# Clustering Art

Kobus Barnard, Pinar Duygulu, and David Forsyth
Computer Division, University of California, Berkeley
{kobus, duygulu, daf}@cs.berkeley.edu

## Abstract

*We extend a recently developed method [1] for learning the semantics of image databases using text and pictures. We incorporate statistical natural language processing in order to deal with free text. We demonstrate the current system on a difficult dataset, namely 10,000 images of work from the Fine Arts Museum of San Francisco. The images include line drawings, paintings, and pictures of sculpture and ceramics. Many of the images have associated free text whose varies greatly, from physical description to interpretation and mood.*

*We use WordNet to provide semantic grouping information and to help disambiguate word senses, as well as emphasize the hierarchical nature of semantic relationships. This allows us to impose a natural structure on the image collection, that reflects semantics to a considerable degree. Our method produces a joint probability distribution for words and picture elements. We demonstrate that this distribution can be used (a) to provide illustrations for given captions and (b) to generate words for images outside the training set. Results from this annotation process yield a quantitative study of our method. Finally, our annotation process can be seen as a form of object recognizer that has been learned through a partially supervised process.*

## 1. Introduction

It is a remarkable fact that, while text and images are separately ambiguous, jointly they tend not to be; this is probably because the writers of text descriptions of images tend to leave out what is visually obvious (the colour of flowers, etc.) and to mention properties that are very difficult to infer using vision (the species of the flower, say). We exploit this phenomenon, and extend a method for organizing image databases using both image features and associated text ([1], using a probabilistic model due to Hofmann [2]). By integrating the two kinds of information during model construction, the system learns links between the image features and semantics which can be exploited for better browsing (§3.1), better search (§3.2), and novel applications such as associating words with pictures, and unsupervised learning for object recognition (§4). The system works by modeling the statistics of word and feature occurrence and co-occurrence. We use a hierarchical structure which further encourages semantics through levels of generalization, as well as being a natural choice for browsing applications. An additional advantage of our approach is that since it is a generative model, it contains processes for predicting image components—words and features—from observed image components. Since we can ask if some observed components are predicted by others, we can measure the performance of the model in ways not typically available for image retrieval systems (§4). This is exciting because an effective performance measure is an important tool for further improving the model (§5).

A number of other researchers have introduced systems for searching image databases. There are reviews in [1, 3]. A few systems combine text and image data. Search using a simple conjunction of keywords and image features is provided in Blobworld [4]. Webseer [5] uses similar ideas for query of images on the web, but also indexes the results of a few automatically estimated image features. These include whether the image is a photograph or a sketch and notably the output of a face finder. Going further, Cascia et al integrate some text and histogram data in the indexing [6]. Others have also experimented with using image features as part of a query refinement process [7]. Enser and others have studied the nature of the image database query task [8-10]. Srihari and others have used text information to disambiguate image features, particularly in face finding applications [11-15].

Our primary goal is to organize pictures in a way that exposes as much semantic structure to a user as possible. The intention is that, if one can impose a structure on a collection that “makes sense” to a user, then it is possible for the user to grasp the overall content and organization of the collection quickly and efficiently. This suggests a hierarchical model which imposes a coarse to fine, or general to specific, structure on the image collection.

## 2. The Clustering Model

Our model is a generative hierarchical model, inspired by one proposed for text by Hofmann [2, 16], and first applied to multiple data sources (text and image features) in [1]. This model is a hierarchical combination of the assymetric clustering model which maps documents into clusters, and the symmetric clustering model which models the joint distribution of documents and features (the "aspect" model). The data is modeled as being generated by a fixed hierarchy of nodes, with the leaves of the hierarchy corresponding to clusters. Each node in the tree has some probability of generating each word, and similarly, each node has some probability of generating an image segment with given features. The documents belonging to a given cluster are modeled as being generated by the nodes along the path from the leaf corresponding to the cluster, up to the root node, with each node being weighted on a document and cluster basis. Conceptually a document belongs to a specific cluster, but given finite data we can only model the probability that a document belongs to a cluster, which essentially makes the clusters soft. We note also that clusters which have insufficient membership are extinguished, and therefore, some of the branches down from the root may end prematurely.

