

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

3-2013

Image collection summarization via dictionary learning for sparse representation

Chunlei YANG

University of North Carolina at Charlotte

Jialie SHEN

Singapore Management University, jlshen@smu.edu.sg

Jinye PENG


Northwest University

Jianping FAN

Northwest University

DOI: <https://doi.org/10.1016/j.patcog.2012.07.011>

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

Citation

YANG, Chunlei; SHEN, Jialie; PENG, Jinye; and FAN, Jianping. Image collection summarization via dictionary learning for sparse representation. (2013). *Pattern Recognition*. 46, (3), 948-961. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/1597

This Journal Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Image collection summarization via dictionary learning for sparse representation

Chunlei Yang^{a,*}, Jialie Shen^b, Jinye Peng^c, Jianping Fan^c

^a Computer Science Department, University of North Carolina at Charlotte, 2993 Penman, Tustin, CA 92782, USA

^b School of Information Systems, Singapore Management University, Singapore 178902, Singapore

^c School of Information Science and Technology, Northwest University, Xian 710069, China

A B S T R A C T

In this paper, a novel approach is developed to achieve automatic image collection summarization. The effectiveness of the summary is reflected by its ability to reconstruct the original set or each individual image in the set. We have leveraged the dictionary learning for sparse representation model to construct the summary and to represent the image. Specifically we reformulate the summarization problem into a dictionary learning problem by selecting bases which can be sparsely combined to represent the original image and achieve a minimum global reconstruction error, such as MSE (Mean Square Error). The resulting “Sparse Least Square” problem is NP-hard, thus a simulated annealing algorithm is adopted to learn such dictionary, or image summary, by minimizing the proposed optimization function. A quantitative measurement is defined for assessing the quality of the image summary by investigating both its reconstruction ability and its representativeness of the original image set in large size. We have also compared the performance of our image summarization approach with that of six other baseline summarization tools on multiple image sets (ImageNet, NUS-WIDE-SCENE and Event image set). Our experimental results have shown that the proposed dictionary learning approach can obtain more accurate results as compared with other six baseline summarization algorithms.

Keywords:

Automatic image summarization
Sparse coding
Dictionary learning
Simulated annealing

1. Introduction

Automatic image summarization, which attempts to select a small set of the most representative images to highlight larger amounts of images briefly, becomes very important to enable interactive navigation and exploration of large-scale image collections [3]. Many multimedia applications can benefit from the results of automatic image summarization: (a) On-line shopping sites generate multiple icon images (i.e., image summary) for each product category by selecting a limited number of the most representative pictures; (b) Tourism websites provide a small set of the most representative photos from large-scale photo gallery and display the photos on their web page to attract visitors, which may further result in low information overload on user navigation; (c) Online image recommendation system learns the user intention in real time and recommends a small amount of most representative images out of a large collection [24]. Such interesting applications have motivated researchers to

develop more effective models and mechanisms for achieving more accurate summarization of large-scale image collections.

For a given image set, most existing summarization techniques follow the same criterion by selecting a small set of the most representative images to highlight all the significant visual properties of the original image set [3]. Thus the task for 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 in large size. If we define $\mathbf{X} \in \mathbb{R}^{d \times n}$ as the original image set in large size and $\mathbf{D} \in \mathbb{R}^{d \times k}, k \ll n, \mathbf{D} \in \mathbf{X}$, as the summary out of the given image set \mathbf{X} , automatic image summarization is to determine the summary \mathbf{D} by minimizing the global reconstruction error in L2-norm:

$$\min_{\mathbf{D}} \|\mathbf{X} - f(\mathbf{D})\|_2^2 \quad (1)$$

The selection of the reconstruction function $f(\cdot)$ is to determine how each image in the original image set \mathbf{X} can be reconstructed by the most representative images in the summary \mathbf{D} . In this paper, we have defined the reconstruction function $f(\cdot)$ as a linear regression model that uses the summary \mathbf{D} to sparsely reconstruct each image in the original set \mathbf{X} . The sparsity means that only limited number of bases will actually be involved in the

* Corresponding author. Tel.: +1 704 491 9489.

E-mail addresses: cyang36@uncc.edu, yangchunlei22@yahoo.com (C. Yang).

reconstruction of an image. The idea of “induced sparsity” has already been introduced in Ma’s work [25], which also learns the sparse coefficients from a given data set. However, Ma’s work fixes the dictionary as the original training set of a given category. In our problem, the dictionary and coefficient matrix are jointly learnt so that the coefficient learning process in [25] can only be considered as an alternative to the sparse coding stage of our proposed work.

From the above description, we now successfully reformulate the task of automatic image summarization into the problem of dictionary learning for sparse representation as shown in Eq. (1). Therefore, two research issues, automatic image summarization and dictionary learning for sparse representation, are linked together according to their intrinsic coherence: both of them try to select a small set of the most representative images that can effectively and sufficiently reconstruct large amounts of images in the original image set.

We have discovered that the image collection summarization problem can be interpreted straightforwardly with the dictionary learning for sparse representation model under the SIFT BoW framework. Therefore, the summarization performance can be directly evaluated by the corresponding value of the reconstruction function. 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 for automatic image summarization has some unique constraints such as the fixed basis selection range, nonnegative and L_0 -norm sparsity of the coefficients. The constraints are critical and differ the proposed framework from most of the existing works. For the basis learning stage, traditional methods such as MOD [35], K-SVD [36], Discriminative K-SVD [37], online dictionary learning [38], all learn or update the basis analytically, which does not restrict the search range. The sparse modeling pipelines introduced in Sapiro’s work [32] propose similar sparse coefficients model, but do not have a restriction on the bases learned either. On the other hand, the summarization problem requires the bases to be chosen from a pool of given candidates, which results in a “selecting” action, rather than “learning”. This observation implies the use of simulated annealing algorithm for discrete bases search, which is the most important difference between the proposed work to other works [32,35–38].

Most existing research work for automatic image summarization evaluate their summarization results subjectively by using user satisfaction and relevancy score. There lacks an objective and quantitative evaluation metric for assessing the performance of various algorithms for automatic image summarization. By reformulating the issue of assessing the quality of summarization results as a reconstruction optimization task, we can objectively evaluate the performance of various algorithms for automatic image summarization in terms of their global reconstruction ability. In addition to the subjective evaluation, the global MSE is defined as the objective evaluation metric to measure the performance of our proposed algorithm for automatic image summarization and compare its performance with that of other 6 baseline methods.

The contributions of this paper reside in three aspects:

- i. The problem of automatic image summarization is reformulated as an issue of dictionary learning for sparse representation. As a result, we can utilize the theoretical methods for sparse representation to solve the problem of automatic image summarization.
- ii. A global optimization algorithm is developed to find the solution of the optimization function for automatic image summarization, which can avoid the local optimum and achieves better reconstruction performance.

3. An interactive image navigation system is designed, which can provide a good platform for users to interactively assess the performance of various algorithms for automatic image summarization.

The rest of this paper is organized as follows: The state-of-the-art techniques for both automatic image summarization and dictionary learning for sparse representation are discussed in Section 2. In Section 3, our proposed algorithm for automatic image summarization is introduced. We present and discuss our experimental results in Section 4. Finally, we conclude this paper in Section 5.

2. Related work

Most existing algorithms for automatic image summarization can be classified 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 are 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 image similarity contexts with a graph structure, where the nodes represent the images and the edges indicate their similarity contexts, finally, the nodes (images) 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 a pre-set number of the most representative images are picked out [6]. Simon et al. [6] have used a greedy method to select the best candidates by investigating the weighted combinations of some important summarization metrics such as likelihood, coverage and orthogonality. Sinha [17] proposed a similar algorithm with the metrics of quality, diversity and coverage. Fan et al. [24] proposed “JustClick” system for image recommendation and summarization which incorporates both visual distribution of the images and user intention discovered during exploration. Wong et al. [15] integrated the dynamic absorbing random walk method to find diversified representatives. The idea is to use the absorbing states to drag down the stationary probabilities of the nearby items to encourage the diversity, where the item with the highest stationary probability in the current iteration is selected. The above greedy methods focus on selecting the current most representative images at each iteration while penalizing the co-occurrence of the similar candidates (images). Our proposed model for automatic image summarization takes the benefit of both two types of approaches, e.g., we use the explicit measurements in the iterative approaches to characterize the property of a summary and we learn the bases (candidates) simultaneously to avoid the possible local optimum solution.

