Cross-Modal Image Clustering via Canonical Correlation Analysis

Cheng Jin, Wenhui Mao, Ruiqi Zhang, Yuejie Zhang, Xiangyang Xue
School of Computer Science, Shanghai Key Laboratory of
Intelligent Information Processing, Fudan University, Shanghai, China
{jc, 13210240099, 12210240075, yjzhang, xyxue}@fudan.edu.cn

Abstract

A new algorithm via Canonical Correlation Analysis (CCA) is developed in this paper to support more effective cross-modal image clustering for large-scale annotated image collections. It can be treated as a bi-media multimodal mapping problem and modeled as a correlation distribution over multimodal feature representations. It integrates the multimodal feature generation with the Locality Linear Coding (LLC) and co-occurrence association network, multimodal feature fusion with CCA, and accelerated hierarchical *k*-means clustering, which aims to characterize the correlations between the inter-related visual features in images and semantic features in captions, and measure their association degree more precisely. Very positive results were obtained in our experiments using a large quantity of public data.

Introduction

With the massive explosion of annotated image data on the Web, methods to seamlessly handle the complex structures of image data to achieve more efficient organization and management have become an important research focus (Datta et al. 2008; Rasiwasia et al. 2010). Usually, an annotated image is exhibited in a multimodal form, that is, both semantic and visual. Thus, integrating multimodal information sources to enable cross-modal image clustering has been the core component. However, due to the semantic gap, there may be significant differences and independence among visual images and textual captions, which leads to the huge difficulty and uncertainty in making full use of the relations between the visual features (in images) and semantic features (in captions) (Fan et al. 2012).

To achieve effective cross-modal image clustering, three inter-related issues should be addressed simultaneously: 1) the valid coding and discovery of valuable multimodal features to characterize visual images and textual captions more reasonably, 2) multimodal feature fusion to identify better multimodal correlations between the visual features in images and semantic features in captions, and 3) appropriate clustering methods to speed up the cross-modal clustering process. To address the first issue, it is very important to leverage large-scale annotated images for robust visual coding and semantic mining to achieve more comprehensive multimodal feature representation. To address the second issue, it is very interesting to develop new algorithms for fusing multimodal features and efficiently exploiting the correlations among attributes of different modalities in annotated images. To address the third issue, it is critical to explore an appropriate clustering mechanism with high efficiency but no sacrifice of clustering accuracy.

Based on the above observations, a novel scheme is developed in this paper to facilitate more effective cross-modal image clustering for large-scale annotated image collections. Our scheme significantly differs from other earlier work as follows. a) The visual feature generation based on Locality Linear Coding (LLC) can achieve a smaller quantization error and better coding stability. The semantic feature generation based on a co-occurrence association network can exploit the inter-term statistical co-occurrence associations to learn the semantic feature representations. b) The multimodal feature fusion based on Canonical Correlation Analysis (CCA) has a strong ability to characterize the multimodal correlations between the visual features in images and semantic features in captions. c) The specified accelerated hierarchical *k*-means clustering is exploited to face scalability problems when scaling up to large-scale annotated images. d) A new cross-modal image clustering framework is built by integrating the above feature representation, fusion, and clustering mechanisms.

How to integrate multimodal features for image clustering is an open issue because it is hard to provide a base to multiple similarities among the images which are calculated from multiple features in different modalities. The main contribution of our work is that we effectively apply CCA to enable cross-modal image clustering, which has provided a more reasonable base for us to integrate visual similarity with semantic similarity by determining an optimal correlated projection space. Thus such cross-modal clustering can be treated as a bi-media multimodal mapping problem, and modeled as a correlation distribution over multimodal feature representations, where the most important task is creating the multimodal association between visual and semantic features and measuring the degree to which they are related. We have obtained very positive results in our experiments to demonstrate our observations.

Related Work

Image clustering is not a novel task, but has been the subject of extensive research in areas such as multimedia information processing and retrieval (Chaudhuri et al. 2009; Yang et al. 2010). Earlier research placed the main emphasis on unimodal approaches, where the clustered images shared a single modality (Pan et al. 2004; Goldberger, Gordon, and Greenspan 2006). However, because of considering only visual or textual information in annotated images, such methods often demonstrate a poor performance (Chen, Wang, and Dong 2010; Jia, Salzmann, and Darrell 2011). Recently, closer attention has been given to methods that rely on both low-level visual and high-level semantic features, that is, grouping together visually similar and semantically related images. Thus, there has been increasing research interest in combining multiple information sources and exploiting the multimodal information in annotated images to support precise image clustering.

In recent years, some related research has used the textual information that accompanies an image. Bekkerman and Jeon (2007) introduced the powerful *Comraf* framework to cluster multimedia collections, in which they exploited the multimodal nature of multimedia collections from multiple views or modalities. Moëllic, Haugeard, and Pitel (2008) studied an image clustering process for descriptors of different natures, using textual information and visual features based on bags-of-SIFT descriptors. Rege, Dong, and Hua (2008) addressed the problem of Web image clustering by the simultaneous integration of visual and textual features from a graph partitioning perspective. Yang et al. (2009) presented image clustering as an example to illustrate how unsupervised learning can be improved by transferring knowledge from auxiliary heterogeneous data. Chen, Wang, and Dong (2010) presented a semi-supervised approach for image co-clustering based on non-negative matrix factorization. Fu et al. (2011) presented a multi-modal constraint propagation approach to exploiting pairwise constraints for constrained clustering tasks on multi-modal datasets. Hamzaoui, Joly, and Boujemaa (2011) proposed a new multi-source shared neighbors scheme applied to multi-modal image clustering.