The model is illustrated further in Figure 1. To the extent that the sunset image illustrated is in the third cluster, as indicated in the figure, its words and segments are modeled by the nodes along the path shown. Taking all clusters into consideration, the document is modeled by a sum over the clusters, weighted by the probability that the document is in the cluster. Mathematically, the process for generating the set of observations D associated with a document d can be described by

$$P(D \mid d) = \sum_{c} P(c) \prod_{i \in D} \left( \sum_{l} P(i \mid l,c) P(l \mid c,d) \right) \tag{1}$$

where c indexes clusters, i indexes items (words or image segments), and l indexes levels. Notice that *D* is a set of observations that includes both words and image segments.

### 2.1. An Alternative Model

Note that in (1) there is a separate probability distribution over the nodes for each document. This is an advantage for search as each document is optimally characterized. However this model is expensive in space, and documents belonging mostly to the same cluster can be quite different because their distribution over nodes can differ substantially. Finally, when a new document is considered, as in the case with the "auto-annotate" application described below, the distribution over the nodes must be computed using an iterative process. Thus for some applications we propose a simpler variant of the model which uses a cluster dependent, rather than document dependent, distribution over the nodes. Documents are generated with this model according to

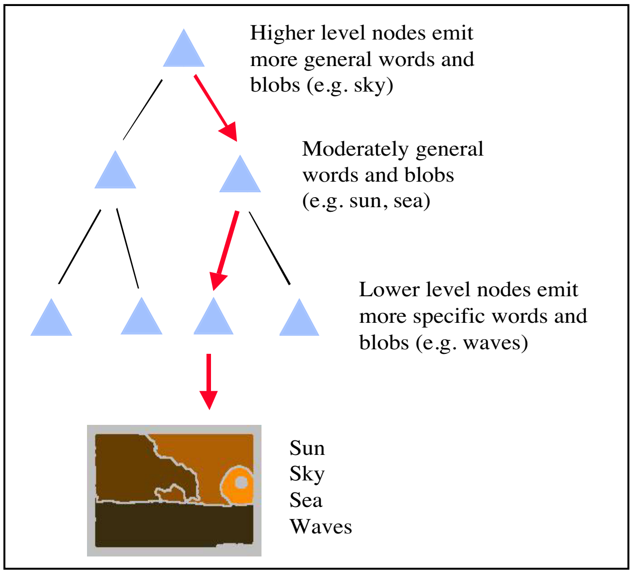

Higher level nodes emit more general words and blobs (e.g. sky)
Moderately general words and blobs (e.g. sun, sea)
Lower level nodes emit more specific words and blobs (e.g. waves)
Sun
Sky
Sea
Waves

Figure 1. Illustration of the generative process implicit in the statistical model. Each document has some probability of being in each cluster. To the extent that it is in a given cluster, its words and segments are modeled as being generated from a distribution over the nodes on the path to the root corresponding to that cluster.

$$P(D) = \sum_{c} P(c) \prod_{i \in D} \left( \sum_{l} P(i \mid l,c) P(l \mid c) \right) \tag{2}$$

In training the average distribution, $P(l \mid c)$, is maintained in place of a document specific one; otherwise things are similar. We will refer to the standard model in (1) as Model I, and the model in (2) as Model II. Either model provides a joint distribution for words and image segments; model I by averaging over documents using some document prior and model II directly.

The probability for an item, $P(i \mid l,c)$, is conditionally independent, given a node in the tree. A node is uniquely specified by cluster and level. In the case of a word, $P(i \mid l,c)$ is simply tabulated, being determined by the appropriate word counts during training. For image segments, we use Gaussian distributions over a number of features capturing some aspects of size, position, colour, texture, and shape. These features taken together form a feature vector X. Each node, subscripted by cluster c, and level l, specifies a probability distribution over image segments by the usual formula. In this work we assume independence of the features, as learning the full covariance matrix leads to precision problems. A reasonable compromise would be to enforce a block diagonal structure for the covariance matrix to capture the most important dependencies.

To train the model we use the Expectation-Maximization algorithm [17]. This involves introducing hidden variables $H_{d,c}$ indicating that training document d is in cluster c, and $V_{d,i,l}$ indicating that item i of document d was generated at level l. Additional details on the EM equations can be found in [2].

We chose a hierarchical model over several non-hierarchal possibilities because it best supports browsing of large collections of images. Furthermore, because some of the information for each document is shared among the higher level nodes, the representation is also more compact than a similar non-hierarchical one. This economy is exactly why the model can be trained appropriately. Specifically, more general terms and more generic image segment descriptions will occur in the higher level nodes because they occur more often.

