Effective Summarization of Large-Scale Web Images

Chunlei Yang
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
cyang36@uncc.edu

Jialie Shen
School of Information Systems
Singapore Management
University
Singapore, Singapore
jlshen@smu.edu.sg

Jianping Fan
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
jfan@uncc.edu *

ABSTRACT

In this paper, we present a novel framework to achieve effective summarization of large-scale web images by treating the problem of automatic image summarization as the problem of dictionary learning for sparse coding, e.g., the summary of a given image set can be treated as a sparse representation of the given image set (i.e., sparse dictionary for the given image set). For a given semantic category (i.e., certain object class or image concept), we build a sparsity model to reconstruct all its relevant images by using a subset of most representative images (i.e., image summary); and a stepwise basis selection algorithm is developed to learn such sparse dictionary (i.e., image summary) by minimizing an explicit optimization function. By investigating their reconstruction ability, the reconstruction Mean Square Error (MSE) is adapted to objectively measure the performance of various algorithms for automatic image summarization. Our experimental results demonstrate that our dictionary learning for sparse representation algorithm can obtain more accurate summary as compared with other baseline algorithms for automatic image summarization.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding - *Perceptual Reasoning.*

General Terms

Algorithms, Measurement, Experimentation

Keywords

Automatic image summarization, sparse coding, dictionary learning

*Area chair: Lexing Xie

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

1. INTRODUCTION

As a fundamental building block to enable interactive image navigation and exploration, automatic image summarization aims to select a small set of the most representative images for highlighting large amounts of images briefly.[3]. Most existing techniques focus on selecting a small set of the most representative images to highlight the most significant visual properties of a given image set in large size [3]. Thus the issue of automatic image summarization can be treated as an optimization problem, e.g., selecting a small set of the most representative images that can be used to reconstruct the original image set more effectively. If we define $\mathbf{X}$ as the original image set in large size and $\mathbf{D}$ as the summary of the given image set $\mathbf{X}$ in small size, automatic image summarization is to determine the summary $\mathbf{D}$ by minimizing the global reconstruction error:

$$\min_{\mathbf{D}} ||\mathbf{X} - f(\mathbf{D})||_F$$

The selection of the reconstruction function $f(\cdot)$ is to determine how each image in the original image set $\mathbf{X}$ can be reinterpreted by the most representative images in the summary $\mathbf{D}$. It is worth noting that all these images in the original image set $\mathbf{X}$ are represented as a set of visual features. In this paper, we define the reconstruction function $f(\cdot)$ as a small set of the most representative images in the summary $\mathbf{D}$ that can sparsely reinterpret all the images in the original image set $\mathbf{X}$. As a result, we can successfully reformulate the issue of automatic image summarization as the task of dictionary learning for sparse representation. Therefore, two research issues (automatic image summarization and dictionary learning for sparse representation) are linked together by their intrinsic coherence: both of them try to select an image subset in small size that can effectively and sufficiently reconstruct large amounts of images in the original image set.

Although automatic image summarization and dictionary learning for sparse representation have intrinsic coherence, we need to clarify that they have significant differences as well, e.g., the optimization function of the former has some unique constraints such as the fixed basis selection range, positive and L_0-norm sparsity of the coefficients. Based on this observation, we proposed a stepwise basis selection algorithm for dictionary selection and sparse coding to solve the reformulated optimization problem for automatic image summarization.

2. RELATED WORK

Existing algorithms can solve image summarization prob-

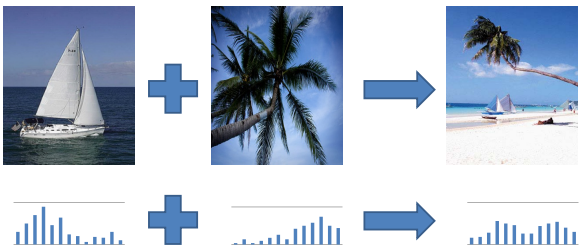

Figure 1: Demonstration for the additivity property of interest point feature.

lem simultaneously. Usually, the global distribution of an image set is investigated and image clustering is involved. In particular, Jaffe *et al.* [1] have developed a Hungarian clustering method by generating a hierarchical cluster structure and ranking the candidates according to their relevance scores. Denton *et al.*[2] have introduced the Bounded Canonical Set (BCS) by using a semidefinite programming relaxation to select the candidates, where a normalized-cut method is used for minimizing the similarity within BCS while maximizing the similarity from BCS to the rest of the image set. Other clustering techniques such as k-medoids [5] and affinity propagation [6] are also acknowledged. The global distribution of an image set can also be characterized by using a graphical model. Jing *et al.* [3] have expressed the relationship between the images with a graph structure, where the edges indicate their similarities and the nodes with the most connected edges are selected as the summary of the given image set.

