

Mining Visual and Textual Data for Constructing a Multi-Modal Thesaurus

Hichem Frigui*

Joshua Caudill†

Abstract

We propose an unsupervised approach to learn associations between continuous-valued attributes from different modalities. These associations are used to construct a multi-modal thesaurus that could serve as a foundation for inter-modality translation, and for hybrid navigation and search algorithms. We focus on extracting associations between visual features and textual keywords. Visual features consist of low-level attributes extracted from image content such as color, texture, and shape. Textual features consist of keywords that provide a description of the images. We assume that a collection of training images is available and that each image is globally annotated by few keywords. The objective is to extract representative visual profiles that correspond to frequent homogeneous regions, and to associate them with keywords. These profiles would be used to build the a multi-modal thesaurus. The proposed approach was trained with a large collection of images, and the constructed thesaurus was used to label new images. Initial experiments indicate that we can achieve up to 71.9% relative improvement on captioning accuracy over the state-of-the-art.

Keywords: Multimedia mining, multi-modal thesaurus, clustering, feature weighting, image annotation

1 Introduction

The advent of digital libraries has made it necessary to develop automated tools for storing, retrieving, organizing, and mining large multimedia databases. Image data offers unique advantages in that it is relatively easy for human to explore and interpret. However, for computer methods, it poses serious challenges. To man-

age image databases, Content-Based Image Retrieval (CBIR) emerged as a new research subject [1, 2]. CBIR involves the development of automated methods that are able to recognize the visual features of the images, and to make use of this information in the indexing and retrieval processes.

The performance of most CBIR systems is inherently constrained by the used low-level features, and cannot give satisfactory results when the user's high level concepts cannot be expressed by low level features. In an attempt to bridge this semantic gap, few approaches that integrate low level visual features and textual keywords have been proposed [3, 4, 5]. Unfortunately, manually labeling each image by a set of keywords is subjective and labor intensive. Moreover, region labeling (as opposed to entire image labeling) may be needed, which makes manual labeling more tedious. To address this issue, few algorithms that can annotate images/regions in an unsupervised (or semi-supervised) manner have been proposed in the past few years [5, 6, 7, 8, 9, 10, 11, 12, 13].

In this paper, we propose a different approach to annotate image regions. Our approach is based on learning associations between low-level visual features and textual keywords through multimedia data mining. These associations would be used to construct a multi-modal thesaurus that relates keywords to visual profiles through frequently co-occurring patterns. A novel algorithm that performs clustering and feature weighting simultaneously is used to learn the associations. First, unsupervised clustering is used to identify representative profiles that correspond to frequent homogeneous regions. The feature discrimination process, embedded in the clustering, would identify the relevant features in each profile. Second, representatives from each cluster and their relevant visual and textual features are used to build a thesaurus. This thesaurus could be used to facilitate many tasks such as *auto-annotation*, *hybrid* searching and browsing, and *query expansion*.

2 Proposed Model

We assume that we have a large collection of images and that each image is annotated by few keywords. We do

*Hichem Frigui is with the Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY 40292, USA (email: h.frigui@louisville.edu).

†Joshua Caudill is with the Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY 40292, USA (email: joshua.caudill@gmail.com).

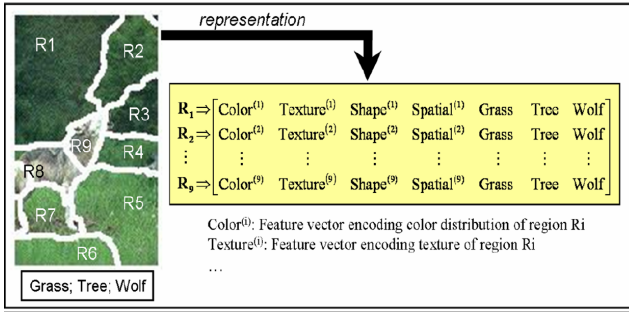


Figure 1: Representation of visual and textual features

not assume that the annotation is complete or accurate. For instance, the image may contain many objects, and we do not have a one to one correspondence between objects and words. This scenario is very common as images with annotations are readily available, but images where the regions themselves are labeled are rare and difficult to obtain.