## 3. Implementation

Previous work [1] was limited to a subset of the Corel dataset and features from Blobworld [4]. Furthermore, the text associated with the Corel images is simply 4-6 keywords, chosen by hand by Corel employees. In this work we incorporate simple natural language processing in order to deal with free text and to take advantage of additional semantics available using natural language tools (see §4). Feature extraction has also been improved largely through Normalized Cuts segmentation [18, 19]. For this work we use a modest set of features, specifically region color and standard deviation, region average orientation energy (12 filters), and region size, location, convexity, first moment, and ratio of region area to boundary length squared.

### 3.1 Data Set

We demonstrate the current system on a completely new, and substantially more difficult dataset, namely 10,000 images of work from the Fine Arts Museum of San Francisco. The images are extremely diverse, and include line drawings, paintings, sculpture, ceramics, antiques, and so on. Many of the images have associated free text provided by volunteers. The nature of this text varies greatly, from physical description to interpretation and mood. Descriptions can run from a short sentence to several hundred words, and were not written with machine interpretation in mind.

### 3.2 Scale

Training on an large image collection requires sensitivity to scalability issues. A naive implementation of the method described in [2] requires a data structure for the vertical indicator variables which increases linearly with four parameters: the number of images, the number of clusters, the number of levels, and the number of items (words and image segments). The dependence on the number of images can be removed at the expense of programming complexity by careful updates in the EM algorithm as described here. In the naive implementation, an entire E step is completed before the M step is begun (or vice versa). However, since the vertical indicators are used only to weight sums in the M step on an image by images bases, the part of the E step which computes the vertical indicators can be interleaved with the part of the M step which updates sums based on those indicators. This means that the storage for the vertical indicators can be recycled, removing the dependency on the number of images. This requires some additional initialization and cleanup of the loop over points (which contains a mix of both E and M parts). Weighted sums must be converted to means after all images have been visited, but before the next iteration. The storage reduction also applies to the horizontal indicator variables (which has a smaller data structure). Unlike the naive implementation, our version requires having both a "new" and "current" copy of the model (e.g. means, variances, and word emission probabilities), but this extra storage is small compared with the overall savings.

## 4. Language Models

We use WordNet [20] (an on-line lexical reference system, developed by the Cognitive Science Laboratory at Princeton University), to determine word senses and semantic hierarchies. Every word in WordNet has one or more senses each of which has a distinct set of words related through other relationships such as hyper- or hyponyms (IS_A), holonyms (MEMBER_OF) and meronyms (PART_OF). Most words have more than one sense. Our current clustering model requires that the sense of each word be established. Word sense disambiguation is a long standing problem in Natural Language Processing and there are several methods proposed in the literature [21-23]. We use WordNet hypernyms to disambiguate the senses.

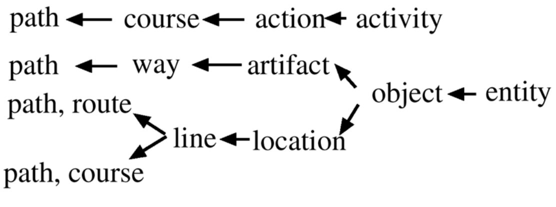


Figure 2: Four possible senses of the word “path”

For example, in the Corel database, sometimes it is possible that one keyword is a hypernym of one sense of another keyword. In such cases, we always choose the sense that has this property. This method is less helpful for free text, where there are more, less carefully chosen, words. For free text, we use shared parentage to identify sense, because we assume that senses are shared for text associated with a given picture (as in Gale et. al's one sense per discourse hypothesis [24]).

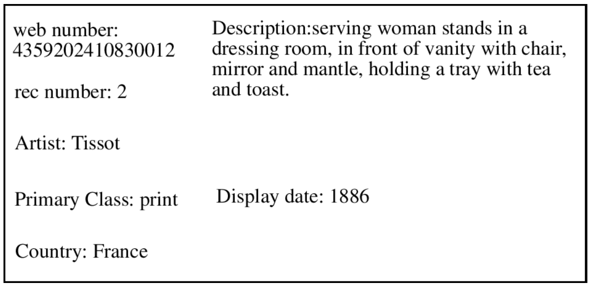
Thus, for each word we use the sense which has the largest hypernym sense in common with the neighboring words. For example, figure 2 shows four available senses of the word path. Corel figure no. 187011 has keywords path, stone, trees and mountains. The sense chosen is path<-way<-artifact<-object.

The free text associated with the museum data varies greatly, from physical descriptions to interpretations and descriptions of mood. We used Brill's part of speech tagger [25] to tag the words; we retained only nouns, verbs, adjectives and adverbs, and only the hypernym synsets for nouns. We used only the six closest words for each occurrence of a word to disambiguate its sense. Figure 3 shows a typical record; we use WordNet only on descriptions and titles. In this case, the word "vanity" is assigned the furniture sense.

