IMAGE CLASSIFICATION USING MULTIMEDIA KNOWLEDGE NETWORKS *

Ana B. Benitez and Shih-Fu Chang

Dept. of Electrical Engineering, Columbia University, New York, NY 10027
{ana, sfchang} @ ee.columbia.edu

ABSTRACT

This paper presents novel methods for classifying images based on knowledge discovered from annotated images using WordNet. The novelty of this work is the automatic class discovery and the classifier combination using the extracted knowledge. The extracted knowledge is a network of concepts (e.g., image clusters and word-senses) with associated image and text examples. Concepts that are similar statistically are merged to reduce the size of the concept network. Our knowledge classifier is constructed by training a meta-classifier to predict the presence of each concept in images. A Bayesian network is then learned using the meta-classifiers and the concept network. For a new image, the presence of concepts is first detected using the meta-classifiers and refined using Bayesian inference. Experiments have shown that combining classifiers using knowledge-based Bayesian networks results in superior (up to 15%) or comparable accuracy to individual classifiers and purely statistically learned classifier structures. Another contribution of this work is the analysis of the role of visual and text features in image classification. As text or joint text+visual features perform better in classifying images than visual features, we tried to predict text features for images without annotations; however, the accuracy of visual + predicted text features did not consistently improve over visual features.

1. INTRODUCTION

In recent years, there has been a major increase in available multimedia and in technologies to access the multimedia. Users often want to retrieve, filter and navigate multimedia at the semantic level (e.g., people). However, current multimedia applications use features at the perceptual level (e.g., color) failing to meet user needs. For example, the study [5] found that less than 20% of the attributes used by humans in describing images for retrieval were related to visual features. In addition, the most popular user operation in the web image search engine WebSEEk [10] was found to be subject hierarchy browsing. This paper focuses on image classification. Image classifiers can be used to annotate images with semantic labels. However, current approaches lack flexibility: they are often constrained to specific domains and trained on limited data sets.

Prior work on image annotation and classification can be reviewed in terms of input features, classifier structure, and class selection. Many methods rely uniquely on perceptual features such as color histogram [4][9][11]; whereas few also consider text features from annotations or captions [7][8]. There are some approaches that only use individual classifiers or joint distributions [1]; while others combine multiple classifiers for improved accuracy [4][9][11]. Finally, experts handpick the classes in many methods [11][7][8][9] to which the classifiers are often fine-tuned. Exceptions are frameworks where "expert" users define their own classes and relations [4], and approaches that associate words annotating images to new images or regions [1]. The most similar prior work is [7] and [9], which learn Bayesian Networks (BNs) with classifiers as nodes. However, the BN is either manually entered by experts or automatically learned using costly statistical methods.

In this paper, we present novel approaches towards image classification using visual and text features. The main contributions of this work are the automatic selection of salient classes, and the combination of multiple classifiers based on knowledge extracted from annotated images. In addition, this work analyses the role of visual and text features in image classification. As text or joint text+visual features perform better than visual features [8], we try to predict text features for images without annotations. We use the term “knowledge classifier” to refer to our image classification framework, and “knowledge network” to a concept network with associated media examples.

Knowledge networks are constructed from annotated images by clustering the images based on visual and text features (perceptual knowledge); and disambiguating the senses of the words in the annotations using WordNet [6] and the image clusters (semantic knowledge) [2]. Visual, statistical and semantic relations are discovered among concepts (e.g., image clusters and words senses). Statistically similar concepts can be merged to reduce the number of concepts in the knowledge network. We propose to build a knowledge classifier for a knowledge network in two steps. First, we train a meta-classifier to predict the presence of each concept in images using visual and text features. A meta-classifier can be the result of combining several classifiers of different types or feature inputs. Then, a Bayesian network is learned using the meta-classifiers and the concept network. The presence of concepts in a new image is first detected using the meta-classifiers and this initial classification is refined using Bayesian inference. Text features are predicted for images without annotations using clustering and statistical approaches based on visual features extracted from the images.

The paper is organized as follows. Section 2 summarizes the knowledge discovery process. Section 3 describes the construction of the knowledge classifier. The way concepts are detected in new images is explained in section 4. Section 5 presents the experimental setup and results. Finally, section 6 concludes with a summary and some future work.

