Multi-modal Clustering for Multimedia Collections

Ron Bekkerman and Jiwoon Jeon
Center for Intelligent Information Retrieval
University of Massachusetts at Amherst, USA
{ronb|jeon}@cs.umass.edu

Abstract

Most of the online multimedia collections, such as picture galleries or video archives, are categorized in a fully manual process, which is very expensive and may soon be infeasible with the rapid growth of multimedia repositories. In this paper, we present an effective method for automating this process within the unsupervised learning framework. We exploit the truly multi-modal nature of multimedia collections—they have multiple views, or modalities*, each of which contributes its own perspective to the collection's organization. For example, in picture galleries, image captions are often provided that form a separate view on the collection. Color histograms (or any other set of global features) form another view. Additional views are blobs, interest points and other sets of local features. Our model, called* Comraf* *(pronounced* Comraf-Star*), efficiently incorporates various views in multi-modal clustering, by which it allows great modeling flexibility. Comraf* is a light-weight version of the recently introduced* combinatorial Markov random field (Comraf). *We show how to translate an arbitrary Comraf into a series of Comraf* models, and give an empirical evidence for comparable effectiveness of the two. Comraf* demonstrates excellent results on two real-world image galleries: it obtains 2.5-3 times higher accuracy compared with a uni-modal k-means.*

1. Introduction

Clustering is the core component in many data management systems. Recent explosion of multimedia information on the WWW demands new clustering techniques that can seamlessly handle complex structures of multimedia data. Multimedia information is intrinsically multi-modal. The word "modality" can be interpreted in many different ways—in this paper, modality means the type of input. For example, image captions and color histograms are different types of input to an image processing system and therefore we consider them as two separate modalities. Each modality gives a different aspect of the data, and each modality has its own dependency relationship with other modalities. For example, image captions tend to describe events captured on the image, while annotations usually list salient objects in the image. Color histograms and texture features convey visual information to the system.

In this paper, we develop an effective multimedia information clustering system based on the recently introduced *combinatorial Markov random field (Comraf)* [4]. Comraf is a special type of a Markov random field (MRF), designed to be a convenient framework for multi-modal learning in general, and for multi-modal clustering in particular. In Comrafs, each data modality is represented by a *single* node, corresponding to a random variable of "rich" structure (called a *combinatorial* random variable); undirected edges are drawn between modalities that stay in a statistical interaction with each other.

We focus on multi-modal clustering of image collections, when multiple views on the collection are available. Image clustering can be a useful component in a retrieval system [7], it can also be a stand-alone application, for example, for constructing semantic groups of image retrieval results [19], or for browsing image collections [1]. Being fully unsupervised, existing clustering methods often demonstrate poor performance. We show that by employing the multi-modal learning paradigm we can significantly improve clustering results. Multiple modalities are a cheap form of supervision: while it is expensive to create a large dataset, each element of which is labeled with its semantic category, it is usually straightforward to obtain another, orthogonal type of labels, over the data modalities.

The idea of clustering images using both low-level image features and surrounding text (i.e. grouping together visually similar and semantically related images) has attracted the close attention of research community. Barnard *et al*. [1] propose a generative hierarchical model for image clustering, in which every node generates words and blobs based on the given probability distributions for that node. Higher level nodes generate more general terms and lower level nodes generate more specific terms. The EM algorithm is used to fit the model. This approach can handle

only two feature types (words, blobs); to handle more types, the model and the learning procedure must be revised.

Cai *et al*. [6] cluster Web image search results using visual, textual and link analysis. They extract text relevant to the image using a vision-based page segmentation algorithm. First, only text and hyperlink data is used to cluster images. The resulting clusters are clustered again using low-level image features. Loeff *et al*. [16] apply a similar approach: they calculate a histogram of gradient magnitude of the pixel values from every *interest point* and then cluster images using these local features with global color histograms and surrounding text. Both Cai *et al*. and Loeff *et al*. use spectral clustering methods, where the affinity scores for every pair of images and every modality must be calculated, resulting in an unrealistically heavy representation.

Bipartite spectral graph partitioning [8] is useful for co-clustering two modalities such as documents and words. Gao *et al*. [11] extend this method to handle one more modality. In their tripartite graph model, nodes are arranged in three layers: words, images and image features. To handle more modalities, Gao *et al*. [12] propose another method that is most closely related to our work: they organize modalities in a *star structure* of interrelationships, where a central modality is connected to all the others. They treat this problem as fusion of multiple pairwise co-clustering problems. Each sub-problem is solved using the bipartite graph partitioning method.

