IMPROVING IMAGE SIMILARITY MEASURES FOR IMAGE BROWSING AND RETRIEVAL THROUGH LATENT SPACE LEARNING BETWEEN IMAGES AND LONG TEXTS

Yoshitaka Ushiku, Tatsuya Harada, Yasuo Kuniyoshi

The University of Tokyo
Graduate School of Information Science and Technology
{ushiku, harada, kuniyoshi}@isi.imi.i.u-tokyo.ac.jp

ABSTRACT

The amount of multimedia data on personal devices and the Web is increasing daily. Image browsing and retrieval systems in a low-dimensional space have been widely studied to manage and view large numbers of images. It is essential for such systems to exploit an efficient similarity measure of the images when searching for them. Existing methods use the distance in a low-level image feature space as the similarity measure, and therefore, images with different content may be treated as similar images. In this paper, we propose a novel method to improve the similarity measures for images by considering the text surrounding the images. If there is text describing the images, similarities can be measured more effectively by taking into account the text streams. The proposed method improves the image similarity measures based on the latent semantics obtained from the combination of image and text. It should be noted that the text does not need to be clear tags; indeed, any generic Web text is applicable. Moreover, our method can effectively improve the similarities even if only a small portion of the images include textual descriptions. Additionally, the proposed method is scalable as it has linear computational complexity based on the number of images. In the experiments, we compare our method with previous methods using an original dataset in which a portion of the images are annotated by long text. We show that the proposed method can retrieve semantically similar images more precisely than existing methods.

Index Terms— Similarity measure, semantic gap, image retrieval and browsing, pCCA, multimodal learning

1. INTRODUCTION

With the advances in information technology, the amount of multimedia data available for research purposes is continually increasing. For the purposes of management and browsing of large image collections, it is useful to be able to search for images based on similarity.

Existing techniques for image retrieval and browsing measure image similarity and then use the similarity scores to generate graph structures, networks, or low-dimensional distributions [1]. However, it is a well-known fact in the CBIR literature that there is a semantic gap between low-level image descriptions and high-level concepts [2]. As a result, the semantic content of the retrieved images often differs significantly from the query image.

To narrow this gap, many researchers have studied the relationship between images and their metadata. In research on image annotation and retrieval, for example, tags are automatically added to images and used in the retrieval process. Machine learning approaches, however, assume the existence of a tagged training database. Thus, it is necessary to prepare labeled images manually. This task requires a great deal of labor-intensive human effort.

Several techniques have been proposed to collect annotated images from the Web. The work of Fan et al. [3] is an example of Web image retrieval/browsing. In this work, a network of topics, generated by using Flickr image tags, is combined with the distributions of the images to enable a topic-based image search. Unfortunately, most Web images do not have such metadata. Instead, images are typically attached either to a long text, or to no data at all. Even if the image is associated with some text, it is difficult to extract important labels. The problem of finding metadata for Web images is common to most existing research in this area.

In this paper, we propose a method for semantic image retrieval, which exploits a large amount of image data and associated text to improve accuracy. Many web images on news sites and tourism pages, for example, are attached to some text. Images with similar semantic meanings tend to be associated with similar textual data. The semantic gap can be eliminated by learning the similarities based on the latent meaning of image descriptors and text. More importantly, even images that do not have associated text can be learned using this improved similarity measure. In fact, the results of combined learning using Web images and text is expected to produce similarity scores for images on personal devices (cameras, etc.) with no attached text, allowing users to search their own image collections more accurately.

2. RELATED WORK

Our goal is to improve image similarity measures for retrieval and browsing. Generally, a system providing image retrieval/browsing incorporates two elementary techniques; searching for similar images in the entire image collection and the layout thereof on a plane with the proper similarities. Our work focuses on the former technique. In this section, we consider the requirements of retrieving images that are semantically similar to each other, and give an overview of related works.

First, we investigate layout techniques for retrieved images, which are required to remedy the bias of the distributions and the overlap of images in an attempt to maintain their distances. One of the traditional approaches attempts to reduce the overlap using a gradient descent method [4]. Another popular approach is Multidimensional Scaling (MDS) [5], which represents the location of images in the desired dimensional space keeping their dissimilarity. More recent approaches include using a Self Organizing Map [6], or describing each image as a Gaussian with the cost function designed via Renyi entropy of the probability functions [7].

