PERCEPTUAL KNOWLEDGE CONSTRUCTION FROM ANNOTATED IMAGE COLLECTIONS

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 and evaluates new methods for extracting perceptual knowledge from collections of annotated images. The proposed methods include automatic techniques for constructing perceptual concepts by clustering the images based on visual and text feature descriptors, and for discovering perceptual relationships among the concepts based on descriptor similarity and statistics between the clusters. The two main contributions of this work lie on the support and the evaluation of several techniques for visual and text feature descriptor extraction, for visual and text feature descriptor integration, and for data clustering in the extraction of perceptual concepts; and on proposing novel ways for discovering perceptual relationships among concepts. Experiments show extraction of useful knowledge from visual and text feature descriptors, high independence between visual and text feature descriptors, and potential performance improvement by integrating both kinds of descriptors compared to using either kind of descriptors alone.

1. INTRODUCTION

The important proliferation of digital multimedia content requires tools for extracting useful knowledge from the content to enable intelligent and efficient multimedia organization, filtering and retrieval. 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, or events in the world (e.g., color pattern and "car"); relationships represent interactions among concepts (e.g., color pattern 1 *visually similar to* color pattern 2 and "sedan" *specialization* of "car").

This paper focuses on the extraction of knowledge representing perceptual information about the world depicted by, or related to an annotated image collection (e.g., color clusters with visual *similarity* relationships among them). Perceptual knowledge is essential for intelligent multimedia applications because it can be extracted automatically and is not inherently limited to a set of words or signs as textual annotations. As many images often have some text directly or indirectly describing their content (e.g., caption or web page of an image, respectively), text can also be integrated into the perceptual knowledge extraction process.

Relevant prior work on perceptual knowledge construction includes visual thesauri approaches [10][11], and joint visual-text clustering and retrieval approaches [1][7]. The Texture Thesaurus is a perceptual thesaurus limited to texture clusters constructed using neural network and vector quantization techniques [10]. The Visual Thesaurus considers and adapts several kinds of visual concepts and visual relationships from typical text thesauri [11]. An approach for building the Visual Thesaurus involves grouping regions within and across images using only visual descriptors, and learning relationships among groupings through user interaction, so it is not a fully automatic system. Barnard et al. clusters images by hierarchically modeling their distributions of words and image feature descriptors [1]. In spite of being an interesting way of modeling the occurrence likelihood of text and visual feature descriptors for image clusters, the hierarchical clustering structure may become a limitation. Grosky et al. uses Latent Semantic Indexing (LSI) and word weighting schemes to reduce the dimensionality of, and to retrieve images using, concatenated feature vectors of visual feature descriptors and category label bits [7]. Limited experiments (50 images and 15 categories) have shown some performance improvement using LSI and weighting schemes.

This paper presents and evaluates new methods for automatically constructing perceptual knowledge from collections of annotated images. In contrast to prior work, this work supports and evaluates several techniques for visual and text feature descriptor extraction, for visual and text feature descriptor integration, and for data clustering in the extraction of perceptual concepts. It also proposes novel ways for discovering perceptual relationships among concepts based on descriptor similarity and statistics between clusters.

These methods are developed and used within the IMKA system [3]. IMKA stands for "Intelligent Multimedia Knowledge Application". The objectives of the IMKA project are to develop methods for extracting knowledge from multimedia content and implementing intelligent applications that use that knowledge. The multimedia knowledge is encoded using MediaNet, a knowledge representation framework that uses multimedia to represent both perceptual and semantic information about the world in terms of concepts and relationships among the concepts [4]. Methods for constructing semantic knowledge from annotated image collections are presented in [2]. Perceptual clusters discovered in this work provide an effective foundation to disambiguate semantic meanings associated with images.

2. PERCEPTUAL KNOWLEDGE EXTRACTION

The proposed approach for extracting perceptual knowledge from an collection of annotated images consists of three steps, as shown in Figure 1: (1) the basic processing of images and textual annotations to generate regions, visual feature descriptors and text feature descriptors; (2) the formation of perceptual concepts by clustering the images based on visual feature descriptors and/or the text feature descriptors; and (3) the discovery of perceptual relationships based on descriptor similarity and statistics between clusters.