For the Corel database, our strategy assigns the correct sense to almost all keywords. Disambiguation is more difficult for the museum data. For example, even though "doctor" and "hospital" are in the same concept, they have no common hypernym synsets in WordNet and if there are no other words helping for disambiguation it may not be possible to obtain the correct sense.

# 5. Testing the System

We applied our method to 8405 museum images, with an additional 1504 used as held out data for the annotation experiments. The augmented vocabulary for this data had 3319 words (2439 were from the associated text, and the remainder were from WordNet). We used a 5 level quad tree giving 256 clusters. Sample clusters are shown in Figure 5. These were generated using Model I. Using Model II to fit the data yielded clusters which were qualitatively at least as coherent.



web number: 4359202410830012

rec number: 2

Artist: Tissot

Primary Class: print

Country: France

Description:serving woman stands in a dressing room, in front of vanity with chair, mirror and mantle, holding a tray with tea and toast.

Display date: 1886

Figure 3: a typical record associated with an image in the Fine Arts Museum of San Francisco collection.

## 5.1. Quality of Clusters

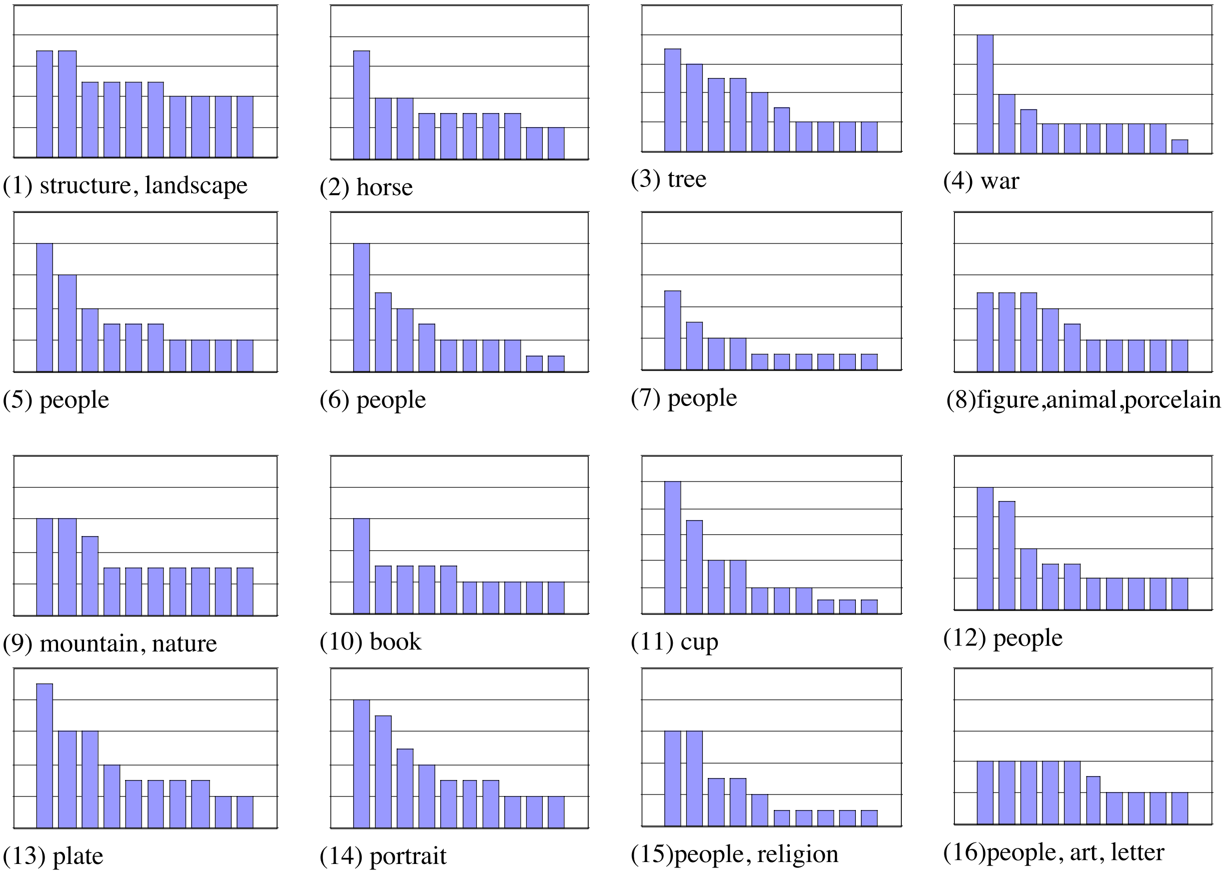
Our primary goal in this work is to expose structure in a collection of image information. Ideally, this structure would be used to support browsing. An important goal is that users can quickly build an internal model of the collection, so that they know what kind of images can be expected in the collection, where to look for them. It is difficult to tell directly whether this goal is met.