On the other hand, there are iterative approaches. Some greedy-like algorithms are applied to select the best summary sequentially until sufficient number of the most representative images are picked out [4]. Simon *et al.* [4] have used a greedy method to select the best candidate by using some important summarization measurements such as likelihood, coverage and orthogonality. The greedy method focuses on selecting the most representative images while penalizing the appearances of the similar images.

Distinguished from the existing approaches, our proposed approach for automatic image summarization characterizes the representativeness of the image summary explicitly by optimizing a reformulated reconstruction function. We will demonstrate that automatic image summarization can be achieved more effectively by solving a problem of dictionary learning for sparse representation.

3. AUTOMATIC IMAGE SUMMARIZATION

The research aims to generate the visual summary for each semantic category (i.e., one certain object class or image concept). Given a semantic category, we extract a set of key-points from all the relevant images and the SIFT [7] descriptors are used for image content representation. All the key-points for the given semantic category are then partitioned into l clusters by using the affinity propagation algorithm and the cluster centroids for all these l clusters are used to construct a l-dimension codeword dictionary. For each relevant image, we further conduct a vector quantization of all its key-points according to the codeword dictionary. As a result, the feature vector for image representation is a normalized l-dimension histogram.

The normalized feature vector indicates the distribution of the codeword for the given image, or the probability of the appearance of the codewords in the given image. From this observation, we assume that the feature vector histogram of one given image can approximately be reconstructed by a weighted linear combination of the feature vector histograms of other images as demonstrated in Figure 1. To obtain the maximum reconstruction power, we try to minimize the overall reconstruction error:

$$\min \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{d}_j \alpha_{ji}||_2 \tag{1}$$

The codewords $\{\mathbf{d}_j\}$ are a set of most representative images that are selected from the original image set. The size of the most representative images $\{\mathbf{d}_j\}$ is determined by a trade-off between concise summarization and accurate reinterpretation: a small size of the most representative images $\{\mathbf{d}_j\}$ means more concise summary but the reinterpretation power may reduce; on the other hand, a large size of the most representative images $\{\mathbf{d}_j\}$ guarantees a better reinterpretation performance but the summary may be verbose. $\{\alpha_{ji}\}$ is a set of the sparse coefficients to guarantee that each image in the same semantic category can be represented by a limited number of the most representative images in the summary. Considering the reasonable complexity of the visual content of an image, bringing the sparsity is necessary (we will elaborate the choice of sparsity in the experiment section). To satisfy this condition, L_0-norm is taken into consideration. Given the image set X for the given semantic category, we formally rewritten Eq.(1) in the matrix form as follows:

$$\begin{aligned} \min_{\mathbf{D},\mathbf{A}} \quad & ||\mathbf{X} - \mathbf{DA}||_F \\ s.t. \quad & ||\alpha_i||_0 \leq t, i = 1, 2, ..., k. \end{aligned} \tag{2}$$

where $\mathbf{X} = \{\mathbf{x}_i | i = 1, 2, ..., n\} \in \mathbb{R}^{l \times n}$ is the original image set, l is the dimension of the feature vector, n is the total number of the images in the given semantic category, $\mathbf{D} = \{\mathbf{d}_i | i = 1, 2, ..., k\} \in \mathbb{R}^{l \times k}$ is the dictionary that we want to learn (i.e., summary for the given semantic category), k is the number of the codewords in this dictionary (i.e., the number of the most representative images), $\mathbf{A} \in \mathbb{R}^{k \times n}$ is the positive coefficient matrix that can be learned jointly, each column of $\mathbf{A} = \{\alpha_i | i = 1, 2, ..., n\}$ is the sparse representation for a given image and t is the corresponding sparsity. If t equals to k, the above formulation for automatic image summarization can be reduced to the k-medoids problem (the discrete form of k-means). The k-medoids algorithm is well known as one of the data summarization methods [5], thus our dictionary learning and sparse representation algorithm for automatic image summarization can be seen as an extension of the k-medoids.

