectively; plain ellipses and plain lines represent semantic concepts and relationships, respectively; and arrow lines represent media representations of concepts. Other figures in this paper follow the same conventions.

3.1 Meta-Classifier Building

In the first step, one or more classifiers are built for each concept and, from these, a meta-classifier per concept. Meta-classifiers are trained to predict the presence of concepts in images or their associated textual annotations based on their visual and text feature descriptors.

A classification algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations [8]. Classifiers basically learn how to predict the class (i.e., the value of the class attribute) of an input (given feature attributes of the input). The IMKA system uses a diverse set of classification algorithms: Naïve Bayes, Support Vector Machine (SVM), Neural Network (NN) and k-Nearest Neighbor (KNN) classifiers. The rationale for selecting each algorithm follows. The Naïve Bayes classifier is a very simple classifier. SVM and NN classifiers are slow at training but quick at classification. The KNN classifier can be trained quickly but it is slow at classification. Finally, the NN classifier requires large training sets whereas the KNN classifier does not.

A classifier is trained to predict the presence of a concept in an image based on a given combination of visual and textual feature descriptors associated with the image or its textual annotations. Therefore, the feature attributes input to each classifier for an image are a subset of the feature descriptors associated with the image. The class attribute that the classifier is trying to predict will have labels such as {presence, no presence} or {strong presence, weak presence, no presence} that indicate different strengths of the presence of a concept in an image. In the case of two-class classifiers (e.g., SVMs), several classifiers are used to learn more than two classes by using the one-per-class coding technique [8]. Multiple classifiers can be trained for the same concept using different combinations of feature descriptors or different classification algorithms. All the classifiers for a concept are combined into a meta-classifier, if needed, using bagging, boosting or stacking techniques [8]

The input feature attributes for building the classifiers of a concept are the visual and text feature descriptors associated with the images in the multimedia knowledge. The IMKA system uses several visual and text feature descriptors [2]. The supported visual feature descriptors are color histogram, Tamura texture, and edge direction histogram globally for images; and mean LUV color, aspect ratio, number of pixels, and position locally for automatically-segmented image regions. The IMKA system also implements two of the most popular schemes for representing textual annotations: 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 feature descriptors can be normalized before being inputted to the classifiers by adjusting the mean and variance of each bin to zero and one, respectively. Feature descriptor normalization is desirable especially when classifiers deal with multiple feature descriptors.

Apart from the feature attributes, each image is associated a score indicating the strength of the presence of each concept in the image. These concept-presence scores are quantized uniformly into a given number of levels, which correspond to the labels of the class attribute for the classifiers. The concept-presence scores are automatically initialized during the multimedia knowledge extraction process, e.g., likelihood that a sense is the real meaning of a word annotating an image [3]. The initial values are propagated along the multimedia knowledge network. For example, if an image contains the concept Dog with a given probability, it also contains the concept Animal with, at least, the same probability because concept Animal is a generalization of concept Dog. In the IMKA system, concept-presence scores can be propagated not only through specialization/generalization relations but also through any relation from the relationship's source to target and/or vice versa given some weights. These propagation relation weights can be either learned or specified by an expert. Common values for propagation relation weights are shown in Table 1.

3.2 Bayesian Network Building

The second step in the multimedia knowledge integration process is to build a Bayesian network using the meta-classifiers constructed in the previous step and the network of multimedia knowledge.

Bayesian Networks (BNs), also known as Belief Networks, are directed graphical models that allow representing joint probability distributions of several random variables in a compact and efficient way [8]. The nodes of a Bayesian network represent the random variables, which are specified by conditional probability distributions. In the case of discrete random variables, the conditional probability distribution of a node is a table that lists the probability that the child node takes on each of its different values for each combination of the values of its parents. Several conditional independence assumptions apply to Bayesian networks. The lack of arcs among nodes represents conditional independence among the nodes. Moreover, a node in a Bayesian network is independent of its ancestors given its parents.

A Bayesian network is fully specified by the topology or structure of the graph, and the parameters of each conditional probability distribution. It is possible to learn both the structure and the parameters of a Bayesian network for a given domain; however, the former is much harder than the latter. Learning the structure of Bayesian networks is especially hard when there is not prior knowledge of the Bayesian network's topology. However, once constructed for a domain, a Bayesian network can be used for probabilistic inference or reasoning about the domain; it can answer arbitrary questions about any conditional or joint probability of one or more of the random variables.