Next, we consider the techniques used to search for similar im-
ages. The requirements for such a system are the following:

**Accuracy** The target images are highly ranked in the retrieved images. The computational cost must be less than or equal to \(O(n)\).

Almost all of the existing works measure similarities using image descriptors. In [4], Principal Component Analysis (PCA) is applied to the features for dimension reduction and similar images are found in the space. In general, however, as images are nonlinearly distributed, retrieval precision is limited. Several works consider these distributions as a low-dimensional manifold in a high-dimensional feature space. Each generates a graph structure and applies either MDS, Stochastic Neighbor Embedding, or Local Linear Embedding in combinations called ISOMAP, ISOSNE, and ISOLLE, respectively [8]. However, these nonlinear multivariate analyses increase the calculation cost and result in a lack of scalability.

In this paper, we focus on the text surrounding images to improve the retrieval precision. As discussed in Section 1, a subset of Web images are described by accompanying text. Image similarities can be improved by considering such text streams. Consequently, we consider obtaining a space based on the latent meanings of such images and accompanying text. A latent space is generated from the low-level features of these images and texts. In this space, image distributions are improved according to the similarities of their corresponding objects. Thus, similar image retrieval can be realized as a neighborhood search.

Consequently, we investigate a probabilistic model that represents the relations between images and accompanying text. We also investigate a machine learning method for application to a model consulting the image annotation/retrieval works, since its goal is to learn the relations between images and labels. There are two types of methods for learning relations between images and labels. One represents the direct relation between them, while the other assumes latent variables and considers two relevancies, that is, between the latent variables and image features and between the latent variables and labels. Unfortunately, the direct manner requires a complicated method, such as a Gaussian mixture model, since the probability distributions are generally complex. Methods of the other type realize learning and recognition with a reduced calculation cost by assuming that two probabilistic relations are independent of each other. Although this is a strong assumption, it usually provides better performance. Consequently, we propose a method based on a model that assumes latent variables above images and text.

We use Probabilistic Canonical Correlation Analysis (PCCA) to obtain the latent space by referring to a related work [9]. PCCA is a probabilistic interpretation of Canonical Correlation Analysis (CCA) and we can obtain a latent space by solving the same generalized eigenvalue problem as CCA. The space generated by CCA regards the distributions of images and text, whilst there are some approaches [10] that consider image/word features as latent variables. On the contrary, there are some approaches [11] that adopt a strict analysis, such as probabilistic Latent Semantic Analysis (pLSA) or Kernel Canonical Correlation Analysis (KCCA). However, we do not use these as they have problems with high computational cost and solution stability. We can obtain a globally-optimized solution faster by CCA, although this is an approximate linear method.

### 3. SIMILAR IMAGE SEARCH

Given image features \(x_1, \ldots, x_n\) and text features \(y_1, \ldots, y_n\), matrices \(A\) and \(B\) found by CCA maximize the correlation between canonical variables \(s_i = A^T x_i\) and \(t_i = B^T y_i\). These are derived from the following generalized eigenvalue problem:

\[
R_{XY} R_{X}^{-1} R_{X} A = R_{X} A \Lambda^{2}, \quad R_{XY} R_{X}^{-1} R_{XY} B = R_{T} B \Lambda^{2},
\]

where \(\Lambda\) is a diagonal matrix of eigenvalues, \(R_{X} = \sum_{i=1}^{n} x_i x_i^T\), \(R_{Y} = \sum_{i=1}^{n} y_i y_i^T\), and \(R_{XY} = \sum_{i=1}^{n} x_i y_i\). The space generated by CCA is called the canonical space. This is also a latent space that considers both image feature and text feature distributions.

We define the metric for the combined latent space denoted by \(z\). Common CCA generates two canonical spaces, denoted by \(s_i\) and \(t_i\). An obvious approach is to search for similar images with their Euclidean distance within the canonical space generated from the image features. However, this loses certain information about the latent meanings as the distance of each dimension is considered without distinction of the canonical correlation. Mixing two spaces according to their correlation makes it possible to search latently-similar data more accurately. To denote the posterior probability of latent variable \(z\), we use \(p(z|x)\) when an image feature \(x\) is given. When both an image feature \(x\) and a text feature \(y\) are given, we use \(p(z|x, y)\). Using a PCCA framework, their optimal representations are obtained as Gaussians, the means and variances of which are described as follows:

\[
E(z|x) = M_x^T A^T (x - \bar{x}), \quad \Phi_x = \text{var}(z|x) = I - M_x M_x^T,
\]

\[
E(z|x, y) = \left( M_x \left( I - \Lambda^2 \right)^{-1} \right)^T \left( \frac{I - \Lambda^2}{I - \Lambda^2} \right) \left( A^T (x - \bar{x}) \right), \quad \Phi_{xy} = \text{var}(z|x, y) = I - \left( M_x \left( I - \Lambda^2 \right)^{-1} \right)^T \left( \frac{I - \Lambda^2}{I - \Lambda^2} \right) \left( M_x \left( I - \Lambda^2 \right)^{-1} \right) ,
\]

where \(M_x\) and \(M_y\) are defined as \(M_x = \Lambda^2\) and \(M_y = \Lambda^{1-\beta}\), respectively, with \(0 < \beta < 1\). To search for neighboring images, the Kullback-Leibler divergence between the input image and learned paired data are evaluated as the distance:

\[
K(p(z|x_j)||p(z|x_i, y_j)) = \left\| \Phi_{xy}^{-\frac{1}{2}}(E(z|x_j) - E(z|x_i, y_j)) \right\|^2 .
\]

This is equivalent to the Euclidean distance between \(z_j = \Phi_{xy}^{-\frac{1}{2}} E(z|x_j)\) and \(z_{xy} = \Phi_{xy}^{-\frac{1}{2}} E(z|x_i, y_j)\). Note that we use text in combination with images in learning, but can search for similar images using images only.

Here we provide a brief overview of the system. First, feature vectors are extracted from the images and associated text. Then, transformation matrices \(A, B\), and canonical correlations \(\Lambda\) are obtained by applying CCA to the features. Finally, coordinate \(z\) for each datum based on the similarities of the latent topics, is computable using \(A, B\), and \(\Lambda\).

Searching can be done simply. First, feature vectors are extracted from images that do not have associated text. Then, the \(z\) coordinates are computed. Similar images can be found via a neighborhood search based on the Euclidean distance between all data and an input datum.

### 4. EXPERIMENTAL RESULTS

In this section, we evaluate the effectiveness of the proposed similarity improvement. First, we describe the dataset of images and texts...
and feature extraction methods. Thereafter, we present the results of an evaluation showing how similar images can be found close by.

### 4.1. New York Times Dataset

For this research, we constructed an original dataset consisting of several classes. Existing datasets containing both images and text include only a dozen or so words per image. The length of this text is much shorter than that usually encountered in Web texts. Consequently, we used the New York Times API and constructed a news dataset comprising five classes of data, namely, American selection, baseball, basketball, terrorism, and weddings. The total number of image/text pairs is 16,152. Some examples of this dataset are shown in Fig. 1. Images differ greatly in size and appearance, while the accompanying text includes 800 words on average.

#### 4.2. Feature Extraction

We adopted the popular image features and merged their feature vectors. For this research, we selected Color-HLAC[12], gist[13], and LBP[14]. Color-HLAC is found by integrating the local autocorrelation between each pixel and the neighboring pixels on the whole image. We extract local autocorrelations at the highest two dimensions from color images. The width of the mask is three pixels. Each image is used at 1/1, 1/2, 1/4, and 1/8 resolution. The dimensions of the extracted vectors are 2,856. Gist is calculated by integrating filtered image-parts. For this research, we use Gabor filters in three scales, with the number of filter directions being eight, four, and four in scale order. Each color image is divided into a grid with 16 regions. The dimensions of the gist features are 960. LBP focuses on the difference between a region and circumjacent regions. We use the standard LBP which divides circumjacent regions into eight areas. The number of dimensions is 256.

Text features must have appropriate weights of hundreds of words in each text. We employ tfidf [15], which is the product of the frequency of each term in a document and the inverse of the appearance frequency of that term in all documents. Tfidf represents how well each word distinguishes the text. The dimension of text features is 103,086, since the dataset has 103,086 words, excluding stop words.

#### 4.3. Space Improvement Evaluation

We used a selection of the 16,152 data items for learning. First, we computed the PCA in the image feature space to give a contribution ratio of 99%. PCA was also computed in the text feature space to ensure that the dimensions are the same as the compressed image features. Subsequently, we distributed all images using the proposed method. We performed 50 fold cross validation over all the 16,152 images and evaluated whether images in the same class were distributed close to one another. One hundred images were searched per input image, with the score represented as the Mean Precision (MP) of the retrieved images belonging to the same class of input data.