Our approach has a few advantages over the others. First, our method has no practical limitation in the number of modalities as long as the pairwise interaction data is available—addition of a modality increases the computational complexity only linearly. Second, our model can cluster multiple modalities while taking into account other modalities, which do not have to be clustered. Third, our information-theoretic clustering method does not rely on hard-to-obtain affinity matrices of individual modalities. Instead, easily computable contingency tables of interacting modalities are used. Overall, our paper proposes a general framework for clustering multimedia collections, which can be straightforwardly applied to video data, sound tracks, hypertext etc. as well as to any of their combinations.

2. Combinatorial Markov Random Fields

Definition 1 *A* combinatorial random variable X^c *is a discrete random variable defined over a combinatorial set (i.e. a set of all subsets, partitionings, partial orderings etc. of a given finite set).*

In this paper, we focus on the task of *hard* clustering, or partitioning, of a given set. For this task, we define a combinatorial random variable over all the possible partitionings of the set. Note that a combinatorial random variable, while being an ordinary discrete random variable with a finite domain, has a unique property: in most real-world cases, the event space of X^c is so large that the distribution $P(X^c)$ cannot be explicitly specified.

For example, consider a discrete random variable X with three values $\{red, green, blue\}$ selected according to a probability mass function $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\}$. A combinatorial random variable X^c can then take five values: $\{\{red, green, blue\}\}$, $\{\{red, green\}, \{blue\}\}$, $\{\{red, blue\}, \{green\}\}$, $\{\{green, blue\}, \{red\}\}$, and $\{\{red\}, \{green\}, \{blue\}\}$. Assume that the underlying probability mass function of X^c is $\{\frac{1}{5}, \frac{26}{125}, \frac{1}{5}, \frac{1}{5}, \frac{24}{125}\}$, which is unknown to the practitioner. If the task is to find the most probable partitioning, a long sampling process should be applied. This process will be infeasibly long if the event space of X consists of, say, one hundred values.

For each partitioning x^c, we find it useful (see Section 2.2) to define a discrete random variable $\tilde{X}$ over clusters in x^c. Continuing our example above, for $x^c = \{\{red, blue\}, \{green\}\}$ the variable $\tilde{X}$ takes two values: $\{red, blue\}$ and $\{green\}$. Note that the probability of a cluster is a sum of probabilities of its elements, so that the probability mass function of $\tilde{X}$ here is $\{\frac{1}{2} + \frac{1}{6}, \frac{1}{3}\}$.

Definition 2 *A* combinatorial Markov random field (Comraf) *is an MRF, at least one node of which is a combinatorial random variable.*

In this paper we consider only Comraf models, *each* node of which is a combinatorial random variable. Because of the uniqueness of combinatorial random variables, it is impossible to apply existing MRF inference techniques to Comrafs, therefore a specially designed inference method is used, which is based on combinatorial optimization.

One of the plausible characteristics of Comraf models is compactness: a useful Comraf model can consist of just a handful of nodes. Compactness makes the model analysis much easier; also, choosing the best Comraf model (for a particular problem) is a manageable task, as the number of possible variations is relatively small. In our future work, we will apply model learning, which is feasible in Comrafs.

2.1. Comraf* model for multi-modal clustering

Multi-modal (hard) clustering is a problem of *simultaneously* constructing m partitionings of m data modalities, e.g. of images, their colors, their interest points, words in their captions etc. By clustering modalities simultaneously, one would overcome statistical sparseness of the data representation, leading to a dense, smooth joint distribution of the modalities, which would result in potentially more accurate clusterings than the ones obtained separately.

Note that in most real-world cases, a practitioner is interested in clustering only one modality (images, in our case), which we call a *target* modality. This implies that not every

modality *has* to be clustered: if a representation of a modality is dense enough, clustering it may cause an underestimation of the joint (an effect known as *oversmoothing*), which may hurt clustering results of the target modality. For example, if images are distributed over 256 colors, it makes no sense to simultaneously cluster images and colors because the distributions are already dense enough.

In this paper, we present a special case of Comraf models, in which only the target modality is clustered, while the representations of all the other modalities are assumed to be dense enough. Each unclustered modality is represented by an *observed* combinatorial random variable. An observed random variable is a variable whose value is preset and fixed (traditionally, such a variable is shaded on an MRF graph). Recall that a combinatorial random variable is defined over all the possible clusterings of a given set. In case of unclustered modalities, the observed value of a corresponding combinatorial random variable is a clustering of *all singleton* clusters. For example, given a set $\{red, green, blue\}$, the observed value of a corresponding combinatorial random variable is $\{\{red\}, \{green\}, \{blue\}\}$.