Unfortunately, none of these approaches have provided good solutions for the following three important issues. (a) **Multimodal Information Exploration for Discovering Multimodal Features** – Most existing image clustering techniques typically focus on the traditional visual feature representation and the oversimplified utilization of the limited textual information. They may not stress an in-depth consideration of more reasonable coding methods for the low-level visual features involved in images or exploit all the information to acquire more comprehensive high-level semantic features contained in captions. These approaches may be restricted to the use of the limited annotation tags corresponding to salient objects in the images, while discarding the wealth of information and its connotations in the captions. An attractive paradigm is to improve the unimodal-based model by using multimodal information and performing a deep exploration for the construction of both a visual and semantic feature space. The captioned images could be well represented in this multimodal space and beneficial to image clustering. (b) **Inter-related Correlation Measure for Multimodal Feature Fusion** – Most existing related approaches focus on exploiting information from different modalities separately, and the inter-related correlations between different modalities are completely ignored. Although some advances have been reported on multimodal image clustering, they have usually involved the fusion of features from different modalities into a single vector, or learning different models for different modalities and fusing their output. Cross-modal correlations could provide helpful hints on mining multimodal information for image clustering. Establishing multimodal associations between low-level visual features and high-level semantic attributes may shed light on the captioned image understanding and clustering. Thus, the explicit modeling of cross-modal correlations between visual and semantic features becomes very important. From the viewpoint of multimodal feature exploitation and fusion, it's a significant way to combine both visual and semantic abstractions for images and captions in a joint space and establish a multimodal joint modeling. **(3) Clustering Optimization for Time Efficiency –** Building an appropriate optimized clustering mechanism to make clustering, especially cross-modal clustering, scalable to large-scale annotated images is an important challenge. However, most existing unimodal or multimodal clustering schemes have no in-depth consideration of such a mechanism. Some accelerated clustering strategies can be employed in controlling both the time and memory cost within a more reasonable range.

To tackle the above obstacles, we have developed a novel framework by integrating the visual feature generation with LLC and semantic feature generation with co-occurrence association network (i.e., mining valuable multimodal feature information), the CCA-based multimodal feature fusion (i.e., bridging the semantic gap between visual contents and semantic annotations), and the accelerated hierarchical *k*-means clustering (i.e., fusing the optimization strategy to improve the clustering efficiency). In our study, we realized that an annotated image usually appears with multiple correlated semantic concepts and spans multimodal associations in both visual and conceptual levels. Our CCA-based cross-modal image clustering aims at exploring valuable multimodal correlations involved in images and their annotations to improve the reasoning ability for clustering. It's a new attempt on exploiting such feature generation, fusion and clustering optimization strategies, especially CCA, on cross-modal image clustering.

Multimodal Feature Generation

The multimodal information is the significant expression and exhibition for image content, that is, the visual and semantic description in each annotated image. Thus to acquire the cross-modal correlation between visual images and textual captions, the multimodal features in annotated images should be detected and represented more precisely.

Visual Feature Generation

Each image can be represented by using the SIFT descriptor under the Bag-of-Features (BoF) framework (Csurka et al. 2004). However, such a kind of feature usually occupies the large storage space and requires much more computational cost in a high-dimensional space. Some coding methods have been proposed to project the higher dimensional features into a sparse presentation, such as Vector Quantization (VQ) (Sivic and Zisserman 2003), Sparse coding (SC) (Yang et al. 2009) and Local Coordinate Coding (LCC) (Yu, Zhang, and Gong 2009). However, these manners may produce larger quantization error and lose the correlations between codes. Thus to achieve more precise visual description of images, it's very useful to establish an efficient coding manner to achieve dimension reduction and reduce the computational cost.

First, each SIFT descriptor is viewed as a visual word, and assigned one or a few feature points in the codebook based on the coding algorithm according to the codebook. Compared to the traditional methods, Locality Linear Coding (LLC) can generate the coding result by searching the *k*-nearest neighbors and require the less computational cost (Wang et al. 2010). As the main strategy for feature coding, LLC is incorporated into our visual feature extraction.

Assume that S is a set of D-dimensional visual words extracted from an image with N entries, $S=[s_1, s_2, ..., s_N]\in R^{D\times N}$. Given a codebook with M entries, $B=[b_1, b_2, ..., b_M]\in R^{D\times M}$. Each word is converted into a M-dimensional code to generate the final image representation by Formula (1), which is the approximated method of LLC for fast coding.

$$\arg\min_{C} \sum_{i=1}^{N} \left\| s_i - c_i b_i \right\|^2, \ s.t.\ 1^T c_i = 1, \forall i \qquad (1)$$

where b_i only contains a small group of local bases that are the nearest neighbors of s_i in the codebook B. The coding result C is constructed quickly and then the global features of images can be obtained, while the local information is ignored. In order to characterize the local features for images more accurately, the LLC is combined with the Spatial Pyramid Matching (SPM), which partitions images into increasingly fine spatial sub-regions and computes the local features from each sub-region (Lazebnik, Schmid, and Ponce 2006). Typically, this coding manner consists of $2^l\times2^l$ subregions in each level, l=0, 1, 2. Thus the similarity based on SPM between two images u and v can be calculated as shown in Formula (2).

$$Sim(u, v) = \sum_{l=0}^{L-1} \sum_{i=0}^{2^{2l}-1} Dist\left(F\left(u_l^i\right), F\left(v_l^i\right)\right) \qquad (2)$$

Where $F(u^i_l)$ and $F(v^i_l)$ denote the coding result of SIFT features, that is, the i^{th} sub-region in the l^{th} sub-region level of two images u and v; and L means the max level of SPM.

Semantic Feature Generation

To obtain more accurate semantic formulation for annotated images, the co-occurrence association network and optimal semantic feature coding are both constructed to assist in getting more coherent semantic feature representations.

a. Co-occurrence Association Network Construction

Our co-occurrence association network is automatically generated from large-scale annotated images to characterize the co-occurrence correlations between a large number of semantic concepts of interest (i.e., annotation tags). Such a network can provide a good environment for: (a) discovering more meaningful co-occurrence associations among annotation tags; (b) acquiring more abundant semantic descriptions for annotated images; (c) assisting the semantic feature coding optimization more effectively.

Our association network consists of: (1) large amounts of semantic concepts of interest; and (2) their inter-concept co-occurrence association relations. Intuitively, semantic concepts of interest related to an annotated image should be some concept terms associated with the prominent objects or scenes in this image and play important roles in making discrimination between different objects and scenes in different images. Thus the selected semantic concepts of interest focus on the key terms with high frequency in image annotations from *Flickr* and class names from *Caltech* 256 commonly used in computer vision.

We hope our association network could fuse the important semantic concept association information involved in annotations. Thus the flat construction criterion is considered to achieve more precise characterization of the inter-concept semantic association relations in large-scale annotated image collections. For two semantic concepts of interest, their association relation consists of the flat association relation because of their co-occurrences in image collections, e.g., the higher co-occurrence probability corresponds to the stronger association relation. Even if there are no obvious semantic relations among some co-occurrence concepts, it's still easy for people to relate these concepts together, such as "*monkey*" and "*banana*". Thus the co-occurrence probability between two semantic concepts of interest SC_i and SC_j can be computed according to Formula (3), which indicates what percentage of images annotated by both SC_i and SC_j in the images with SC_i or SC_j.

$$Pco\left(IMG\_Set_{SC_i}, IMG\_Set_{SC_j}\right) = \frac{\left|IMG\_Set_{SC_i} \cap IMG\_Set_{SC_j}\right|}{\left|IMG\_Set_{SC_i} \cup IMG\_Set_{SC_j}\right|} \qquad (3)$$

where IMG_Set_{SCi} and IMG_Set_{SCj} denote two image sets that contain SC_i and SC_j as one annotation tag respectively.