Bayesian networks are used during the multimedia knowledge integration process to learn statistical dependencies among concepts in the multimedia knowledge. Two reasons prompted the selection of Bayesian networks for this task. First, there are algorithms to learn statistical dependencies among the nodes in a Bayesian network by learning the structure of a Bayesian network. If the nodes in a Bayesian network represent concepts, then, the algorithms are actually learning statistical relationships among the concepts. The second reason is that once built, the Bayesian network can answer arbitrary probabilistic questions about the concepts, thus functioning as a knowledge classifier in itself.

A Bayesian network is built for multimedia knowledge that needs to be integrated as follows. The nodes of the Bayesian network are the meta-classifiers built as described in section 3.1; each node is thus indirectly representing a concept in the multimedia knowledge. The values of the nodes are the class labels of the meta-classifiers. The topology of the Bayesian network is initialized to the topology of the multimedia knowledge network; this is the best guess for the network topology based on prior knowledge. The initial multimedia knowledge from an image collection could be, for example, the perceptual and semantic knowledge directly extracted from the collection [2,3] or some multimedia knowledge summary. Bayesian networks cannot have directed cycles so certain arcs in the initial network may need to be removed to avoid directed cycles. The IMKA system uses the Markov Chain Monte Carlo (MCMC) algorithm called Metropolis-Hastings (MH) [10] to learn the topology of the Bayesian network. The training data for learning the Bayesian network is obtained by classifying the images in the multimedia knowledge using all the meta-classifiers.

3.3 Relationship Learning

The third step in the multimedia knowledge integration process is to add the newly learned statistical relationships among concepts to the multimedia knowledge.

The learned topology of the Bayesian network basically reveals important statistical relationships among the concepts in the multimedia knowledge. These relationships are compared with the known relationships among the concepts in the multimedia knowledge. A

statistical relationship is added to the multimedia knowledge for each arc between two concepts in the Bayesian network that does not already have a corresponding relationship in the initial multimedia knowledge. New statistical relationships could be added to the multimedia knowledge for each arc in the learned Bayesian network; however, some of these statistical dependencies are likely to be caused by already known relationships among the concepts.

4. MULTIMEDIA KNOWLEDGE SUMMARIZATION

This section presents techniques for automatically summarizing arbitrary multimedia knowledge by reducing the knowledge size in grouping similar concepts together. During this process, the number of concepts and relationships in the multimedia knowledge is reduced by grouping similar concepts into super-concepts and collapsing the relationships among the concepts in two super-concepts into a super-relationship.

The proposed approach for multimedia knowledge summarization consists of three steps, as shown in Figure 3: (1) obtaining the distances among the concepts in the multimedia knowledge; (2) clustering concepts based on the concept distances; and (3) reducing the concepts and the relationships in the multimedia knowledge based on the concept clusters. This section discusses each step in detail. In a preliminary stage, the least frequent concepts can be discarded from the multimedia knowledge and weights can be assigned to concepts for personalized knowledge summarization.

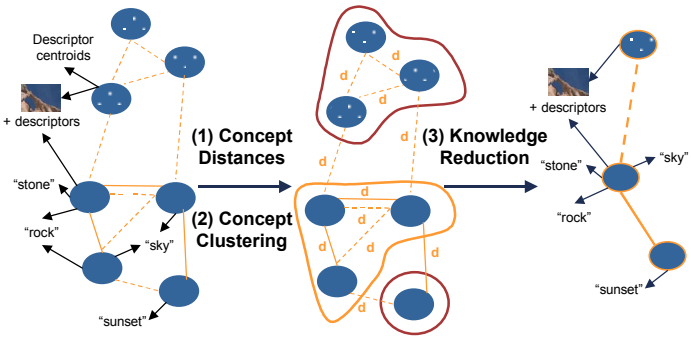

Figure 3. Multimedia knowledge summarization process.

4.1 Concept Distances

The first step in summarizing multimedia knowledge is to calculate the distances among concepts in the multimedia knowledge. Concept distances are calculated based on the concept statistics and the topology of the multimedia knowledge.

There are many proposed methods for calculating semantic distance or similarity among concepts in semantic concept networks such as WordNet. Some methods rely uniquely on the hierarchical specialization/ generalization relationships among concepts [12,13] whereas others take into account all the semantic relations [19]. There are methods that use exclusively the concept network topology [13,19] while others combine both concept network topology information and text corpus statistics (e.g., concept probabilities) [12]. The most commonly used concept network for calculating semantic relatedness is WordNet [12,13,19]. Recent work evaluated five semantic distance measures using WordNet [6], including [12] and [13], in a real-word spelling error correction system in which [12] was found to outperform the rest.