Each observed combinatorial random variable of an unclustered modality is connected by an edge with a *hidden* combinatorial random variable of the target modality. Observed nodes are not connected to each other because they are statistically independent by definition. Hence, the resulting topology of the Comraf model is an *asterisk* with the target modality in the center. We call such a model *Comraf**. Examples of Comraf* graphs are given in Figure 1. Despite that only one modality is clustered in Comraf*, it is still a model for *multi-modal* clustering, as multiple modalities are involved in the clustering process.

The general Comraf model, however, takes care of any number of dense *and* sparse random variables. In Section 3.2 we present a Comraf model for simultaneously clustering images and their local features, while incorporating other (unclustered) modalities. Since the simultaneous clustering can be computationally hard, we show how to reduce the computational burden by translating such a Comraf model into a number of Comraf* models, each of which is then optimized sequentially.

2.2. Inference in Comraf and Comraf*

Given a Comraf model over m combinatorial random variables $\mathbf{X^c} = \{X_0^c, X_1^c, \dots, X_{m-1}^c\}$, where X_0^c corresponds to the target modality, the task is to find the most probable instantiation of $\mathbf{X^c}$ (this task is commonly referred to as the *Most Probable Explanation*, or MPE): $\mathbf{x^c}_{MPE} = \arg\max_{\mathbf{x^c}} P(\mathbf{x^c})$.

According to the Hammersley-Clifford theorem [5], the joint distribution $P(\mathbf{x^c})$ is a Gibbs distribution: $P(\mathbf{x^c}) = \frac{1}{Z_\mathbf{f}} \exp \sum_j f_j(\mathbf{x^c})$, where $f_j(\mathbf{x^c})$ are arbitrary potential functions defined over cliques in the Comraf graph, and $Z_\mathbf{f}$ is a normalization factor called a partition function. Let us consider only the smallest cliques, i.e. edges E. If the potential functions are preset and fixed for each edge, then the partition function becomes a constant and thus the MPE problem is solved with:

$$\mathbf{x^c}_{MPE} = \arg\max_{\mathbf{x^c}} \sum_{e_{ij} \in E} f_{ij}(x_i^c, x_j^c). \tag{1}$$

The simplicity of this model allows to choose complex, theoretically justified potential functions. Here we adopt the idea presented in [3] in a similar setting: as a potential function f_j, we choose (weighted) mutual information between variables $\tilde{X}_i$ and $\tilde{X}_j$, which are defined over x_i^c and x_j^c respectively (see the beginning of Section 2). Thus, our objective function is:

$$(x_0^c)_{MPE} = \arg\max_{x_0^c} \sum_{e_{ij} \in E} w_{ij} I(\tilde{X}_i; \tilde{X}_j), \tag{2}$$

subject to $|\tilde{X}_i| = k_i$ and $|\tilde{X}_j| = k_j$. Mutual information between a clustering and another, interacting random variable has a long history of being used in various unsupervised settings, starting with the Information Bottleneck [18], and including image clustering [13]. For a short review, see [4]. Weights w_{ij} are by default set to 1; non-unity weights can be used to bring widely ranging mutual information terms to the same scale (see an example in Section 3.2).

In Comraf*, where all the edges are attached to X_0^c and all the leaves are observed combinatorial random variables, Equation (1) is transformed into:

$$\begin{aligned} &(x_0^c)_{MPE} = \\ &= \arg\max_{x_0^c} \sum_{j=1}^{m-1} f_j(x_0^c, x_j^c) = \arg\max_{x_0^c} \sum_{j=1}^{m-1} w_j I(\tilde{X}_0; X_j), \end{aligned} \tag{3}$$

since $\tilde{X}_j = X_j$ for the unclustered modalities.