Each semantic concept of interest on the co-occurrence association network can be linked with all the other associated concepts on the network. Eliminating the weak inter-concept links can allow concentrating on the most significant association relations. Thus each concept is linked with the most relevant concepts with stronger flat semantic association relation. The co-occurrence association network for our whole image database is shown in Figure 1.

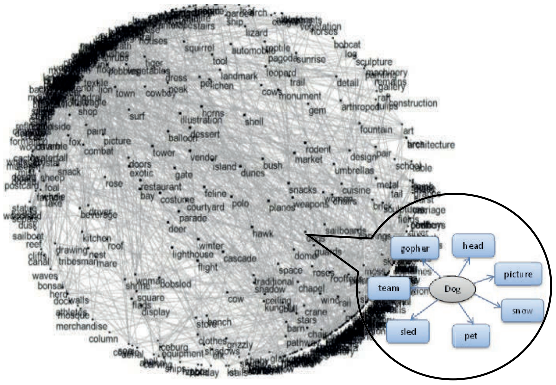


Figure 1. The co-occurrence association network for our image database.

b. Optimal Semantic Feature Coding

To obtain the better quantization for semantic features, the statistical method based on *tf-idf* is incorporated to construct the optimal semantic feature coding, by which the annotation tags (i.e., semantic concepts of interest) could have different importance degrees in different images.

Assume S is an image dataset with the size of N, which contains D different annotation tags in total, Q^i_j means the quantization result of the i^{th} annotation tag Tag_i in the j^{th} image Img_j. Considering that the annotation tags in different images have different importance metrics, we let Q^i_j=0 if $Tag_i \notin Img_j$. Otherwise, if $Tag_i \in Img_j$, we achieve coding the image annotation set according to Formula (4).

$$Q_i^j = \frac{1}{\sum_{k=0}^{D} BL_{\mathrm{Im}\,g_j}(Tag_k)} \cdot \log \frac{N}{\sum_{p=1}^{N} BL_{\mathrm{Im}\,g_p}(Tag_i)} \tag{4}$$

where $BL_{Imgi}(Tag_k)$ is 1 when Tag_k belongs to Img_j, otherwise is 0; and $BL_{Imgp}(Tag_i)$ has the similar definition.

The above algorithm can be considered as the basic model for the initial *tf-idf*-based semantic feature generation. However, the common fact is that there are limited annotation tags in an image caption. If only considering such annotation tags to construct the semantic feature representation, there may cause the restricted or even sparse semantic feature description for an annotated image. It's worth noting that the frequent co-occurrence tags related to the original tags in an image caption can be very beneficial to the semantic information enhancement or supplement for constructing more complete semantic expression for the image. The desirable situation is that more strongly related tags can be explored to the semantic description for an image as possible. Thus, to further acquire more accurate semantic feature representation, the initial semantic feature coding optimization can be implemented based on the co-occurrence associations among different annotation tags on our co-occurrence association network, as shown in Formula (5). The main purpose of such an optimization strategy is to exploit both the explicit semantic information (i.e., the original annotation tags in image caption) and the implicit semantic information (i.e., the related tags with the strong co-occurrence association relations on our co-occurrence association network) for ensuring more accurate and comprehensive formulation of semantic feature.

$$\arg\min_{C^i_j} \sum_{j=1}^{N} \sum_{i=1}^{D} \left(\left\| C^i_j - Q^i_j \right\|^2 + \sum_{k=1}^{D} w_{ik} \left\| C^i_j - C^k_j \right\|^2 \right), s.t., 0 \le C^i_j \le 1 \tag{5}$$

where Q_j can be understood as the initial *tf-idf*-based semantic feature vector for the j^{th} image, Q^i_j is the i^{th} feature in Q_j; C_j can be understood as the optimized semantic feature vector for the j^{th} image, C^i_j is the i^{th} feature in C_j; $||C^i_j\text{-}Q^i_j||^2$ means that the feature coding optimization should keep the characteristics of the initial *tf-idf*-based feature, and the difference between the optimized feature and the initial feature should as small as possible; $||C^i_j\text{-}C^k_j||^2$ means that if an annotation tag has the strong co-occurrence association with the original annotation tag on our co-occurrence association network, it can contribute to the semantic feature representation for the corresponding image and the discrepancy between the semantic features for these two tags should as small as possible; and w_{ik} indicates the weight between two annotation tags (i.e., semantic concepts of interest) of SC_i and SC_k on the co-occurrence association network, and defined as:

$$w_{ik} = \begin{cases} P_{CO}(IMG\_Set_{SC_i}, IMG\_Set_{SC_k}), & P_{CO}(IMG\_Set_{SC_i}, IMG\_Set_{SC_k}) > \tau \\ 0, & P_{CO}(IMG\_Set_{SC_i}, IMG\_Set_{SC_k}) \le \tau \end{cases} \tag{6}$$

where τ is a predefined threshold value.

To guarantee the valid solution for Formula (5), it can be transformed into a simplified form for each image in Formula (7), which is a convex optimization problem.

$$\arg\min_{C^i} \sum_{i=1}^{D} \left(\left\| C^i - Q^i \right\|^2 + \sum_{k=1}^{D} w_{ik} \left\| C^i - C^k \right\|^2 \right), s.t., 0 \le C^i \le 1 \tag{7}$$

Through computing the derivative with respect to C^i and set it to 0, we can obtain Formula (8) after some algebra.

$$C^i = \frac{1}{\left(1 + \sum_{k=1}^{D} w_{ik}\right)} * Q_i + \frac{1}{\left(1 + \sum_{k=1}^{D} w_{ik}\right)} * \sum_{k=1}^{D} w_{ik} * C^k \tag{8}$$

Thus for all the annotation tags in an annotated image, the following Formula (9) can be obtained.

