RELATING WORDS AND IMAGE SEGMENTS ON MULTIPLE LAYERS FOR EFFECTIVE BROWSING AND RETRIEVAL

Andrea Kutics[†‡*] *Akihiko Nakagawa*[‡*] *Shoji Arai*[*] *Hiroyuki Tanaka*[‡] *Sakuichi Ohtsuka*[‡]

[†]School of Media Science, Tokyo University of Technology, Tokyo, Japan
[‡]Media Handling Technology Group, NTT Data Corporation, Tokyo, Japan
[*]Japan System Co. Ltd., Tokyo, Japan
E-mail: andi@media.teu.ac.jp, bs-andi@bs.rd.nttdata.co.jp

ABSTRACT

This work proposes a new method for relating words and image segments by finding semantic coherence between these two cues on multiple layers. The method is based on the matching of visual segment clusters with words on various levels of abstraction. Our purpose here is to ease two main problems encountered in content-based image retrieval, namely, lack of semantic information captured by visual feature-based indexing, and difficulty of handling subjectivity of user queries. The method is very promising for effective browsing and retrieval in large image data sets. It supports both target- and category-type browsing and searching schemes as well as textual and/or visual query specifications. Results of experiments on a wide, non-specific image domain suggests that step by step semantic inference on consecutive layers of image - word association helps to improve accuracy of retrieval and browsing.

1. INTRODUCTION

In recent years, limitations of content–based image retrieval methods based on extraction and multidimensional indexing of visual features have become widely understood and the focus of research has shifted to bridging the semantic gap between low-level visual features and high-level semantic concepts. A number of methods of integrating textual and visual features have been proposed to overcome this problem. Relevance feedback-based methods [1] and various approaches drawn from the field of information retrieval, and document processing, such as methods using latent semantic indexing [2], statistical learning techniques based on mixture models [3], as well as HMM-based methods [4] and several neural network-based approaches [5] have been developed. In most of these methods, low-level visual features like color histograms, wavelet-based texture descriptions, etc. are extracted and then unified to create a composite feature vector or vectors representing the whole image or evenly divided image blocks. These features are then directly connected to words coming from various annotations and/or training data by estimating the joint distribution or co-occurrence of textual and visual features by using various learning approaches. One of the most promising approaches, described in [3], uses image segmentation to capture visual semantics and matches the detected objects with words by using a joint clustering approach on both feature spaces based on unsupervised learning using EM. Even though the reported results of these methods are very appealing, they carry a number of limitations, namely, (1) they are very dependent on the training data which is sometimes difficult and tedious to obtain, (2) only a limited number of semantic concepts can be learned by these methods, (3) it is difficult to apply them on an image domain different from the one they are originally trained on.

This work presents a method for relating words and image segments by finding semantic coherence between these two cues on multiple layers. Our purpose is to ease two main problems encountered in content-based image retrieval, namely, lack of semantic information captured by visual feature-based indexing, and difficulty of handling the subjectivity of user specific queries. We emphasize effective searching and browsing of images in large data sets without using any specific annotations added for training purposes or any domain-specific information. It is shown in [3, 6], that a hierarchical model is very suitable for browsing, which is a very desirable function when searching in large image data sets. Therefore, in this work we also apply a hierarchical model, but we define it on a different basis. We relate image segments and words by matching clusters of visual segment features with textual ones in consecutive steps representing three semantic layers, namely, visual feature-related words, concepts with visual coherence and hierarchies of abstract concepts.

2. RELATING IMAGE SEGMENTS AND WORDS

2.1. Visual segment features

As a first step, we carry out an image segmentation using the nonlinear inhomogeneous diffusion model based on both color and texture properties, which we defined in one of our previous works [7]. Here, we apply a small refinement on this model, such as using a combination of edge histograms and Gabor features for tuning the calculation of the diffusivity coefficient with a more adequately defined texture gradient to obtain accurate segment boundaries. The reader is referred to [7] for a detailed explanation of the above inhomogeneous diffusion model and that of the image segmentation process. An online demonstration of the image segmentation can be found on the URL: http://www.rd-image.com/retrieval. In further processing steps, we select segments that are most relevant to retrieval: mainly 4-6 segments per image. These are determined on the

0-7803-8554-3/04/$20.00 ©2004 IEEE.

basis of their size, geometric properties, and layout features, or on the saliency of their features. Next, we create a visual description of each segment by calculating MPEG-7 compliant features for color and texture, such as dominant colors, edge histograms (EHD) and homogeneous texture features (Gabor features) for each segment. Moment invariants and contour-based descriptors including the curvatures of the boundaries are calculated to obtain a shape description. We also calculate layout properties such as center coordinates, lengths of the main axes and their projections. A hierarchical structure of the main segments considering the entire image as root is determined.