However, we can obtain some useful indirect information. In a good structure, clusters would "make sense" to the user. If the user finds the clusters coherent, then they can begin to internalize the kind of structure they represent. Furthermore, a small portion of the cluster can be used to represent the whole, and will accurately suggest the kinds of pictures that will be found by exploring that cluster further.

In [1] clusters were verified to have coherence by having a subject identify random clusters versus actual clusters. This was possible at roughly 95% accuracy. This is a fairly basic test; in fact, we want clusters to "make sense" to human observers. To test this property, we showed 16 clusters to a total of 15 naïve human observers, who were instructed to write down a small number of words that captured the sense of the cluster for each of these clusters. Observers did not discuss the task or the clusters with one another. The raw words appear coherent, but a better test is possible. For each cluster, we took all words used by the observers, and scored these words with the number of WordNet hypernyms they had in common with other words (so if one observer used "horse", and another "pony", the score would reflect this coherence). Words with large scores tend to suggest that clusters are "make sense" to viewers. Most of our clusters had words with scores of eight or more, meaning that over half our observers used a word with similar semantics in describing the cluster. In figure 4, we show a histogram of these scores for all sixteen clusters; clearly, these observers tend to agree quite strongly on what the clusters are "about".

## 5.2. Auto-illustration

In [1] we demonstrated that our system supports "soft" queries. Specifically, given an arbitrary collection of query words and image segment examples, we compute the probability that each document in the collection generates those items. An extreme example of such search is auto-illustration, where the database is queried based


(1) structure, landscape

(2) horse

(3) tree

(4) war

(5) people

(6) people

(7) people

(8)figure,animal,porcelain

(9) mountain, nature

(10) book

(11) cup

(12) people

(13) plate

(14) portrait

(15)people, religion

(16)people, art, letter

Figure 4. Each histogram corresponds to a cluster and shows the score (described in the text) for the 10 words with highest score used to describe that cluster by human observer in that cluster. The scales for the histograms are the same, and go in steps of 2; note that most clusters have words with scores of eight or above, meaning that about half of our 15 observers used that or word with similar semantics to describe the cluster. Number of total words for each cluster varies between 15-35.

on, for example, a paragraph of text. We tried this on text passages from the classics. Sample results are shown in Figure 6.

### 5.3. Auto-annotation

In [1] we introduced a second novel application of our method, namely attaching words to images. Figure 7 shows an example of doing so with the museum data.

## 6. Discussion

Both text and image features are important in the clustering process. For example, in the cluster of human figures on the top left of figure 5, the fact that most elements contain people is attributable to text, but the fact that most are vertical is attributable to image features; similarly, the cluster of pottery on the bottom left exhibits a degree of coherence in its decoration (due to the image features; there are other clusters where the decoration is more geometric) and the fact that it is pottery (ditto text). Furthermore, by using both text and image features we obtain a joint probability model linking words and images, which can be used both to suggest images for blocks of text, and to annotate images. Our clustering process is remarkably successful for a very large collection of very diverse images and free text annotations. This is probably because the text associated with images typically emphasizes properties that are very hard to determine with computer vision techniques, but omits the "visually obvious", and so the text and the images are complementary.

We mention some of many loose ends. Firstly, the topology of our generative model is too rigid, and it would be pleasing to have a method that could search topologies. Secondly, it is still hard to demonstrate that the hierarchy of clusters represents a semantic *hierarchy*.

Our current strategy of illustrating (resp. annotating) by regarding text (resp. images) as conjunctive queries of words (resp. blobs) is clearly sub-optimal, as the elements of the conjunction may be internally contradictory; a better model is to think in terms of robust fitting. Our system produces a joint probability distribution linking image features and words. As a result, we can use images to predict words, and words to predict images. The quality of these predictions is affected by (a) the mutual information between image features and words under the model chosen and (b) the deviance between the fit obtained with the data set, and the best fit. We do not currently have good estimates of these parameters. Finally, it would be pleasing to use mutual information criteria to prune the clustering model.