The semantic measure described in [12] only considers the specialization/generalization concept hierarchy in WordNet. The weight or distance of the relationship between a child concept c and a parent concept par(c) is the Information Content (IC), as defined in information theory, of the child concept given the parent concept, i.e., of encountering an instance of the child concept c given an instance of the parent concept par(c), as follows:

$$\begin{aligned} \text{dist}(c, \text{par}(c))_{\text{Jiang}} &= \text{IC}(c/\text{par}(c)) = -\log(p(c/\text{par}(c))) \\ &= -\log(p(c)) + \log(p(\text{par}(c))) \end{aligned} \quad (1)$$

where p(c) is the probability of encountering an instance of concept c. It is important to note that an instance of a child concept is always an instance of the parent concept and, therefore, p(c∩par(c)) = p(c). Then, the distance between any two concepts c and c' in the concept hierarchy reduces to the following expression:

$$\text{dist}(c, c')_{\text{Jiang}} = 2 * \log(p(\text{dcp}(c, c'))) - (\log(p(c)) + \log(p(c'))) \quad (2)$$

where dpc(c, c') is the deepest common ancestor of both concepts c and c'.

The IMKA system uses a novel concept distance measure that also uses concept statistics but is not limited to specialization/generalization concept relationships. The proposed concept distance measure generalizes measure [12] to an arbitrary concept network with different relations among concepts similar to measure [19]. Assuming binary relations, the distance of a relationship r between concept c and concept c' is the summation of the information content of concept c given concept c' and relationship r, and of the information content of concept c' given c and relationship r, as follows:

$$\begin{aligned} dist(c,c',r) &= IC(c/c',r) + IC(c'/c,r) \\ &= -\log(p(c/c',r)) - \log(p(c'/c,r)) \end{aligned} \quad (3)$$

where p(c) is still the probability of encountering an instance of concept c; p(c/c',r) is the probability of encountering an instance of concept c given an instance of concept c' through relationship r. The intuition behind Equation (3) is the following: if a relationship makes two concepts almost interchangeably, i.e., p(c/c',r) and p(c'/c,r) are close to 1, the concepts are very similar given that relationship; if not, they are dissimilar. The distance between any two concepts is calculated as the total distance of the shortest distance path between the two concepts in the concept network. Therefore, the proposed concept distance satisfies the non-negative and inequality properties of a distance function.

If the concept network is a specialization/generalization concept hierarchy, the proposed concept distance measure (see Equation (3)) simplifies to the semantic distance measure [12] (see Equation (2)). In this case, concept c' is the parent of concept c, c' = par(c), and r is the specialization/generalization relationship among them. The proof is straight forward realizing that an instance of concept c is always an instance of the parent concept par(c) and, therefore, log(p(par(c)/c,r) is zero.

There are different approaches toward calculating the probabilities of concepts such as WordNet's senses in a text corpus. The approach often used in conjunction with Equation (2) obtains the frequency of each concept c as follows:

$$freq(c)_{Richardson} = \sum_{w \in words(c)} \frac{freq(w)}{|concepts(w)|} \quad (4)$$

where words(c) is the set of words representing all the descendants of concept c in the generalization concept hierarchy including concept c, freq(w) is the frequency of concept w in the text corpus (i.e., word occurrence), and concepts(w) is defined as the set of concepts represented by word w [18]. As for WordNet's senses, this approach assumes concepts are represented by one or more words (e.g., "rock, stone"), and that the same word can represent more than one concept at the same time (e.g., "rock, stone" and "rock, candy"). Concept probabilities are then calculated from the concept frequencies as follows:

$$p(c)_{Richardson} = \frac{freq(c)}{N} \quad (5)$$

where N is the total number of distinct words representing, at least, one concept. Please, note that a concept that is an ancestor for all the rest of the concepts will have a probability of exactly 1.

Another way to understand this approach is that, first, strict concept frequencies are found for each concept without taking into account the specialized concepts or descendants; then, concept frequencies are propagated recursively through the specialization/generalization concept hierarchy from child concepts to direct parent concepts; and, finally, concept probabilities are calculated using Equation (5). In formulistic terms, this means that Equation (4) can be also expressed as follows:

$$freq(c)_{Richardson} = \sum_{c' \in descendants(c)} freq'(c') \quad (6)$$

given

$$freq'(c)_{Richardson} = \sum_{w \in words'(c)} \frac{freq(w)}{|concepts(w)|} \quad (7)$$

where words'(c) is defined as the set of words strictly representing concept c, without considering the words of the descendants of concept c.