Developing a learning algorithm using the above data is a challenging task. First, the training data is incomplete as the words are not specified for the different regions. Second, different types of features need to be extracted and combined. Third, the number of keywords is too large to treat the problem as a standard classification problem where each word corresponds to one class. Last, different visual features are not equally important in characterizing different image regions. Highly relevant features for one group of regions may be completely irrelevant for another group.

2.1 Feature extraction and vector representation of images. First, the image is segmented into homogeneous regions based on color and/or texture features. We do not require accurate segmentation as subsequent steps can tolerate missing and over-segmented regions. Then, each region would be described by visual features such as color, texture, shape, and a set of keywords. Let $\{f_{j_1}^{(i)}, \dots, f_{j_{k_j}}^{(i)}\}$ be a k_j dimensional vector that encodes the j^{th} visual feature set of region R_i of a given image. For the keywords, we use the standard vector space model with term frequencies as features [14]. Let $\{w_1, w_2, \dots, w_p\}$ be the representation of the keywords describing the given image (not region-specific). An image that includes n regions (R_1, \dots, R_n) would be represented by n vectors of the form:

$$\underbrace{f_{11}^{(i)}, \dots, f_{1k_1}^{(i)}}_{\text{visual feat 1 of } R_i}, \dots, \underbrace{f_{C1}^{(i)}, \dots, f_{Ck_C}^{(i)}}_{\text{visual feat C of } R_i}, \underbrace{w_1, \dots, w_p}_{\text{Keywords}}, i = 1 \dots n.$$

Fig. 1 illustrates our image representation ap-

proach. We should note here that since the keywords are not specified per region, we duplicate them for each region representation. Our assumption is that, if word w describes a given region R_i , then a subset of its visual features would be present in many instances across the image database. Thus, an association rule among them could be mined. On the other hand, if none of the words describe R_i , then these instances would not be consistent and will not lead to strong associations.

2.2 Learning associations between visual features and keywords.

Using the above image representation, a large collection of images could be mined to extract associations between the different feature sets. For instance, we can extract association rules of the form:

"If color is green and texture is regular, fine, with dominant orientation at 90° then keyword is grass." (shape and spatial location features are not relevant).

Several algorithms could be used to extract association rules from the proposed data representation [15]. However, due to the uncertainties in the images/regions representation (duplicated words, incorrect segmentation, irrelevant features, ...), standard association rule extraction algorithms may not provide acceptable results. In this paper, we present a different approach that is based on simultaneous clustering and feature weighting. Clustering would be used to group similar regions and identify prototypical visual profiles. The feature weighting component would guide the clustering process to identify meaningful clusters with subsets of relevant features.

3 Clustering and feature discrimination

In [16], we proposed an algorithm that performs Simultaneous Clustering and Attribute Discrimination (SCAD). SCAD was designed to search for the optimal clusters' prototypes and the optimal relevance weight for each feature of each cluster. However, for high dimensional data, learning a relevance weight for each feature may lead to overfitting. To avoid this situation, we use a coarse approach to feature weighting (called SCAD_c). Instead of learning a weight for each feature, we divide the set of features into logical subsets, and learn a weight for each feature subset.

Let $\mathcal{X} = \{\mathbf{x}_j \in \mathbb{R}^p | j=1, \dots, N\}$ be a set of N feature vectors. Let $\mathbf{B} = (\beta_1, \dots, \beta_c)$ represent a C -tuple of prototypes each of which characterizes one of the C clusters. Each β_i consists of a set of parameters. Let u_{ij} represent the membership of \mathbf{x}_j in cluster β_i . The

$C \times N$ fuzzy C -partition $\mathbf{U}=[u_{ij}]$ satisfies [17]:

$$(3.1) \quad u_{ij} \in [0, 1], \forall i \quad \text{and} \quad \sum_{i=1}^C u_{ij} = 1 \quad \forall j.$$

Assume that the p features have been partitioned into K subsets: FS^1, FS^2, \dots, FS^K , and that each subset, FS^s , includes k^s features. Let d_{ij}^s be the partial distance between \mathbf{x}_j and cluster i using the s^{th} feature subset. Let $\mathbf{V} = [v_{is}]$ be the relevance weight for FS^s with respect to cluster i . The total distance, D_{ij} , between \mathbf{x}_j and cluster i is then computed by using

$$(3.2) \quad D_{ij}^2 = \sum_{s=1}^K v_{is} (d_{ij}^s)^2.$$

SCAD_c minimizes

$$(3.3) \quad J = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m \sum_{s=1}^K v_{is} (d_{ij}^s)^2 + \sum_{i=1}^C \delta_i \sum_{s=1}^K v_{is}^2,$$

subject to (3.1) and

$$(3.4) \quad v_{is} \in [0, 1] \quad \forall i, s; \quad \text{and} \quad \sum_{s=1}^K v_{is} = 1, \quad \forall i.$$