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

2. DISCOVERING KNOWLEDGE NETWORKS

The discovery of knowledge from annotated images consists of four steps (see [2] for details): basic image and text processing, perceptual knowledge extraction, semantic knowledge extraction, and knowledge summarization. The result is a network of concepts with associated image and text examples.

First, images and annotations are processed separately. Images are segmented into regions with homogenous color and edge. Then, features are extracted from images and regions such as color histogram and size, respectively. Similarly, words in annotations are stemmed down to their base form and tagged with their part-of-speech (e.g., verb). After discarding stopwords and rare words, words are represented as vectors using word-weighting schemes such as tf * idf and log tf * entropy.

Perceptual knowledge is discovered by grouping images into clusters based on their visual and text features. We use well-known clustering algorithms: k-means, k-nearest neighbors, and self-organizing map algorithms, among others. Relationships among clusters are found based on centroid proximity and cluster statistics. For example, a cluster is considered to have similarity relationships with its k-nearest cluster neighbors based on their centroids' distances. Clusters and cluster relations are concepts and concept relations in the knowledge network.

Semantic knowledge is extracted by disambiguating the senses of words in annotations using WordNet and image clusters. WordNet is a dictionary that organizes English words into sets of synonyms (e.g., "rock, stone") and connects them with semantic relations (e.g., generalization) [6]. We assume images in the same cluster are often related semantically. The words annotating the images in each cluster are matched to the definitions of the possible senses of each word using word-weighting schemes. Disambiguated senses are added as concepts to the knowledge network. Relationships and intermediate senses in the paths connecting disambiguated senses are found in WordNet and added to the knowledge network.

Finally, the knowledge network can be summarized by merging similar concepts (e.g., image clusters and word senses). Merged concepts inherit all relations from individual concepts except for relations whose two vertices belong to the same merged concept. The distance among concepts in a knowledge network is calculated using a novel technique based on both concept statistics and network topology. The distance of a relationship between two concepts increases with the concepts' probabilities but decreases with the concepts' conditional probabilities through that relationship. The distance between any two concepts is the distance of the shortest distance path between them. Figure 1 shows examples of a concept network and a summarized concept network.

3. BUILDING KNOWLEDGE CLASSIFIERS

A knowledge classifier is built for a knowledge network in two steps: training meta-classifiers to predict the presence of concepts in images, and building a Bayesian network using the meta-classifiers and the concept network.

First, one or more classifiers are trained to predict the presence of a concept in images based on visual and text features. The class labels indicate concept presence strength such as {presence, weak presence, absence}. For image clusters, the labels are the presence or absence of an image in the cluster; for word senses, the quantized disambiguation scores. We use well-known classifiers including Naïve Bayes (NB) and Support Vector Machine (SVM). Several two-class classifiers can learn more than two classes using the one-per-class coding technique. If multiple classifiers are built for a concept (e.g., for different features), the classifiers are combined into a meta-classifier using techniques like stacking and majority voting.

Bayesian Networks (BNs) are directed graphical models that allow the efficient and compact representation of joint probability distributions for multiple random variables. We propose two approaches to combine meta-classifiers using Bayesian networks (see Figure 1). In the first approach (BN:MC), the nodes of the BN are the meta-classifiers; each node is thus indirectly representing a concept. The topology of the BN is set to that of the concept network after removing cycles. Each relation is assigned a direction in accordance with the cause-effect dependencies of a BN, if applicable (e.g., specialization: dog -> animal). Cycles are solved by removing all relations between the first two adjacent concepts (i.e., connected by a relationship) in a cycle. In the second approach (BN:MC+RC), the BN has meta-classifiers and real concepts as nodes; where a real concept node directly represents the presence of a concept. The arcs connecting real concept nodes in the BN are the relations in the concept network minus cycles. In addition, real concept nodes have incoming arcs from the meta-classifier nodes associated to adjacent concepts in the concept network. In both approaches, the parameters and the structure of the BN can be learned using standard statistical methods.

4. CLASSIFYING IMAGES

Once trained, the knowledge classifier uses the meta-classifiers to predict the presence of concepts in images. This initial prediction is refined using Bayesian inference.