$$\begin{bmatrix} C^1 \\ \vdots \\ C^i \\ \vdots \\ C^D \end{bmatrix} = \begin{bmatrix} \frac{Q^1}{\left(1+\sum_{k=1}^{D} w_{1k}\right)} \\ \vdots \\ \frac{Q^i}{\left(1+\sum_{k=1}^{D} w_{ik}\right)} \\ \vdots \\ \frac{Q^D}{\left(1+\sum_{k=1}^{D} w_{Dk}\right)} \end{bmatrix} + \begin{bmatrix} \frac{w_{11}}{\left(1+\sum_{k=1}^{D} w_{1k}\right)} & \cdots & \frac{w_{1i}}{\left(1+\sum_{k=1}^{D} w_{1k}\right)} & \cdots & \frac{w_{1D}}{\left(1+\sum_{k=1}^{D} w_{1k}\right)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{w_{i1}}{\left(1+\sum_{k=1}^{D} w_{ik}\right)} & \cdots & \frac{w_{ii}}{\left(1+\sum_{k=1}^{D} w_{ik}\right)} & \cdots & \frac{w_{iD}}{\left(1+\sum_{k=1}^{D} w_{ik}\right)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{w_{D1}}{\left(1+\sum_{k=1}^{D} w_{Dk}\right)} & \cdots & \frac{w_{Di}}{\left(1+\sum_{k=1}^{D} w_{Dk}\right)} & \cdots & \frac{w_{DD}}{\left(1+\sum_{k=1}^{D} w_{Dk}\right)} \end{bmatrix} * \begin{bmatrix} C^1 \\ \vdots \\ C^i \\ \vdots \\ C^D \end{bmatrix} \tag{9}$$

The above operation can be iteratively performed until a certain condition is met. The final (C^1, …, C^i, …, C^D) is the optimized feature representation for the given image.

Multimodal Feature Fusion

The multimodal feature fusion based on Canonical Correlation Analysis (CCA) is implemented to exploit multiple features of annotated images and explore the multimodal associations between visual property features and semantic expression features. To make a clear presentation, we first introduce the canonical correlation model, and then develop our CCA-based multimodal feature fusion mechanism.

Revisit of Canonical Correlation Analysis (CCA)

The CCA algorithm is a classic statistical method to multi-view and multi-scale analysis for multiple data sources, which has received much attention in the field of cross-media/cross-modal processing (Cao et al. 2009; Gordoa et al. 2012). It aims at finding linear projections for different types of data with the maximum correlation, which can provide a linear function learning method by projecting different types of data into a high dimension feature space (Gong et al. 2012). The difference between CCA and other closely related approaches is that CCA learns two separate encodings with the objective that the learned encodings are as correlated as possible, and such different objectives may have advantages in different settings. Thus CCA, which could obtain more accurate highly abstract expression of real data via the complex linear transformation, is introduced to make a better solution for multimodal feature fusion in cross-modal image clustering. We emphasize on building a novel cross-modal feature representation for image clustering, which combines canonical correlations of multimodal (visual and semantic) features.

Integrating Multimodal Features with CCA

For integrating multi-modal features to cross-modal features with CCA, it aims at projecting the multimodal features with different modalities on multiple views into a common subspace and making sure that the correlation between visual and semantic features could be maximized.

Let *IMG* be the set of annotated images that consists of *N* samples; $V \in R^{D_V * N}$ is the centered visual feature matrices for *IMG* and $S \in R^{D_S * N}$ represents the centered semantic feature matrices, D_V and D_S are the dimensionality values for these two matrices, generally $D_V \neq D_S$. The following projection can be considered, as shown in Formula (10).

$$\begin{aligned} CCA_{VV} = VV^T, \quad CCA_{SS} = SS^T, \\ CCA_{VS} = VS^T, \quad CCA_{SV} = SV^T \end{aligned} \tag{10}$$

To find a projection relation between the visual feature space and semantic feature space that could maximize the correlation between different feature views, the following Formula (11) is adopted to achieve such a goal.

$$\begin{aligned} &\underset{p \in R^{DV},\, q \in R^{DS}}{\arg\max}\; p^T CCA_{VS} q, \\ &s.t.,\; p^T (CCA_{VV} + \rho I) p = 1,\; q^T (CCA_{SS} + \rho I) q = 1 \end{aligned} \tag{11}$$

where the small regularization factor ρ is used to avoid the numerically ill-conditioned problem; p and q are the projection directions that force the data from V and S into the common space; and I denotes an identity matrix. Hence, we can obtain the set of the projection matrix $P=\{p_1, p_2, \ldots, p_R\}$ and $Q=\{q_1, q_2, \ldots, q_R\}$ by transforming it to a symmetric eigenvalue problem. The i^{th} component p_i and q_i can be calculated by the following Formula (12).

$$\begin{aligned} (CCA_{VV} + \rho I)^{-1} CCA_{VS} (CCA_{SS} + \rho I)^{-1} CCA_{SV}\, p_i = \lambda_i^2\, p_i \\ (CCA_{SS} + \rho I)^{-1} CCA_{SV} (CCA_{VV} + \rho I)^{-1} CCA_{VS}\, p_i = \varphi_i^2\, q_i \end{aligned} \tag{12}$$

where $\lambda^2{}_i$ and $\varphi^2{}_i$ represent the i^{th} eigenvalue. We can project the feature vectors of V and S into a common space based on the matrices of $P \in R^{D_V * R}$ and $Q \in R^{D_S * R}$. Thus we can embed the multi-modal features (i.e., visual features and semantic features) into a subspace that can generate the final cross-modal feature *MF* as shown in Formula (13).

$$MF = \alpha V^T * P + (1 - \alpha) S^T * Q \tag{13}$$

where α is a constant parameter in [0, 1]. Compared to the common feature concatenation, such a linear weighting can achieve the lower dimensional feature representation without the sacrifice of the clustering performance.

In fact, for the visual and semantic features involved in each annotated image, they belong to different feature spaces with different dimensions, and have no direct associations. Our CCA-based multimodal feature fusion can provide a relatively perfect feature representation with associations between various features in different modalities for cross-modal image clustering, which can mitigate the problem of semantic gap to a certain extent and achieve the better clustering with multimodal feature associations.

Accelerated Hierarchical *K*-means Clustering

Since our cross-modal image clustering orients to large-scale annotated image collections, the number of image class may reach up to hundreds, thousands, or even tens of thousands. To describe multimodal attributes for each annotated image, hundreds of feature points of interest can be extracted from each image to compose the associated high-dimensional feature vector representation. Thus for facilitating the better solution of cross-modal image clustering, it's meaningful to build an efficient optimized clustering mechanism with the ability to work with large dataset, high-dimensional data, reasonable time complexity, etc.