Minimization of J , with respect to \mathbf{V} yields

$$(3.5) \quad v_{is} = \frac{1}{K} + \frac{1}{2\delta_i} \sum_{j=1}^N (u_{ij})^m \left[D_{ij}^2 / K - (d_{ij}^s)^2 \right].$$

The first term in (3.5), $(1/K)$, is the default value if all K feature subsets are treated equally, and no discrimination is performed. The second term is a bias that can be either positive or negative depending on the relative compactness of s^{th} feature subset.

Minimization of J with respect to \mathbf{U} yields

$$(3.6) \quad u_{ij} = \frac{1}{\sum_{k=1}^C \left(D_{ij}^2 / D_{kj}^2 \right)^{\frac{1}{m-1}}}.$$

Minimization of J with respect to the prototype parameters depends on the choice of d_{ij}^s . For instance, if d_{ij}^s is an Euclidean distance, we get the following update equation for the centers of subset s

$$(3.7) \quad \mathbf{c}_i^s = \frac{\sum_{j=1}^N u_{ij}^m \mathbf{x}_j^s}{\sum_{j=1}^N u_{ij}^m}.$$

SCAD_c is an iterative algorithm that starts with an initial partition and alternates between the update equations of u_{ij} , v_{is} , and \mathbf{c}_i^s .

4 Experiments

4.1 Data set. The proposed approach is validated using a subset of the Corel image collection. We used a total of 6,000 images, where each image is manually labeled by 1 to 7 keywords. A total of 100 words were used. Each image is coarsely segmented by clustering the color distribution. The Competitive Agglomeration (CA)[18] was used to cluster each image into an optimum number of regions. Segmentation of all the images resulted in 31,215 regions. Each region is then characterized by a RGB color histogram (64-dim), HSV color moments (9-dim), edge histogram descriptor [19] (5-dim), wavelet texture (20-dim), shape (5-dim), position (6-dim), and keywords (100-dim).

4.2 Region Clustering. The region feature vectors with the 7 feature subsets were clustered using SCAD_c. For feature subset 1, we use the quadratic distance [20], and for feature subset 7, we use a cosine based distance. For all other features, we use the Euclidean distance. In this application, finding the optimum number of clusters (C) is not critical as long as it is large enough to avoid lumping different profiles into one cluster. Here, we report the results when $C=400$.

Fig. 2 displays six representative regions from two sample clusters. As expected, SCAD_c was successful in partitioning the data into clusters of homogeneous regions. Moreover, SCAD_c identified relevance weights for the different feature subsets in the different clusters. In Fig. 2, for each region we show the keywords that were used to annotate the images from which the region was extracted. As it can be seen, not all words are valid. However, some of the words (e.g. "Sky" in the first cluster) would be more consistent across all the regions of the cluster. Consequently, these words will be the dominant terms in the textual feature set.

For each cluster, we use its visual prototype (closest image to centroid), the features of its centroid, the relevance weights for each feature subset, and the dominant keywords from the textual feature set to form one visual profile. The visual profiles of all clusters constitute the multi-modal thesaurus. Fig. 3 displays the extracted visual profiles of the clusters displayed in Fig. 2. These visual profiles could be treated as a multi-modal thesaurus that could be used to translate from one modality to another.

4.3 Image Annotation. Let c_i^1, \dots, c_i^7 , be the centers of the feature subsets FS^1, \dots, FS^7 of cluster i , and let v_i^1, \dots, v_i^7 be the feature relevance weights of these subsets. These representative features and their relevance weights are used to annotate regions as follows. Given a test image, we first segment it into homoge-