The IMKA system generalizes this procedure of obtaining concept probabilities to an arbitrary concept network with several types of relationships among concepts. First, strict concept frequencies are found for each concept without taking into account related concepts. The multimedia knowledge contains the information of which concepts are instantiated in which images, and how many times a concept is instantiated in an image. For example, images are assigned to the concepts corresponding to the senses of all the words in the associated textual annotations, with the same frequency. The strict frequency of concept c is calculated as follows:

$$freq'(c) = \sum_{i \in images(c)} freq(c,i) \quad (8)$$

where freq(c,i) is the number of times concept c is instantiated in image i. As an example, the concept House would have a frequency of five for an image whose textual annotations contain the word "house" five times.

In the second step, the concept frequencies are propagated in the concept network recursively through the relationships among concepts. Considering a relationship r that connects concepts c and c', a different fraction of the frequency of concept c will be added to the frequency of concept c' based on relationship r, and vice versa. As an example, for the specialization/generalization relation, if concept c specializes concept c', the frequency of concept c is added in full to the frequency of concept c', but zero in the opposite direction. The propagation weights for each relation could be specified by an expert or learned automatically using machine learning techniques. In formulistic terms, the total frequency of concept c in the image collection is calculated as follows:

$$freq(c) = freq'(c) + \sum_{c' \in neighbors(c)} \sum_{r \in relations(c,c')} w(r) * freq(c') \quad (9)$$

where neighbors(c) is the set of concepts directly connected to concept c through relationships, relations(c,c') is the set of relationships connecting concepts c and c', and w(r) is the propagation weight for relationship r (see Table 1 for examples). To avoid loops, concepts are only allowed to contribute once to the frequency of another concept. The relations in the multimedia knowledge affect the concept frequencies and, therefore, the distances among the concepts through w(r).

Finally, the concept probabilities are calculated based on the concept frequencies using the following formula:

$$p(c) = \min\left(1, \frac{freq(c)}{\sum_{c \in concepts(K)} freq'(c)}\right) \quad (10)$$

where K is the multimedia knowledge being summarized and concepts(K) is the set of concepts in multimedia knowledge K. The concept frequencies are not exclusive that is the reason for dividing by the summation of strict concept frequencies instead of the summation of total concept frequencies. Also, due to the propagation of concept frequencies through relations other than specialization/generalization relations, the total frequency for some concepts may be larger than the summation of strict concept frequencies.

4.2 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 a given number of clusters, the desired number of concepts in the multimedia knowledge summary.

The IMKA system supports several data clustering algorithms such as the k-means algorithm, the Ward algorithm, the k-Nearest-Neighbor algorithm (KNN), the Self-Organizing Map algorithm (SOM) and the Linear Vector Quantization algorithm (LVQ). A modified KNN clustering that generates a given number of clusters is selected for clustering the concepts. The KNN clustering algorithm was selected to cluster concepts in multimedia knowledge because of the continuity and the non-globular shape of the resulting clusters. Moreover, the KNN clustering algorithm does not use or require a specific distance function. The input of the KNN clustering algorithm [11] is the number of shared neighbors k_t, and the k nearest neighbors, in order from closest to farthest, for each data item to be clustered. The algorithm groups every pair of data items that have at least k_t shared neighbors. The vote of shared neighbors can be weighted according to their positions in the ordered k nearest neighbors (e.g., sharing the second neighbor counting twice as much as sharing the third neighbor). In the KNN clustering algorithm, the number of resulting clusters is determined indirectly by the value of k_t.

The KNN clustering algorithm is modified slightly to generate a given number of clusters. Whereas the KNN clustering algorithm merges the clusters of two data items with at least k_t shared neighbors, 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. Weighting of shared neighbors is also supported 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., it has a higher weight), then, the data item will have fewer shared neighbors and be clustered with fewer other data items; it will tend to maintain its own identity. 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.

The concepts in the multimedia knowledge are clustered using the modified KNN clustering algorithm as follows. The input to the clustering algorithm is the desired number of concepts in the multimedia knowledge summary, and the k nearest concepts for each concept. Different shared neighbor weighting schemes [11] can be selected as well as individual weights for the concepts during clustering. The result of the concept clustering is a set of concept clusters and a centroid for each cluster.

4.3 Knowledge Reduction

The final step in the multimedia knowledge summarization process consists of generating the multimedia knowledge summary using the concept clusters and distances among concepts.

