Image Clustering Using Multimodal Keywords

Rajeev Agrawal [1,2], William Grosky [3], Farshad Fotouhi [2]
1 Kettering University,
Flint, MI 48504, USA
2 Wayne State University,
Detroit, MI 48202, USA
3 The University of Michigan – Dearborn,
Dearborn, MI 48128, USA

Abstract

- Extending our previous work on visual keywords, we use the concept of template-based visual keywords using MPEG-7 color descriptors. MPEG-7, also called the *Multimedia Content Description Interface*, has been a standard for many years. These color descriptors have the ability to characterize perceptual color similarity and need relatively low complexity operations to extract them, besides being scalable and interoperable. We then demonstrate the power of these visual keywords for image clustering, when used in tandem with textual keyword annotations, in the context of latent semantic analysis, a popular technique in classical information retrieval which has been used to reveal the underlying semantic structure of document collections.

Introduction

- Use of low level color and texture features
- Segmentation: weak segmentation, strong segmentation
- Feature extraction: based on the entire image or on regions of the image resulting from a segmentation process. Clustering: k-means, hierarchical agglomerative clustering, or a learning-based approach.

Introduction

- Problems: capturing semantics and formulating queries.
- Solution:
  - Annotate images with keywords manually,
  - Query on these keywords.
  - The quality of this method => dependent on the perception of the person annotating the images.

Introduction

- Our Approach: use both low-level image features, in the form of *visual keywords*, and text annotation to cluster the images.
- What is a Visual Keyword?
- Idea is similar to Classical information retrieval.
- Subdivide an image using templates of certain sizes.
- Visual keyword represents similar sub-images in entire collection.

Introduction

- Motivation: Organizing a large text collection. Use of a *term-document matrix;* rows represent the textual keywords and columns represent the documents. Then, techniques such as latent semantic analysis (LSA) can be used to discover the latent relationships between correlated words and documents.
- In our approach, consider each image as a document and each template region as a word (visual keyword). Hence, each image is represented by multiple template regions. These regions are called *tiles*.

MPEG-7 Descriptors

- MPEG-7, formally called the *Multimedia Content Description Interface:* standard for describing multimedia content data that supports some degree of interpretation of semantics determination, MPEG-7 compatible data include still pictures, graphics, 3D models, audio, speech, video, and composition information about how these elements are combined in a multimedia presentation.
- Color descriptors: color space, color quantization, dominant colors, scalable color, color layout, color structure, and GoF/GoP color.
- color spaces: monochrome, RGB, YCrCb, HSV, HMMD, and monochrome (intensity only).

MPEG-7 Descriptors

- Scalable color descriptor:
  - global color histogram, encoded by a Haar transform
  - image-to-image matching and retrieval.
  - number of bits: 16, 32, 64, 128 or 256.
- Color layout descriptor:
  - spatial color information.
  - matching functionality with high retrieval efficiency at very small computational costs.
- Color structure descriptor:
  - both color content and the structure of this content.
  - image-to-image matching and for still image retrieval.
  - distinguish two images in which a given color is present in identical amounts but where the structure of the groups of pixels having that color is different in the two images.
  - number of bins can be 32, 64, 128 or 256.

Overview of the proposed approach

- Extracting and Clustering Visual Keywords
- Creating a term-document matrix using textual keywords
- Combining Visual Keywords and Textual Keywords Information
- Evaluating the MPEG-7 visual keyword Model

Extracting and Clustering Visual Keywords

Input: A set of images $I = \{I_1, I_2, \ldots, I_n\}$.

Output: Visual keyword-image matrix

Algorithm:

1. Divide each image I_i into non-overlapping tiles *ti* of the fixed template size.
2. Extract MPEG-7 descriptors (SCD, CLD, CSD) to form a tile vector $t_{i,j}$ for each tile tj of image I_i.
3. Generate a tile matrix *V*, where each $t_{i,j}$ above is a row vector of *V.*
4. Normalize *V* and then apply SVD to reduce the dimension.
5. Apply a clustering algorithm to create *C* clusters out of all the tiles.
6. Compute the visual keyword-image matrix, having one column for each image and one row for each cluster, where the (i,j) th element of this matrix is the number of times tiles from the i th cluster appear in the j th image.