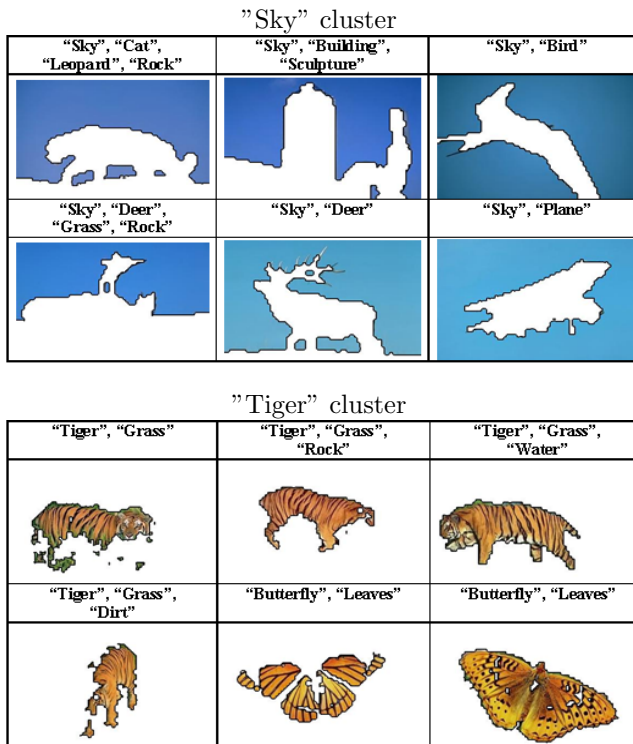


Figure 2: Representative regions from sample clusters.

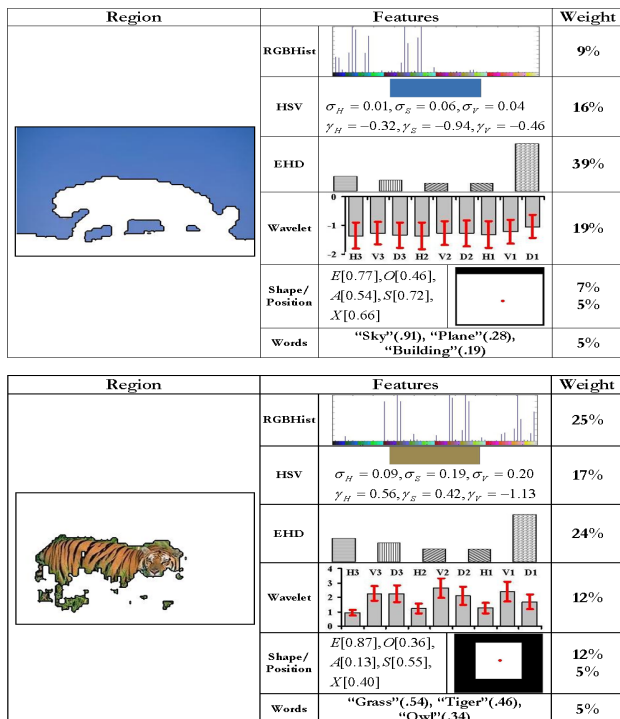


Figure 3: Visual profiles of the clusters in Fig. 2.

neous regions. Then, for each region, R_k , we extract its visual feature, $R_k^f, f = 1, \dots, 6$ (not including words, which are what we are trying to learn), and compare it to the clusters' representatives using

$$D_{ik} = \sum_{f=1}^6 v_i^f \times \text{dist}(R_k^f, c_i^f), i = 1, \dots, C.$$

Based on the distances D_{ik} and the distribution of the clusters, several ways could be used to annotate R_k and assign a confidence value to each label. In this paper, we present a fuzzy labeling approach that uses fuzzy membership functions. In particular, we use an FCM-type membership function [17], where the membership of R_k in cluster i is computed using:

$$(4.8) \quad \mu_i(R_k) = \frac{1}{\sum_{p=1}^C (D_{ik}/D_{pk})^{2/(m-1)}}.$$

The keyword components of the prototypes, c_i^7 , are biased by more frequent words. The standard approach to overcome this bias in text document classification is to weigh the term frequencies by the inverse document frequencies (IDF) [14]. Similarly, we define the inverse cluster frequency (ICF) of word j as

$$(4.9) \quad ICF(w_j) = \log(1 + \frac{C}{C_j}),$$

where C_j is the number of clusters that include the word w_j with a significant frequency. Then, the word frequencies in each cluster would be scaled using

$$(4.10) \quad \tilde{c}_i^7 = ICF \times c_i^7.$$

The confidence value of assigning word w_j to region R_k is computed using

$$(4.11) \quad \text{Conf}(w_j^k) = \sum_{i=1}^C \mu_i(R_k) \times \tilde{c}_{ij}^7$$

Fig. 4 displays samples of test images that were segmented and labeled. For each region, we only show one annotating word that has the highest confidence value. In some instances, the second word has a confidence value that is very close to the top one. In this case, we show both words. As it can be seen, some of these images have a good segmentation and a correct label was assigned to each region. Other images do not have a good segmentation and many of their objects are fragmented. For instance, for the first image, the sky was split into 3 regions. Two of these regions were labeled correctly ('Sky'), and one was labeled incorrectly ('Snow').