The most popular method for image clustering is the *k*-means algorithm, which has been widely applied in the field of image processing (Elkan 2004; Arai and Barakbah 2007; Hamerly 2010). It has the relatively fast computation speed, and can gradually converge to the best situation with the execution of iteration. The superiority of *k*-means is that even if the convergence process for clustering need execute thousands of iterations, the solution close to the final convergence results can be obtained just after dozens of iterations. However, k-means also has the obvious defect, that is, the convergence effect is very sensitive to the selection of initial starting points, which leads to the difficulty to reach global optimum, but only in local minimum. The initial cluster is generated randomly in k-means, thus the unique clustering results cannot be guaranteed. Meanwhile, due to the fact that most distance calculations in k-means are redundant, if a point is far away from a cluster center, it's not necessary to calculate the exact distance between this point and the center in order to confirm that this point should not be assigned to the center. Such redundant calculations may usually consume more computational cost. Thus in our cross-modal image clustering, the Accelerated Hierarchical K-means Clustering (AHKMC) algorithm is adopted to determine the initial centers for the accelerated k-means, which combines the idea of hierarchical clustering and accelerated k-means to satisfy various restrictions under the condition of huge amounts of data and make the suitable optimization for cutting down the computational cost. The basic framework of AHKMC is shown as follows.

```
Algorithm 1 AHKMC(multimodal feature point set MFS)
1:  for $p = 1$ to $P$ do
2:    $CK_p \leftarrow$ random($MFS$, $K$)
3:    $IC_p \leftarrow$ Accelerated K-means Clustering AKMC($MFS$, $CK_p$)
4:    $FC \leftarrow$ all elements in $IC_p$, $p \in [1, P]$
5:  while $|FC| > K$ do
6:    Merge($FC_i$, $FC_j$), $\arg\min_{i \neq j \text{ and } i,j \in [1,|FC|]} CDist(FC_i, FC_j)$
    //CDist( , ) is a distance metric defined on center-center pairs
7:  return AKMC($MFS$, $FC$)
```

```
Algorithm 2 AKMC(multi-modal feature point set MFS, initial center set C)
1:  for $i$=1 to $|MFS|$ do
2:    $Center(i) \leftarrow \arg\min\{Dist(MFS_i, C_j)\}$
    //Center(i) is the index of the center to which $MFS_i$ is assigned
    //Dist( , ) is a distance metric defined on point-center pairs
3:  while not converged do
4:    $C \leftarrow$ new centers $C'$ except the first iteration; UPDATE($Center(i)$, $C$)
5:    for $i$=1 to $|MFS|$ do
6:      if $Dist(C_{Center(i)}, MFS_i)+Dist(C'_{Center(i)}, C_{Center(i)})<BFC_{Center(i)}$, continue
        //$BFC_j$ denotes the first bound of centers
7:      if $Dist(C_{Center(i)}, MFS_i)<BSC_{Center(i)}-1/2*Dist(C'_{Center(i)}, C_{Center(i)})$, continue
        //$BSC_j$ denotes the second bound of centers
8:      for $j$=1 to $|C|$ do
9:        if $Dist(C'_{Center(i)}, MFS_i)<1/2*Dist(C'_{Center(i)}, C'_j)$, continue
10:       else if $Dist(C'_{Center(i)}, MFS_i)>Dist(C'_j, MFS_i)$, $Center(i) \leftarrow j$
11: return $C'$
```

```
Algorithm 3 UPDATE(Center(i), C)
1:  for $j$=1 to $|C|$ do
2:    $C' \leftarrow avg_{Center(i)=j}\{MFS\}$
3:  for $j$=1 to $|C|$ do
4:    $BFC_j \leftarrow 1/2 * \min_{j \neq j' \text{ and } j,j' \in [1,|C'|]} \{Dist(C'_j, C'_{j'})\}$, $BSC_j \leftarrow 1/2 * \min_{j \neq j' \text{ and } j \in [1,|C|], j' \in [1,|C'|]} \{Dist(C_j, C'_{j'})\}$
```

It's important to note that our AHKMC algorithm can achieve the significant acceleration effect through filtering out some avoidable redundant calculations with the characteristic of triangle inequality. The first bound for the first triangle inequality in Line 6 of Algorithm 2 is set based on Lemma 1 in (Elkan 2004; Hamerly 2010). The second bound is a novel boundary check condition for the second triangle inequality in Line 7, which is proposed under consideration for further refining effective calculations. Its derivation process is described in detail as follows. According to the principle of triangle inequality, for any three points x, y and z, $d(x, z)<d(x, y)+d(y, z)$ and $d(x, z)>d(x, y)-d(y, z)$, where $d(\ ,\ )$ is a distance metric function. Assume in Algorithm 2, x is a point, a and b are two centers in C (i.e., the center in the previous iteration), a' and b' are two centers in C' (i.e., the center in the current iteration), as shown in Figure 2. Thus for the triangle $\Delta axb'$, $d(x, b')>d(a, b')-d(x, a)$; and for $\Delta axa'$, $d(a, a')+d(x, a)>d(x, a')$. If $d(a, b')-d(x, a)>d(a, a')+d(x, a)$, i.e., $d(x, a)<1/2*(d(a, b')-d(a, a'))$, then $d(x, b')>d(x, a')$.

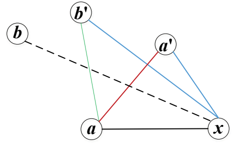


Figure 2. An instantiation for the second bound in Line 7 of Algorithm 2.