As baseline methods, we selected PCA on image descriptors and plain CCA. We also evaluated the influence of the amount of paired data using the proposed method and plain CCA. In particular, we conducted experiments with 10 and 90 percent of the images having added text. We examined the best dimension of CCA from ten dimensions, since only the upper space of the generated canonical space is usually used when plain CCA is applied. The parameter $\beta$ in the proposed method varied between 0.1 and 0.9 in steps of 0.1. The MP scores are given in Table 1. We also show the MP scores when the input data were limited to images without associated text. This shows that the corresponding text contributes to an improvement in the image distributions by the proposed method. Additionally, even if no text is added to the input images, the performance of retrieval is still superior to the traditional image search.

Next, we evaluated the proposed method compared with popular nonlinear methods such as ISOLLE [8]. Learning all 16,152 images takes a long time, and hence we formulated the following setup. We used a selection of images, 90 percent of which were described by associated text. We improved the distributions of 1,616 image-only data using the proposed method, and thereafter searched for 100 images from these 1,616 images only using 10 fold cross validation. For comparison, we selected PCA and ISOMAP. PCA was applied to all 16,152 images, and the same 1,616 images used for the evaluation with 10 fold cross validation. The accuracy of the neighborhood search with ISOLLE and ISOSNE, which we reviewed in Section 3 as popular nonlinear techniques, is dependent on the accuracy of the neighborhood search with ISOMAP. LLE and SNE provide the role of dimensional compression. Therefore, we evaluated the performance of ISOMAP in image retrieval from the 1,616 images. The resulting MP scores are shown in Table 2. The performance of PCA is lower than that in Table 1, because the smaller number of data causes a decrease in the accuracy of the neighborhood search. Under these circumstances, the proposed method achieves a better distribution of the images based on the similarities in latent meanings. In particular, the proposed method using associated text achieves better performance than the nonlinear methods such as ISOMAP.

<table>
<thead>
<tr>
<th>Class</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>American</td>
<td><img src="image1.png" alt="American Images" /></td>
</tr>
<tr>
<td>Baseball</td>
<td><img src="image2.png" alt="Baseball Images" /></td>
</tr>
<tr>
<td>Terrorism</td>
<td><img src="image3.png" alt="Terrorism Images" /></td>
</tr>
<tr>
<td>Wedding</td>
<td><img src="image4.png" alt="Wedding Images" /></td>
</tr>
</tbody>
</table>

Fig. 1. Image examples from the dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>PCA</th>
<th>CCA 0%</th>
<th>CCA 90%</th>
<th>proposed 0%</th>
<th>proposed 90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>0.443</td>
<td>0.409</td>
<td>0.507</td>
<td>0.498</td>
<td>0.596</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>ISOMAP</th>
<th>proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>0.338</td>
<td>0.342</td>
</tr>
</tbody>
</table>

Table 1. MP performance using all data.

Table 2. MP performance using only part of the data.
4.4. Evaluation through Distribution Visualization

To show the advantage of the proposed method visually, we present the distributions of images without layout techniques such as in [7] and [5]. Fig. 2 shows the distributions of the first and second principal component obtained by applying PCA to the image feature space, whilst Fig. 3 shows the distributions of the first and second canonical variables obtained by applying the proposed method to the image feature space and text feature space. From the comparisons of pairs of images, images in the wedding class, in particular, are located close to one another. Note that these figures only show the improvement in the distributions. Existing layout techniques are applicable when higher dimension distributions are visualized as an interface for image browsing/retrieval.

5. CONCLUSION

An image browsing and retrieval system that operates in a low-dimensional space (two or three dimensions) is useful for searching large image databases. Text that is attached to images can be used to bridge the semantic gap. We proposed a novel retrieval method for a semantic image search. This method is based on a neighborhood search in the space of improved image distributions constructed from the latent meanings obtained from images together with associated text. In the proposed method, the metric learning and image retrieval is done in linear time. Additionally, existing image layout techniques can be adopted. We have shown that the accuracy of the proposed method, in terms of similarity between the query image topic and the topics of the retrieved images, is higher than existing methods that use image features only.

In future work, we aim to consider faster neighborhood search methods and to evaluate the proposed method using data from general web sites.

6. REFERENCES