In our reformulation, two aspects differ from the traditional sparse representation: 1) From the description of additivity of the local property we need to make sure the coefficients $\{\alpha_{ji}\}$ must be non-negative; 2) the dictionary $\mathbf{D}$ is selected from a group of the original images in $\mathbf{X}$ rather than a weighted combination of arbitrary images.

The optimization problem in Eq.(2) with L_0-norm penalty over the coefficients$\{\alpha_{ji}\}$ is "NP-hard"[9], thus the plan is to iteratively update the bases (codewords) and the coefficients while decreasing the reconstruction error until it stops at a stationary point. Our stepwise basis selection algorithm can

be divided into two stages: (a) sparse coding stage; and (b) codeword updating stage.

Sparse coding stage: The dictionary $\mathbf{D}$ is fixed and the coefficients are updated as follows: The coefficients are learned individually for each image. For a given image x_i in the same semantic category, we initialize all its coefficients equal to zero and then find the basis of d_m which has the biggest inner product with x_i (the smallest angle between the two vectors). We gradually increase the coefficients of d_m at the positive sign direction until another basis d_o has a competitive inner product with the residual r as d_m, then we increase the coefficients at the joint direction of d_o and d_m. We repeat these operations until t coefficients are found or r decreases to zero.

Codeword updating (basis selection) stage: When the dictionary $\mathbf{D}$ and the coefficient matrix $\mathbf{A}$ are fixed, we sequentially release one of the k bases and update the released basis and the corresponding coefficient. The updating criteria is to choose the pair of $\mathbf{d}_o$ and α_o from all the $n-k+1$ candidates which can minimize the global reconstruction error e. e is calculated as :

$$\begin{aligned} e &= \sqrt{\sum_i (r_i - d_i\alpha)^2} \\ &= \sqrt{lr_i^2 + (\sum d_i^2)\alpha^2 - 2(\sum r_i d_i)\alpha} \end{aligned}$$

The minimum of $e = \sqrt{lr_i^2 - (\sum r_i d_i)^2}$ can be reached when $\alpha = \sum (r_i d_i) \backslash \sum d_i^2 = \sum (r_i d_i)$

After all the k codewords are updated, we fix the dictionary and go back to the sparse coding stage to update the coefficients. Such iteration process stops when the objective reaches a stationary point (no basis is being updated).

After the codeword updating (basis selection) stage, we have already got a set of coefficients that are good enough. The method used in sparse coding stage is greedy-like method and won't guarantee the decrease of the objective function. So we further perform another method which can guarantee the updated coefficients to decrease the objective function. This method starts matching pursuit from a given set of coefficients. It changes the values of the non-zero coefficients rather than the positions of these coefficients. The coefficient matrix A is updated by the result of two methods whichever reaches a smaller reconstruction error.

After the dictionary $\mathbf{D}$ is learned from the original image set $\mathbf{X}$ for the given semantic category, the codewords (basis) in the dictionary $\mathbf{D}$ are then treated as a good summary of the original image set $\mathbf{X}$ for the given semantic category.

4. ALGORITHM EVALUATION

To evaluate the proposed algorithm, we collected images from 1000 visual text terms and about 3000 relevant images are kept after image cleaning for each category. We then constructed a codeword dictionary of size 200. Then we apply the given codeword dictionary (200 dimensions) to quantize the key-points for each image and construct a 200-dimensional histogram of the feature vectors for each image, which can be the representation of $\mathbf{x}_i$ and $\mathbf{d}_i$ as appeared in Eq. (2). We denote the number of images in the given semantic category by N, the number of codewords by K and the sparsity by T. $\mathbf{D}$ represents the dictionary (summary) and $\mathbf{A}$ represents the coefficient matrix.

Three algorithms are introduced as the baseline algorithms in our experiments: k-medoids [5], affinity propagation [6] and Sparsifying Dictionary Selection (SDS) algorithm [8]. Our proposed algorithm are compared with the above three baseline algorithms objectively and subjectively.