For a new image, visual (and text) features are extracted from the image (and its annotations, if any). The features are inputted to the meta-classifiers. In BN:MC, the concept labels predicted by the best meta-classifiers are entered as observed values of the corresponding nodes in the BN (phase MC). An expert decides the number of best meta-classifiers. The performance of a meta-classifier is the concept detection accuracy in training images. The labels of the other concepts are inferred using the Bayesian network (phase MC+BN). Unconnected concepts are labeled using only the meta-classifiers. In addition, new concept labels for concepts detected using meta-classifiers can be refined or found using Bayesian inference (phase MC+2BN). In BN: MC+RC, the output labels of the meta-classifiers are observed values of the associated nodes in the BN. The presence of all the real concepts is then predicted using Bayesian inference. In both cases, senses disambiguated in the annotations of new images can be entered as observed values for corresponding meta-classifier and real concept nodes in BN:MC and BN:MC+RC, respectively.

Text or joint text+visual features perform better than visual features in image classification [8]. If a new image does not have annotations, we try to predict the text features based on visual features in order to label the image using knowledge classifiers that use text features. We propose to estimate the text features by clustering the training images based on text features

and modeling the visual features of the images within each cluster using a Gaussian model (clustering approach). We predict the text features for an image as the center of the cluster associated with the most likely Gaussian model given the visual features of the image. We also adopt the statistical approach proposed for handling missing and unreliable acoustic data in [3]. This technique models the distribution of features for the images of a given class using a mixture of Guassian models with diagonal-only covariance. The predicted text features for a new image are the mean text features conditioned on the visual features of the image given a class.

5. EVALUATION

From a collection of 2706 nature images with annotations, 2437 were used to train knowledge classifiers with different parameters. The remaining 269 were used to test the performance of the classifiers in terms of classification accuracy.

5.1. Experimental setup

The collection of 2706 nature images was taken from the Berkeley's CalPhotos collection (http://elib.cs.berkely.edu/photos). The images in CalPhotos are labeled as *plants* (857), *animals* (818), *landscapes* (660) or *people* (371). We use a few keywords from the annotations describing the main objects or people depicted on the pictures (e.g., "plant, flower").

A knowledge network was constructed using the 2437 training images. Color histogram (166 bins) was extracted from the images; and log tf * entropy (125 bins with latent-semantic analysis) from the annotations. Color histogram has been proven to be effective in retrieving natural images; in addition, it is widely accepted that log tf * entropy outperforms other word-weighting schemes in Information Retrieval. A concept network was then constructed using the senses of words in the annotations. The initial network of 52 semantic concepts, 47 specialization relations and 2 aggregation relations was summarized into 16 concepts and 13 specification relations. See Table 1 for a list of the most frequent words in the annotations, and concepts in the summarized knowledge network. Knowledge classifiers were then built for different classifiers, features, and structures, among others. We used the mean classification accuracy (for 16 concepts) to compare the resulting classifiers. For a concept, the accuracy is the percentage of testing images to which the concept is correctly assigned. Concept accuracies were weighted by $1 - p \log(p)$, where p is the probability of a concept in the training annotations. Common and rare concepts are given less importance. The first author of this paper generated the ground truth of correct senses for words in all the image annotations.

5.2. Experimental results

Table 2 lists the mean classification accuracy of knowledge classifiers built for (1) different features: color histogram (CH), log tf * entropy (LE), predicted log tf * entropy using the clustering (CPLE) and statistical (SPLE) approaches, and combinations of these; (2) different meta-classifiers (or classifiers): SVM and NB; (3) different structures for the Bayesian network: the meta-classifiers with no BN (MC no BN), BN of meta-classifiers (BN:MC) and BN of meta-classifiers and real concepts (BN:MC+RC); (4) and learning the parameters (PA), and also the structure (+ST) of the BN. The accuracies for BN:MC correspond to the best knowledge classifier at phase MC+BN or MC+2BN using 2 or 8 meta-classifiers in phase MC. In addition, we include results for disambiguating senses in annotations and activating the corresponding nodes in the BN (+O), another way of using annotations in the classification. For baseline comparison, randomly deciding the presence of concepts in images resulted in accuracies of about 50%.