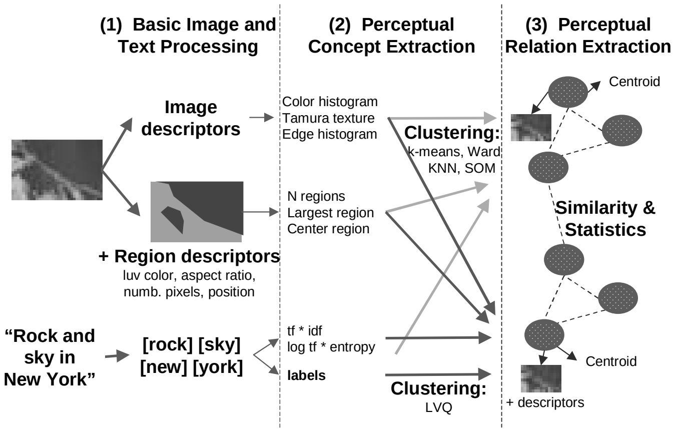


Figure 1: Perceptual knowledge extraction process. Ellipses and the dash lines represent perceptual concepts and relationships, respectively.

2.1. Basic images 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 [12]. The IMKA system uses a merge-and-split region segmentation using color and edge information of pixels [13]. This method is part of the semantic video object segmentation system AMOS and 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 [9][12]; and mean LUV color, aspect ratio, number of pixels, and position locally for segmented regions [13]. This feature descriptor set covers important visual features; moreover, each descriptor has been independently shown to be effective in retrieving visually similar images or videos from visual databases [9][12][13].

In the basic text processing, the textual annotations are broken down into words and punctuation signs. The words are tagged with their part-of-speech information (e.g., noun and preposition) and stemmed down to their base form (e.g., both "burns" and "burned" are reduced to "burn"). 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 (i.e., whose global frequency is below a threshold) are discarded.

The remaining words for an image are then represented as a feature vector using word-weighting schemes [6]. These schemes assign weights to words that reflect the discriminating power of each word in a text document collection. The IMKA system implements two of the most popular schemes: 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 [6]. The two text feature descriptors are computed as described in [6] considering the annotations of each image as a text document.

2.2. Perceptual concept extraction

The second step in the perceptual knowledge extraction process is to form perceptual concepts by clustering the images based on the visual and/or the text feature descriptors. Latent Semantic Analysis (LSI) is used to integrate and to reduce the dimensionality of feature descriptors before clustering. Each cluster is considered a perceptual concept.

Clustering is the process of discovering natural patterns in data by grouping data items based on some similarity criteria [8]. 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 and discover similarity relationships among clusters. Besides, LVQ enables to treat image annotations as labels for driving the image clustering; this is one way to integrate textual annotations and visual features in the clustering.

The IMKA system clusters images based on each kind of 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 each bin of a concatenated feature vector are normalized to zero and one, respectively. The dimension of text and concatenated feature vector can be reduced using Latent Semantic Indexing (LSI) [5]. Applying LSI has the effect of uncorrelating feature vector bins. This is the second way of integrating descriptors in the IMKA system.

2.3. Perceptual relationship extraction

The third step is to discover perceptual relationships among the perceptual concepts. Some clustering algorithms, such as SOM/LVQ and Ward clustering, already provide similarity and specialization relationships among the clusters, respectively. Additional relations of similarity, equivalence, co-occurrence, and specialization/generalization among clusters are discovered in novel ways by analyzing the descriptor similarity and the statistics between the clusters.

Each cluster is said to have s*imilar* relationships with its k nearest cluster neighbors. The distance between two clusters is calculated as the distance between the clusters' centroids. The number of neighbors could be taken in the range from 2 to 4, as that is the cluster neighbor range for SOM and LVQ clusters. *Equivalent*, *specialize*, *generalize*, *co-occur*, and *overlap* relationships between clusters are discovered based on cluster statistics as summarized in Table 1, where c1 and c2 are two clusters, FD(c) is the feature descriptor(s) used to generate cluster c, P(c1/c2) is the probability of an image to belong to c1 if it belongs to c2, and α is a positive real number smaller but close to one, and β is positive real number smaller than α. 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*. If the same clusters use different feature descriptors, they are said to *co-occur* on the same images.

	FD (c1) = FD(c2)	**FD (c1) ≠ FD(c2)**
P(c1/c2), **P(c2/c1) > α**	c1 equivalent to c2 c2 equivalent to c1	c1 co-occurs with c2 c2 co-occurs with c1
0 < P(c1/c2) < β **P(c2/c1) > α**	c1 specializes c2 c2 generalizes c1	-- c2 co-occurs with c1
0 < P(c1/c2), **P(c2/c1) < α**	c1 overlaps c2 c2 overlaps c1	-- --

Table 1: Relationship discovery rules based on conditional cluster probabilities.