2.3. Comraf* optimization procedure

To compute the weighted sum of pairwise mutual information from Equation (3), the following procedure is used. The input of the procedure is an (empirical) joint distribution $P(X_0, X_j)$ of the underlying data of each interacting pair (X_0^c, X_j^c). For a given partitioning x_0^c, the distribution $P(\tilde{X}_0, X_j)$ is computed using the cumulative rule $P(\tilde{x}_0; x_j) = \sum_{x_0 \in \tilde{x}_0} P(x_0, x_j)$. Marginals $P(\tilde{X}_0)$ and $P(X_j)$ are obtained through the marginalization $P(\tilde{x}_0) = \sum_{x_j} P(\tilde{x}_0, x_j)$ and $P(x_j) = \sum_{\tilde{x}_0} P(\tilde{x}_0, x_j)$. Now we have all the ingredients to calculate the mutual information:

$$I(\tilde{X}_0; X_j) = \sum_{\tilde{x}_0, x_j} P(\tilde{x}_0, x_j) \log \frac{P(\tilde{x}_0, x_j)}{P(\tilde{x}_0) P(x_j)}.$$

In most real-world cases, it is infeasible to find the global maximum in Equation (3) because the number of possible

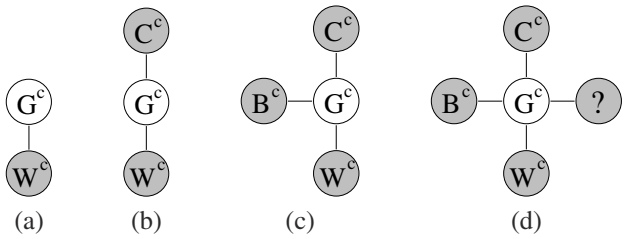


Figure 1. Comraf* models: (a) for images G^c and words W^c from their captions; (b) for images, words and colors C^c; (c) for images, words, colors and blobs B^c; (d) straightforward generalization to any number of modalities.

partitionings x_0^c is exponentially large in the size of the X_0's domain. We apply a local optimization procedure, in which we start with some clustering and greedily optimize the objective while exploring a local neighborhood of the initial configuration. The local search is performed by iteratively transferring each value of X_0 from its current cluster into a cluster such that the objective is maximized.

Note that we can apply Equation (3) only subject to $|\tilde{X}_0| = k_0$, i.e. only when the number of clusters is fixed, otherwise such an optimization can result in a degenerative clustering of all singleton clusters. This does not imply, however, that hierarchical clustering is not applicable to our case. Clusters can be arbitrarily merged or split, after which the number of clusters should be fixed and then it is safe to optimize Equation (3). When the greedy optimization converges to a local maximum, we may again merge or split clusters, after which we proceed with the next iteration of the optimization routine etc.

Thus, we can employ both top-down and bottom-up clustering procedures. In the top-down procedure, we start with one cluster that contains all the values of X_0 and split it until the required number of clusters is obtained (while interleaving with the optimization routine). In bottom-up clustering, we start with all singleton clusters and merge them until, again, reaching the required number of clusters.

Computational complexity of the top-down algorithm is $O(l|X_0|\sum_{j=1}^{m-1}|X_j|)$, and of the bottom-up algorithm $O(l|X_0|^2\sum_{j=1}^{m-1}|X_j|)$, where l is a (fixed) number of clustering iterations. Note that an arbitrary number of leaves (unclustered modalities) can be incorporated into the Comraf* model, while adding new modalities increases the complexity only linearly.

3. Modalities

In this work, along with images, we consider three other modalities. The first one is words from image captions. We remove stopwords and apply simple 's'-stemming (removal of plural suffixes). A joint probability of an image g and a word w is $P(g,w) = \frac{N_{w\in g}}{|W|}$, where $N_{w\in g}$ is the number of occurrences of w in g's caption, $|W|$ is the total number of words. Another modality is colors appearing in images. The joint probability distribution of colors and images is obtained from color histograms, as a number of pixels of color c in image g divided by the total number of pixels in all images. The third modality is blobs, as described below.

3.1. Rectangular blobs

Blobs (or *visual terms*) are a special type of image content representation based on a *fixed vocabulary*. To generate blobs, images are first segmented into regions, which are then clustered across all images. Blobs are the resulting region clusters. Each image is mapped onto the set of blobs which leads to in a representation analogous to the bag-of-words (BOW) in text processing.

Barnard and Forsyth [2] and Duygulu *et al*. [9] segment images into semantically coherent regions using Blobworld and Normalized-Cuts algorithms. Unfortunately, these algorithms do not always produce segmentations accurate enough for further use. Jeon and Manmatha [15] and Feng *et al*. [10] use a rectangular grid to segment images and report better results on an image retrieval task. We apply the same set of blobs as in [10], built using the following procedure. Images are first segmented to regions using a 6 by 4 grid. Then, for each region, a feature vector is constructed that contains texture and color information: Gabor texture filters with 4 orientations and 3 scales are used to construct 12 dimensional texture features; the mean, standard deviation and skewness of RGB and LAB components are computed to build 18 dimensional color features. The resulting 30 dimensional feature vectors are clustered using k-means.