The classifiers in Table 2 use the correct senses of words in annotations during the knowledge network and classifier construction. We do this for the purpose of decoupling classification and disambiguation errors. If senses were disambiguated automatically, as described in section 2, only 65% of the words were disambiguated correctly. However, classification accuracies still reached 90% and 80% for SVM and NB, respectively, using color histogram + log tf * entropy and log tf * entropy features. In addition, for both, correct and automatically disambiguated senses, we observed similar trends in the results for the same features, classifiers, etc.

As shown in Table 2, if *annotations are available* for new images, the best performing systems use (1) the individual SVM meta-classifiers (MC no BN) and (2) the BN of SVM meta-classifiers and real concepts (BN: MC+RC), using either text features (LE) or text and visual features (CH+LE). The differences in accuracy of these systems are not significant. When *annotations are not available* for classification (i.e., only color histogram inputs to meta-classifiers), the highest accuracy is achieved again for (1) the individual SVM meta-classifiers and (2) the BN of SVM meta-classifiers and real concepts. In both cases, *using and not using annotations*, having real concepts in the BN outperformed the BN of meta-classifiers alone (BN:MC) by up to 15%. Although the improvements for the BN of meta-classifiers and real concepts are insignificant with respect to no combination of classifiers for SVM, gains of up to 15% in accuracy where obtained for NB. These are good indications of the *importance of including nodes corresponding to real concepts in the BN*. In addition, combining classifiers using a BN can offer *significant performance gains that are not affected by specific choices of features and classifiers*.

Other conclusions can be drawn from Table 2. First, the structure of the *knowledge network discovered from annotated images using WordNet helps in classifying images*. BNs of meta-classifiers (and, especially, of real concepts) whose structures were based on discovered knowledge networks consistently outperformed BNs with purely statistically learned structures by up to a 15%. In addition, *observing values of nodes in the BN based on disambiguated senses in annotations improves the accuracy and robustness* of knowledge classifiers even with text feature inputs (+O). As an example, the most accurate NB-based knowledge classifier used color histogram inputs and +O. Finally, *predicting text features using visual features did not improve the most accurate* knowledge classifier with color histogram inputs and the SVM classifier. However, it improved the results of MC no BN and BN:MC for the NB classifier. Based on the results, a better way to improve the classification of images without annotations would be to do Bayesian inference using predicted concepts labels as observed values of nodes in the BN (+O with predicted concept labels).

6. CONCLUSIONS

This paper presents novel methods for classifying images based on knowledge discovered from annotated images. The main novelty of this work is to automatically use the extracted knowledge to discover salient classes, and to combine multiple classifiers for improved performance. Experiments have shown that combining classifiers based on knowledge discovered and summarized from annotated images using WordNet results in superior (up to 15%) or comparable accuracy to individual classifiers and purely statistically learned classifier structures. Another contribution of this work is the analysis of the role of visual and text features in image classification. As text or joint text+visual features perform better in classifying images than visual features, we tried to predict text features for images without annotations; however, the accuracy of visual + predicted text features did not consistently improve over visual features. Directions for future work are discovering knowledge from and classifying image regions, determining concepts that are accurately detected using trained classifiers, and distinguishing concepts that are applicable to image and/or regions. We envision the use of this information to refine discovered knowledge networks.

7. REFERENCES

[1] Barnard, K., P. Duygulu, D. Forsyth, et al., "Matching Words and Pictures", *JMLR*, in press.

[2] Benitez, A.B., and S.-F. Chang, "Automatic Multimedia Knowledge Discovery, Summarization and Evaluation", submitted to *IEEE Trans. on Multimedia*.

[3] Cooke, M., P. Green, L. Josifovski and A. Vizinho, "Robust Automatic Speech Recognition with Missing and Uncertain Acoustic Data", *Speech Communication*, 2001.

[4] Jaimes, A., and S.-F. Chang, "Learning Structured Visual Detectors From User Input at Multiple Levels", *IJIG, Special Issue on Image and Video Databases*, Aug. 2001.

[5] Jörgensen, C, "Attributes of Images in Describing Tasks", *Information Processing & Managem.*, Vol. 34, No. 2/3, 1998.

[6] Miller, G.A., "WordNet: A Lexical Database for English", *Comm. of the ACM*, Vol. 38, No. 11, pp. 39-41, Nov. 1995.