The k-medoids is a typical clustering-based algorithm for automatic image summarization, k is the number of clusters (the size of the dictionary) and the medoids of each cluster is taken as the selected basis. The clustering process aims to partition the images in the original set into k clusters which can minimize the within-cluster sum of squares:

$$\min_{\mathbf{S}} \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in \mathbf{S}_i} ||\mathbf{x}_j - \mathbf{d}_i||^2 \tag{3}$$

The affinity propagation algorithm partition the image set into multiple partitions. The number of partitions is affected by the preference value and thus can be binary searched. The element that has the minimum average distance to the rest of the cluster is selected as the basis.

$$\mathbf{d} = \arg \min_{\mathbf{d}_i \in \mathbf{C}} \sum_{\mathbf{x}_j \in \mathbf{C} \backslash \mathbf{d}_i} ||\mathbf{x}_j - \mathbf{d}_i||^2 \tag{4}$$

The SDS algorithm, on the other hand, represents a series of greedy algorithms which iteratively select the current best basis. Krause *et al.* suggested in [10] that if the data collection satisfy the submodular condition then the local optimal derived by greedy algorithm is a near-optimal solution. The greedy algorithm starts with an empty dictionary $\mathbf{D}$, and at every iteration i adds a new element by:

$$\mathbf{d}_i = \arg \min_{\mathbf{d} \in \mathbf{X} \backslash \mathbf{D}} F(\mathbf{D}_{i-1} \cup \mathbf{d}) \tag{5}$$

where F is the reconstruction function. The SDS algorithm is modified to satisfy our positive coefficients constraint.

Evaluation metric: We evaluate four algorithms for automatic image summarization based on their reconstruction abilities. Specifically, the performance is evaluated by the global reconstruction error in terms of Mean Square Error (MSE):

$$\begin{aligned} MSE := \quad & \tfrac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{D}\alpha_i||_2 \\ s.t. \quad & ||\alpha_i||_0 \leq T \end{aligned} \tag{6}$$

Smaller MSE value represents better reconstruction ability. Thus MSE is treated as the objective metric for evaluating the performance of various algorithms for automatic image summarization.

Objective evaluation : (1) The convergence of our proposed algorithm for automatic image summarization is evaluated in terms of the number of bases being updated and the change of MSE value. We set up the experiment with the size of the dictionary equals to 20 and the sparsity equals to 6. We have observed that there are large drops of the MSE curve in the steps of 11, 21 in the sparse coding stage. During each of the sparse coding stage, the decrease of the MSE value is caused by basis updating. After 3 or 4 iterations, no basis is being updated and the corresponding sparse coefficients gradually become stable. We repeated the experiment with random initials for over 20 times and observed similar results, thus our SBS algorithms can quickly converge after around 4 or 5 iterations.

(2) Our proposed SBS algorithm performs better than the other three baseline algorithms with a lower MSE value

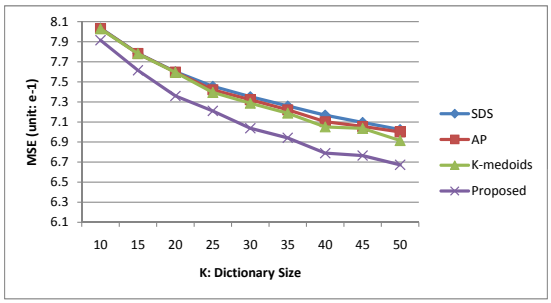

Figure 2: MSE performance evaluation in comparison with three baseline algorithms.

throughout the different choices of dictionary sizes. When the dictionary size is small (K=10), our proposed model performs similar to the baseline algorithms (with an improvement of 1.5%); as the K increases, the greedy-based algorithm performs the worst and our proposed SBS algorithm performs the best (with an improvement of 5.3%).

(3) We have also confirmed our assumption in section 3: as the dictionary size increases, the reconstruction error decreases and the summarization becomes less concise. The choice of K actually depends on the requirement of specific task.

(4) As explained in section 3, the sparsity in the image summarization problem indicates the use of a limited number of basic local visual attributes to reconstruct an image and we assumed that an image can only be composed by a limited number of salient local visual attributes. We increase the sparsity of T from 2 and observed that when T increases above 6, the reconstruction performance for the LPAR algorithm would not improve significantly, which means that most of the images in the same semantic category can sufficiently be reinterpreted by no more than 6 bases. This observation explains why we make T equals to 6 in the experiments.