Experiment and Analysis

Dataset and Evaluation Metrics

Our dataset is established based on two benchmark datasets of *Corel*30*k* and *Nus-Wide*. To evaluate the performance, we employ the benchmark metric of *Normalized Mutual Information* (*NMI*) and two metrics of *Average Clustering Accuracy* (*ACA*) and *Average Clustering Entropy* (*ACE*). The higher value of *NMI* or *ACA* appears better clustering, but the lower *ACE* indicates the excellent clustering.

$$NMI(\Omega, C) = \frac{I(\Omega, C)}{(H(\Omega) + H(C))/2} \tag{14}$$

where Ω is the cluster set after clustering; C is the ground-truth cluster set; I=(Ω, C) represents the mutual information; and H(Ω) and H(C) denote the entropy values.

$$ACA(\Omega) = \frac{\sum_{i=1}^{|\Omega|} CA(\omega_i)}{|\Omega|}, CA(\omega_i) = \frac{\sum_{k=0}^{|\omega_i|} \delta(TagSet(Img_i), Topic(\omega_i))}{|\omega_i|} \tag{15}$$

$$\delta(TagSet(Imgi), Topic(\omega_i)) = \begin{cases} 1, & iff\ Topic(\omega_i) \in TagSet(Imgi) \\ 0, & ff\ Topic(\omega_i) \notin TagSet(Imgi) \end{cases}$$

where $CA(\omega_i)$ indicates the clustering accuracy for the cluster ω_i, $TagSet(Img_i)$ is the tag set for the image Img_i; $Topic(\omega_i)$ is a concept that has the highest correlation with ω_i,; and $\delta(TagSet(Img_i), Topic(\omega_i))$ is to judge whether the annotation for an image in the cluster contains the most important concept that can be representative of this cluster.

$$ACE(\Omega) = \frac{\sum_{i=1}^{|\Omega|} CE(\omega_i)}{|\Omega|}, CE(\omega_i) = -\sum_{l=1}^{M} (P(tag_l) \ln P(tag_l)) \tag{16}$$

where M is the number of different tags in the same cluster; and $P(tag_l)$ is the occurrence probability of the l^{th} tag in ω_i.

Experiment on Parameter Setting

To show the effect of the important parameter α in Formula (13) on the whole clustering performance, we compare the performance rising speeds for different settings of α, as shown in Figure 3. It can be observed that on *Corel*30*k*, the *NMI*, *ACA* and *ACE* have no significant fluctuations under different settings of α. This implies the smaller correlation gap between the visual and semantic feature spaces obtained on this dataset. However, the different situation appears on *Nus-Wide*. The *NMI* and *ACA* decrease with *α* increasing while the *ACE* rises, which means the weak consistency between two modalities in such a dataset. Thus we set *α* as 0.3 to make a trade-off for better performance.

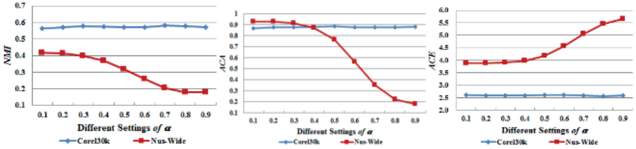


Figure 3. The experimental results for different settings of *α*.

Experiment on Cross-modal Image Clustering

Our modal is created by integrating Multimodal Feature Generation (MFG) [Visual/Semantic Feature Generation (VFG/SFG)], Multimodal Feature Fusion (MFF) and AHKMC. To investigate the effect of each part, we introduce three evaluation patterns: 1) *Baseline*(*MFG_VFG*) [*BMV*]; 2) *Baseline*(*MFG_SFG*) [*BMS*]; and 3) *Baseline*(*MFG_VFG&SFG*)+*MFF* [*BMV&S+MFF*]. The experimental results are shown in Figure 4, 5 and 6.

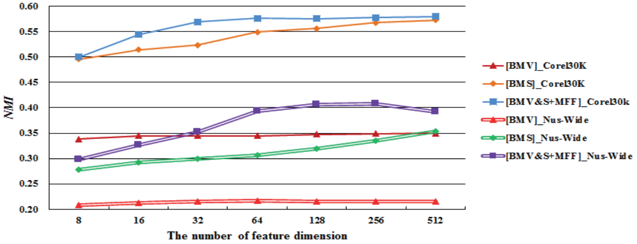


Figure 4. The experimental results for *NMI* on *Corel*30*k* and *Nus-Wide*.

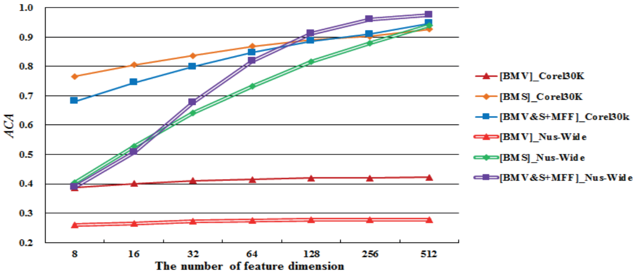


Figure 5. The experimental results for *ACA* on *Corel*30*k* and *Nus-Wide*.

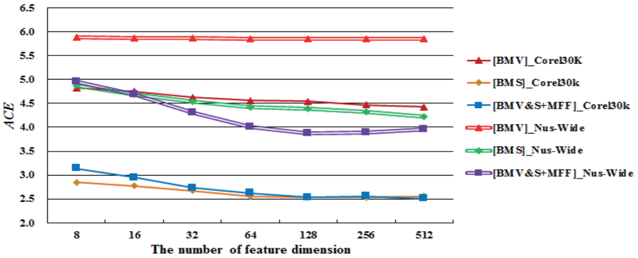


Figure 6. The experimental results for *ACE* on *Corel*30*k* and *Nus-Wide*.

It can be seen from Figure 4 that on *Corel*30*k* and *Nus-Wide*, we can obtain the best *NMI* value of 0.5802 in the evaluation pattern of fusing *Baseline* with *MFG_VFG*, *MFG_SFG* and *MFF*. In comparison with the baseline model using unimodal visual information, the performance could be greatly promoted by successively adding *MFG_SFG* and *MFF*, which confirms the obvious advantage of our cross-modal clustering. Compared two baseline models using unimodal visual and semantic feature information, the performance with semantic feature generation appears better, which shows the beneficial effect of semantic feature on image clustering. Compared two models with unimodal semantic and multimodal feature information, our cross-modal clustering can still gain the significant advantage for the performance on both *Corel*30*k* and *Nus-Wide*. Compared the results for different numbers of feature dimension, we can observe that with the dimension increasing, the performance gradually becomes pretty good. These results are consistent with what we expect considering more valuable multimodal feature information. Compared the results on *Corel*30*k* and *Nus-Wide*, the results on *Nus-Wide* appear less performant due to the differences between these two datasets. *Corel*30*k* is a relatively small-scale dataset with normative and consistent annotations and the images in the same cluster have the higher visual similarity, while *Nus-Wide* from *Flickr* is a relatively large-scale dataset with more obvious characteristics of social annotated images, such as noisy tag, redundant tag, missing tag, etc. As shown in Figure 5 and 6, the same conclusions as above can be drawn from the *ACA* and *ACE* on *Corel*30*k* and *Nus-Wide*, which show the consistence of our model on different indicators. All these observations indicate that the CCA-based multimodal feature fusion have more positive effect on image feature description, which can exactly yield the significant performance improvement more than only considering the unimodal feature representation. It can be found that our cross-modal manner is obviously superior to the traditional unimodal methods and more suitable for the optimized clustering in the complicated environment of social annotated images.