Most existing techniques for dictionary learning and sparse coding use machine learning techniques to obtain more compact representations, 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], which 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 style 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 our proposed approach for automatic image summarization because our model only takes discrete bases rather than numerical outputs from SVD. Besides it, the sparse coefficient could be either positive or negative, which guarantees a smaller reconstruction error, but the bases learned do not have a practical meaning to be considered as the summary. The methods of Matching Pursuit (MP) [11] and Lasso (forward stepwise regression and least angle regression) are widely accepted for sparse coding. These methods could provide us with some ideas on the design of an appropriate sparse coding algorithm. 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 in the case that the original data set satisfies the submodular condition. However, most of the real-world image sets do not satisfy the submodular condition which makes Krause’s algorithm less convincing for automatic image summarization application. All these existing algorithms may fall into the traps of the local optimums, thus the simulated annealing algorithm is adopted in our proposed approach to achieve global optimum with a high probability when enough search steps are performed.

3. Automatic image summarization

In this section, we first define the criterion for assessing the quality of an image summary (i.e., whether the image summary are good enough to effectively reconstruct all the images in the original image set), where the problem of automatic image summarization is reformulated as the issue of dictionary learning under sparsity and diversity constraints. We then point out the significant differences between our reformulation of dictionary learning for automatic image summarization with traditional formulation of dictionary learning for sparse coding.

3.1. Problem reformulation

The BoW (Bag-of-Word) model serves as a basic tool for visual analytic tasks, such as object categorization. The summarization problem, which tries to generalize the major visual components that appear in a collection, will therefore, utilize the BoW model

very well. The choice of local descriptor in BoW model is application dependent: the use of both texon descriptors [29,30] and SIFT (Scale Invariant Feature Transform) [12] descriptors [27,28] is widely observed. Considering the fact that texon descriptors are suitable for scene image categorization, and SIFT descriptor has a much wider range of usage, we have chosen the SIFT descriptor as the feature to construct BoW model.

Each image, in a given set, is represented with BoW model. The “visual words” in BoW model are iconic image patches or fragments which are learned by clustering methods, and therefore represents prominent visual perspectives of the entire collection. The feature vector is represented in a histogram fashion, with each bin value represents the frequency of the corresponding visual word occurrence. We can presume that the major visual contents of an image will be reflected by a large value on the corresponding bins of the feature vector; while other bins will have close-to zero values, which implies non-existence of the corresponding “visual words” in the image. Therefore, the BoW vector of an image can be understood as the distribution of the occurrence probability of the visual words or visual patterns. If we assume the visual patterns appear independently in the images and we will observe the additivity property of the BoW model, which is, one feature vector and be represented by the weighted summation of several other vectors; or the accumulated probability of the appearance of visual patterns. One visual pattern should either present or not present in an image, which implies positive and zero weights respectively. A negative weight for a vector does not have practical meaning in illustrating the additivity property of the BoW model. Therefore, sparse coefficients applied on the dictionary should be nonnegative. Such restriction is unique for summarization problem and BoW model. We also observe similar design in face recognition applications [25], which allows negative coefficients but without providing a practical explanation.

By treating the problem of automatic image summarization as the issue of dictionary learning, each image in the original image set can be approximately reconstructed by a nonnegative weighted linear combination of the summary images, or in other words, represented by accumulated probability of the appearances of various visual words (visual patterns) as shown in Fig. 1. The summary images “beach” and “palmtree” will jointly reconstruct the image which has both two visual objects, and such linear correlation is reflected by the corresponding feature histograms. The above linear reconstruction model illustrates the foundation of how each image can be reconstructed by the exemplars or bases. Also from Fig. 1, one can observe that the richness of the visual content in an image is limited, thus one

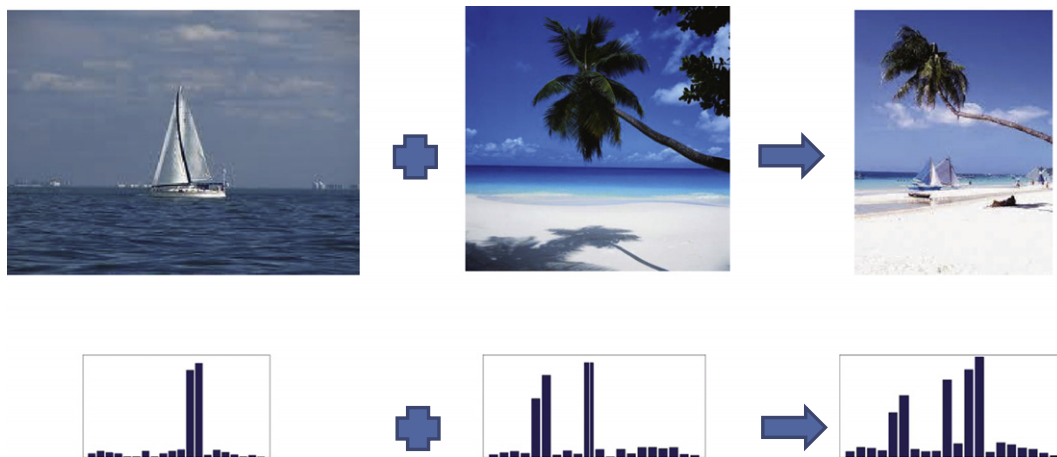


Fig. 1. Demonstration for the additivity property of BoW feature.

image can only be “sparsely” represented by the bases of a dictionary. The definition of sparsity in this work is different from the dictionary selection model such as in [22], our proposed approach for automatic image summarization considers the dictionary to “sparsely” represent the images in the original image set. Based on our new definition of the reconstruction function, automatic image summarization is achieved by minimizing the overall reconstruction error in L2-norm:

$$\min \sum_{i=1}^n \left\| \mathbf{x}_i - \sum_{j=1}^k \mathbf{d}_j \alpha_{ji} \right\|_2^2 \quad (2)$$

where $\mathbf{x}_i, \mathbf{d}_j \in \mathbb{R}^d$, $\alpha_{ij} \in \mathbb{R}_0^+$. \mathbf{x}_i and \mathbf{d}_j are data items from the original collection; α_{ij} is the nonnegative weight for the corresponding \mathbf{d}_j .

For the problem of automatic image summarization, $\{\mathbf{d}_j\}$ is the set of the most representative images that we want to learn, and $\{\mathbf{d}_j\}$ should come from the original image set. The size of $\{\mathbf{d}_j\}$ (summary) is a trade-off between concise summarization of the original image set and accurate reinterpretation of the original image set: a small size of $\{\mathbf{d}_j\}$ means more concise summarization of the original image set but its reinterpretation power for the original image set may reduce; on the other hand, a large size of $\{\mathbf{d}_j\}$ guarantees a better reinterpretation power but the summarization could be verbose.

The idea of this proposed reconstruction model (for automatic image summarization) is similar to nonnegative matrix factorization which learns the prominent objects or major components of an image set. In our problem for automatic image summarization, the summary (which is learned in this manner) is inclined to be composed by the salient visual components of the original image set. If we heavily penalize on the sparsity term α (such as $\|\alpha\|_0 = 1$) which is used for determining the number of bases for reinterpretation, our proposed model for automatic image summarization can be reduced to k -medoids (the discrete form of k -means). The k -medoids algorithm is well known as one of the effective methods for collection summarization [7]. Thus, our proposed approach for automatic image summarization via dictionary learning for sparse representation can be treated as an extension of the k -medoids. Consequently, considering that the richness of the visual content of an image is limited, it is necessary to bring in the sparsity constraint to the objective function for guaranteeing that only a limited number of bases may take effect in the reconstruction. Hence, only the bases with non-zero coefficients are used to reconstruct the images in the original image set. Meanwhile, the bases should be diverse; each basis represents one type of principal visual patterns and all these bases should be different from each other. Thus the diversity constraint should be included in the objective function for dictionary learning. We rewrite Eq. (2) as follows by adding both the sparsity constraint and the diversity constraint.

$$\min_{\mathbf{D}, \mathbf{A}} \sum_i \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \sum_i \|\alpha_i\|_0 + \beta \max_{j \neq k} \text{corr}(\mathbf{d}_j, \mathbf{d}_k) \quad (3)$$

The problem of automatic image summarization is reformulated as the optimization problem in Eq. (3), which can be jointly optimized with respect to the dictionary \mathbf{D} (a small set of most representative images) and the nonnegative coefficient matrix $\mathbf{A} = [a_1^T, \dots, a_n^T]^T$, $a_i \in \mathbb{R}^{1 \times k}$. The diversity constraint is determined by the maximized correlation score rather than the average correlation, or the mean distance to the centroids [26]. Because the diversity (quality of the bases set) is determined by the least different base pairs; while the mean value measurements do not guarantee that the member of any pair differs from each other to some degree.