[7] Paek, S., and S.-F. Chang, "The Case for Image Classification Systems Based on Probabilistic Reasoning", *ICME-2000*, New York, NY, USA, July/Aug 30-2, 2000.

[8] Paek, S., C.L. Sable, V. Hatzivassiloglou, A. Jaimes, et. al., "Integration of Visual and Text Based Approaches for the Content Labeling and Classification of Photographs", *SIGI-1999* Berkeley, CA, 1999.

[9] Naphade, R.M., and T.S. Huang, "A Probabilistic Framework for Semantic Video Indexing, Filtering, and Retrieval", *IEEE Trans. on Multimedia*, Vol. 3, No. 1, March 2001.

[10] Smith, J.R., and S.-F. Chang, "An Image and Video Search Engine for the World-Wide Web", *IS&T/SPIE-1997*, San Jose, CA, 1997.

[11] Szummer, M., and R. Picard, "Indoor-Outdoor Image Classification", *IEEE Workshop in Content-Based Access to Image and Video Databases*, Bombay, India, Jan. 1998.

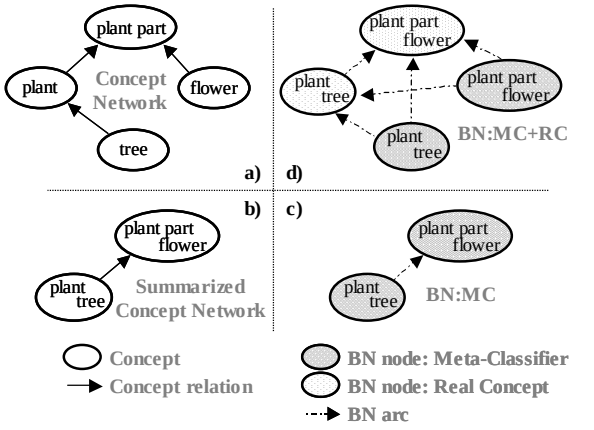

Figure 1: Examples of (a) a concept network, (b) a summarized concept network, (c) a BN of meta-classifier nodes, and (b) a BN of meta-classifier and real concept nodes.

Words		Concepts	
Plant	15.88	Plant flora, vine, tree	18.66
Animal	15.08	Animal, beast, fauna	14.96
Flower	13.30	Natural object, plant part, flower	13.19
Habitat	12.19	Society, people, group, culture	12.66
Landscape	12.19	Vicinity, country, landscape	12.09
People	6.85	Habitat, geographic area, region	12.09

Table 1: Most frequent words in annotations and concepts in knowledge summary with occurrence probabilities (%).

	CH							CH + CPLE			CH+SPLE		
	MC no BN	BN:MC			BN:MC+RC			MC no BN	BN: MC	BN: MC+RC	MC no BN	BN: MC	BN: MC+RC
		PA	+ST	+O	PA	+ST	+O						
SVM	84.31	81.93	83.00	84.36	82.30	80.40	94.27	79.30	79.19	79.40	43.40	66.32	39.90
NB	65.45	65.3	60.56	68.89	81.33	80.40	94.96	70.35	77.24	78.94	60.79	75.68	77.16

	CH+LE							LE						
	MC no BN	BN:MC			BN:MC+RC			MC no BN	BN:MC			BN:MC+RC		
		PA	+ST	+O	PA	+ST	+O		PA	+ST	+O	PA	+ST	+O
SVM	99.58	95.59	95.58	99.48	99.61	83.09	99.62	99.66	95.72	99.66	99.56	99.74	83.12	99.81
NB	82.76	82.10	81.10	88.29	86.04	80.47	92.40	85.52	83.87	83.88	87.74	90.35	83.31	95.48

Table 2: Mean classification accuracy for different classifiers (SVM: Support Vector Machines, NB: Naïve Bayes), different input feature features (CH: color histogram, LE: log tf * entropy, CPLE: LE predicted using clustering approach, SPLE: LE predicted using statistical approach), different structures of the BN (MC: only meta-classifiers, BN:MC: BN of meta-classifiers, BN:MC+RC: BN of meta-classifiers and real concepts). Columns PA and + ST are results for learning the parameters, and also the structure of the BN, respectively. Column +O are results from observing nodes in the BN+PA for senses disambiguated in annotations.