Image Collection Summarization via Dictionary Learning for Sparse Representation

Chunlei Yang\textsuperscript{a}, Jinye Peng\textsuperscript{b}, Jianping Fan\textsuperscript{b,a}
\textsuperscript{a} Department of Computer Science, University of North Carolina, Charlotte, NC 28223, USA
\textsuperscript{b} School of Information Science and Technology, Northwest University, Xi’an 710069, China

Abstract

In this paper, a novel framework is developed to achieve effective summarization of large-scale image collection by treating the problem of automatic image summarization as the problem of dictionary learning for sparse representation, e.g., the summarization task can be treated as a dictionary learning task (i.e., the given image set can be reconstructed sparsely with this dictionary). For image set of a specific category or a mixture of multiple categories, we have built a sparsity model to reconstruct all its images by using a subset of most representative images (i.e., image summary); and we adopted the simulated annealing algorithm to learn such sparse dictionary by minimizing an explicit optimization function. By investigating their reconstruction ability under sparsity constrain and diversity constrain, we have quantitatively measure the performance of various summarization algorithms. Our experimental results have shown that our dictionary learning for sparse representation algorithm can obtain more accurate summary as compared with other baseline algorithms.

1. Introduction

Automatic image summarization, which attempts to select a small set of the most representative images for highlighting large amounts of images briefly, becomes very important to enable interactive image navigation and exploration [3]. Many multimedia applications can benefit from automatic image summarization: (a) on-line shopping sites can generate multiple icon images (i.e., image summary) for a category of product by selecting a limited number of the most representative pictures from a collection of several thousands; (b) tourism websites can generate a small set of the most representative photos from large-scale photo gallery and display on their front page to attract visitors while resulting in low information overload on navigation. Such interesting applications have motivated researchers to develop more effective models and mechanisms for achieving accurate summarization of large-scale image collection.

Most existing techniques for automatic image summarization follow the same criteria as selecting a small set of representatives to highlight all the significant visual properties of a given image collection 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 best reconstruct the original image set. If we define \(X\) as the original image set in large size and \(D\) as the summary of the given image set \(X\) in small size, automatic image summarization is to determine the summary \(D \in X\) by minimizing the global reconstruction error:

\[
\min_D ||X - f(D)||^2_2
\]  

(1)

The selection of the reconstruction function \(f(\cdot)\) is to determine how each image in the original image set \(X\) can be reinterpreted by the most representative images in the summary \(D\). In this paper, we define the reconstruction function \(f(\cdot)\) to take summary \(D\) as input, which then sparsely reinterpret each image in the original set \(X\). As a result, we can successfully reformulate the task of automatic image summarization as the problem of dictionary learning for sparse representation as shown in Eqn 1. 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 a subset that can effectively and sufficiently reconstruct large amounts of items in the original 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, nonnegative and \(L_0\)-norm sparsity of the coefficients. Thus most existing methods for dictionary learning and sparse coding can not be directly adopted for solving the optimization problem for automatic image summarization. Based on this observation, an adopted simulated annealing algorithm is developed for dictionary selection and
a multivariate algorithm is developed for nonnegative sparse coding to solve the reformulated optimization problem for automatic image summarization. Most existing research works for automatic image summarization evaluate their results (summaries) subjectively by using user satisfaction and relevancy score. There lacks an objective and quantitative evaluation metric for evaluating the performance of various algorithms for automatic image summarization. By reformulating the problem of automatic image summarization as a problem of reconstruction optimization, we can objectively evaluate the performance of various algorithms in terms of their global reconstruction ability. Thus, in addition to the subjective evaluation, we leveraged the global Mean Square Error (MSE) as the objective evaluation metric to measure the performance of our proposed algorithm for automatic image summarization and compare the result with other baseline methods.

The paper is organized as follows: We discuss the state-of-the-art techniques in the area of both automatic image summarization and dictionary learning for sparse representation in section 2. In section 3, we will introduce the proposed algorithm for automatic image summarization. We present and discuss our experiment results in section 4. We conclude this paper in section 5.

2. Related Work

Most existing algorithms for automatic image collection summarization can be categorized into two categories: (a) simultaneous summarization approach; and (b) iterative summarization approach.

For the simultaneous summarization approach, the global distribution of an image set is investigated and image clustering techniques is usually involved [1, 2]. 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 [7], affinity propagation [8] and SOM [16] are also widely acknowledged. The global distribution of an image set can also be characterized by using a graphical model. Jing et al. [3] have expressed the similarities of 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 a given image set.