3. EVALUATION

Perceptual knowledge was constructed for a collection of images with associated category labels and textual annotations. An entropy-based criterion was used to evaluate both the purity of the categories within each cluster and the spread of each category over the clusters.

3.1. 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 current 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 short annotations in the form of keywords or well-formed phrases, respectively (see Figure 2).

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 were discarded for the extraction of the text feature descriptors, whose dimensionality was further reduced from 2,000 to 500 and 125 using LSI. Clustering was done using different algorithms -k-means, SOM and LVQ- and 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. The SOM and LVQ maps were made square. The labels used for the LVQ clustering algorithms were the category labels listed above.

An entropy-based criterion was used to evaluate the resulting clusters. If L = {l_1, … l_m} and C = {c_1, … c_n} are the sets of category labels and clusters, respectively, the evaluation measure was the harmonic mean of one minus the mean entropies of the categories within each cluster (INH(C)), and of each category over the clusters (INH(L)), normalized from 0 to 1, as follows:

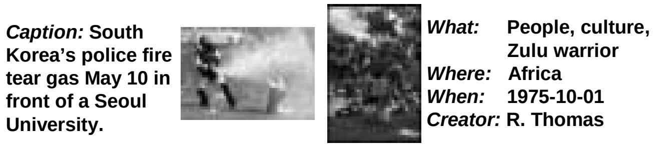
$$M(C,L) = \frac{2 * INH(C) * INH(L)}{INH(C) + INH(L)}$$

where

$$INH(C) = 1 - \sum_{j=1}^{n} P(c_j) * \sum_{i=1}^{m} P(l_i / c_j) * \log(P(l_i / c_j)) \Big/ \log(m)$$

$$INH(L) = 1 - \sum_{i=1}^{m} P(l_i) * \sum_{j=1}^{n} P(c_j / l_i) * \log(P(c_j / l_i)) \Big/ \log(n)$$

Note that the closer M(C,L) is to one, the better the clusters (C) are with respect to the labels (L). M(C,L) is an adaptation of the F-score used in Information Retrieval for combining precision and recall results.



Caption: **South Korea's police fire tear gas May 10 in front of a Seoul University.**

What: **People, culture, Zulu warrior**
Where: **Africa**
When: **1975-10-01**
Creator: **R. Thomas**

Figure 2: Example of a *news* image (left) and a *nature* image (right) with corresponding textual annotations.

3.2. Experiment results

Figure 3 shows the entropy results, M(C,L), of the best clustering algorithm for each feature descriptor excluding local (region) visual feature descriptors, which provided the worst results. Figure 3.(a) and Figure 3.(b) correspond to the entropy measure values for the set of primary categories {*nature, news*}, and of secondary categories {*plant, animal, landscape, people, struggle, politics, disaster, crime, other*}, respectively. Figure 3 also displays the 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. Please, note that 500-bin log tf * entropy descriptor was not integrated with the visual feature descriptors although it provides the best results. For baseline comparison, the figure also includes results for randomly generated clusters.

Interesting conclusions can be drawn from Figure 3. Both *text and visual feature descriptions are useful in obtaining some meaningful clusters*, in the sense of containing some useful knowledge, because their results are well above the baseline random behavior. As expected text feature descriptors are more powerful than visual feature descriptors, in generating more semantic-like clusters, and the log tf*entropy descriptor outperforms the tf*idf descriptor. The performance of the different visual feature descriptors is similar although edge histogram seems slightly better. The poorer results obtained with local visual feature descriptors are most likely due to the use of the Euclidean metric in the evaluated clustering algorithms because these descriptors work best with specialized metrics [13]. Regarding the performance of the different clustering algorithms, the neural network algorithms, in particular, LVQ with the labels {*nature*, *news*}, consistently provided much better results than the k-means algorithm for visual feature descriptors. However, all the clustering algorithms performed comparably for the text feature descriptors with the k-means algorithm providing some minor improvements over the rest.

It is important to note that *visual and text feature descriptors are highly uncorrelated*. This can be concluded from the minor dimensionality reduction of the concatenated visual-text feature descriptor when applying LSI with no coefficient dropping for all three visual feature descriptors (e.g., 166-bin color histogram + 125-bin log tf * entropy resulted in a 278-bin instead of a 281-bin feature vector), and the negligible differences of the results in using or not using LSI after bin normalization in concatenated visual-text feature descriptors. The independence between visual and text feature descriptors indicates that *both kinds of descriptors should be integrated* in the knowledge extraction process for best results in providing different kinds of useful knowledge. As an example, the concatenated and normalized visual-text feature descriptors slightly outperformed the individual text feature descriptor for the primary categories, probably because these categories are more visually separable. This was not the case, however, for the secondary categories. Although not shown in Figure 3, the trend of INH(C) and INH(L) is monotonically increasing and decreasing, respectively.