3.2. Blobs constructed by Comraf models

As discussed in Section 3.1 above, a clustering process is involved in constructing blobs from rectangular regions, represented by color and texture features. Naturally, since Comrafs are models for multi-modal clustering, an intrinsic Comraf model can be used for simultaneously clustering images and their regions. Co-clustering of images and features has been recently described in literature [17], however, Comrafs have an additional power over co-clustering methods: Comrafs can incorporate multiple modalities, both sparse (that are to be clustered) and dense (that are not).

Figure 2 (left) shows a Comraf model for clustering images G simultaneously with their regions R, taking into account color C and texture T information of the regions, as well as the colors and caption words W of the images. Obviously, more edges and nodes can be added to the model, depending on the data availability.

In Section 2.3 we mentioned that the input of a Comraf inference procedure is a set of pairwise probability tables $P(X_i, X_j)$ for each edge in the Comraf graph. An interesting case is the (G^c, R^c) edge between image and region

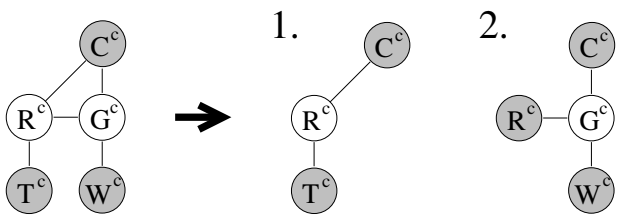


Figure 2. (left) A Comraf model for simultaneously clustering images G^c and their rectangular regions R^c, while taking into account words W^c from image captions, colors C^c and texture data T^c; (right) a translation of this model into a two-step Comraf*: the first Comraf* is for clustering regions into blobs, whereas the second Comraf* is for clustering images based on these blobs.

combinatorial random variables in Figure 2 (left). Unlike colors and caption words, each region is unique, so for each region r and each image g, their joint probability is:

$$P(r,g) = \begin{cases} \frac{1}{|R|}, & \text{if } r \in g \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where $|R|$ is the total number of regions in the dataset. Such a probability mass function is useless for clustering regions, because only regions that belong to the same image can be clustered together. A possible way to resolve this problem would be to estimate this probability by giving a portion of its mass to $P(r,g)$ even if $r \notin g$. Such an estimation can be made based on computing similarities between regions of various images, which is computationally hard: $O(|R|^2|r|)$, where $|r|$ is the size of any region.

Comrafs offer an elegant solution to this problem: since regions are clustered *not only* based on images, but also based on colors and texture, neither of which has this problem, we still can use Equation (4). As long as images are clustered in parallel with regions, Equation (4) allows grouping together regions that belong to the same image *cluster*, as desired. Therefore, we apply the Comraf model from Figure 2 (left) as it is. We choose to cluster images bottom-up and regions top-down. In our objective function (2), we cope with the fact that $I(\tilde{R};\tilde{G})$ is two orders of magnitude larger than $I(\tilde{R};\tilde{C})$ and $I(\tilde{R};\tilde{T})$, by setting the weights of the latter two terms to 100.

The simultaneous clustering of images and regions is a time consuming process: its complexity is $O(|R|\,|G|\,(|C|+|T|+|W|))$. We propose a light-weight version of this model, in which inference is done in *two steps*: first, regions are clustered based on their color and texture features, and then images are clustered based on colors, caption words and region clusters. Such a model is equivalent to a set of two Comraf* models presented in Figure 2 (right). This model's complexity is plausible: $O(|R|\,(|C|+|T|)+|G|\,(|C|+|\tilde{R}|+|W|))$, where $|\tilde{R}|$ is the number of region clusters. Generalizing this setting, it is easy to see that any Comraf can be translated into a series of Comraf* models.

Category	# of images	Category	# of images
Birds	152	Christianity	191
Desert	172	Islam	96
Flowers	165	Judaism	187
Trees	190	Personalities	188
Food	187	Symbols	130
Housing	165	OVERALL:	1823

Table 1. Categories (and their sizes) of the IsraelImages dataset.

4. Experimentation

We experiment with a variety of particular Comraf* models (see examples in Figure 1), as well as with the general Comraf models from Figure 2. The experiments are conducted using our open-source Comraf clustering tool.[1] In all our models, images are clustered agglomeratively. All our results are averaged over 10 independent runs, with the standard error reported. As a baseline, we use the k-means algorithm (`SimpleKMeans` implementation of WEKA[2]), where images are represented as BOW of their captions. Also, our 2-node Comraf* model is equivalent to the hard-clustering version of Information Bottleneck (IB) [18] (see [4] for discussion), hence we use it as our baseline as well.