2.2. Visual adjectives and nouns

On the first layer, we map the image segments to adjectives and nouns in order to transform their features to the textual domain. We accomplish this by applying clustering on each visual feature space such as color, texture, shape and layout of the segments and match the obtained clusters to psychophysically predetermined word clusters. In this process, we first determine dominant colors of the segments defined over a vector quantized HSV space and map them to color names by using a naming metric $D_{HSV}(s,i)=\left(\left|V_s-V_i\right|\right)^2+(S_s)^2+(S_i)^2-2S_sS_i\cos\left(\left|H_s-H_i\right|\right)^{1/2}$ to determine distances to prototypes of color name categories proposed in [8]. We also determine a hierarchy of texture feature clusters by using learning vector quantization and repeated vector quantizations (GLA) on the texture features of the segments. In order to obtain an initial codebook, prototype images representing these clusters are selected from the "Texture" categories of the Corel collection and their edge histogram, and Gabor features are calculated. Next, we determine texture-related adjectives by using the mapping of these texture clusters to texture-related keyword clusters (eleven main clusters) proposed as the texture lexicon in [9] and to several subclusters determined on the basis of psychophysical studies. We also determine shape clusters by using agglomerative clustering on segment shape features and heuristically determined shape-related words. At the end of this process each segment is assigned by a couple of words: mainly adjectives and a few nouns. Note that this is not a one-to-one cluster matching of visual and textual clusters, except for texture features, as segments can possess multiple dominant colors especially when texture is present, and also a shape can be a member of multiple textual shape categories and vice versa.

2.3. Image segments and words with visual coherence

On the second layer, we determine joint clusters of segments mapped to visual adjectives and nouns, and words with visual coherence. It has already been noted by various researchers that some words are directly related or have stronger coherence with specific visual features like "grass, tulip, cheetah, lake, etc.", while others on a higher abstraction level like "work, sports, age, religion, reflection etc." have little or no visual coherence. In this process, we use only those words that were parsed from naturally attached annotations and were not assigned for direct training purposes. Next, we try to select and eject the ones with more abstract meaning. This 'word selection' is simply done by using lexical definitions. In our case, we use the WordNet Lexical Database [10] and we drop words with a higher level of abstraction by determining their sense hierarchy (superordinate tree) and looking up visual or visual-related adjectives and nouns in their senses. As the WordNet contains only very short definitions, here we also utilize a lexical dictionary of pictures.

In order to obtain joint segment-word clusters, we apply a soft vector representation $(\vec{V}_{r_C},\vec{V}_{r_T},\vec{V}_{r_S})$ of the visual-related words where $v_{rj}\in[0,1]$ and define the probability of which word (j) expresses image segment (r). These visual word probabilities can be defined by calculating the weighted normalized distance $(\alpha_C D_{j_C},\beta_T D_{j_T},\gamma_S D_{j_S})$ from the given visual cluster center. The weights $(\alpha_C,\beta_T,\gamma_S)$ express the relations between the corresponding visual and textual clusters. Next, we apply a vector quantization on these soft vectors to determine clusters of image segments that are similar in their visual properties. Then we estimate the conditional probability that a word with visual coherence ($item_i$) belongs to a given segment cluster ($cluster_j$). These probabilities are obtained by calculating the frequencies of these words or their direct superordinates, depending on the level of hierarchy, over the clusters and this can be expressed by the following equations:

$$P(item_i \mid cluster_j)=\frac{P(cluster_j \mid item_i)P(item_i)}{\sum_{l=1}^{I}P(cluster_j \mid item_l)P(item_l)}=\frac{\text{count}(item_i)\in cluster_j}{\sum_{l=1}^{I}\left(\text{count}(item_l)\in cluster_j\right)} \quad (1)$$

Finally, we assign a maximum of eight words to segments determined on the basis of their probabilities ranked over the clusters.