4. CONCLUSIONS

This paper proposes novel techniques for automatically extracting perceptual knowledge from annotated image collections. The evaluation of the perceptual concept extraction technique has shown that both visual and text feature descriptors are very uncorrelated but useful in extracting useful perceptual knowledge from annotated images. Certain clustering algorithms tend to work better with certain feature descriptors. Besides, the integration of visual and text feature descriptors has potential to improve performance compared to the individual descriptors.

Our current work is focused on evaluating the automated discovery of perceptual relationships and, more generally, of arbitrary knowledge, and discovering interactions among knowledge at different abstraction levels. For example, how to interrelate the perceptual knowledge discovered in this paper and the semantic knowledge discovered in [2], and use such interrelations for knowledge summarization, image classification, and automated concept illustration.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Barnard, K. and D. Forsyth, "Learning the Semantics of Words and Pictures", *ICVPR*, Vol. 2, pp. 408-415, 2001.

[2] Benitez, A.B., and S.-F. Chang, "Semantic Knowledge Construction From Annotated Image Collections", *ICME-2002*, Lausanne, Switzerland, Aug 26-29, 2002; also Columbia University ADVENT Technical Report #002, 2002.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Grosky, W.I., and R. Zhao: "Negotiating the Semantic Gap: From Feature Maps to Semantic Landscapes", *SOFSEM-2001*.

[8] 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.

[9] Kumar, R., and S.-F. Chang, "Image Retrieval with Sketches and Coherent Images", ICME-2000, New York, Aug. 2000.

[10] 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.

[11] Picard, R.W., "Toward a Visual Thesaurus", *Workshops in Computing (MIRO-1995)*, Glasgow, UK, Sep. 1995.

[12] 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.

[13] Zhong, D., and S.-F. Chang, "An Integrated Approach for Content-Based Video Object Segmentation and Retrieval", *IEEE Trans. on CSVT*, Vol. 9, No. 8, pp. 1259-1268, 1999.

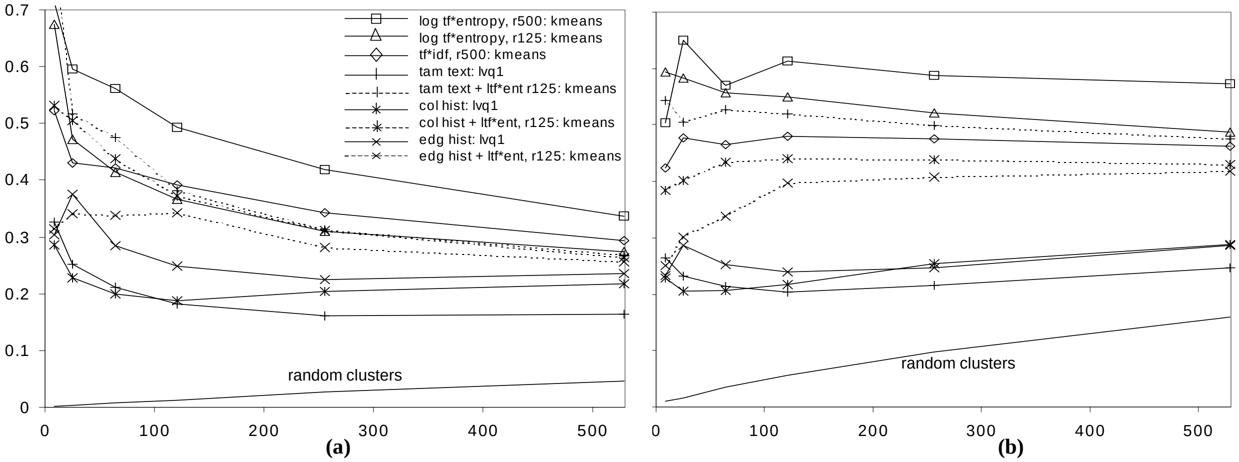


Figure 3: Entropy results (y axis) per number of clusters (x axis). (a) corresponds to the two primary categories {*nature*, *news*}; (b) corresponds to the nine secondary categories listed in Section 3.1. **"col hist" is "color histogram", "tam text" is "Tamura texture", "edg hist" is "edge direction histogram", "ltf*ent" is "log tf*entropy", "lvq1" is "LVQ with primary category labels", and r125/500 is LSI for reduced dimensionality of 125/500 bins.**