Once the clusters of concepts have been obtained, the multimedia knowledge summary is generated as follows. Each cluster becomes a super-concept in the summary and inherits the text and image representations of the cluster members. The most important text representation of the super-concept is the one of cluster centroid. If all the members of a cluster are semantic concepts, the super-concept will be labeled a semantic concept; otherwise, it will be labeled as a perceptual concept. The type of the super-concept is set to the type of the cluster centroid (e.g., visual concept based on color histogram similarity). Super-relationships are created between pairs of super-concepts based on the relationships between their cluster centroids in the original multimedia knowledge. The type of the super-relationship between two super-concepts is set to the type of the largest-distance relationship between the cluster centroids (e.g., generalization), as a worst-case scenario. Another possible approach for setting the type of a super-relationship would be selecting the most dominant

relationship (e.g., the one that appears most often between the concepts grouped by the two super-concepts).

5. MULTIMEDIA KNOWLEDGE EVALUATION

This section proposes several automatic application-independent techniques for evaluating the goodness of multimedia knowledge based on information and graph theory notions. These follow criteria used to manually evaluate and assess semantic ontologies and knowledge bases [9]. In contrast, many multimedia applications evaluate the quality of their multimedia knowledge by assessing the performance of complete applications using that knowledge, for example, automatic annotation performance of images [1].

A review on previous work on ontology evaluation has identified five criteria for the manual evaluation and assessment of semantic ontologies [9]. These criteria are the following: consistency, completeness, conciseness, expandability and sensitiveness. Expandability refers to the efforts required to add a new definition to an ontology, without altering the properties in the ontology. Sensitiveness relates to how small changes in a definition alter the set of well-defined properties guaranteed in an ontology. These two criteria are dependent on the way the knowledge is constructed, entered and maintained in the ontology so they are not considered in this section. This section proposes automatic ways for measuring the other three criteria -consistency, completeness and conciseness- for multimedia knowledge.

5.1 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 evaluated 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. The larger the distance spread among concepts, the more inconsistent or contradictory the different paths connecting the concepts.

In formulistic terms, the proposed way to measure the inconsistency of multimedia knowledge K is as follows:

$$ICST(K) = \log\left(\frac{\sum_{c,c' \in concepts(K)} \sum_{i=1}^{i=k} (d(c,c',i) - d(c,c',1))^2}{|concepts(K)|^2 * k} + 1\right) \quad (11)$$

where concepts(K) is the set of concepts in multimedia knowledge K, k is the number of shortest distance paths considered between 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 at to i = k. The lower ICST(K) for multimedia knowledge K, the more consistent the multimedia knowledge.

5.2 Completeness

Completeness refers to the completeness of both the ontology and the definitions in the ontology. The two proposed ways of evaluating the completeness of multimedia knowledge try to quantify the uniformity of the multimedia knowledge using entropy and graph density. The more uniform the multimedia knowledge, the more complete.

The first proposed way to calculate the uniformity of multimedia knowledge is by calculating the entropy of concepts, as follows:

$$CPT\_H(K) = -\sum_{c \in concepts(K)} p(c) * \log(p(c)) \quad (12)$$

where p(c) is the probability of concept c obtained as described in section 4.1. The higher CPT_H(K) for multimedia knowledge K, the more complete the multimedia knowledge.

The second proposed way to calculate the uniformity of multimedia knowledge adapts the formula for graph density to weighted relationships, as follows:

$$CPT\_D(K) = \frac{\sum_{r \in relations(K)} weight(r)}{|concepts(K)| * (|concepts(K)| - 1)} \quad (13)$$

where relations(K) is the set of relationships in multimedia knowledge K, and weight(r) is the weight of relationship r. If d(r) is the distance of relationship r and dmax is the maximum distance for a relationship, the weight of relationship r is obtained as follows:

$$weight(r) = \frac{dmax - d(r)}{dmax} \quad (14)$$

The higher CPT_D(K) for multimedia knowledge K, the more complete the multimedia knowledge.

Another way to measure the completeness of the semantic part of multimedia knowledge would be to compare it with an existing ontology or thesaurus, preferably, in the same domain for which the multimedia knowledge was constructed (e.g., News or Nature). However, thesauri do not exist for every domain. Comparing the semantic knowledge with general-purpose thesaurus such as WordNet is also not desirable because these generic thesauri often treat different domains with different

degrees of detail (e.g., good coverage of Animal species but limited coverage of News-related concepts in WordNet).

5.3 Conciseness

Conciseness refers to whether all the information in the ontology is precise, necessary and useful. The conciseness of multimedia knowledge can be evaluated by applying Single-Value Decomposition (SVD) to the concept distance matrix to find the rank of the matrix. The number of non-null eigen values is compared with the number of concepts. The closer the number of non-null eigen values to the number of concepts, the more concise the multimedia knowledge.