For evaluation of our clustering results, we use image clustering *accuracy*. Let $\mathcal{C}$ be the set of ground truth categories. For each image cluster $\tilde{g}$, let $\mu_{\mathcal{C}}(\tilde{g})$ be the maximal number of elements of $\tilde{g}$ that belong to one category. Then, the precision of $\tilde{g}$ with respect to $\mathcal{C}$, is defined as $Prec(\tilde{g},\mathcal{C}) = \mu_{\mathcal{C}}(\tilde{g})/|\tilde{g}|$. The micro-averaged precision of the entire clustering g^c is: $Prec(g^c,\mathcal{C}) = \sum_{\tilde{g}} \mu_{\mathcal{C}}(\tilde{g}) / \sum_{\tilde{g}} |\tilde{g}|$, which is the portion of documents that belong to the dominant categories. For all our experiments, we fix the number of clusters to be equal to the number of categories, thus $Prec(g^c,\mathcal{C})$ equals clustering *accuracy*.

4.1. Datasets

We demonstrate the performance of our clustering methods on two datasets: a subset of the benchmark Corel dataset and a new multimedia dataset, which we refer to as *IsraelImages*, collected by us especially for this work.

The Corel subset[3] has already been used in various previous research projects [9, 14, 10]. The dataset consists of 5,000 images from 50 Corel Stock Photo CDs. Each CD contains 100 images on the same topic, such as "Sunrises and Sunsets", "Mountains of America" and "Wild Animals". Every image has a caption and an annotation. The caption is a brief description of the scene and the annotation is a list of objects that appear in the image. An example of an image caption is "Man And Boy Fishing Mountain

[1] `comraf.sourceforge.net`
[2] `cs.waikato.ac.nz/ml/weka`
[3] `kobus.ca/research/data/eccv_2002`

Method	Accuracy
k-means: images over caption words	22.0%
IB: images/caption words	$44.2 \pm 1.0\%$
IB: images/colors	$24.4 \pm 0.2\%$
Comraf*: images/words/colors	$54.2 \pm 0.9\%$
Two-step Comraf*: Figure 2 (right)	$69.0 \pm 0.6\%$
General Comraf: Figure 2 (left)	$68.6 \pm 1.0\%$

Table 2. Clustering results on the IsraelImages dataset. All IB/Comraf results are averaged over 10 independent runs with the standard error of the mean reported after the '$\pm$' sign.

Lake", while "Tree People Mountain Water" is an annotation for this image. Overall 371 words are used to annotate the collection. The original dataset has 4,500 training images and 500 test images. Since our model does not require training, we use 4,500 training images for our experiments and save the remaining 500 images for future use.

The second dataset consists of 1823 images downloaded from IsraelImages.com. The images reflect main aspects of Israel scenery/society and are grouped into 11 categories (see Table 1). Each image is 375 by 250 pixels and has a 1 to 18 words long caption. This dataset is available to the research community.[4]

4.2. Results and discussion

Our results on the IsraelImages dataset are reported in Table 2. Adding the color modality to the caption BOW improves the clustering result by 10% (on an absolute scale), whereas adding the regions (in a 2-step Comraf* scheme) leads to an additional 15% improvement. These findings demonstrate the value of multi-modal setting in image clustering. The general Comraf model from Figure 2 (left) is not able to outperform the 2-step Comraf*. This is probably due to the fact that color and texture information is more important for clustering regions than the correspondence between regions and image clusters.

We also experiment with various levels of color granularity in a 3-node Comraf* setting (from Figure 1b)—the results are presented in Figure 3 (left). As can be seen, if the color information is detailed enough (above 216 colors), the difference in the results is statistically insignificant. Figure 3 (center) shows the results of the 2-step Comraf* over various numbers of colors for clustering regions. Generally, less colors are needed for clustering regions than for clustering images: 216 colors appear to be too many.

A summary of our results on the Corel dataset is presented in Table 3. It shows surprisingly similar trends as for IsraelImages. On a 3-node setup with caption words and blobs we obtain 59.4% accuracy, which is especially impressive given that a random assignment of images into 50 clusters would lead to 2% accuracy (our result is 30 times above random). Adding the color modality improves this result only insignificantly (as expected, since blobs already incorporate the color information, among with texture). The success of 3-node and 4-node Comraf* clustering models is also supported by the fact that they outperform a 2-node *supervised* clustering model, in which images are clustered with respect to their annotations assigned by human experts.