There are two different aspects between our formulation of sparse coding for automatic image summarization and traditional formulations of dictionary learning for sparse representation: (1) the coefficients $\{\alpha_{ji}\}$ have to be non-negative; (2) the dictionary \mathbf{D} is selected from a group of given candidates (original images) \mathbf{X} rather than their combinations. This can be explained briefly: Firstly, from our description of the accumulated appearance probability of various visual patterns, we know that each image may contain certain types of visual patterns (positive coefficients) or do not contain these visual patterns (zero coefficients). It does not make sense that any type of visual patterns contributes negatively (negative coefficients) to an image in the original image set. Thus Eq. (3) has to satisfy the constraint that α has non-negative elements. Secondly, the purpose for automatic image summarization via dictionary learning is to get a small set of the most representative images from the original image set, thus the dictionary for automatic image summarization should be selected from the original image set rather than learning analytically (such as the combination or variation of the original images).

3.2. Dictionary learning and sparse coding

The optimization problem defined in Eq. (3) is NP-hard (i.e. the search space is discrete and can be transformed to k -medoids problem which is known NP-hard [13]), and most existing algorithms are inevitable to fall into the traps of the local optimums, such as our previous work in [31]. In contrast, the simulated annealing algorithm is suitable for solving the global optimization problem, which can locate a good approximation of the global optimum of a given function in a large search space.

The basic idea of exploiting the simulated annealing algorithm for dictionary learning is to avoid the local optimum by efficiently searching the solution space to obtain the global optimum solution. It is well known that the greedy algorithms seek for the local optimal solution and the final results of the AP and k -medoids algorithms largely depend on the initial inputs. During each iteration, the simulated annealing algorithm searches the neighborhood space for all the possible candidates, which is based on the *Metropolis criterion* and can effectively avoid the local traps, e.g., the candidate that does not decrease the objective function still has a chance to be accepted for the next iteration. The current global best solution will be recorded for future reference. When enough search iterations are performed, the region for the global minimum can be found with a high probability. We follow the idea of simulated annealing to design our algorithm by introducing the major components as below:

Cooling schedule: The cooling schedule is used to decide when the searching process will stop. The canonical annealing schedules is defined as below:

$$T_k = \frac{T_0}{\log(k_0 + k)} \quad (4)$$

where k is the iteration index. The temperature T_k decreases faster during the computational expensive initial steps and slower during the later steps. The temperature can be used to determine the search range and the acceptance probability, the temperature decreases monotonically to make sure that the search will terminate in a limited number of iterations.

Acceptance probability density function: The improvement of reconstruction ability is measured by the difference of the objective function, as defined in Eq. (3), between two consecutive selections of the bases of the dictionary. The scale of the measurement decreases as temperature increases and it is

compared with a random threshold as below:

$$\exp\left(-\frac{R(D_{k+1})-R(D_k)}{\alpha T_k}\right) > U \quad (5)$$

where $R(\cdot)$ is the reconstruction function as defined in Eq. (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 when the above inequity holds. The candidates, that decrease the objective function, are definitely 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

$$s_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \quad (6)$$

Then we sort the columns of the similarity matrix in decreasing order. For each new basis, we randomly search in its neighborhood in terms of similarity as defined above. The search range is restricted by $\exp(T_k - T_0/T_0) \cdot |X|$ which defines the maximum index that can be searched in the sorted column. During the basis update stage, each of these K bases is updated in parallel according to the above criterion. A total number of $MaxTries$ dictionaries are selected in this stage and can be filtered by the acceptance function as defined in Eq. (5). The accepted dictionaries can form a candidate set and be used as the input for next iteration.

Sparse coding stage: Every time we have found a set of candidate dictionaries with the above operation, we will need to calculate a set of coefficients that can minimize the optimization function. As we have discussed before, the coefficient matrix satisfies L0-norm constraint. Given the tractability of L1-norm problem (P1) and the general intractability of the L0-norm problem (P0), it has been proved that the solutions for P1 dictionaries are the same as the solutions for P0 dictionaries when they are sufficiently sparse [18]. As discussed above about the highly sparsity of our proposed model for automatic image summarization, we can replace the L0-norm by L1-norm and seek for analytical solution. Furthermore, during the sparse coding stage, the dictionary is fixed, hence, we can reduce the objective function to the following form which overlooks the diversity constraint $\beta \max_{j \neq k} corr(d_j, d_k)$.

$$R: \min_{\mathbf{A}} \sum_i \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \sum_i \|\alpha_i\|_1 + C \quad (7)$$

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

The above formulation is similar to the 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 function is non-increasing under the update rule:

$$\mathbf{A}^{t+1} = \mathbf{A}^t \cdot * (\mathbf{D}^T \mathbf{X}) \cdot / (\mathbf{D}^T \mathbf{D} \mathbf{A}^t + \lambda \mathbf{1}) \quad (8)$$

where $\cdot *$ and $\cdot /$ denote element-wise multiplication and division (respectively). \mathbf{A} is updated by simply multiplying nonnegative factors during the update stage, so that the elements of \mathbf{A} are guaranteed to be nonnegative under this update rule. As long as the initial values of \mathbf{A} are chosen strictly positive ($1/k$ in our case), the iteration is guaranteed to reach the global minimum.

Diversity function: The diversity metric is measured by the correlation between two distributions rather than their Euclidean distance or cosine distance. Because the correlation of two variables is known to be both scale invariant and shift invariant when compared to Euclidean distance and cosine distance. Thus, it is more appropriate for the additive appearance property of the bag-of-visual-words model. The correlation between two images is calculated as follows:

$$corr(\mathbf{d}_i, \mathbf{d}_j) = \frac{(\mathbf{d}_i - \bar{\mathbf{d}}_i)(\mathbf{d}_j - \bar{\mathbf{d}}_j)}{\sigma_i \sigma_j} \quad (9)$$

where $\bar{\mathbf{d}}$ is the mean value of the vector and σ is the standard deviation.

The dictionary and the coefficients are updated in turns. In practical implementation, the current optimal combination is always saved as (A_{opti}, D_{opti}) which keeps $R(A_{opti}, D_{opti})$ in the current minimum. The annealing process stops when the temperature reaches T_{stop} or the R_{opti} is not being updated for $MaxConseRej$ (number of maximum consecutive rejection) times of iterations. Then, we go to the iterative basis selection stage, which strictly decreases 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 the basis b_j for those i whose corresponding coefficient α_{ij} is not zero, we fix all

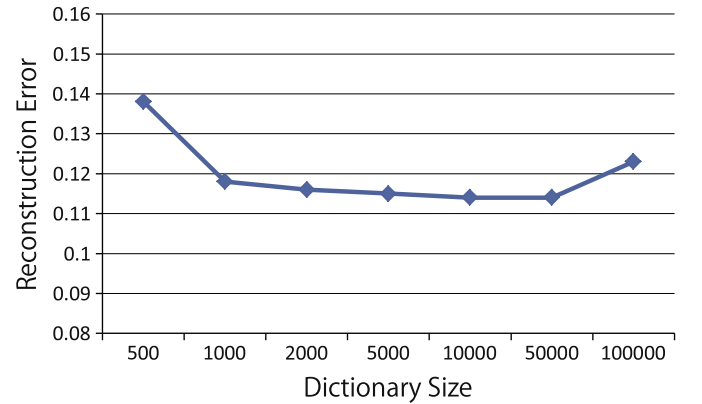


Fig. 2. MSE performance in terms of different dictionary size on a mixture data set with summary size equals to 9.

Table 1
ImageNet data collection statistics: 13 object categories with different size of summary.

	Bakery	Banana	Bridge	Church	Cinema	Garage	Libra
Number of images	1214	1409	1598	1329	1392	1291	1305
Size of summary	10	26	33	34	25	20	16
	Monitor	Mug	Pajama	Schoolbus	Skyscraper	Mix	
Number of images	1399	1573	900	1303	1546	1759	
Size of summary	31	27	18	22	37	31	

the other $k-1$ bases and calculate the residue as:

$$E_i = \sum \left\| \mathbf{x}_i - \sum_{p \neq j} \mathbf{d}_p \alpha_{ip} \right\|^2 \quad (10)$$

Then a new \mathbf{d}_j^* , which can maximally approximate the current residue $\sum E_i \mathbf{d}_j^*$, is found and it is equivalent to

$$\mathbf{d}_j^* = \arg \min \langle \sum E_i, \mathbf{d}_j^* \rangle \quad (11)$$

which means \mathbf{d}_j^* is the closest point to the center of all the nonzero E_i . Then we check whether \mathbf{d}_j^* decreases the objective function or not. After all the K bases are updated, we calculate the coefficient matrix by using the method which is introduced in the sparse coding stage and repeat the updating process until convergence. The algorithm stops when no basis is being updated. The purpose for this stage is to make sure that our proposed algorithm can converge to some points. The algorithm is summarized as Algorithm 1.