In formulistic terms, the proposed way to calculate the inconsistency of multimedia knowledge K is as follows:

$$\mathrm{ICCS(K)} = \frac{|\mathrm{concepts(K)}| - \mathrm{rank(M)}}{|\mathrm{concepts(K)}|} \qquad (15)$$

where M is the concept distance matrix, and rank(M) is the rank of the matrix M. The lower ICCS(K) for multimedia knowledge K, the more concise the multimedia knowledge.

6. EXPERIMENTS

Semantic and perceptual multimedia knowledge was integrated and summarized for a collection of images with associated textual annotations. The semantic and perceptual multimedia knowledge was generated for the annotated image collection using the techniques described in [2] and [3], respectively. The proposed multimedia knowledge evaluation measures were used to compare the proposed approaches with respect to several baseline approaches. The knowledge evaluation measures were also evaluated in these experiments by comparing their values for knowledge extracted from the image collection with the ones for random knowledge.

6.1 Experiment Setup

The test set was a collection of 25 images of plants from the Berkeley's CalPhotos collection (http://elib.cs.berkeley.edu/photos/). The images had short annotations in the form of keywords or well-formed phrases, as the example shown in Figure 4.

Perceptual knowledge was extracted by clustering the images using the k-means clustering algorithm based on the color histogram of the images, the log tf*entropy of the textual annotations and an integrated feature vector with both descriptors, and by finding relationships among the concepts based on statistical relations among the clusters [2]. Semantic knowledge was constructed by disambiguating the sense of the words in the textual annotations using WordNet and the image clusters [3]. Relationships among the semantic concepts were discovered based on the relationships among words senses in WordNet. The resulting multimedia knowledge had 75 semantic concepts, 15 perceptual concepts, 67 generalization relations, 16 aggregation relations and 15 association relations.

What: Plant, flower, orchid, western coralroof
Where: Montana, United States
When: 1959-05-07
Creator: C. Webber

Figure 4. Example of a plant image with corresponding textual annotations.

Summaries of different sizes were generated from the extracted multimedia knowledge using the propagation relation weights shown in Table 1, among others. Additional statistical relationships were discovered for one of the multimedia knowledge summaries using different classifiers – Naïve Bayes, SVM and 3-Nearest Neighbors (3NN) classifiers – trained on the integrated color histogram/log tf * entropy feature descriptor. The concept-presence scores were quantized into two values representing the presence and the absence of concepts in images, respectively.

Table 1: Propagation weights for some relations from source to target and vice versa.

Relation	Source to Target	Target to Source
Equivalence	1.0	1.0
Generalization	0.0	1.0
Aggregation	0.5	0.5
Statistical	0.25	0.25

The criteria to evaluate the multimedia knowledge integration and summarization were ICST(K), CPT_H(K), CPT_D(K) and ICCS(K) obtained as described in section 5. The performance of the proposed methods was compared to several baseline approaches. The baseline approach for multimedia knowledge summarization used the semantic distance [12] instead of the proposed concept distance. For multimedia knowledge integration, the baseline approach used the ZeroR classifier (which predicts the majority class). The four measures for multimedia knowledge evaluation were also evaluated by comparing the results obtained for the multimedia knowledge extracted from the image collection and for a randomized version of the multimedia knowledge.

6.2 Experiments Results

Table 2, Table 3 and Table 4 show the values for ICST(K), CPT_H(K) and CPT_D(K) obtained in the experiments evaluating the proposed techniques for evaluation, summarization and integration of multimedia knowledge, respectively. The values of ICCS(K) have been omitted because they were zero in all the instances.

Table 2 shows the results for the multimedia knowledge generated from the image collection using the proposed concept distance (dist(c,c'), see Equation (3)) and the semantic distance [12] ($dist(c,c')_{Jiang}$, see Equation (2)), and a random version of this multimedia knowledge. The random multimedia knowledge was generated by randomly changing the vertices of the relationships in the knowledge maintaining the types of the vertices. For example, if relationship r connected concept c and image i in the original multimedia knowledge, relationship r would connect any randomly chosen concept and image in the random multimedia knowledge. As expected, the random multimedia knowledge provides higher entropy than the extracted multimedia knowledge. On the other hand, the results for the distance spread and graph density of the extracted multimedia knowledge were better using the proposed concept distance. The semantic distance [12] did not perform very well because it is very conservative in calculating distances among concepts using only specialization/generalization relations.

Table 2: Inconsistency and completeness results for extracted multimedia knowledge using the proposed concept distance and the semantic distance [12], and for random multimedia knowledge.