Annotation should be seen as a form of object recognition. In particular, a joint probability distribution for images and words is a device for object recognition. The mutual information between the image data and the words gives a measure of the performance of this device. Our work suggests that unsupervised learning may be a viable strategy for learning to recognize very large collections of objects.

## 8. Acknowledgements

This project is part of the Digital Libraries Initiative sponsored by NSF and many others. Kobus Barnard also receives funding from NSERC (Canada), and Pinar Duygulu is funded by TUBITAK (Turkey).

## 9. References

[1] K. Barnard and D. Forsyth, "Learning the Semantics of Words and Pictures," *Proc. International Conference on Computer Vision*, pp. II:408-415, 2001.

[2] T. Hofmann, "Learning and representing topic. A hierarchical mixture model for word occurrence in document databases," *Proc. Workshop on learning from text and the web*, CMU, 1998.

[3] D. A. Forsyth, "Computer Vision Tools for Finding Images and Video Sequences," *Library Trends*, vol. 48, pp. 326-355, 1999.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using Expectation-Maximization and its application to image querying," IEEE Transactions on Pattern Analysis and Machine Intelligence *IEEE Transactions on Pattern Analysis and Machine Intelligence*, available in the interim from http://HTTP.CS.Berkeley.EDU/~carson/papers/pami.html.

[5] C. Frankel, M. J. Swain, and V. Athitsos, "Webseer: An Image Search Engine for the World Wide Web," U. Chicago TR-96-14, 1996,

[6] M. L. Cascia, S. Sethi, and S. Sclaroff, "Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web," *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1998.

[7] F. Chen, U. Gargi, L. Niles, and H. Schütze, "Multi-modal browsing of images in web documents," *Proc. SPIE Document Recognition and Retrieval*, 1999.

[8] P. G. B. Enser, "Query analysis in a visual information retrieval context," *Journal of Document and Text Management*, vol. 1, pp. 25-39, 1993.

[9] P. G. B. Enser, "Progress in documentation pictorial information retrieval," *Journal of Documentation*, vol. 51, pp. 126-170, 1995.

[10] L. H. Armitage and P. G. B. Enser, "Analysis of user need in image archives," *Journal of Information Science*, vol. 23, pp. 287-299, 1997.

[11] R. Srihari, Extracting Visual Information from Text: Using Captions to Label Human Faces in Newspaper Photographs, SUNY at Buffalo, Ph.D., 1991.

[12] V. Govindaraju, A Computational Theory for Locating Human Faces in Photographs, SUNY at Buffalo, Ph.D., 1992.

[13] R. K. Srihari, R. Chopra, D. Burhans, M. Venkataraman, and V. Govindaraju, "Use of Collateral Text in Image Interpretation," *Proc. ARPA Image Understanding Workshop*, Monterey, CA, 1994.

[14] R. K. Srihari and D. T. Burhans, "Visual Semantics: Extracting Visual Information from Text Accompanying Pictures," *Proc. AAAI '94*, Seattle, WA, 1994.

[15] R. Chopra and R. K. Srihari, "Control Structures for Incorporating Picture-Specific Context in Image Interpretation," *Proc. IJCAI '95*, Montreal, Canada, 1995.

[16] T. Hofmann and J. Puzicha, "Statistical models for co-occurrence data," Massachusetts Institute of Technology, A.I. Memo 1635, 1998,

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1-38, 1977.

[18] J. Shi and J. Malik., "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888-905, 2000.

[19] Available from http://dlp.CS.Berkeley.EDU/~doron/software/ncuts/.

[20] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: an on-line lexical database," *International Journal of Lexicography*, vol. 3, pp. 235 - 244, 1990.

[21] D. Yarowski, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," *Proc. 33rd Conference on Applied Natural Language Processing*, Cambridge, 1995.

[22] R. Mihalcea and D. Moldovan., "Word sense disambiguation based on semantic density," *Proc. COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, 1998.

[23] E. Agirre and G. Rigau, "A proposal for word sense disambiguation using conceptual distance," *Proc. 1st International Conference on Recent Advances in Natural Language Processing*, Velingrad, 1995.

[24] W. Gale, K. Church, and D. Yarowski, "One Sense Per Discourse," *Proc. DARPA Workshop on Speech and Natural Language*, New York, pp. 233-237, 1992.