2.4. Generating concept hierarchies

Finally, on the third layer we determine a hierarchy of abstract concepts. This can be simply accomplished by assigning global sense or superordinate hierarchies by using WordNet, and selecting a concept hierarchy on the basis of word frequency, co-occurrence and polysemy counts (word familiarity). Another possible way to determine these concept hierarchies is by assigning the words with visual coherence together with their superordinate trees on the previous level. However, this would require a subtree matching or a very complicated training process. Concept or word hierarchies can also be determined by applying a hierarchical combination of asymmetric and symmetric clustering of words and segments on the second level as proposed in [3]. In this way, segments and words with higher probabilities are put on the higher level. However, here we argue that global semantic concept hierarchies are more suitable for browsing in general data sets, even if these are sometimes unbalanced. That is, there exist nodes with few or no membership. An overview of the word - segment relating process is illustrated in Figure 1.

3. BROWSING AND RETRIEVAL

The method is very useful for target or category type browsing and searches. The user can present a query by using keywords, phrases, example images or image segments by using the segmentation tool. Except for a very specific target-type search when the user has a given image or image object prepared for upload, users tend to present their queries for searching and also for browsing by specifying keywords as a starting step. In this

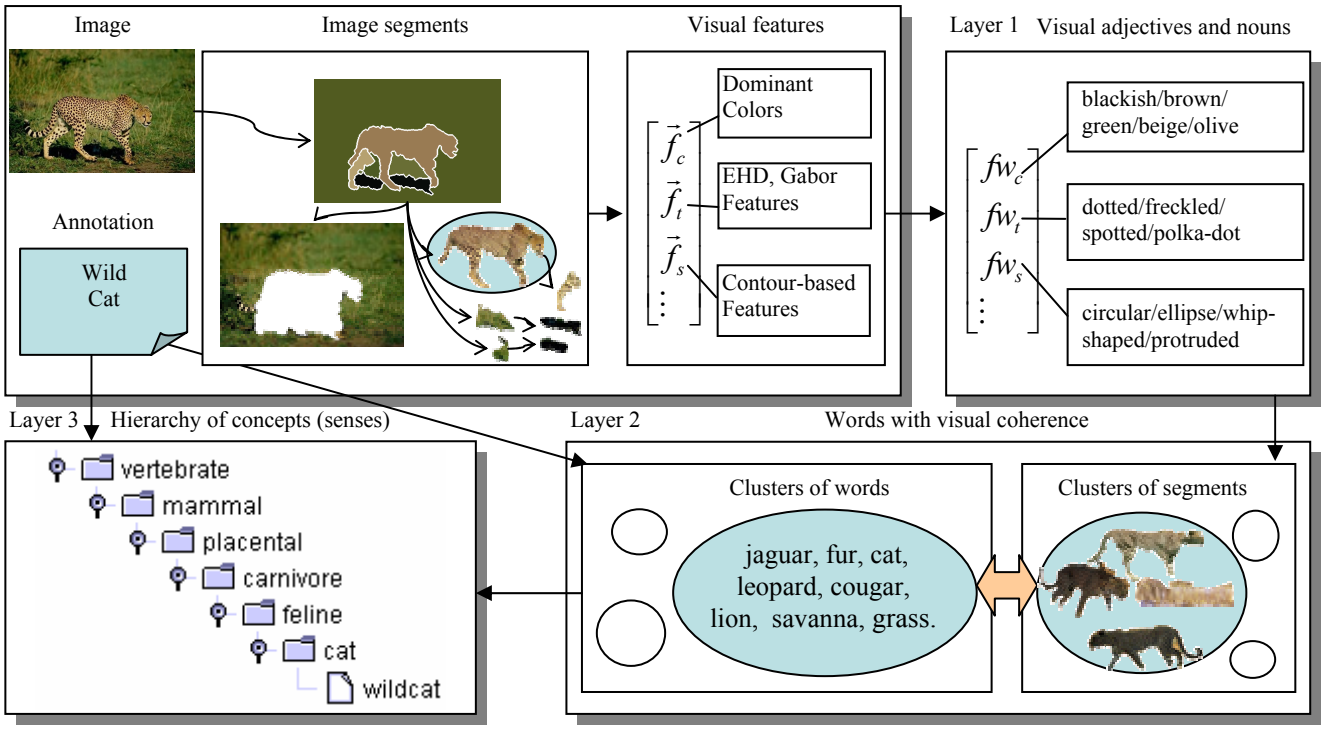
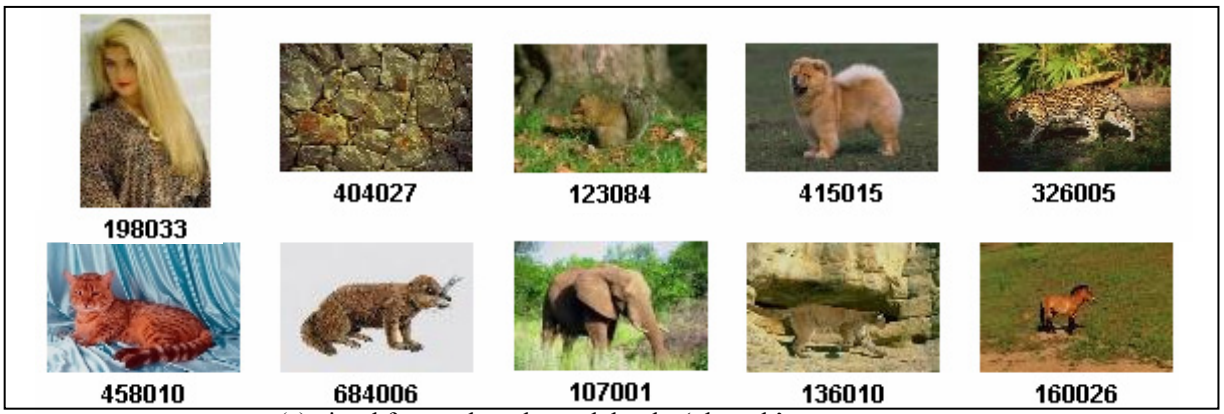
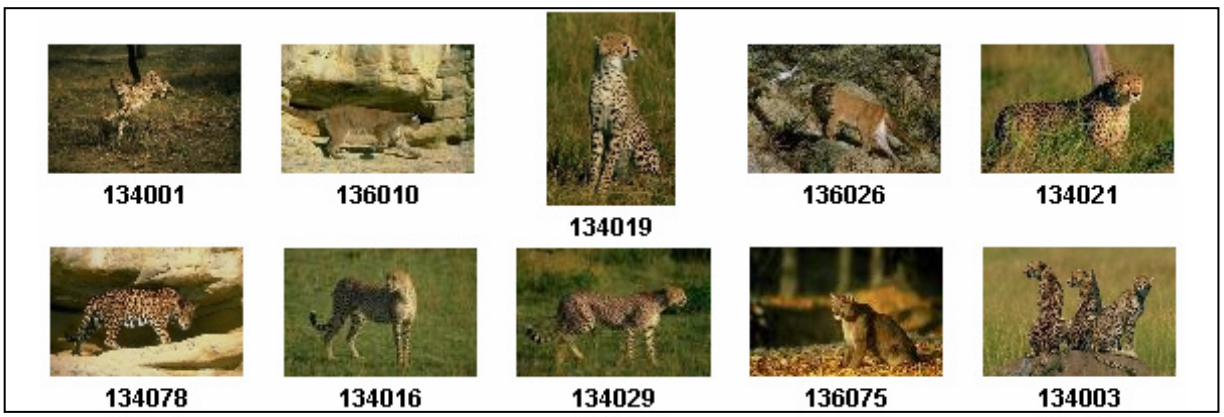