[4] www.cs.umass.edu/~ronb/image_clustering.html

Method	Accuracy
k-means: images over caption words	22.0%
IB: images/caption words	$46.6 \pm 0.5\%$
IB: images/colors	$22.5 \pm 0.2\%$
IB: images/blobs (see Section 3.1)	$24.7 \pm 0.3\%$
Comraf*: images/words/colors	$55.3 \pm 0.5\%$
Comraf*: images/words/blobs	$59.4 \pm 0.5\%$
Comraf*: images/words/colors/blobs	$60.1 \pm 0.3\%$
Two-step Comraf*: Figure 2 (right)	$61.2 \pm 0.4\%$
IB: images/annotation words	$58.6 \pm 0.3\%$

Table 3. Clustering results on the Corel dataset. All IB/Comraf results are averaged over 10 independent runs with the standard error of the mean reported after the '$\pm$' sign.

The 2-step Comraf* shows some further (minor) improvement over the 1-step Comraf* models. Here, in contrast to IsraelImages, 8 colors are enough for clustering regions, and adding more colors causes a significant drop in the performance. We suspect that the Corel dataset is "too simple": it contains many images that are almost identical to each other, therefore more advanced clustering models lead to no (or minor) gain over the simpler ones.

Analogously to our IsraelImages experiment with various sizes of color sets, we test various numbers of blobs on Corel. In previous work [9, 14], the number of blobs is set to 500, to (roughly) correspond to the number of annotation keywords. Here we show that 500 blobs are not enough for clustering: when moving from 1000 to 2000 blobs, a significant boost in the system's performance can be seen.

Figures 4 and 5 are illustrations of the quality of multi-modal setup: unrelated groups of images are mixed together when the clustering is based only on caption words, whereas they are nicely separated when a visual modality is added.

5. Conclusion

In this paper, we have introduced the powerful Comraf framework to clustering multimedia collections. We have also proposed a family of lightweight Comraf models called Comraf*, which demonstrate excellent performance on clustering two real-world data collections. To further improve the results, a semi-supervised Comraf setting [4] may be used, in which a few labeled examples are taken into account in the clustering process. We plan to experiment with this setting in our future work.

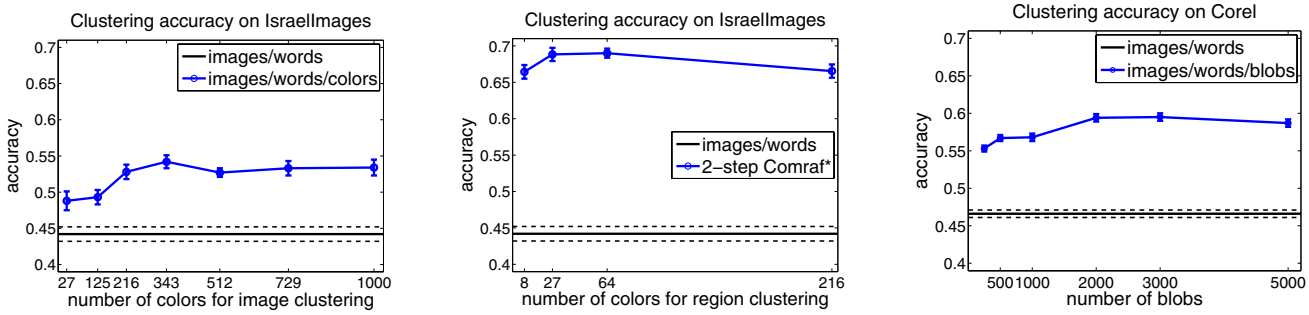


Figure 3. Experimentation with various numbers of: (left) colors on IsraelImages in a 3-node images/words/colors Comraf*; (center) colors for clustering regions in the 2-step Comraf* on IsraelImages; (right) blobs on Corel in a 3-node images/words/blobs Comraf*. Our baseline is the 2-node images/words clustering result. Left and right graphs show the same trend: after reaching a certain number of colors (256) or blobs (2000), the results vary only insignificantly. The central graph, however, shows that too many colors for clustering regions can hurt.

Designing general Comraf models for image clustering (in flavor of the model shown in Figure 2 left) is an ongoing process. Various setups should still be tested, various modalities should be incorporated. For example, *interest points* of images is an important modality not to be ignored. Extensive experimentation will lead to discovering the optimal Comraf setting for clustering multimedia collections.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023.