[25] E. Brill, "A simple rule-based part of speech tagger," *Proc. Third Conference on Applied Natural Language Processing*, 1992.

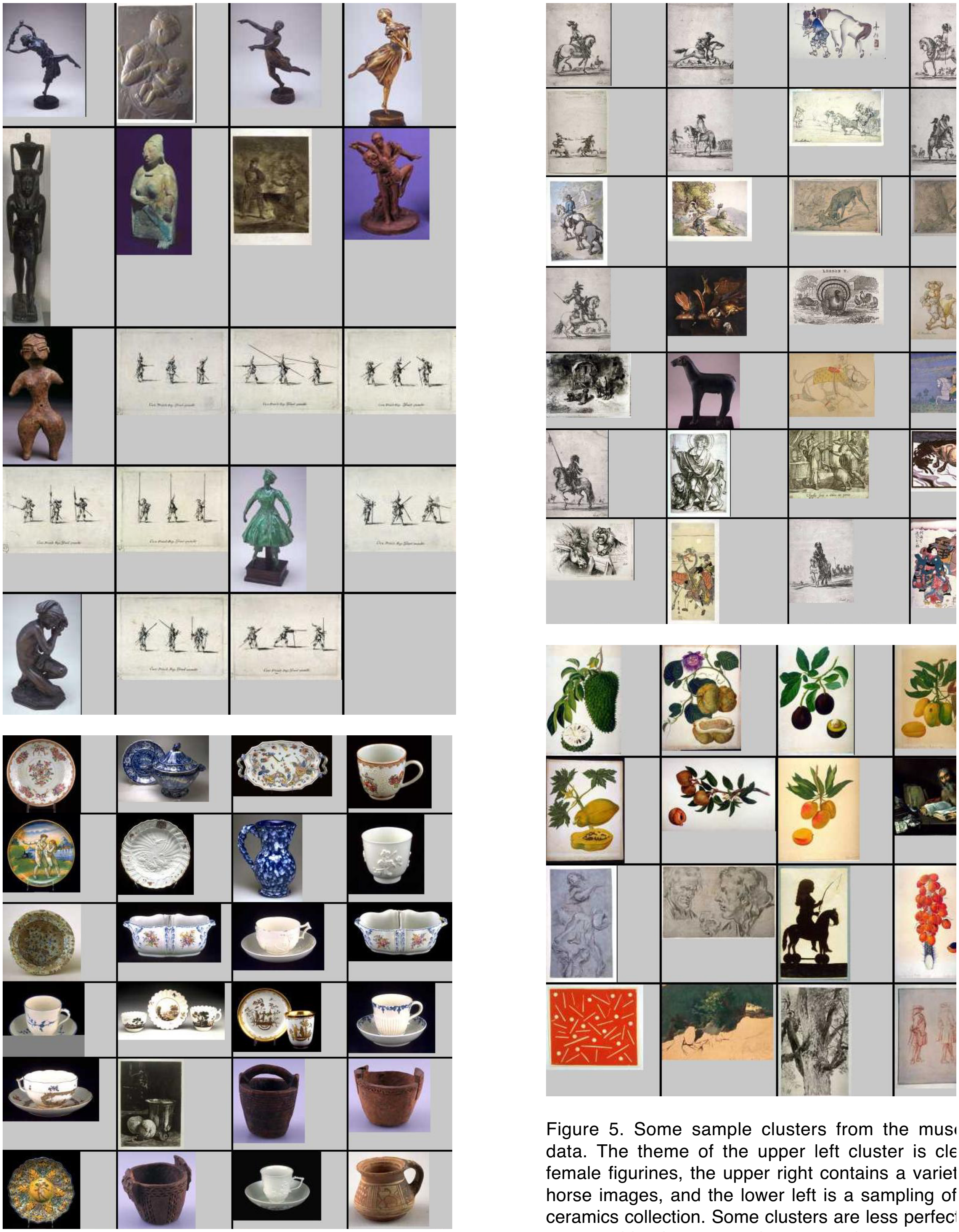
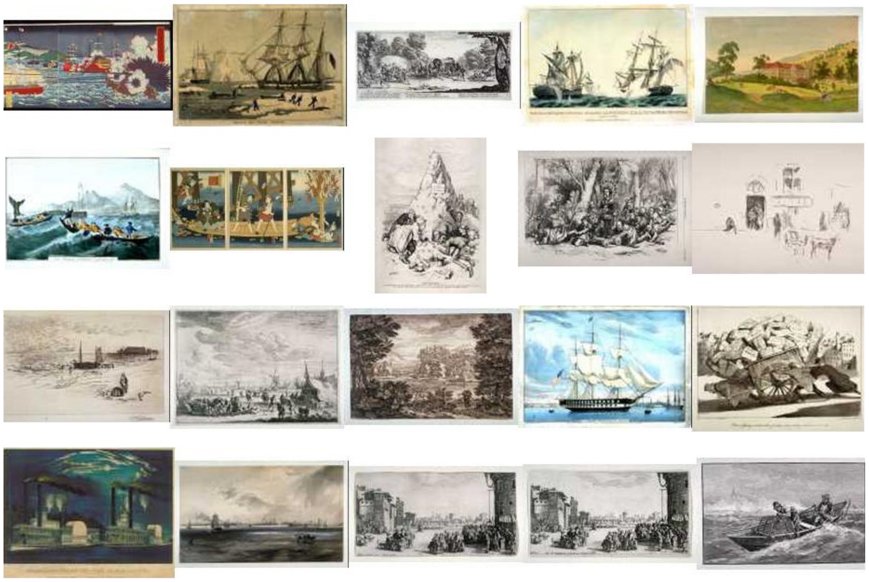