Algorithm 1. Proposed Dictionary Learning.

Input: Original image set $\mathbf{X} \in \mathbb{R}^{d \times n}$.

Output: Optimized dictionary $\mathbf{D}_{opti} \in \mathbb{R}^{d \times k}, k \ll n, \mathbf{D}_{opti} \in \mathbf{X}$.

Initialization: Initial dictionary is appointed by random selection of k bases from \mathbf{X} .

Basis Update:

while $T^k > T_{stop}$ and $Rej < MaxConseRej$ **do**

$T^{k+1} = Update_T(T^k)$

for each d in D^k **do**

$d' = Update_D(d)$

if $accept(d', T^k)$ **then**

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

end if

end for

$A = Sparse_Coding(\mathbf{X}, D^k, T^k)$

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

$R_{opti} = R(\mathbf{X}, A, D^k)$

$D_{opti} = D^k$

else

$Rej = Rej + 1$

end if

end while

Iterative Selection:

while not converge **do**

for $i=1$ to k **do**

$Update(d_i)$

end for

$A = Sparse_Coding(\mathbf{X}, D^k, T^k)$

$R_{opti} = R(\mathbf{X}, A, D^k)$

$D_{opti} = D^k$

end while

4. Experiment setup and algorithm evaluation

In this section, we report our experimental setup and algorithm evaluation results. The experiments are designed to acquire both the objective performance and the subjective performance of our proposed algorithm as compared with other six baseline algorithms such as SDS (spasifying dictionary selection) [9], K-medoids [7], AP (Affinity Propagation) [8], Greedy (Canonical View) [14], ARW (Absorbing Random Walk) [15] and K-SVD.

4.1. Experiment setup

Image sets: The image sets used in this work are collected from ImageNet [19], NUS-WIDE-SCENE [21] and Event image set [23].

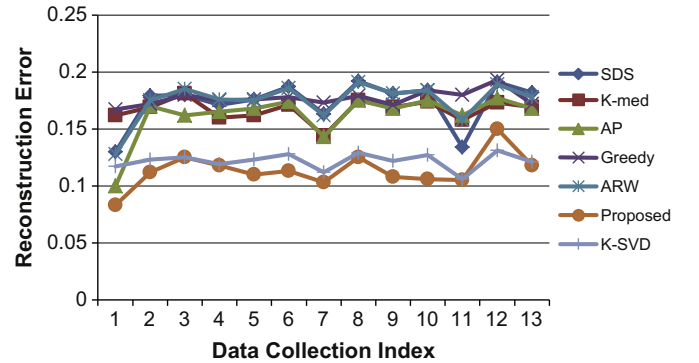


Fig. 3. MSE comparison among all these six algorithms on ImageNet with different summary sizes.

Table 2

Performance comparison of the proposed algorithm with six other baseline algorithms in terms of reconstruction error; 13 object categories selected from ImageNet with different summary sizes.

	Bakery	Banana	Bridge	Church	Cinema	Garage	Library
SDS	0.13	0.179	0.180	0.171	0.176	0.187	0.163
K-med	0.162	0.169	0.181	0.160	0.162	0.171	0.144
AP	0.100	0.170	0.162	0.165	0.168	0.174	0.143
Greedy	0.167	0.172	0.180	0.175	0.176	0.178	0.173
ARW	0.128	0.175	0.185	0.176	0.175	0.186	0.162
Proposed	0.083	0.112	0.125	0.118	0.11	0.113	0.103
K-SVD	0.117	0.123	0.125	0.119	0.123	0.128	0.112
Size	10	26	33	34	25	20	16
	Monitor	Mug	Pajama	Sch-bus	Skyscrap	Mix	Avg.
SDS	0.192	0.181	0.184	0.134	0.191	0.182	0.173
K-med	0.175	0.169	0.174	0.158	0.173	0.169	0.167
AP	0.175	0.168	0.175	0.162	0.177	0.168	0.162
Greedy	0.179	0.171	0.184	0.180	0.193	0.172	0.177
ARW	0.191	0.181	0.184	0.158	0.189	0.178	0.174
Proposed	0.125	0.108	0.106	0.105	0.15	0.118	0.114
K-SVD	0.129	0.122	0.127	0.106	0.131	0.121	0.121
Size	31	27	18	22	37	31	N/A

Best performance values are shown in bold.

ImageNet is an image collection which is organized according to the WordNet hierarchy. The majority of the meaningful concepts in WordNet are nouns (80,000+) which are called "synset". There are more than 20,000 such synsets/subcategories in ImageNet and we have downloaded only partial of this large-scale image set and reported our summarization results on 13 object categories of *bakery*, *banana*, *bridge*, *church*, *cinema*, *garage*, *library*, *monitor*, *mug*, *pajama*, *school bus*, *skyscraper*, and *mix*.

The algorithms for automatic image summarization should work on image collections with various sizes and visual variety, so we have integrated the images from ImageNet to construct a new image category called *mix* by mixing the images from multiple object categories to strengthen the visual diversity and enlarge the size of image category. For each of these 13 categories used in our experiments, the number of images ranges from 900 to 1800 and the predefined size of image summary is reported in Table 1.

The NUS-WIDE database consists of 269,648 images which are collected from Flickr. We focused on a subset called NUS-WIDE-SCENE which covers 33 scene concepts with 34,926 images in total. We have collected 11 scene concepts which are *beach* (449 images), *building* (451), *clouds* (317 images), *hillside* (466 images), *lakes* (383 images), *plaza* (425 images), *running* (302 images), *skyline* (147 images), *sunrise* (111 images), *weather* (225 images) and *zoos* (448 images).

The Event image set contains eight sport event categories: *rowing* (250 images), *badminton* (200 images), *polo* (182 images), *bocce* (137 images), *snowboarding* (190 images), *croquet* (236 images), *sailing* (190 images) and *rock climbing* (194 images). The images in the Event image set are closer to personal photo album which focuses on the presence of people or ongoing activities.

Each of these three image sets covers different visual aspects: ImageNet focuses on object categories, NUS-WIDE-SCENE focuses on natural scene categories, and Event image set focuses on event categories.

Experimental specification: We extract interest points and calculate their SIFT descriptors for image representation. A universal codebook with 1000 visual words is constructed, where the *k*-means algorithm is performed on 10 million interest points as introduced in Section 3 for codebook (dictionary) learning. We have investigated how the size of dictionaries will affect the reconstruction performance. The affection of the dictionary size is evaluated on the reconstruction

performance on the mixture data set and we observed that too small size (less than 500) or too large size (larger than 100,000) dictionary will all reduce the reconstruction performance, as shown in Fig. 2, so that we choose size 1000 for the purpose of computation efficiency. The image representation (1000-dimensional histogram of code words) is obtained by quantifying all the interest points in the images into the codeword dictionary. In our experiments, we have found that our 1000-dimensional codebook can produce good representations of the images. In the following, without special indication, we denote the number of images in the given category by *N* and the number of codewords by *K*.

Baseline algorithms: We have selected six baseline algorithms for comparison.

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 the medoid of each cluster is selected as one basis. The clustering algorithm aims to partition the original image set into *k* clusters which can minimize the within-cluster sum of the square errors:

$$\min_S \sum_{i=1}^K \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \mathbf{d}_i\|^2$$

The *SDS algorithm* [9] represents a series of greedy algorithms which iteratively select the current best basis. Krause et al. suggested in [9] that the local optimal derived by the greedy

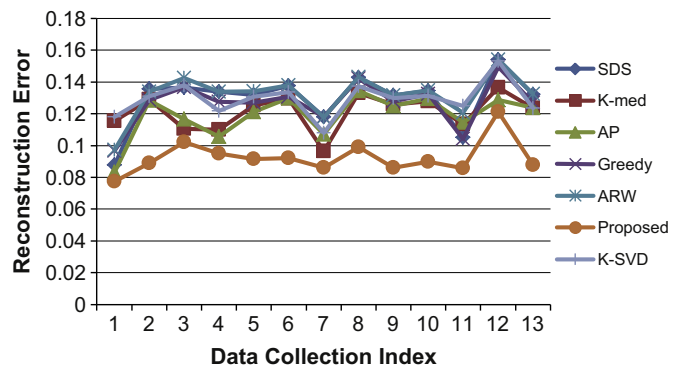


Fig. 4. MSE comparison among the algorithms on ImageNet with equal summary size of 9.