For the iterative summarization approach, some greedy-fashion algorithms are applied to select the best summary sequentially until preset number of the most representative images are picked out [6]. Simon et al. [6] have used a greedy method to select the best candidate by investigating the weighted combination of some important summarization metrics such as likelihood, coverage and orthogonality. Sinha [17] proposed a similar algorithm with metrics of quality, diversity and coverage. Wong et al. [15] integrates a dynamic absorbing random walk method to find diversified representatives. The idea is to use absorbing states to drag down the stationary probabilities of the nearby items, thus encourage the diversity. The current item with the highest stationary probability is iteratively selected. The above greedy methods focus on selecting the current most representative image at each iteration while penalizing the co-occurrence of similar candidates. Our model takes the benefit of both two types of approaches in that, we use explicit measurements to characterize the property of a summary and we learn the bases simultaneously and avoid the possible local optimum solution.

Most existing techniques for dictionary learning and sparse coding use machine learning techniques to obtain more compact representation, such as PCA, the Method of Optimal Direction (MOD) [4] and K-SVD [10]. The MOD algorithm is derived directly from Generalized Lloyd Algorithm (GLA) [5], iteratively updates the codebook and the codewords are updated as the centroids from a nearest neighbor clustering result. The K-SVD algorithm follows the same pattern by updating the bases iteratively and the new basis is generated directly from the SVD calculation result. The K-SVD method is not applicable to the proposed model because our model only takes discrete bases rather than numerical outputs from SVD. The methods of Matching Pursuit (MP) [11] and Lasso (forward stepwise regression and least angle regression) are widely accepted for sparse coding. Recently, Krause et al. [9] have proposed the submodular dictionary selection method for sparse representation and have proved that the dictionary (which is selected greedily) is close to the global optimum solution if the original data set satisfied the submodular condition. However, most of the real-world image sets do not satisfy the submodular condition which makes Krause’s output less convincing for automatic image summarization application. All the above algorithms cannot avoid falling into the trap of local optimum, so we adopted the simulated annealing algorithm which is a combinatorial algorithm and is proved to achieve global optimum with a high probability after enough number of iterations.

3. Automatic Image Summarization

In this section, we define the criteria to evaluate the summarization of image collections and reformulate the problem of automatic image summarization via dictionary learning under sparsity and diversity constrains. The problem of automatic image summarization is therefore re-casted into the problem of dictionary learning with various constraints.
We firstly point out the differences of our reformulation of dictionary learning for automatic image summarization with the traditional formulation, then we introduce our solution to address the problem.

3.1. Problem Reformulation

Given an image set, we represent the images with the normalized bag-of-visual-words model. Local feature is described by the SIFT [12] descriptor and the image feature is derived by vector quantization. Thus each image feature vector indicates the distribution of the codewords that appeared in the given image. In other words, the distribution is reflected by the accumulated probability of the appearance of each codeword. From this observation, we can assume that one feature vector can be approximately reconstructed by a nonnegative weighted linear combination of other feature vectors which also reflects the accumulated probability of the appearance of the codewords, hence is defined in section 1 is a linear nonnegative weighted function. This is the foundation how each image can be reconstructed by exemplars or bases. To obtain the biggest reconstruction power, we try to minimize the overall reconstruction error given as:

$$\min\sum_{i=1}^{n} ||x_i - \sum_{j=1}^{k} d_j \alpha_{ji} ||_2^2$$

(2)

For image collection summarization problem, the collection of are the most representative images that we want to learn, so it should come from the original image set. The determination of the size of is a tradeoff between concise summarization and accurate reinterpretation: a small size of means a more concise summarization but the reinterpretation power will reduce; on the other hand, a large size of guarantees a better reinterpretation performance but the summarization will be verbose. The idea of this reconstruction model is similar to nonnegative matrix factorization which learns the prominent objects or major components of a collection. In image summarization problem, the summary learned in this manner are inclined to be composed by the salient visual components of the collection. If we heavily penalize on the sparsity parameter such as , which determines the number of bases can be used in a reinterpretation, the model can be reduced to k-medoids (the discrete form of k-means). The k-medoids algorithm is well known as one of the major collection summarization methods [7]. Our sparse representation summarization algorithm can be seen as an extension to the k-medoids. Consequently, considering the limited richness of the visual content of an image, it is necessary to bring in sparsity constraint to the objective function so that limited number of coefficients will take effect in the reconstruction. Hence, only the bases with non-zero coefficients will be involved in the reconstruction of an image. Meanwhile, the bases learned should be diverse; each basis represents for an individual visual aspect and should be as different from each other as possible. So we also include the diversity constraint to the objective. We rewrite Eqn. 2 as follows with sparsity and diversity constraints.