Figure 5. Some sample clusters from the mus
data. The theme of the upper left cluster is cle
female figurines, the upper right contains a variet
horse images, and the lower left is a sampling of
ceramics collection. Some clusters are less perfect
illustrated by the lower right cluster where a variet
images are blended with seven images of fruit.



large importance attached fact old dutch century more command whale ship was per son was divided officer word means fat cutter time made days was general vessel whale hunting concern british title old dutch official present rank such more good american officer boat night watch ground command ship deck grand political sea men mast way professional superior

“The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship was not wholly lodged in the person now called the captain, but was divided between him and an officer called the Specksynder. Literally this word means Fat-Cutter; usage, however, in time made it equivalent to Chief Harpooneer. In those days, the captain's authority was restricted to the navigation and general management of the vessel; while over the whale-hunting department and all its concerns, the Specksynder or Chief Harpooneer reigned supreme. In the British Greenland Fishery, under the corrupted title of Specksioneer, this old Dutch official is still retained, but his former dignity is sadly abridged. At present he ranks simply as senior Harpooneer; and as such, is but one of the captain's more inferior subalterns. Nevertheless, as upon the good conduct …“

Figure 6. Examples of auto-illustration using a passage from Moby Dick , half of which is reproduced to the right of the images. Below are the words extracted from the passage used as a conjunctive probabilistic query.

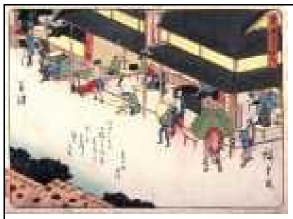
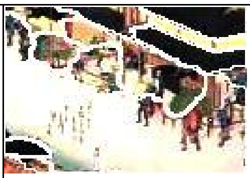
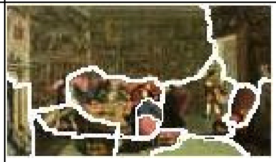
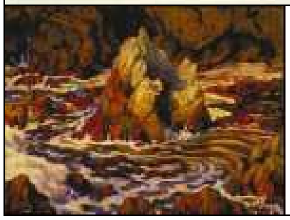
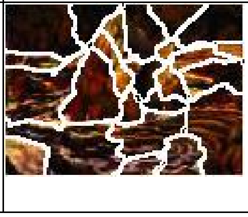
| | | |
|---|---|---|
|  |  | Associated Words<br>KUSATSU SERIES STATION TOKAIDO TOKAIDO GOJUSANTSUGI PRINT HIROSHIGE<br>Predicted Words (rank order)<br>tokaido print hiroshige object artifact series ordering gojusantsugi station facility arrangement minakuchi sakanoshita maisaka a |
| |  | Associated Words<br>SYNTAX LORD PRINT ROWLANDSON<br>Predicted Words (rank order)<br>rowlandson print drawing life_form person object artifact expert art creation animal graphic_art painting structure view |
|  |  | Associated Words<br>DRAWING ROCKY SEA SHORE<br>Predicted Words (rank order)<br>print hokusai kunisada object artifact huge process natural_process district administrative_district state_capital rises |

Figure 7. Some annotation results showing the original image, the N-Cuts segmentation, the associated words, and the predicted words in rank order. The test images were not in the training set. Keywords in upper-case are in the vocabulary. The first two examples are excellent, and the third one is a typical failure.