Table 3

Performance comparison of the proposed algorithm with six other baseline algorithms in terms of reconstruction error; 13 object categories selected from ImageNet image set with equal summary size of 9.

	Bakery	Banana	Bridge	Church	Cinema	Garage	Library
SDS	0.088	0.136	0.137	0.134	0.132	0.138	0.118
K-med	0.115	0.129	0.110	0.109	0.125	0.129	0.096
AP	0.083	0.128	0.116	0.105	0.121	0.129	0.107
Greedy	0.096	0.128	0.136	0.127	0.127	0.130	0.117
ARW	0.097	0.133	0.142	0.133	0.134	0.137	0.117
Proposed	0.077	0.088	0.102	0.094	0.091	0.092	0.086
K-SVD	0.118	0.130	0.137	0.121	0.130	0.133	0.108
	Monitor	Mug	Pajama	Sch-bus	Skyscrap	Mix	Avg.
SDS	0.143	0.132	0.134	0.105	0.154	0.132	0.129
K-med	0.132	0.125	0.127	0.113	0.136	0.123	0.121
AP	0.134	0.125	0.129	0.114	0.128	0.123	0.119
Greedy	0.142	0.127	0.132	0.104	0.149	0.126	0.126
ARW	0.142	0.131	0.134	0.120	0.153	0.132	0.131
Proposed	0.099	0.086	0.089	0.085	0.121	0.087	0.092
K-SVD	0.137	0.129	0.131	0.124	0.153	0.123	0.129

Best performance values are shown in bold.

algorithm is a near-optimal solution when the data collection satisfy the submodular condition. The greedy algorithm starts with an empty dictionary \mathbf{D} , and at every iteration i adds a new element (basis) via

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

where F is the evaluation function. The SDS algorithm is modified to satisfy our positive coefficient constraint.

The *Affinity Propagation algorithm* [8] updates the availability function and the responsibility function in turns as below:

$$r(i, k) \leftarrow s(i, k) - \max_{k' : s(i, k') \neq k} \{a(i, k') + s(i, k')\}$$

$$a(i, k) \leftarrow \min \left\{ 0, r(k, k) + \sum_{i' : i' \neq (i, k)} \max\{0, r(i', k)\} \right\}$$

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

The *Greedy algorithm* [14] follows Simon’s definition of the quality function as written below. The image, which maximally increases the quality function at each iteration, is added to the basis set D . The algorithm terminates when the quality function reduces below zero or the preset number of bases is reached. We tune the penalty weight α to ensure the required number of bases can be selected automatically.

$$Q(D) = \sum_{\mathbf{x}_i \in \mathbf{X}} (\mathbf{x}_i \cdot \mathbf{D}_{d(i)}) - \alpha |\mathbf{D}| - \beta \sum_{\mathbf{d}_i \in \mathbf{D}} \sum_{\mathbf{d}_j > \mathbf{d}_i \in \mathbf{D}} (\mathbf{d}_i \cdot \mathbf{d}_j)$$

The *ARW algorithm* [15] turns the selected items to the absorbing state by setting the transition probability to 0 (from the current item to other items), and 1 when it transits to itself. The item, which has the largest expected number of visits in the current iteration, is selected. The average expected number ν is

calculated as follows, and N is the so-called fundamental matrix

$$\nu = \frac{\mathbf{N}^T \mathbf{e}}{n - |\mathbf{D}|} \mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$$

The *K-SVD algorithm* [10] is flexible, and works in conjugation with any sparse coding algorithms. In order to incorporate the K-SVD algorithm into the proposed framework, we learn the sparse coefficient matrix under the non-negative constraint. In the dictionary update stage, we follow the same SVD decomposition operation and update the basis iteratively. After the dictionary learned from K-SVD, we will assign each basis in the dictionary to its nearest neighbor in the original set and construct the final summarization.

For automatic image summarization, our proposed algorithm is compared with all these six baseline algorithms objectively and subjectively. We compare all these algorithms (our proposed algorithms and six baseline algorithms) on their reconstruction abilities under the sparsity and diversity constraints as defined in Eq. (3), specifically, in terms of mean square error (MSE). Smaller MSE value indicates better reconstruction ability.

4.2. Experimental results and observations

MSE performance on ImageNet: The MSE value is calculated for all these six algorithms (our proposed algorithm and six baseline algorithms) on 13 object categories where the size of image summary is predefined as shown in Table 3. We have observed that: (a) Our proposed algorithm has the best performance in terms of the reconstruction ability on 12 out of 13 object categories. The results are reported in Table 2 and Fig. 3. For our proposed algorithm, its improvement on the reconstruction ability is insignificant when compared with K-SVD, but is significant as compared with other five baseline algorithms. (b) The simultaneous summarization algorithms like AP and k -medoids performed slightly better than the iterative summarization algorithms like Greedy, SDS and ARW. (c) The performance improvement on the *mix* category is especially significant, which implies that the proposed algorithm has better summarization ability on more visually diverse data collections.

The improvement comes from two aspects: (1) our proposed algorithm considers both the sparsity constraint and the diversity constraint while other baseline algorithms do not have such

Table 4

Performance comparison of the proposed algorithm with six other baseline algorithms in terms of reconstruction error; 11 scene categories selected from NUS=WIDE-SCENE with equal summary size of 9.

	Beach	Building	Clouds	Hillside	Lakes	Plaza
SDS	0.134	0.127	0.125	0.124	0.123	0.114
K-med	0.124	0.121	0.105	0.135	0.116	0.111
AP	0.125	0.116	0.115	0.109	0.119	0.103
Greedy	0.123	0.117	0.122	0.121	0.123	0.110
ARW	0.140	0.123	0.121	0.135	0.127	0.120
Proposed	0.119	0.106	0.106	0.107	0.108	0.097
K-SVD	0.131	0.122	0.107	0.135	0.109	0.112
	Running	Skyline	Sunrise	Weather	Zoos	Avg.
SDS	0.125	0.125	0.130	0.148	0.107	0.127
K-med	0.121	0.110	0.126	0.131	0.097	0.120
AP	0.113	0.108	0.118	0.130	0.099	0.116
Greedy	0.118	0.110	0.109	0.130	0.110	0.119
ARW	0.130	0.123	0.127	0.137	0.116	0.128
Proposed	0.104	0.102	0.110	0.127	0.090	0.109
K-SVD	0.123	0.107	0.119	0.124	0.108	0.118

Best performance values are shown in bold.

complete consideration of a good summary; (2) the simulated annealing algorithm is adapted to seek for the global optimum solution while all the other five algorithms seek the local optimum solutions. When the same size of image summary is used, we have also compared their performance in terms of MSE values as shown in Table 3 and Fig. 4. The performance is similar to the predefined size of summary experiment.

MSE performance on NUS-WIDE-SCENE: The MSE value is calculated for all these six algorithms on 11 scene categories in NUS-WIDE-SCENE image set when the size of image summary is fixed. Similar performance is obtained as what we have got in ImageNet, however, the performance improvement for the proposed algorithm, and also among all these six algorithms is not as significant as we have observed in ImageNet data set, and the proposed algorithm is outperformed by other algorithms on two categories as shown in Table 4 and Fig. 5. The absolute MSE value and the difference among the baseline algorithms are also smaller as compared with the object categories in ImageNet. The result demonstrates that the images in the scene categories are more evenly distributed and our proposed algorithm does not have as distinguish performance as we have obtained in the object categories.

MSE performance on Event image set: The MSE value is calculated for all these six algorithms on eight categories in the Event image set with an equal summary size of 9. We have observed that the MSE curves are more consistent as compared with the MSE curves for ImageNet and NUS-WIDE-SCENE and the

difference is very consistent and relatively small as shown in Table 5 and Fig. 6. The reason is that the images for the Event image set is organized much better and more consistent on visual content as compared with ImageNet and NUS-WIDE-SCENE.

Discussion: We will discuss how the major components and parameters will affect the performance of the proposed algorithm.

The *spatial information* is believed to be discarded with the proposed SIFT BoW model, which is one of the major drawbacks for BoW model. However, for the image collection summarization applications, the spatial distribution or organization of objects within a certain image is not critical. The critical property is the existence of an object or visual component in an image, and the distribution of the occurrence probability of the visual words within an image. Under such interpretation, the MSE measure should be enough to serve for image collection summarization task evaluation, compared to measuring metric such as SSIM [33].