In addition, to explore the efficiency advantage of our framework with AHKMC, we focus on making a comparison of the time cost between the accelerated clustering in AHKMC and the general *k*-means clustering under the same settings of initial condition. The curves of clustering time are shown in Figure 7. With the number of feature dimension increasing, the general *k*-means manner can get the fast clustering within 313 and 16,304 seconds on *Corel*30*k* and *Nus-Wide* respectively. In contrast, the accelerated clustering time in AHKMC can be dramatically decreased to 63% (196 seconds) and 81% (13,200 seconds) of those with the general *k*-means clustering on *Corel*30*k* and *Nus-Wide* respectively, which gives no side-effects on the performance and contributes the better efficiency.

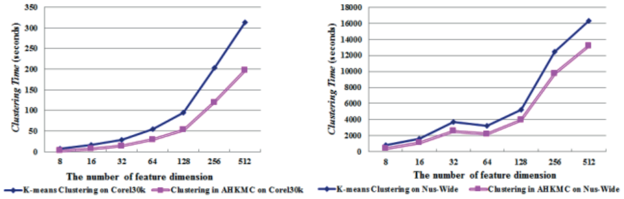
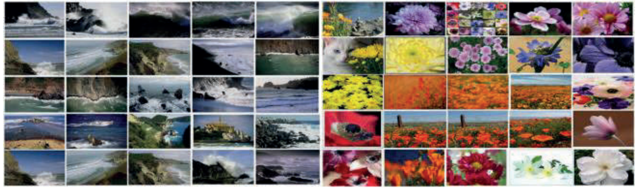


Figure 7. The comparison results for clustering time.

It's worth noting that in our optimized clustering algorithm, the in-depth multimodal analysis is available and presents more impactful ability for discovering the meaningful multimodal features and correlations. Our framework can not only significantly improve the clustering performance via CCA, but also greatly reduce the time cost via AHKMC. The same conclusions can be drawn from the different sets of *Corel*30*k* and *Nus-Wide*, which show the consistence of our model on different data sources. An instantiation of some clusters is shown in Figure 8.



waves *flowers*

Figure 8. An instantiation of some cross-modal clusters.

Comparison with Existing Approaches

Compared to the common image clustering methods in recent years, our approach is a new exploration for taking full advantage of CCA to integrate multimodal information in clustering large-scale annotated images. To give full exhibition to the superiority of our model, we have performed a comparison between our and the other classical approaches for the past few years. Rege et al. (2009) (*Rege*) and Fu et al. (2011) (*Fu*) are analogous with ours, and were accomplished on the same dataset of *Corel*30*k* and a tailored subset of *Nus-Wide* (26,965 images, selecting one image every other ten images). This is because it must take a massive memory to implement *Rege&Fu* on the original *Nus-Wide*. The results are presented in Table 1, which reflect the difference of power among three approaches.

Dataset	Approaches		Feature Dimensions	Evaluation Metrics		
				NMI	***ACA***	***ACE***
Corel*30*k	**Rege *et al.* (2009) (*Rege*)**		**-**	**0.4724**	**0.7096**	**3.3587**
	Fu *et al.* (2011) (*Fu*)		**320**	**0.2314**	**0.4488**	**3.2375**
	Our Approach	***Baseline* [*BMV*]**	**512**	**0.3495**	**0.4219**	**4.4281**
		***Baseline* [*BMS*]**	**512**	**0.5727**	**0.9259**	**2.5578**
		BMV&S* [*MFG*] +*MFF*+*AHKMC	8	0.5004	0.6809	3.1387
			16	0.5443	0.7447	2.9469
			32	0.5692	0.7992	2.7302
			64	0.5766	0.8468	2.6235
			128	0.5755	0.8855	2.5343
			256	0.5780	0.9106	2.5539
			512	**0.5802**	**0.9463**	**2.5041**
Nus-Wide	**Rege *et al.* (2009) (*Rege*)**		**-**	**0.2542**	**0.1258**	**5.2656**
	Fu *et al.* (2011) (*Fu*)		**704**	**0.1365**	**0.2807**	**4.1142**
	Our Approach	***Baseline* [*BMV*]**	**512**	**0.2155**	**0.2768**	**5.8589**
		***Baseline* [*BMS*]**	**512**	**0.3627**	**0.9388**	**4.2252**
		BMV&S* [*MFG*] +*MFF*+*AHKMC	8	0.2973	0.3879	4.9599
			16	0.3269	0.5078	4.6864
			32	0.3515	0.6779	4.3076
			64	0.3937	0.8192	4.0111
			128	0.4066	0.9116	3.8776
			256	0.4081	0.9591	3.8956
			512	**0.3927**	**0.9757**	**3.9612**

Table 1. The comparison results between our and the other approaches.

It can be found from Table 1 that the best performance can be acquired on *Corel*30*k* and *Nus-Wide* by our approach. Compared the results of *Rege*/*Fu* and our baseline model with *VFG*, we can find the better performance appears in *Rege*/*Fu*. However, when integrating *MFG_VFG*, *MFG_SFG*, *MFF* and *AHKMC*, we can acquire the obviously better performance than those of *Rege*/*Fu* with lower feature dimensions. As the numbers of dimension are set as 16 and 64 on *Corel*30*k* and *Nus-Wide* respectively, all the values of *NMI*, *ACA* and *ACE* based on our approach have been dramatically higher than those based on *Rege*/*Fu*. With the feature dimension increasing, our approach reveals more significant advantage. This indicates that our approach is superior to *Rege*/*Fu*, and further confirms the prominent roles of *MFG* and *MFF* in cross-modal clustering, which implies that our model is exactly a better way for determining multimodal associations among images.