$$\min_{D,A} \sum_{i} ||x_i - D \alpha_i ||_2^2 + \lambda \sum_{i} ||\alpha_i ||_0 + \beta \max_{j\neq k} corr(d_j, d_k)$$

s.t. $\alpha_i \geq 0$

(3)

The summarization problem is reformulated into the optimization problem as Eqn. 3, which can be jointly optimized with respect to $D$ and coefficient matrix $A$. The diversity constraint is determined by the maximum correlation score rather than the average correlation, because diversity quality of the bases set is determined by the least different bases pairs only.

There are two aspects differs the above formulation to the traditional sparse representation: 1) the coefficients have to be non-negative; 2) the dictionary $D$ is selected from a group of given candidates $X$. This can be explained briefly: Firstly, from our description of the accumulated appearance probability of the local features, we know that one image may contain certain type of local features (positive coefficients) or do not contain them (zero coefficients). It does not make sense if one type of visual attributes contributes negatively (negative coefficients) to an image. So Eqn. 3 has to satisfy the constraint that is non-negative element. Secondly, the purpose we learn the dictionary is to get a summarization of the original image set so that the dictionary cannot be learned analytically but can only be selected from the original image set.

3.2. Dictionary Learning and Sparse Coding

The optimization problem defined in Eqn 3. is NP-hard (i.e. the search space is discrete) and most existing algorithms will be inevitable to fall into the trap of local optimum. In contrast, the simulated annealing algorithm is suitable for global optimization problem, such as locating a good approximation to the global optimum of a given function in a large search space.

The basic idea of exploiting the simulated annealing algorithm is to avoid local optimum and efficiently search the solution space for global optimum solution. It is well known that the greedy algorithms seek for local optimal solution and the final result of algorithms like AP and k-medoids, largely depends on the initial condition. Simulated annealing algorithm search the neighborhood solution space for possible candidates, and based on the Metropolis criterion, the candidate that does not decrease the objective still has a chance to be accepted to participate in the next iteration, which can largely avoid falling into the local minimum. Given enough search iterations, it is likely with high probability that the local region include the global minimum can
found. We have followed the idea of simulated annealing and design our algorithm as follows:

Cooling schedule: Traditional cooling schedule is set in a simple way as $T_{k+1} = \alpha T_k$ with $\alpha \in (0,1)$. We will be using the canonical annealing schedules defined as below:

$$T_k = \frac{T_0}{\log(K_0 + k)}$$

where the temperature decrease faster during the initial steps and slower during the later steps. This will reduce the computation cost, because the search space and number of candidate are much larger in the initial steps. The temperature will be used to determine the search range and acceptance probability and it will decrease monotonically during the process to make sure the search will end in a limited number of iterations.

Acceptance probability density function: The reconstruction error gain is measured by the difference of the objective between two consecutive selection of bases. The scale of measurement decreases with the temperature and is compared with a random threshold as below,

$$\exp\left(-\frac{R(D_{k+1}) - R(D_k)}{\alpha T_k}\right) > U$$

where $R(\cdot)$ is the reconstruction function as defined in Eqn. 3. $T_k$ is the current temperature in the $k$th iteration. $U \in [0,1)$ is randomly chosen as the acceptance threshold at each test and new selection is accepted if the above inequity holds. The candidates that decrease the objective are guaranteed to be accepted; while the other candidates are accepted with a probability proportional to the current temperature.

Basis update stage: We iteratively update each basis by searching from its neighborhood in the similarity matrix $S$. The similarity is defined as

$$s_{ij} = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$$

Then we sort the columns of the similarity matrix in decreasing order. For each basis, we randomly search in its neighborhood, in terms of similarity defined above, for new basis. The search range is restricted by $\exp\left(\frac{T_k - T_0}{T_0}\right) \cdot |X|$ which defines the maximum index that can be searched in the sorted column. During the basis update stage, each of the $K$ basis is updated in parallel based on the above criteria. A total number of MaxTries dictionaries are selected in this stage and will be filtered by the acceptance function as defined before. The dictionaries accepted will construct a candidate set and used as the input for next iteration.