As for the *choice of the feature* and the *size of the image*, we further conduct another experiment on the eight scene categories of Torralba dataset [34], whose images have the same size (256 by 256). Although GIST feature [34] does not have a straightforward interpretation ability as SIFT BoW feature, we still consider it appropriate for scene image representation and use this feature for summarization task. The complete result is reported in Fig. 7, and can be briefly concluded as: the average MSE value for the proposed algorithm is 0.3833, with K-SVD followed closely by 0.3871. The other five algorithms performs relatively poor in the range between 0.4 and 0.44. Similar to NUS-WIDE-SCENE and

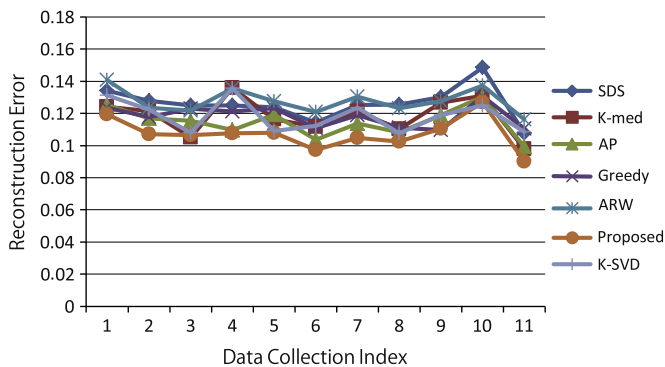


Fig. 5. MSE comparison among the algorithms on NUS-WIDE-SCENE with equal summary size of 9.

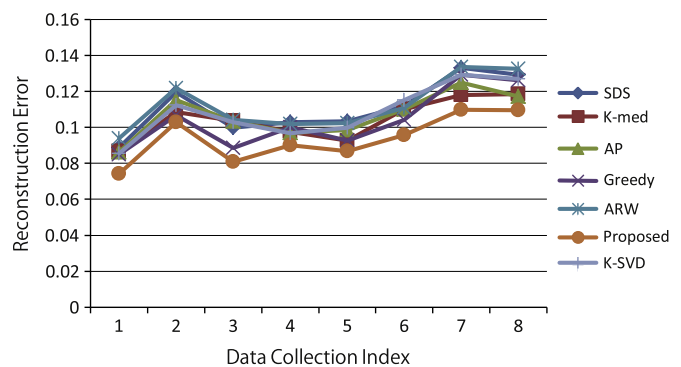


Fig. 6. MSE comparison among the algorithms on Event image set with equal summary size of 9.

Table 5

Performance comparison of the proposed algorithm with six other baseline algorithms in terms of reconstruction error; eight event categories selected from Event image set with equal summary size of 9.

	Rockclimb	Badminton	Bocce	Croquet	Polo
SDS	0.088	0.119	0.099	0.102	0.103
K-med	0.086	0.108	0.103	0.097	0.092
AP	0.086	0.115	0.103	0.096	0.098
Greedy	0.084	0.106	0.088	0.099	0.092
ARW	0.093	0.121	0.104	0.101	0.102
Proposed	0.074	0.102	0.08	0.09	0.086
K-SVD	0.085	0.112	0.102	0.096	0.099
	Rowing	Sailing	Snowboard	Avg.	
SDS	0.111	0.133	0.129	0.111	
K-med	0.11	0.117	0.118	0.104	
AP	0.109	0.124	0.117	0.106	
Greedy	0.104	0.129	0.126	0.104	
ARW	0.110	0.133	0.132	0.112	
Proposed	0.095	0.109	0.109	0.093	
K-SVD	0.115	0.129	0.126	0.108	

Best performance values are shown in bold.

Event image dataset, the Torralba 8 scene category dataset is also consistent on visual content, the performance improvement of the proposed algorithm is not as significant as with object datasets. For this data collection, both K-SVD and the proposed algorithm can achieve close to optimal results. In conclusion, the consistency of visual content is the critical factor for summarization task, rather than the spatial layout or size of the image.

The *initial choice of k random bases* does not affect the final reconstruction performance. Our proposed algorithm has consistent reconstruction value with different inputs. We have used the clustering results from AP or k -medoids as the initial inputs and no significant difference is observed as compared with random initial inputs.

We also observed that the *$L1$ -norm sparse coding scheme* can be used to replace the $L0$ -norm sparse coding scheme. The coefficients are very sparse, and the majority of the weights concentrate on a few number of bases (two or three bases in general;

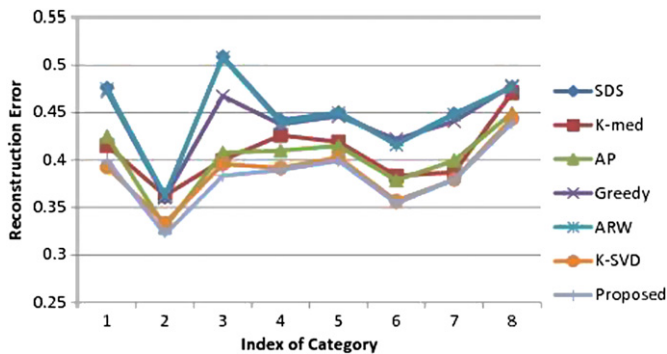


Fig. 7. MSE comparison among all the six algorithms on the Torralba-8 dataset with GIST feature.

Number of iteration	40
Diversity weight β	0.05
Number of different initials	3
MaxConseRej	20
MaxTries	40
Temperature decrease rate	0.9

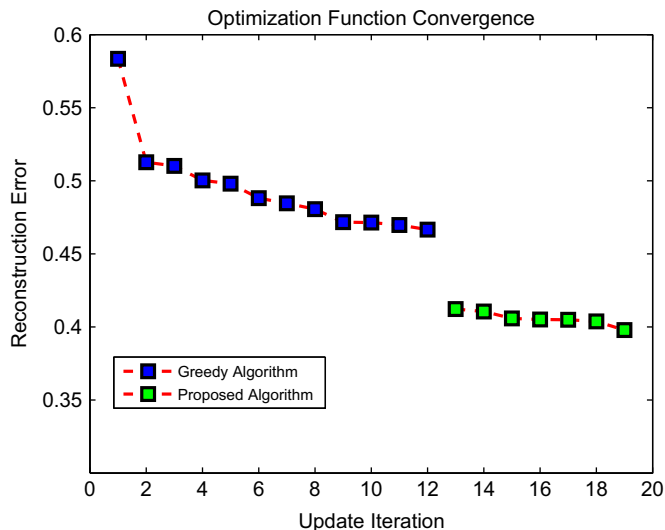


Fig. 8. Optimization comparison in terms of reconstruction error: the blue dot is the greedy algorithm; the red dot is the SA algorithm. Only the updated steps are shown in this figure, thus the curve is not smooth. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

extremely small as compared with the size of the dictionary), which coincide with our assumption.

The *sparsity penalty weight α* and the *diversity weight β* may also affect the reconstruction value. We have tuned these two parameters, so that two constraint terms can contribute equally to the reconstruction function. We have tuned these two parameters under the following rules: (a) the sparsity penalty weight α is determined first to make sure that each image is represented sparsely enough by the dictionary; (b) we tune the diversity weight of β , so that the MSE curve decreases when the summary size is increased. The MSE curves under different β values are shown in Fig. 9. The value of $\beta = 0.05$ (the middle curve in Fig. 9) produces a balanced diversity term while other β values lead to unbalanced diversity terms. We also observed that the MSE curve (y -axis in Fig. 9) decreases when the summary size increases (x -axis in Fig. 9). This observation coincides with our assumption in Section 3 that the reconstruction ability will increase as the size of the summary increases. We also observed that most of the results strictly decrease the objective as the size of the dictionary increases, but there are still some outliers that do not fit the curve well. The reason is that the simulated annealing algorithm does not guarantee that the global optimum is found every time

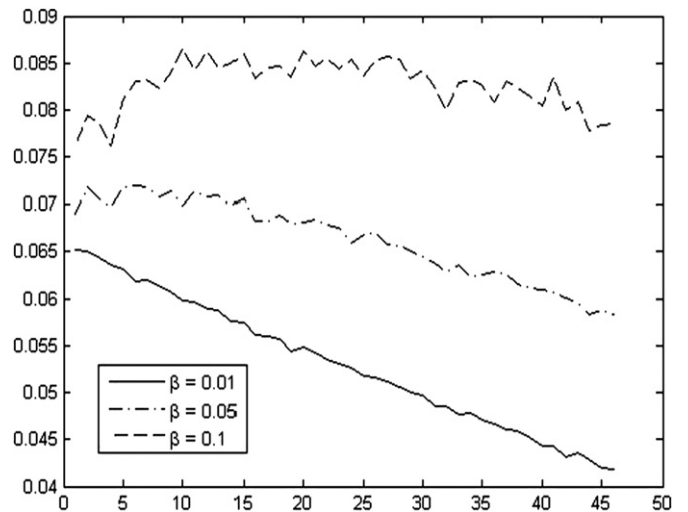


Fig. 9. The MSE curve under different β value; x -axis represents the size of the summary, y -axis represents the MSE value.