Figure 4: Samples of segmented and labeled images.

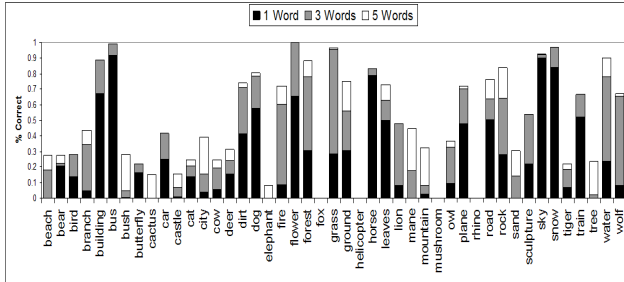


Figure 5: Accuracy of labeled regions for 45 test words using top 1, 2, and 3 labels.

4.4 Validation. To validate the proposed method two aspects need to be evaluated: how accurate is the algorithm in labeling regions and how does it compare to existing methods.

To verify the quality of labels assigned to each region 45 words were chosen that range in frequency throughout the data. For each word, 25 test regions were manually extracted from images that conceptually represent the label. These test regions depict diverse words. Fig. 5 shows the *region accuracy* when one, three, and five words caption a region. *Region accuracy* is defined as the percentage of "correct" captions.

The results, given in Fig. 5, vary with words based on their respective frequency. The frequent words are presented more in the clustering and, as such, have a higher correct percentage. Additionally there are also some words that are simply un-predictable; they are either never used or always used in the wrong region.

In the second experiment, we compare our approach to *CorrSvd* [21]. For the proposed approach, we select M_r words per region, and all words were combined to form the image captions. This means that the number of captions returned per image can vary, as the number of regions per image vary. The *CorrSvd* method returns a fixed amount of captions, and as such, the average amount of words returned in *RegionLabel* is used as the *CorrSvd* M value. In testing the average number of regions is five, so all reference to M will be the number of words returned per region, M_r , multiplied by five for

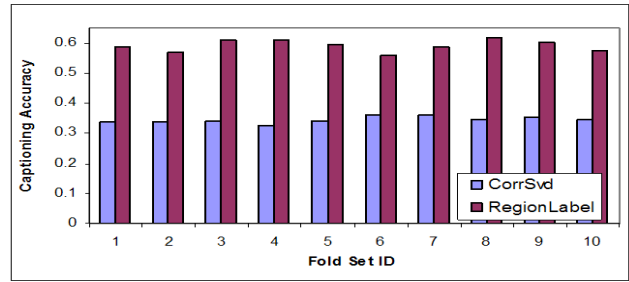


Figure 6: Captions accuracy of proposed method vs *CorrSVD* one word per region (5 average words)

CorrSvd.

This experiment is performed using 10-fold cross validation where each fold is trained on 5,400 images and tested with the remaining 600 images. The *captioning accuracy* for a test image is measured as

$$(4.12) \quad S = m_C / m_T,$$

where m_C is the number of correctly captioned terms in an image and m_T is the total number of truth terms for the image.

Fig. 6 compares the proposed *RegionLabel* with the *CorrSvd* algorithm. The proposed method achieves an improvement around 24.7% absolute accuracy or 71.9% relative improvement over the *CorrSvd* for $M = 1$. Experiments were also conducted for $M = \{2, 3\}$ and showed similar results.

Another measurement of the performance between methods is the recall and precision values for each word (Fig. 7). While *CorrSvd* can better represent highly frequent words, *RegionLabel* can predict a higher percentage with more consistency.