Analysis and Discussion

Through the analysis for the clustering results, it can be found that our cross-modal clustering quality is highly related to the following aspects. (1) The clustering effect is closely associated with the preprocessing for annotated images. It's easier to introduce error or noisy detections for visual and semantic feature information, which will seriously affect the whole clustering performance. (2) There is abundant information connotation involved in visual image. It's empirically realized that only using visual features is not sufficient for well formulating the distinguishability among image classes. The intensive visual feature expression can be utilized to further improve the clustering effectiveness and stableness. (3) There are different multimodal attributes among different annotated images. Although more obvious performance superiority has been exhibited via our CCA-based multimodal feature fusion, it's very beneficial to exploit an adaptive fusion between visual and semantic feature for each annotated image. (4) Some annotated images present an extreme vision with wrong or even without any valid annotations. With very limited useful annotation information and too much noises, it's hard for such images to successfully implement precise cross-modal clustering. This may be the most stubborn problem.

Conclusions and Future Work

A new framework is implemented to exploit multimodal correlations among annotated images to enable more effective cross-modal clustering. The in-depth feature analysis is established for characterizing the multimodal attributes for annotated images. The CCA-based multimodal feature fusion is introduced to acquire the specified multimodal feature expressions. The AHKMC manner is exploited to solve the efficiency for optimized clustering. Our future work will focus on making our system available online, so that more Internet users can benefit from our research.

Acknowledgments

This work is supported by National Science & Technology Pillar Program of China (No. 2012BAH59F04), National Natural Science Fund of China (No. 61170095), Zhuoxue Program of Fudan University, and Shanghai Leading Academic Discipline Project (No. B114). Yuejie Zhang is the corresponding author.

References

Arai, K., and Barakbah, A.R. 2007. Hierarchical K-means: An Algorithm for Centroids Initialization for K-means, Reports of the Faculty of Science and Engineering, 36(1):25-31.

Bekkerman, R., and Jeon, J. 2007. Multi-modal Clustering for Multimedia Collections. In *Proceedings of CVPR 2007*, 1-8.

Cao, L.L.; Yu, J.; Luo, J.B.; and Huang, T.S. 2009. Enhancing Semantic and Geographic Annotation of Web Images via Logistic Canonical Correlation Regression. In *Proceedings of MM 2009*, 125-134.

Chaudhuri, K.; Kakade, S.M.; Livescu, K.; and Sridharan, K. 2009. Multiview Clustering via Canonical Correlation Analysis. In *Proceedings of ICML 2009*, 129-136.

Chen, Y.H.; Wang, L.J.; and Dong, M. 2010. Non-negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1459-1474.

Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; and Bray, C. 2004. Visual Categorization with Bags of Keypoints. In *Proceedings of ECCV Workshop on Statistical Learning in Computer Vision*, 59-74.

Datta, R.; Joshi, D.; Li, J.; and Wang, J.Z. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys* (*CSUR*) 40(2), Article 5.

Elkan, C. 2004. Using the Triangle Inequality to Accelerate K-means. In *Proceedings ICML 2003*, 3:147-153.

Fan, J.P.; He, X.F.; Zhou, N.; Peng, J.Y.; and Jain, R. 2012. Quantitative Characterization of Semantic Gaps for Learning Complexity Estimation and Inference Model Selection. *IEEE Transactions on Multimedia* 14(5):1414-1428.

Fu, Z.H.; Ip, H.H.S.; Lu, H.T.; and Lu, Z.W. 2011. Multi-modal Constraint Propagation for Heterogeneous Image Clustering. In *Proceedings of MM 2011*, 143-152.

Goldberger, J.; Gordon, S.; and Greenspan, H. 2006. Unsupervised Image-Set Clustering Using an Information Theoretic Framework. *IEEE Transactions on Image Processing* 15(2):449-458.

Gong, Y.; Lazebnik, S.; Gordo, A.; and Perronnin, F. 2012. Iterative Quantization: A Procrustean Approach to Learning Binary codes for Large-scale Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(12):2916-2929.

Gordoa, A.; Rodriguez-Serrano, J.A.; Perronnin, F.; and Valveny, E. 2012. Leveraging Category-level Labels for Instance-level Image Retrieval. In *Proceedings of CVPR 2012*, 3045-3052.

Hamerly, G. 2010. Making K-means Even Faster. In *Proceedings of SIAM 2010*, 130-140.

Hamzaoui, A.; Joly, A.; and Boujemaa, N. 2011. Multi-source Shared Nearest Neighbours for Multi-modal Image Clustering. *Multimedia Tools and Applications* 51(2):479-503.

Jia, Y.Q.; Salzmann, M.; and Darrell, T. 2011. Learning Cross-modality Similarity for Multinomial Data. In *Proceedings of ICCV 2011*, 2407-2414.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of CVPR 2006*, 2:2169-2178.

Moëllic, P.A.; Haugeard, J.E.; and Pitel, G. 2008. Image Clustering based on a Shared Nearest Neighbors Approach for Tagged Collections. In *Proceedings of CIVR 2008*, 269-278.

Pan, J.Y.; Yang, H.J.; Faloutsos, C.; and Duygulu, P. 2004. Automatic Multimedia Cross-modal Correlation Discovery. In *Proceedings of SIGKDD 2004*, 653-658.

Rasiwasia, N.; Pereira, J.C.; Coviello, E.; Doyle, G.; Lanckriet, G.R.G.; Levy, R.; and Vasconcelos, N. 2010. A New Approach to Cross-modal Multimedia Retrieval. In *Proceedings of MM 2010*, 251-260.

Rege, M.; Dong, M.; and Hua, J. 2008. Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering. In *Proceedings of WWW 2008*, 317-326.

Sivic, J., and Zisserman, A. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of ICCV 2003*, 2:1470-1477.

Wang, J.J.; Yang, J.C.; Yu, K.; Lv, F.J.; Huang, T.; and Gong, Y.H. 2010. Locality-constrained Linear Coding for Image Classification. In *Proceedings of CVPR 2010*, 3360-3367.

Yang, Q.; Chen, Y.Q.; Xue, G.R.; Dai, W.Y.; and Wan, W.G. 2009. Image Co-clustering with Multi-modality Features and User Feedbacks. In *Proceedings of MM 2009*, 1-9.

Yang, Y.; Xu, D.; Nie, F.P.; Yan, S.C.; and Zhuang, Y.T. 2010. Image Clustering Using Local Discriminant Models and Global Integration. *IEEE Transactions on Image Processing* 35(12):2761-2773.

Yang, J.C.; Yu, K.; Gong, Y.H.; and Huang, T. 2009. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In *Proceedings of CVPR 2009*, 1794-1801.

Yu, K.; Zhang, T.; and Gong, Y.H. 2009. Nonlinear Learning Using Local Coordinate Coding. In *Proceedings of NIPS 2009*, 1-8.