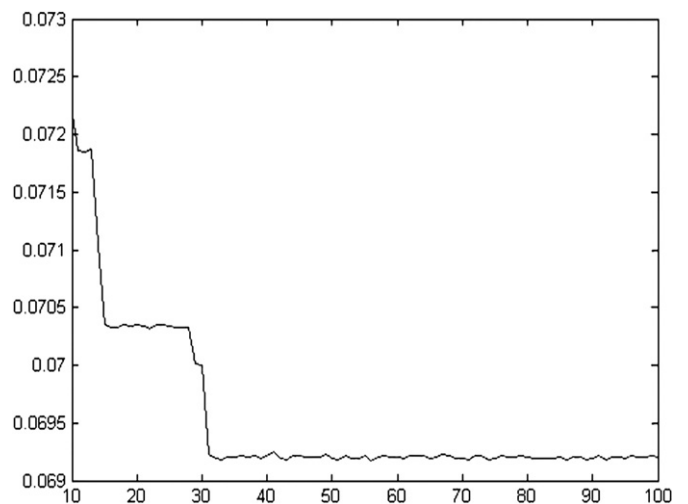


Fig. 10. The MSE curve under different number of iterations; x -axis is the number of iterations of the proposed algorithm, y -axis represents the MSE value.



Fig. 11. Screen shot of the system interface of category “clouds”; the algorithm and category names are not hidden in this case.

Table 6

Subjective evaluation of the proposed algorithm with six other baseline algorithms in terms of user grading on ImageNet with different summary size.

	Bakery	Banana	Bridge	Church	Cinema	Garage	Library
SDS	6	6.9	7.5	7.3	7	6.9	6.3
K-med	6.2	7.6	7.4	7.2	7.4	7.5	6.9
AP	6.7	7.5	7.8	7.1	7.6	7.5	7.1
Greedy	5.8	6.3	6.9	6.6	6.7	7.1	5.9
ARW	5.7	6.3	7.3	6.9	7.4	7.8	6
Proposed	6.6	7.8	8.1	7.9	8	8.2	7.7
K-SVD	6.6	6.7	7.1	7.1	7.2	7.5	6.9
Size	10	26	33	34	25	20	16
	Monitor	Mug	Pajama	Schoolbus	Skyscraper	Mix	Avg.
SDS	7.1	6.7	6.4	7.7	7.9	7	7
K-med	6.9	6.6	7.3	7.1	8.4	7.2	7.2
AP	7.2	7.2	7.2	6.9	8.4	7.2	7.3
Greedy	6.8	6.5	6.8	7	8	7	6.7
ARW	6.1	7.1	6.6	6.6	7.7	6.9	6.8
Proposed	7.4	7.7	7.8	7.5	8.5	7.3	7.7
K-SVD	6.5	6.6	7.2	7.3	7.2	7	7
Size	31	27	18	22	37	31	N/A

Best performance values are shown in bold.

Table 7

Subjective evaluation of the proposed algorithm with six other baseline algorithms in terms of user grading on ImageNet with equal summary size of 9.

	Bakery	Banana	Bridge	Church	Cinema	Garage	Library
SDS	6.4	5.8	5.9	6.7	6	5.8	5.3
K-med	7.2	5.9	6.7	8.9	6	6.2	5.1
AP	7.3	5.7	8.6	5.4	7.8	7	7.9
Greedy	8.7	6.2	5.6	5.7	6.6	5.4	6
ARW	8.5	6.8	5.9	8.6	7.3	6.1	8.2
Proposed	7.4	8.3	6.2	6.7	7.4	7.6	8.7
K-SVD	7.3	7.7	5.9	6.1	6.6	6.7	7.2
	Monitor	Mug	Pajama	Schoolbus	Skyscraper	Mix	Avg.
SDS	6.9	7	6.4	5.3	5.4	8.5	6.0
K-med	6.8	7.4	8.5	7.7	8.1	5.7	7.0
AP	7.1	6.5	8.1	7.8	8.6	7.9	7.3
Greedy	7.3	5.9	8.9	6	7.6	6.3	6.6
ARW	5.9	6.9	5.1	6.3	6.9	7.7	6.8
Proposed	8.8	7.7	8.6	5.5	7.8	5.1	7.5
K-SVD	8.2	7.2	5.9	6.3	6.2	6.1	6.7

Best performance values are shown in bold.

Table 8

Subjective evaluation of the proposed algorithm with six other baseline algorithms in terms of user grading on NUS-WIDE-SCENE with equal summary size of 9.

	Beach	Building	Clouds	Hillside	Lakes	Plaza
SDS	7	8.2	6.9	5.2	8.2	7.8
K-med	6.9	7.3	5.6	7.7	8.2	7.5
AP	7.4	8.5	7	7	7.6	6.7
Greedy	8.6	5.7	8.9	5.1	7.8	8.2
ARW	7.4	5.9	7.8	5.2	5.5	6.8
Proposed	8.4	8.6	6.8	5.3	7	8.3
K-SVD	7.4	8.2	7.1	7.5	6.7	7.7
	Running	Skyline	Sunrise	Weather	Zoos	Avg.
SDS	5.3	5.2	6.1	6.4	5.2	6.5
K-med	5.5	6.5	6.7	5.7	7.9	6.8
AP	8.3	7.6	5.6	7.8	7.1	7.3
Greedy	5.6	7.1	5	6.9	6	6.8
ARW	6.5	6.6	6.9	6.3	6.6	6.5
Proposed	8.2	8.5	7.4	8.6	8.7	7.8
K-SVD	5.2	5.9	6.8	6.6	6.4	6.8

Best performance values are shown in bold.

(although it is close to the global optimum). If we can sacrifice the efficiency and repeat the learning process with more iterations, we can have a much higher probability to achieve the global optimum. In other words, the curve in Fig. 9 can prove that our proposed algorithm finds the close-to-global optimum solution with high probability.

We will discuss about the *convergence* of the simulated annealing algorithm in this task. The use of annealing schedule is to make it possible to avoid local optima, and terminates the basis update stage in a limited number of steps. The newly accepted updates do not critically decrease the reconstruction error; the solution, which does not decrease the optimization function, still has a chance to be accepted, which makes it possible to jump out of a local minimum neighborhood. We have compared the proposed algorithm with greedy algorithm in terms of reconstruction error on a given data set with GIST feature. The result can be found in Fig. 8. We observed that after the greedy algorithm converges at a local optimum position (blue dot), the SA algorithm (green dot) could still jump out of the local optimal neighborhood and find a better optimal solution. The reconstruction error curve is not smooth because the solution space is not continuous. Some important factors such as iteration number, number of attempts with different initials, cooling schedule, would all affect the convergence result. The optimal parameters are given as below:

Number of iteration	40
Diversity weight β	0.05
Number of different initials	3
MaxConseRej	20
MaxTries	40
Temperature decrease rate	0.9

We further tested how the *number of iterations* may affect our proposed algorithm and reported the summarization result for the category of “clouds” as in Fig. 10. We have repeated the algorithm for 90 times and each time with different number of iterations. We have observed that the optimization function can find close to optimum solution when a certain amount of iteration is guaranteed.

Our proposed approach treated each image in the image collection equally for getting the summary, so there does not

Table 9

Subjective evaluation of the proposed algorithm with 6 other baseline algorithms in terms of user grading on Event image set with equal summary size of 9.

	Rockc	Badm	Bocce	Croq	Polo	Rowi	Saili	Snowb	Avg.
SDS	6.2	5.7	8.5	5.6	6.9	6	7.9	8.2	6.8
K-med	7.8	5.5	7.6	8.4	5.4	6.1	6.3	6	6.6
AP	6.6	7.7	5.1	6.8	5.7	6.5	8.2	8.6	6.9
Greedy	7.6	8.9	5.7	7.5	7.3	7.4	7.3	7.3	7.3
ARW	7.1	5.6	6.4	6.5	5.9	6	5.4	5	5.9
Proposed	8.9	7.6	7.2	8.9	6.7	7.3	8.5	8.5	7.9
K-SVD	8.4	7.7	5.6	6.7	7.2	6.9	8.1	6.8	7.2

Best performance values are shown in bold.

exist so-called “outliers” (a group of similar images that are far different from the rest of the data set). As a result, the summarization result may not coincide with the human perception of that image categories. For example, the “bakery” category in ImageNet contains a bunch of blank images which maybe the result of a broken download link. So the summarization results for our proposed approach can always include a blank image which are usually eliminated by some other algorithms.