5 Conclusions

We have presented an unsupervised approach that extracts representative visual prototypes from large collections of images through a process of clustering and unsupervised feature selection. This approach consistently outperforms the state-of-the-art correlation translation in captioning accuracy. In addition to summarizing the large number of regions by few visual prototypes, the identified clusters could be used to reveal inter- and intra-modality correlations. In particular, the inter-modality correlation could be used to extract associations between visual profiles and textual keywords. These associations, along with the cluster-dependent feature relevance weights, could be used to build a multi-modal thesaurus that could serve as a foundation for inter-modality translation, and for hybrid navigation

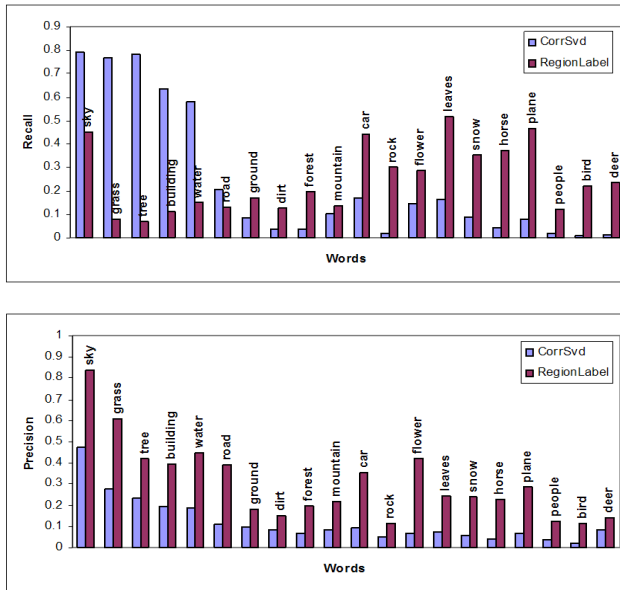


Figure 7: Recall and precision of the top 20 frequent words using $M=1$.

and search in content-based image retrieval. For instance, a textual query using the terms "grass" could be expanded to include the associated visual features. Thus, allowing the user to use keywords to query unlabeled images. Future work will include the construction of a much larger thesaurus and demonstration of its application to hybrid query and navigation in CBIR.

Acknowledgments

This material is based upon work supported by the NSF under Grant No. IIS-0514319.

References

- [1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 22, no. 12, 2000.
- [2] A. Del Bimbo, *Visual Information Retrieval*, Morgan Kaufmann, 1999.
- [3] S. Sclaroff, M. Cascia, and S. Sethi, "Unifying textual and visual cues for content-based image retrieval on the world wide web," *Computer Vision and Image Understanding*, vol. 75, no. 1/2, pp. 86–98, 1999.
- [4] 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.
- [5] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, "Matching words and pic-

- tures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.
- [6] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence & wordnet," in *ACM Multimedia*, 2005, pp. 706–715.
- [7] Y. Mori, H. Takahashi, and R. Oka, "Image to word transformation based on dividing and vector quantizing images with words," in *First Int. Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
- [8] Jia Li and James Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, 2003.
- [9] Jianping Fan, Yuli Gao, Hangzai Luo, and Guangyuo Xu, "Automatic image annotation by using concept-sensitive salient objects for image content representation," in *ACM SIGIR*, 2004, pp. 361–368.
- [10] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *ACM SIGIR*, 2003.
- [11] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *NIPS*, 2003.
- [12] F. Monay and D. Gatica-Perez, "On image auto-annotation with latent space models," in *ACM Multimedia*, 2003, pp. 275–278.
- [13] W. Liu and X. Tang, "Learning an image-word embedding for image auto-annotation on the nonlinear latent space," in *ACM Multimedia*, 2005, pp. 451–454.
- [14] G. Salton and M.J. McGill, *An Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
- [15] T. P. Hong and C. S. Kuo nd S. C. Chi, "Mining association rules from quantitative data," *Intelligent Data Analysis*, vol. 3, pp. 363–376, 1999.
- [16] Hichem Frigui and Olfa Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognition Journal*, vol. 37, pp. 567–581, 2004.
- [17] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.
- [18] H. Frigui and R. Krishnapuram, "Clustering by competitive agglomeration," *Pattern Recognition*, vol. 30, no. 7, pp. 1223–1232, 1997.
- [19] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG 7: Multimedia Content Description Language*, John Wiley, 2002.
- [20] J. Hafner, H. Sawhney, W. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. Patt. Analysis Mach. Intell.*, vol. 17, pp. 729–736, 1995.
- [21] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in *ICME Conf.*, 2004.