Figure 1. Overview of the word - segment relating process.



(a) visual feature-based search by the 'cheetah' segment query.



(b) combined search by the proposed method using both 'cheetah' segment (Fig. 1.) and keyword queries.

Figure 2. Retrieval results.

case, a concept hierarchy and corresponding images can be retrieved by the method, enabling the user to freely browse in the concept tree, find target images and execute a combined search. For queries ($qitem_{1...K}$) of both keyword(s) and/or image/segment example(s), segments/images ($seg_{1...I}$) are retrieved by calculating query-item probabilities over each cluster ($cluster_j$) weighted by the probability of the cluster with respect to the given segment(seg_i):

$$P(qitem_{1...K} \mid seg_i) = \sum_j \left(P(cluster_j \mid seg_i) * \prod_{k=1}^{K} P(qitem_k \mid cluster_j) \right) \quad (2)$$

The user can also refine the retrieval results by specifying relevant and/or irrelevant images among the retrieved images. Via this relevance feedback, not only the probabilities of both visual and conceptual words are updated, but user-specific query words are also saved in a "user dictionary" to enable user subjectivity to be handled more effectively. Retrieval result examples are illustrated in Figure 2. The results shown in (a) were obtained via a traditional visual feature-based search by specifying a cheetah object query. Figure 2. (b) illustrates the results obtained by the proposed method by specifying the keyword "cheetah", and a "cheetah" segment extracted from the image depicted in Figure 1.