Sparse coding stage: Given the tractability of L1-norm problem (P1) and the general intractability of the L0-norm problem (P0), it has been approved that for certain dictionaries solutions, when sufficiently sparse, are the same as the solutions to P0 [18]. As we have discussed above, the highly sparsity of the proposed summarization model, we consider to replace the L0-norm by L1-norm and seek for analytical solution. Furthermore, during the sparse coding stage, the dictionary is fixed, hence, we reduce the objective to the following form which overlook the diversity constraint $\beta \max_{j \neq k} \text{corr}(d_j, d_k)$.

$$f : \min_A \sum_i ||x_i - D\alpha_i||^2_2 + \lambda \sum_i ||\alpha_i||_1 + C$$

$$\forall i \in [1..N], \alpha_i \geq 0$$

The above formulation is similar to nonnegative matrix factorization and nonnegative sparse coding, so we can make use of the multiplicative algorithm [20] to solve the above convex optimization problem. The objective is nonincreasing under the update rule:

$$A^{t+1} = A^t \cdot (D^T X)/(D^T DA^t + \lambda I)$$

where .*$ and ./ denote elementwise multiplication and division (respectively). The coefficient matrix $A$ is updated by simply multiplying nonnegative factors during the update stage, so the elements of $A$ is guaranteed to be nonnegative under this update rule. As long as the initial values of $A$ are all chosen strictly positive (1/k in our case), iteration of this update rule is in practice guaranteed to reach the global minimum to any required precision.

Diversity function: The diversity metric is defined by the correlation between two distributions instead of using Euclidean distance or cosine distance. Because correlation is known to be both scale and shift invariant, which is more appropriate for the accumulated appearance probability property of the bag-of-visual-words model. The correlation between two items is calculated as follows:

$$\text{corr}(d_i, d_j) = \frac{(d_i - \bar{d}_i)(d_j - \bar{d}_j)}{\sigma_i \sigma_j}$$

where $\bar{d}$ is the mean value of the vector and $\sigma_{i,j} \in R$ is the standard deviation.

The dictionary and the coefficients are updated iteratively. When we update one component, the other one is fixed and objective function is calculated as we have discussed above. In actual implementation, the current optimal combination is always saved as $(A_{\text{opti}}, D_{\text{opti}})$ which keeps $R(A_{\text{opti}}, D_{\text{opti}})$ the current minimum. The annealing process stops when the temperature reaches $T_{\text{stop}}$ or the $R_{\text{opti}}$ is not being updated for MaxConseRej(maximum consecutive rejection) times of iteration. Then we go to the iterative basis selection stage, which strictly decrease the reconstruction function until convergence.

Iterative basis selection stage: In this stage, the basis is updated iteratively and the reconstruction function is strictly
decreased during each iteration. Suppose we are updating basis \( b_{ij} \), for those \( i \), whose corresponding coefficient \( \alpha_{ij} \) is not zero, then we fix all the other \( k - 1 \) bases and calculate the residue as:

\[
E_i = \sum_{p \neq j} |x_i - \sum_{p \neq j} d_p \alpha_{ip}|^2 \tag{9}
\]

Then we find a new \( d_j^* \), which will maximally approximate the current residue \( \sum E_i d_j^* \). It is equivalent to

\[
d_j^* = \arg \max \sum E_i d_j^* > \tag{10}
\]

which means \( d_j^* \) is the closest point to the center of all the nonzero \( E_i \). Then we calculate whether \( d_j^* \) decrease the objective function. After \( K \) bases are all updated, we calculate the coefficient matrix based on the method introduced in the sparse coding stage and repeat the update process until convergence. The algorithm stops when no basis is being updated. The purpose for this stage is to make sure the proposed algorithm converge to some point. The whole algorithm is summarized as Algorithm 1.

**Algorithm 1 Proposed Dictionary Learning**

Input: Original image set \( X \in \mathbb{R}^{l \times n} \).
Output: Optimized dictionary \( D \in \mathbb{R}^{l \times k} \)
Initialization: Initial dictionary is appointed by random selection of \( k \) bases from \( X \).