Subjective evaluation: Because of the semantic gap, the MSE value may not be effective for evaluating the quality of image summary. Thus a user study is performed to evaluate the effectiveness of our proposed image summarization approach with the baseline approaches. The subjective evaluation metric is measured by the users' feedback on how well the summarization results represent the original images for the given semantic category. Our survey consists the following components : (1) 30 users (graduate students) are involved in this survey to investigate the summarization results for 15 semantic categories as listed in Table 1. In order to ease the burden from the user, the category size is reduced to 500. (2) The users are allowed to explore the image set and pick their own summary of 10 images. The summary from all the user are collected and counted, and the top 10 choices are selected as the ground truth. (3) The output from the 4 algorithms are compared with the ground truth and number of matches are reported. (4) The users will also give a relevance score scaled from 0 to 5 (5 being the best) to the four outputs based on their own judgement of relevancy. The algorithm name is hidden from the user during the process to avoid biased choice. The results reported in Table 1 indicates that our proposed dictionary learning approach has more matches and higher relevance score as compared with the baseline algorithms.

	K-med	AP	SDS	Proposed
coast	1/2.8	0/2.1	0/1.2	**5/4.8**
bridge	1/1.7	1/2.3	0/1.3	**3/3.8**
tower	1/2.2	1/2.3	0/1.0	**3/4.2**
pyramid	**2**/2.8	0/1.0	1/2.1	**2/3.6**
zoo	2/2.7	0/1.3	0/1.1	**3/4.3**
valley	0/1.5	**2/2.9**	1/1.7	1/2.2
trafficlight	0/1.4	0/1.2	1/2.1	**3/3.7**
sailboat	**2**/2.7	0/1.5	0/1.1	**2/3.2**
mountain	1/1.5	1/1.8	1/1.8	**2/3.5**
sunset	2/2.6	2/3.1	0/1.1	**4/4.7**
airport	1/1.7	**2/2.5**	0/1.1	1/2.1
streetview	0/1.5	0/1.4	1/2.3	**3/4.5**
skyscraper	0/1.2	0/1.3	1/1.9	**2/3.9**
stonehenge	**2/3.1**	1/1.6	1/2.1	1/2.7
waterfall	1/1.5	**2/2.8**	0/1.8	1/2.2
average	1.07/2.06	0.8/1.94	0.47/1.58	**2.4/3.56**

Table 1: Subjective evaluation results for 15 categories: (# of matches with the ground truth/relevance score from user)

5. CONCLUSION

Most existing algorithms for automatic image summarization have not formulated the problem explicitly and lack an objective evaluation metric. In this paper, we have discovered intrinsic coherence between the issue of automatic image summarization and the problem of dictionary learning for sparse representation, e.g., their ability to use an image subset in small size to sparsely reinterpret the original image set in large size. We utilize the knowledge from the research areas of dictionary learning and sparse coding, and explicitly reformulate the problem of automatic image summarization with a sparse representation model and solve an optimization problem by using our proposed stepwise basis selection algorithm. Our proposed algorithm outperforms the baseline algorithms in both objective and subjective evaluations.

6. REFERENCES

[1] E. Jaffe, M. Naaman, T. Tassa, M. Davis "Generating summaries for large collections of geo-referenced photographs", *Proc. Int. Conf. on World Wide Web*, 853-854, 2006.

[2] T. Denton, M. Demirci, J Abrahamson, A. Shokoufandeh, S. Dickinson, "Selecting Canonical Views for View-Based 3-D Object Recognition", *ICPR*, 2004.

[3] Y. Jing, S. Baluja, H. Rowley "Canonical Image Selection from the Web", *CIVR*, 2007.

[4] I. Simon, N. Snavelv, S. Seitz "Scene Summarization for Online Image Collections", *Intl. Conf on Computer Vision*, ICCV, pp. 1-8, 2007.

[5] Y. Hadi, F. Essannouni, R. Thami "Video Summarization by k-medoid Clustering", *ACM Symposium on Applied Computing*, SAC, 2006.

[6] B. Frey, D. Dueck "Clustering by Passing Messages Between Data Points", *Science*, 315, 972-977, 2007.

[7] D. Lowe, "Distinctive Image Features from Scale Invariant Keypoints", *Intl Journal of Computer Vision*, vol.60, pp.91-110, 2004.

[8] A. Krause, V. Cevher, "Submodular Dictionary Selection for Sparse Representation", *Proc. International Conference on Machine Learning*, ICML, 2010.

[9] B. Natarajan, "Sparse Approximate Solutions to Linear Systems", *SIAM J. COMPUT.* Vol. 24, No.2, pp.277-234, 1995.