Creating a term-document matrix using textual keywords

- Create an initial term-document matrix.
- Porter’s stemming algorithm.
- Normalized to unit-length.

Combining Visual Keywords and Textual Keywords

- Visual keywords are based on MPEG-7 color descriptors derived for each tile cluster of the image. (Tvis)
- Textual keywords: annotations about the image. (Ttex)
- Advantage of both visual keywords and textual keywords.
- *Tvis*, *Ttex* are concatenated to create a single matrix, *Tvis-tex*.
- Apply LSA on this combined visual and textual space and learn co-occurrence relations among textual keywords and visual keywords.
- In summary, we extract the semantic relationship between text to text, image to image, and text to image in this step.

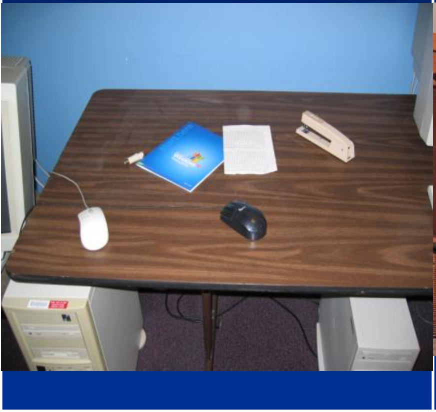
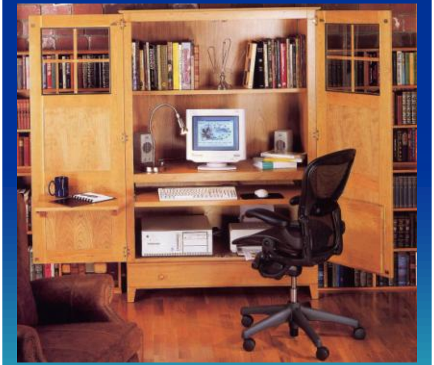
Evaluating the MPEG-7 visual keyword Model

- Cluster the images using the visual + textual keyword model and compare it with both the visual keyword model and the textual keyword model using the template concept and the template-as-entire-image concept.
- K-means: unsupervised learning algorithm.
- First k centroids, one for each cluster, are defined, and then each data point is assigned to one of these clusters. K-Means minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances.
- The Adjusted Rand Index is a technique for measuring similarity between two data clusters. It has a value between 0 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

Experiments

- *LabelMe dataset*, (MIT AI Lab).
- 658 images from 15 categories.
- Boston street scene (152), cars parked in the underground garage (39), kitchen (14), office (24), rocks (41), pumpkins (58), apples (11), oranges (18), conference room (28), bedroom (14), dining (63), indoor home (59), home office (19), silverware (81), and speaker (37).