**Basis Update:**

while \( T \) > \( T_{stop} \) and \( R \) < \( MaxConseRej \) do

\[ T^k + 1 = Update(T^k) \]

if accept(\( d' \), \( T^k \)) then

\[ D^{k+1} = \{D^k \cup d'\} \]

\[ A = Sparse(D^k, T^k) \]

if \( R(A, D^k) < R_{opti} \) then

\[ R_{opti} = R(A, D^k) \]

\[ D_{opti} = D^k \]

end if

end if

end while

**Iterative Selection:**

for \( i = 1 \) to \( k \) do

\[ Update(d_i) \]

end for

### 4. Experiment and Algorithm Evaluation

In this section, we report the experimental setup and the evaluation results. The experiment is designed to acquire both objective and subjective performance for the proposed algorithm compared with another 5 baseline algorithms of SDS (sparsifying dictionary selection) [9], K-medoids [7], AP (Affinity Propagation) [8], Greedy (Canonical View)[14] and ARW (Absorbing Random Walk) [15].

#### 4.1. Experimental Setup

**Image Set:** The image set used in this work are collected from ImageNet [19]. We download partial of the entire ImageNet collection and report the summarization result on 13 categories (both scene and object categories) of bakery, banana, bridge, church, cinema, garage, library, monitor, mug, pajama, schoolbus, skyscraper, and mix. The summarization algorithms should work on image collections with any scale of visual variety, so we purposely construct the category of mix by mixing images from multiple categories to strengthen the visual variety of the collection. For each of the categories used in this paper, the number of images ranges from 900 to 1800 and the size of the summarization predefined is reported in Table 1.

**Experimental Specification:** We extract interest points and calculate the SIFT descriptors for all the images we collected. We constructed a codebook of size 1000 with \( k \)-means algorithm as introduced in section 3. Then we apply vector quantization according to the codeword dictionary to all the interest points in the images and construct a 1000-dimensional feature vector for each image, which will be the representation of \( x_i \) and \( d_i \) as appeared in Eqn. 3. It has been proved that the codebook of size 1000 produce good representation of the images. As for the notations used in the following contents, without special indication, we denote the number of images in the given category by \( N \), the number of codewords by \( K \) and the sparsity by \( T \).

**Baseline algorithms:**

The \( k \)-medoids algorithm [7] is a typical clustering-based image summarization algorithm, \( k \) is the number of clusters or the size of the dictionary and basis is selected as the medoid of each cluster. The clustering aims to partition the original set into \( k \) clusters which can minimize the within-cluster sum of square errors:

\[
\min_{S} \sum_{i=1}^{K} \sum_{x_j \in S_i} ||x_j - d_i||^2 \tag{11}
\]

The SDS algorithm [9] represents for a series of greedy algorithms which iteratively select the current best basis. Krause et al. [9] suggested in [9] 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 \( D \), and at every iteration \( i \) adds a new element via

\[
d_i = \arg \min_{d \in \mathcal{X} \cup D} F(D_{i-1} \cup d) \tag{12}
\]

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

The **Affinity Propagation algorithm** [8] updates the availability and responsibility functions in turns. The re-
sponsibility and availability are defined as

\[ r(i, k) \leftarrow s(i, k) - \max_{k' \neq k} \{a(i, k') + s(i, k')\} \quad (13) \]

\[ a(i, k) \leftarrow \min\{0, r(k, k) + \sum_{i' \neq i, i' \neq k} \max\{0, r(i', k)\}\} \quad (14) \]

where \( s(i, k) \) is the similarity between two points. The number of exemplars detected is determined by the choice of preference which is usually set to be median of the input similarities. Algorithms like AP and Greedy does not require a preset number of bases. If this number is required, we can make it by tuning the value of preference. Instead, we fix the value of preference, generate a set of bases with AP, then make sure other algorithms generate same number of bases for same category as shown in Table 1.