4. EXPERIMENTS

We selected 12,500 natural images representing various categories (animals, toys, vehicles etc.) of the Corel Gallery collection to evaluate the method. We did not use any additional annotations for training purposes, but only the group names (1 name/100 images) parsed from the image group titles that were assigned by Corel. These are very short, like "Spectacular Landscapes", "Reflective Effects", etc., are determined very subjectively, and are redundant semantically. In order to produce a query image set, we randomly chose 100 test images from 10 different, additional category groups of Corel, regardless of the existence of well-defined relevant objects. We also used 6-12 test query words assigned to each test image by 5 persons representing different genders, generations (2) and cultural backgrounds (3 countries). We obtained the highest retrieval precision for combined queries (queries presented by both keyword(s) and image(s) or image segment(s)). About 70 percent of the most relevant objects or pictures were retrieved for each test person in the set of 24 result images in the first retrieval cycle and 95 percent retrieval precision was obtained after five relevance feedback cycles. We compared these results to those obtained by applying traditional visual feature-based searches carried out using the same test images or image segments, and the retrieval precision dropped to an average of 55 percent. Retrieval failures occur for two main reasons: (1) according to segmentation errors, which errors have the most severe effect on shape features, and mostly occur on images containing areas of inhomogeneous texture features (2) according to word mapping errors on the second or third layers. In these latter cases, the method often fails to assign relevant words with visual coherence or fails to determine proper superordinate hierarchies. These errors can be avoided by applying more user interaction, i.e., by inventing more browsing steps into the retrieval or applying a more precise statistical model for concept matching, which is an ongoing research topic in our laboratory.

5. CONCLUSIONS

In this work, we proposed a new approach for relating words and image segments by trying to establish semantic inference using visual and textual cues for efficient image retrieval. The presented method uses a non-linear segmentation framework to detect image objects which themselves express some semantics on the visual layer. The method relates image segments with words on three layers, starting by assigning visual adjectives and nouns, then creating joint clusters of segments and words with visual coherence and finally determining a hierarchy of higher level concepts. The advantage of the method is that it supports both target- and category-type browsing and searching schemes as well as textual and/or visual query specifications. By comparing the results of experiments conducted on a large, non-specific image domain with reported results of traditional visual feature-based CBIR methods or other direct image - word matching approaches, it can be shown that higher retrieval precision and browsing effectiveness can be achieved by using step-by-step semantic inference on consecutive layers of image - word association.

6. REFERENCES

[1] X. S. Zhou and T. S. Huang, "Unifying Keywords and Visual Contents in Image Retrieval", *IEEE Multimedia*, Vol. 9, No. 2, pp. 23-33, 2002.

[2] R. Zhao, et al., "Negotiating the semantic gap: from feature maps to semantic landscapes", *Pattern Recognition*, Vol. 35, pp. 593-600, 2002.

[3] K. Barnard, et al., "Matching Words and Pictures," *Journal of Machine Learning Research*, Vol. 3, pp 1107-1135, 2003.

[4] J. Z. Wang and J. Li, "Learning-based linguistic indexing of pictures with 2-D MHMMs," *Proc. ACM Multimedia*, pp. 436-445, 2002.

[5] J-H. Lim, Q. Tian, P. Mulhem, "Home Photo Content Modeling for Personalized Event-Based Retrieval", *IEEE Multimedia*, Vol. 9, No. 2, pp. 28-37, 2003.

[6] T. Hofmann, "Learning and Representing Topic. A Hierarchical Mixture Model for Word Occurrences in Document Databases", *Proc. of CONALD*, Pittsburgh, 1998.

[7] A. Kutics, et al., "An object-based image retrieval system using an inhomogeneous diffusion model", *Proc. of the ICIP'99*, Vol. II, pp. 590-594, 1999.

[8] A. Mojsilovic, "A method for color naming and description of color composition in images", *Proc. of the ICIP2002*.

[9] N. Bhusnan, et al., "The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images", *Cognitive Science*, Vol. 21(2), pp. 219-246, 1997.

[10] C. Fellbaum, et al., *WordNet: An Electronic Lexical Database*, MIT Press, May 15, 1998.