	ICST	CPT_H	CPT_D
Extracted			
dist(c,c')	16.32	9.14	0.0122
$dist(c,c')_{Jiang}$	16.68	6.65	0.0084
Random	16.50	13.77	0.0119

Table 3 shows the results in summarizing the extracted multimedia knowledge into different number of concepts (i.e., knowledge summaries of 3, 9 and 18 concepts) using the proposed concept distance and the semantic distance [12]. Comparing the results in Table 2 and Table 3, the summarization of multimedia knowledge seems to increase the graph density and reduce the concept entropy. The summaries obtained using the proposed concept distance seem to consistently provide better overall results. As an example, although the graph density is higher for the summary of size 3 using semantic distance [12], the entropy of this summary is very small; the contrary seems to happen for the summary of size 18. Interestingly, the results for the summaries generated using semantic distance [12] show important oscillations compared to the ones obtained with the proposed concept distance, which are more stable.

Table 3: Inconsistency and completeness results in summarizing extracted multimedia knowledge into different number of concepts using the proposed concept distance and the semantic distance [12].

	Distance	ICST	CPT_H	CPT_D
3	dist(c,c')	15.82	0.14	0.1666
	$dist(c,c')_{Jiang}$	1.95	0.08	0.4998
9	dist(c,c')	15.92	1.79	0.0833
	$dist(c,c')_{Jiang}$	0.00	1.10	0.0000
18	dist(c,c')	16.43	1.04	0.2157
	$dist(c,c')_{Jiang}$	14.87	2.53	0.0196

Finally, Table 4 shows the results obtained in integrating the multimedia knowledge summary of nine concepts (whose results are in the second row of Table 3) using different classification algorithms. The table also includes the number of new statistical relationships discovered using each classifier. The results for the ZeroR classifier (which predicts the majority class) are provided for baseline comparison. The tendency seems to be the following: the fewer statistical relationships are added to the multimedia knowledge, the larger the entropy and the distance spread, and the smaller the graph density of the integrated knowledge. The Naïve Bayes and SVM classifiers seem to provide the best overall results, which consistently range from average to good. It is also important to note the different effects of using different classifiers in the knowledge quality. For example, Naïve Bayes improves upon the non-integrated multimedia knowledge in all measures (second row of Table 3). The general tendency seems to be for the distance spread to decrease importantly, the entropy to decrease slightly, and the graph density to increase slightly when adding the new statistical relationships.

Table 4: Inconsistency and completeness results in integrating the multimedia knowledge summary of nine concepts using different classifiers. Column Rels is the number of new statistical relationships discovered using each classifier.

	ICST	CPT_H	CPT_D	Rels
Naïve Bayes	1.47	1.59	0.2500	12
SVM	1.23	0.64	0.2777	14
3NN	16.26	1.93	0.1250	3
ZeroR	1.24	0.07	0.3194	17

Some global conclusions that can be drawn from the experimentation follows. First, all the knowledge evaluation measures are useful in comparing different

multimedia knowledge, concept distance measures and classifiers, among others, except for the inconsistency measure. The inconsistency measure was not very useful for the multimedia knowledge in these experiments because it lacked equivalence relationships among concepts. However, the large variation of the results especially observed for knowledge summaries of different size seem to indicate the need to review the definitions of some of these measures. Second, the discovery of new statistical relationships using classifiers and Bayesian networks usually improves the quality of the knowledge. However, the use of different classifiers has different effects on the results, which might be due to the fact that the Bayesian network is learned for the meta-classifiers and not the concepts themselves. The Bayesian network could be learned using both the meta-classifiers and the concepts (i.e., the actual presence or absence of a concept in the images); however, this would require the unfeasible task of generating the ground truth of which concepts appear in which images. Third, summarizing multimedia knowledge seems to increase the graph density and decrease the concept entropy. The use of different concept distances in the knowledge summarization process seems to have a very important impact in the quality of the resulting summaries. The proposed concept distance seems to provide fairly consistent results for different summary sizes during knowledge summarization and different classifiers during knowledge integration.

7. CONCLUSIONS

This paper has presented novel techniques for automatically integrating, summarizing and evaluating arbitrary multimedia knowledge. In particular, it has proposed (1) a novel way to integrate classifiers and Bayesian networks to discover statistical relationships among concepts; (2) a new technique for calculating distances among concepts used by a modified KNN algorithm to cluster concepts with the purpose of generating summaries of multimedia knowledge; and (3) automatic ways of measuring the quality of multimedia knowledge in terms of consistency, completeness and conciseness. Experiments have shown the potential of knowledge integration techniques for improving the knowledge quality, the importance of good concept distance measures for clustering and summarizing knowledge, and the usefulness of automatic measures for comparing the effects of different processing techniques on multimedia knowledge.