The Greedy algorithm [14] follows Simon’s definition of quality function as listed below. The image which maximally increase the quality function at each iteration is added to the basis set \( D \). The algorithm terminates when quality function reduces below zero or the preset number of bases are selected. We tune the penalty weight \( \alpha \) to ensure the required number of bases can be selected.

\[ Q(D) = \sum_{x_i \in X} (x_i \cdot D_{d(i)}) - \alpha |D| - \beta \sum_{d_i \in D} \sum_{d_j > d_i \in D} (d_i \cdot d_j) \quad (15) \]

The ARW algorithm [15] turns the selected items to absorbing state by setting the transition probability from this item to other items to 0; and 1 if transit to itself. We select the item with the largest expected number of visits in current iteration. The average expected number \( v \) is calculated as follows, and \( N \) is the so-called fundamental matrix

\[ v = \frac{N^T e}{n - |D|} \quad (16) \]

\[ N = (I - Q)^{-1} \quad (17) \]

Our proposed algorithm can be compared with the above five baseline algorithms both objectively and subjectively. We evaluate all the algorithms by their reconstruction abilities under sparsity and diversity constrain as defined in Eqn 3, specifically, in terms of mean square error (MSE). Smaller MSE value represents for better reconstruction ability.

**4.2. Experimental Results and Observations**

MSE performance: The MSE value is calculated for all the six algorithms on 13 categories with different dictionary size as shown in Table 1. We have the observation that: 1) The proposed algorithm performs the best in terms of reconstruction ability on all the 13 categories. The results are reported in Table 2. and Fig 2. The improvement of reconstruction ability is significant when compare with other baseline algorithms. 2). The simultaneous summarization algorithms like AP and \( k \)-medoids performed slightly better than iterative summarization algorithms like Greedy, SDS and ARW. The improvement of the proposed algorithm comes from two aspect: firstly, the proposed algorithm consider both the sparsity constraints and diversity constraints; while other baseline algorithms do not have such complete consideration of a good summary; secondly, the adopted simulated annealing algorithm seeks for global optimum solution while all the other five algorithms seeking local optimum solutions.

**Discussion:** The initial choice of \( k \) random basis doesn’t affect the final reconstruction performance. The proposed algorithm has consistent reconstruction value with different input. We have tested to use the clustering result, such as AP and \( k \)-medoids, as the initials as most of other methods will do, and observed no significant difference when compared with completely random initial inputs.

We also observed that the L1-norm sparse coding result can fully replace the L0-norm sparse coding scheme. The coefficients learned with the proposed rule are very sparse, and the majority of the weights concentrate on few basis (2 or 3 basis in general), which coincide with our original assumption.

The sparsity penalty weight \( \alpha \) and diversity weight \( \beta \) will also affect the reconstruction value. We have tune these two parameters so that the the two constraints will affect the reconstruction but not dominate the reconstruction. We can observe the MSE performance in terms of different size of dictionary as in Fig 2. The MSE curve are ap-
<table>
<thead>
<tr>
<th>Method</th>
<th>reconstruction error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDS</td>
<td>0.061</td>
</tr>
<tr>
<td>K-med</td>
<td>0.064</td>
</tr>
<tr>
<td>AP</td>
<td>0.065</td>
</tr>
<tr>
<td>Greedy</td>
<td>0.066</td>
</tr>
<tr>
<td>ARW</td>
<td>0.067</td>
</tr>
<tr>
<td>Our</td>
<td>0.068</td>
</tr>
</tbody>
</table>

Table 2. Performance comparison of the proposed algorithm with 5 other baseline algorithms in terms of reconstruction error; smaller reconstruction error means better reinterpretation ability.

![Figure 2](image_url)

Figure 2. The reconstruction error under different size of dictionary.

approximately smoothly decrease as the size of the dictionary increase. This is the same as we had assumed in section 3. We also observed from Fig. 4 that most of the results strictly decrease as the objective as the size of the dictionary increase, but there are still some outlier that does not fit the curve well. That’s because the simulated annealing algorithm doesn’t guarantee the global optimum is found every time. If we can repeat the learning process multiple times and take the best result we got, then we will have a much higher probability to avoid local optimum. On the other hand, the curve in Fig 2 proved that the proposed algorithm finds close-to-global optimum solution with high probability.

**Computation efficiency:** The computation cost of the proposed algorithm is majorly affected by the annealing schedule which determines the number of iterations the algorithm will take. During each iteration, the most time consuming operation is the non-negative sparse coefficient learning. In actual implementation, the simulated annealing stage terminates after 30 to 40 iterations and the overall computation time is around 2 to 3 minutes. By contrast, simultaneous summarization learning algorithms such as AP and k-medoids computes the fastest in around 30 to 40 seconds. ARW and Greedy algorithms has similar computation cost as compared to the proposed algorithm. The SDS algorithm runs the slowest because it needs to examine the reconstruction performance for every item in the collection during each iteration. All the above experiments are conducted in a 2.6G CPU and 4G memory computation environment.