Experiments

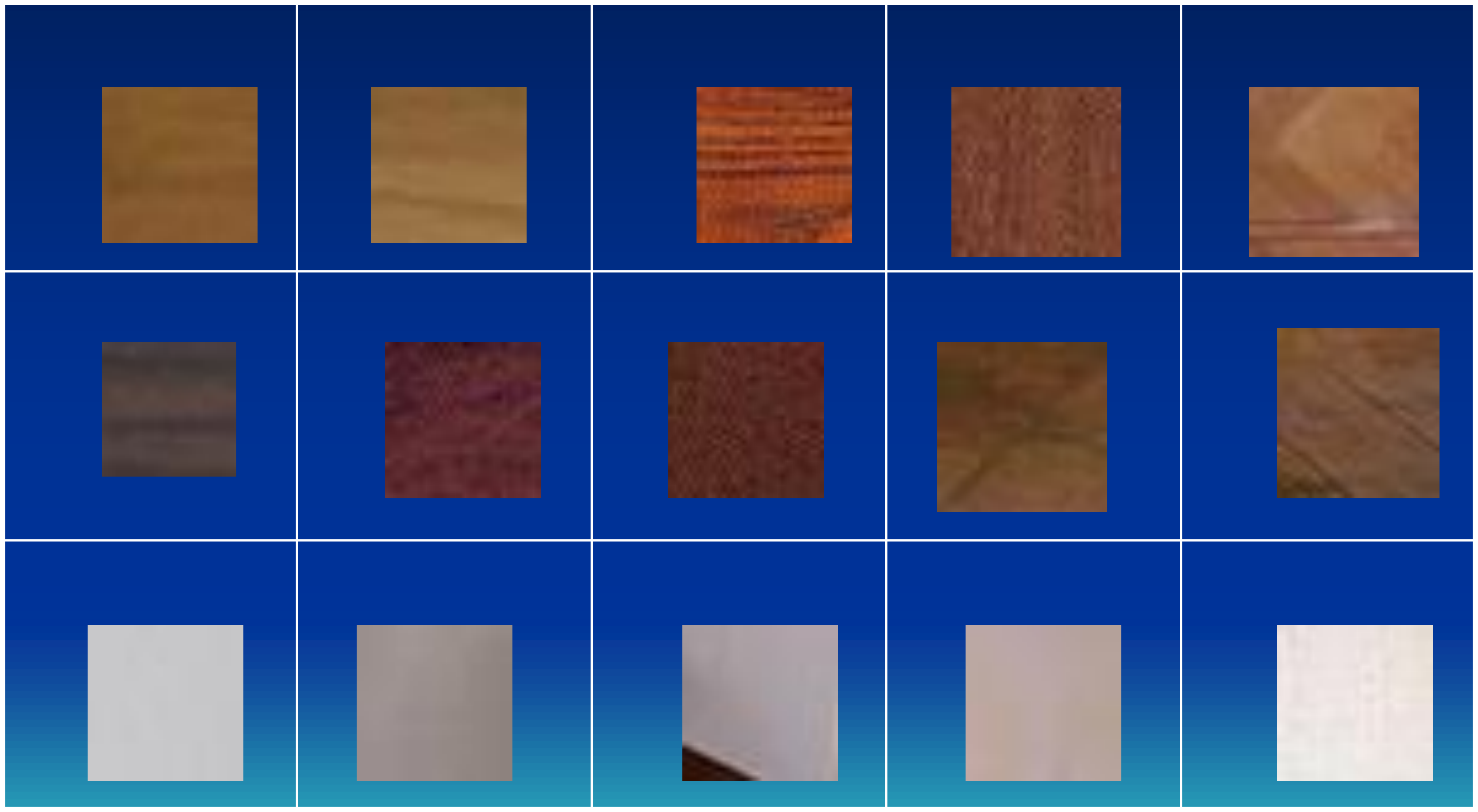
Experiments

- Template size of 32 pixels * 32 pixels
- The images are resized to 640 pixels * 480 pixels if they are bigger to restrict the number of tiles to a fixed limit; however the smaller images are left in their original sizes.
- The maximum number of tiles an image can have is 300; the total number of tiles of 658 images is 165750.
- 1500 visual keyword clusters => 658 images * 1500 clusters.
- The textual keyword matrix => 658 images * 506 words
- Final matrix: 658 images * 2006 keywords,
- LSA/SVD is then applied to select only 200 principal components (coefficients), which results in a matrix of 658 images * 200 concepts.

Experiments: Data set

- Full size image (*mpfs*):
- Full-size image and textual keywords (*mpfstk*):
- Tiles of each image (*mpts*):
- Tiles of each image and textual keywords (*mptstk*):
- Only the textual keywords:

Experiments: Tiles



Experiments: Results

Dataset	ARI
Mpfs	.32
Mpfstk	.39
Mpts	.34
Mptstk	.51
Text keywords only	.26

- K is used as, 15; actual number of classes

Conclusions and Future Work

- Image clustering model using MPEG-7 color descriptors to represent template-based visual keywords
- LSA on the visual keywords and textual keywords of the images.
- Visual keywords and text annotations, if used together, can improve the quality of the clusters.
- LSA helps in establishing the relationship between visual and textual keywords.
-
- The text annotations for each image range only from 1 to 10. extend this list to include words from other synsets using Wordnet.
- Usage of other color and texture descriptors, which can be examined.
- Scaling of the different color descriptors to the same length.
- The number of visual keywords is more than the number of textual keywords.
- Finding optimal number of visual and text keywords in an annotated image collection.

Thanks