Current work is focused on extending the evaluation of these techniques to more images, evaluation measures, classification algorithms and propagation relation weights, among others. Other important current work aims at improving the efficiency of the implementation of these techniques in terms of processing time and memory usage as well as the scalability of these methods for a large number of images and concepts by developing heuristic approximations of some of proposed knowledge integration and summarization techniques. Future work will consist of implementing and evaluating applications that use the constructed multimedia knowledge for image classification and retrieval, automated concept illustration, and multimedia knowledge browsing, as well as, proposing a complexity-constraint framework for personalizing the quality values of the multimedia knowledge including complexity to specific user applications. Some of the remaining open issues are the extraction of multimedia knowledge from dynamic content such as video and audio, and the dynamic update of the knowledge based on user feedback or other external knowledge resources.

ACKNOWLEDGMENTS

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

REFERENCES

1. Barnard, K., P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M.I.Jordan, "Matching Words and Pictures", submitted to Special Issue on Text and Images, JMLR, 2002; also available at http://www.cs.berkeley.edu/~kobus/research/publications/JMLR/JMLR.pdf, 2002.

2. Benitez, A.B., and S.-F. Chang, "Perceptual Knowledge Construction From Annotated Image Collections", *International Conference On Multimedia & Expo (ICME-2002)*, Lausanne, Switzerland, Aug 26-29, 2002; also Columbia University ADVENT Technical Report #001, 2002.

3. Benitez, A.B., and S.-F. Chang, "Semantic Knowledge Construction From Annotated Image Collections", *International Conference On Multimedia & Expo (ICME-2002)*, Lausanne, Switzerland, Aug 26-29, 2002; also Columbia University ADVENT Technical Report #002, 2002.

4. Benitez, A.B., S.-F. Chang, and J.R. Smith, "IMKA: A Multimedia Organization System Combining Perceptual and Semantic Knowledge", *ACM International Conference on Multimedia (ACM MM-2001)*, Canada, Ottawa, Sep 30-Oct 5, 2001.

5. Benitez, A.B., J.R. Smith, and S.-F. Chang, "MediaNet: A Multimedia Information Network for Knowledge Representation", *SPIE Conference on Internet Multimedia Management Systems (IS&T/SPIE-2000)*, Vol. 4210, Boston, MA, Nov 6-8, 2000.

6. Budanitsky, A., and G. Hirst, "Semantic Distance in WordNet: An Experimental, Application-oriented

Evaluation of Five Measures", *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, Pittsburgh, PA, June 2001.

7. 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.
8. Duda, R.O., P.E. Hart, D.G. Stork, "*Pattern Classification*", John Wiley & Sons, Second Edition, United States of America, 2001.
9. Gomez-Perez, A., "Evaluation of Taxonomic Knowledge in Ontologies and Knowledge Bases", *Workshop on Knowledge Acquisition (KAW-1999)*, Alberta, Canada, Oct. 16-21, 1999.
10. Hastings, W.K., "Monte Carlo Sampling Methods Using Markov Chains and their Applications", *Biometrika*, Vol. 57, No. 1, pp. 97-109, 1970.
11. 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.
12. 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.
13. Leacock, C., and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification", *Fellbaum*, pp. 265-283, 1998.
14. Mihalcea, R., and D. Moldovan, "Automatic Generation of a Coarse Grained WordNet", *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, Pittsburgh, PA, June 2001.
15. Miller, G.A., "WordNet: A Lexical Database for English", *Communications of the ACM*, Vol. 38, No. 11, pp. 39-41, Nov. 1995.
16. MPEG Requirements Group, "MPEG-7: Context, Objectives and Technical Roadmap, V.12", ISO/IEC JTC1/SC29/WG11 MPEG99/N2861, Vancouver, July 1999.
17. Paek, S., and S.-F. Chang, "The Case for Image Classification Systems Based on Probabilistic Reasoning", *IEEE International Conference on Multimedia and Expo (ICME-2000)*, New York, NY, USA, July/Aug 30-2, 2000.
18. 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.
19. Sussna, M., "Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network", *International Conference of Information and Knowledge Management (CIKM-1993)*, pp. 67-74, 1993.
20. Szummer, M., and R. Picard, "Indoor-Outdoor Image Classification", *IEEE International Workshop in Content-Based Access to Image and Video Databases*, Bombay, India, Jan. 1998.
21. Tansley, R., "The Multimedia Thesaurus: Adding A Semantic Layer to Multimedia Information", Ph.D. Thesis, Computer Science, University of Southampton, Southampton UK, August 2000.
22. Vailaya, A., A. Jain, and H.J. Zhang, "On Image Classification: City vs. Landscape", *IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara, CA, USA, June 1998.