References

[1] K. Barnard, P. Duygulu, and D. A. Forsyth. Clustering art. In *Proceedings of CVPR*, pages 434–441, 2001. 1

[2] K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In *Proceedings of ICCV-8*, pages 408–415, 2001. 4

[3] R. Bekkerman, R. El-Yaniv, and A. McCallum. Multi-way distributional clustering via pairwise interactions. In *Proceedings of ICML-22*, pages 41–48, 2005. 3

[4] R. Bekkerman, M. Sahami, and E. Learned-Miller. Combinatorial Markov Random Fields. In *Proceedings of ECML-17*, 2006. 1, 3, 5, 6

[5] J. Besag. Spatial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36(2):192–236, 1974. 3

[6] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the 12th international conference on Multimedia*, pages 952–959, 2004. 2

[7] Y. Chen, J. Z. Wang, and R. Krovetz. Clue: Cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing*, 14(8):1187–1201, 2005. 1

[8] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of SIGKDD-7*, pages 269–274, 2001. 2

[9] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of ECCV-7*, 2002. 4, 5, 6

[10] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of CVPR*, pages 1002–1009, 2004. 4, 5

[11] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the 13th international conference on Multimedia*, 2005. 2

[12] B. Gao, T.-Y. Liu, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering. In *Proceeding of SIGKDD-11*, pages 41–50, 2005. 2

[13] J. Goldberger, S. Gordon, and H. Greenspan. Unsupervised image-set clustering using an information theoretic framework. *Transactions on Image Processing*, 15, 2006. 3

[14] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of SIGIR-26*, pages 119–126, 2003. 5, 6

[15] J. Jeon and R. Manmatha. Using maximum entropy for automatic image annotation. In *Proceedings of the 5th International Conference on Image and Video Retrieval*, pages 24–32, 2004. 4

[16] N. Loeff, C. O. Alm, and D. A. Forsyth. Discriminating image senses by clustering with multimodal features. In *Proceedings of COLING/ACL*, pages 547–554, 2006. 2

[17] G. Qiu. Image and feature co-clustering. In *Proceedings of ICPR-17*, pages 991–994, 2004. 4

[18] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method, 1999. Invited paper to the 37th Allerton Conference on Communication, Control, and Computing. 3, 5

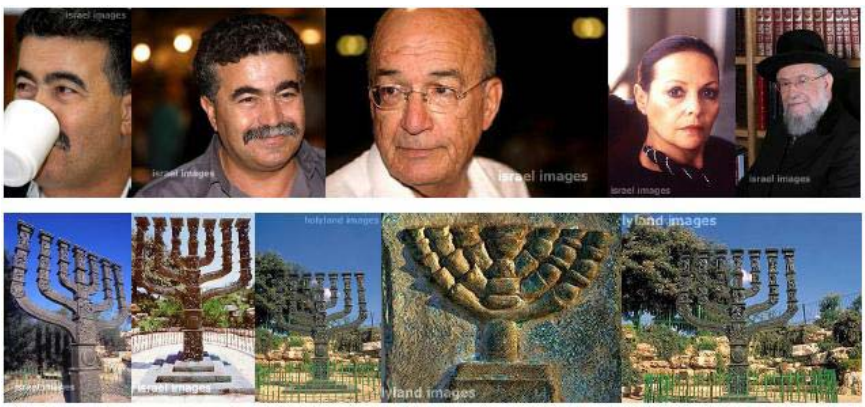
[19] Y. Uematsu, R. Kataoka, and H. Takeno. Clustering presentation of web image retrieval results using textual information and image features. In *Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications*, pages 217–222, 2006. 1

(a) Clustering results using only caption words, Corel dataset

(b) Clustering results using words and blobs, Corel dataset

Figure 4. **Corel dataset.** The first row shows clustering results using only words. Swimmers and swimming tigers are clustered together because they share common terms like "water" and "swim". The second and the third rows show clustering results using both words and blobs. The swimmers and the swimming tigers are now in two different clusters with other similar images.

(a) Clustering results using only caption words, IsraelImages dataset



(b) Clustering results using words and color histograms, IsraelImages dataset

Figure 5. **IsraelImages dataset.** People portraits and pictures of the menorah monument are clustered together using caption words because they have a word 'Knesset' (the Israeli parliament) in common: the individuals are Knesset members, while the menorah monument is placed in front of the Knesset building. The problem is resolved after the color modality is added.