**Subjective evaluation:** Summarization problem is often task-specific, so the subjective results from user study are meaningful and also inevitable. We have designed and performed user study to evaluate the effectiveness of the proposed dictionary learning based approach with the baseline approaches. The evaluation metric is measure by the users’ feedback on how well the summarization results can recover the overall visual aspects reflected by the original data set. Our survey consists the following components: 1) 30 users (graduate students) are involved in this survey to investigate the summarization result of 13 categories as listed in Table 1. 2) The system interface is shown in Fig 3. The users should be able to explore the the category list (left:treeview), the data set (right:panel), and summarization results as given in the middle blob (summary size may vary based on user’s demand) for all six algorithms. 3) In actual survey, the category names are hidden from the users because we do not want to distract the users’ judgment by involving their semantic understanding of that category. The judgment should rely on the given visual collection of images only. The algorithm names are also hidden from the users to avoid biased opinion. 4) The average scores are reported in Table 3. The results indicate that the proposed dictionary learning approach has higher average appropriateness score compared to the baseline algorithms, which coincide with the objective evaluation result. However, considering the appropriateness score for each individual category, there isn’t an one-sided view of the best summarization approach.

5. Conclusion

Most recent algorithms developed for image collection summarization either lack explicit formulation or quantitative evaluation metric. We had discovered the intrinsic coherence between summarization problem and dictionary learning for sparse representation problem by their ability to use a subset to sparsely reinterpret the original set. We utilize the knowledge from the research areas of dictio-
Table 3. Subjective evaluation of the proposed algorithm with 5 other baseline algorithms in terms of user grading

<table>
<thead>
<tr>
<th></th>
<th>bake</th>
<th>bana</th>
<th>brid</th>
<th>chur</th>
<th>cine</th>
<th>gara</th>
<th>libr</th>
<th>moni</th>
<th>mug</th>
<th>paja</th>
<th>schb</th>
<th>skys</th>
<th>mix</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDS</td>
<td>6</td>
<td>6.9</td>
<td>7.5</td>
<td>7.3</td>
<td>7</td>
<td>6.9</td>
<td>6.3</td>
<td>7.1</td>
<td>6.7</td>
<td>6.4</td>
<td>7.7</td>
<td>7.9</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>K-med</td>
<td>6.2</td>
<td>7.6</td>
<td>7.4</td>
<td>7.2</td>
<td>7.4</td>
<td>7.5</td>
<td>6.9</td>
<td>6.9</td>
<td>6.6</td>
<td>7.3</td>
<td>7.1</td>
<td>8.4</td>
<td>7.2</td>
<td>7.2</td>
</tr>
<tr>
<td>AP</td>
<td>6.7</td>
<td>7.5</td>
<td>7.8</td>
<td>7.1</td>
<td>7.6</td>
<td>7.5</td>
<td>7.1</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
<td>6.9</td>
<td>8.4</td>
<td>7.2</td>
<td>7.3</td>
</tr>
<tr>
<td>Greedy</td>
<td>5.8</td>
<td>6.3</td>
<td>6.9</td>
<td>6.6</td>
<td>6.7</td>
<td>7.1</td>
<td>5.9</td>
<td>6.8</td>
<td>6.5</td>
<td>6.8</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>6.7</td>
</tr>
<tr>
<td>ARW</td>
<td>5.7</td>
<td>6.3</td>
<td>7.3</td>
<td>6.9</td>
<td>7.4</td>
<td>7.8</td>
<td>6</td>
<td>6.1</td>
<td>7.1</td>
<td>6.6</td>
<td>6.6</td>
<td>7.7</td>
<td>6.9</td>
<td>6.8</td>
</tr>
<tr>
<td>Proposed</td>
<td>6.6</td>
<td>7.8</td>
<td>8.1</td>
<td>7.9</td>
<td>8</td>
<td>8.2</td>
<td>7.7</td>
<td>7.4</td>
<td>7.7</td>
<td>7.8</td>
<td>7.5</td>
<td>8.5</td>
<td>7.3</td>
<td>7.7</td>
</tr>
</tbody>
</table>

Figure 3. Screen shot of the system interface of category “bridge”; the algorithm and category name are not hidden in this case.

6. Acknowledgement

This work is partly supported by National Science Foundation of China under Grant 61075014, by Doctoral Program of Higher Education of China (Grant No. 20096102110025, 20116102110027)

References