Computation efficiency: The computation cost of our proposed algorithm is largely affected by the annealing schedule which is used to determine the number of iterations. During each iteration, the most time consuming operation is to learn the non-negative sparse coefficients. In practical implementation, the simulated annealing stage terminates after 30 to 40 iterations and the overall computation time is around 2 to 3 min for each image set (with around 1400 images). By contrast, the simultaneous summarization learning algorithms such as AP and *k*-medoids take around 30 to 40 s. The ARW, K-SVD and Greedy algorithms has similar computation cost as compared with our proposed algorithm. The SDS algorithm runs slowest because it needs to examine the reconstruction performance for every image in the image set during each iteration. All these experiments are carried out in a 2.6G CPU and 4G memory computation environment.

Subjective evaluation: Image summarization is often task-specific, so the subjective results from user study are meaningful and also inevitable. We have performed user study to evaluate the effectiveness of our proposed approach and compared with other baseline approaches. The evaluation metric is measured by the users’ feedback on how well the summarization results can

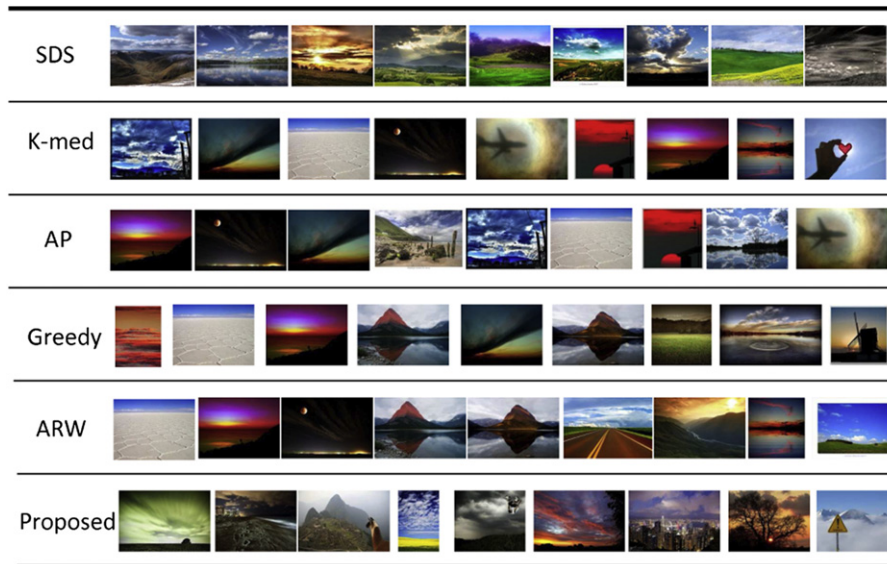


Fig. 12. Summarization results for category “clouds” for the proposed algorithm and 5 other baseline algorithms (without K-SVD); a tile-view illustration of Fig. 11.

recover the overall visual aspects for the original image set.¹ Our survey consists the following components : (1) 30 users (graduate students) are involved in this survey to investigate the summarization results for three image sets. (b) The system interface is shown in Fig. 11. A tile-view of the example summarization result, as shown in Fig. 11, can be found in Fig. 12. The users should be able to explore the image category list (left: treeview), the image set (right: panel), and summarization results as given in the middle blob (summary size may vary according to user’s demand) for all six algorithms (our proposed algorithm and other six baseline algorithms). (c) In actual survey, the category names are hidden from users because we do not want to distract users’ judgment by involving their semantic understanding of that image category. The judgment should rely only on the visual aspects of the images. The algorithm names are also hidden from users to avoid biased opinion. (d) The average scores are reported in Tables 6–9. The results indicate that our proposed approach (via dictionary learning) has higher average appropriateness score as compared with other baseline algorithms, which coincides with the objective evaluation results.

5. Conclusion

Most existing algorithms for image summarization lack either explicit formulation or quantitative evaluation metric. We had discovered that there is an intrinsic coherence between the problem of image collection summarization and the issue of dictionary learning for sparse representation, which both focus on selecting a small set of the most representative images to sparsely reinterpret the original image set in large size. We have explicitly reformulated the problem of automatic image summarization by using a sparse representation model and the simulated annealing algorithm is adopted to solve the optimization function more effectively. The reconstruction ability in terms of the MSE are used to objectively evaluate various algorithms for automatic image summarization. Our proposed algorithm outperformed the six baseline algorithms both objectively and subjectively on three different image sets.

¹ The score ranges from 0 to 10, with 10 represents that all the visual aspects can be discovered by the summarization result. Visual aspects usually means salient objects or major scenes that are reflected in the original image set.

Acknowledgment

This project is supported by National Science Foundation of China under 61272285.

References

- [1] E. Jaffe, M. Naaman, T. Tassa, M. Davis Generating summaries for large collections of geo-referenced photographs, in: Proceedings of the International Conference on World Wide Web, 2006, pp. 853–854.
- [2] T. Denton, M. Demirci, J. Abrahamson, A. Shokoufandeh, S. Dickinson, Selecting canonical views for view-based 3-D object recognition, in: International Conference on Pattern Recognition, 2004.
- [3] Y. Jing, S. Baluja, H. Rowley, Canonical image selection from the web, in: Conference on Image and Video Retrieval, 2007.
- [4] K. Engan, S. Aase, J. Husoy, Method of Optimal Directions for Frame Design, in: Proceedings of ICASSP’99, 1999.
- [5] A. Gersho, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Boston, 1992.
- [6] I. Simon, N. Snavely, S. Seitz, Scene summarization for online image collections, in: International Conference on Computer Vision, ICCV, 2007, pp. 1–8.
- [7] Y. Hadi, F. Essannouni, R. Thami, Video summarization by k-medoid clustering, in: ACM Symposium on Applied Computing, SAC, 2006.
- [8] B. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–977.
- [9] A. Krause, V. Cevher, Submodular dictionary selection for sparse representation, in: Proceedings of International Conference on Machine Learning, ICML, 2010.
- [10] M. Aharon, M. Elad, A. Bruckstein, K-SVD: design of dictionary for sparse representation, in: Proceedings of SPARS, 9–12, 2005.
- [11] S. Mallat, Z. Zhang, Matching pursuit with time-frequency dictionaries, IEEE Transactions on Signal Processing 41 (12) (1993) 3397–3415.
- [12] D. Lowe, Distinctive image features from scale invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.
- [13] B. Natarajan, Sparse approximate solutions to linear systems, SIAM Journal of Computing 24 (2) (1995) 234–277.
- [14] I. Simon, N. Snavely, S. Seitz, Scene summarization for online image collections, in: International Conference on Computer Vision, 2007.
- [15] J. Wang, L. Jia, X. Hua, Interactive browsing via diversified visual summarization for image search results, Multimedia Systems 17 (5) (2011) 379–391.
- [16] D. Deng, Content-based image collection summarization and comparison using self-organizing maps, Journal of Pattern Recognition 40 (2) (2007).
- [17] P. Sinha, Summarization of archived and shared personal photo collections, in: 20th International World Wide Web Conference, 2011.
- [18] D. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization, in: Proceedings of National Academy of Science USA, 2003.
- [19] F. Li, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Proceedings of CVPR, 2005.
- [20] P. Hoyer, Non-negative sparse coding, in: Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, 2002.

- [21] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng, NUS-WIDE: A Real-World Web Image Database from National University of Singapore, in: ACM International Conference on Image and Video Retrieval, Greece, July 8–10, 2009.
- [22] Y. Cong, J. Yuan, J. Luo, Towards scalable summarization of consumer videos via sparse dictionary selection, *IEEE Transactions on Multimedia* 99 (2011).
- [23] L. Li, F. Li, What, where and who? Classifying event by scene and object recognition, in: *IEEE International Conference in Computer Vision (ICCV)*, 2007.
- [24] J. Fan, D. Keim, Y. Gao, H. Luo, Z. Li, JustClick: personalized image recommendation via exploratory search from large-scale Flickr images, *IEEE Transactions on Circuits and Systems for Video Technology* 19 (2) (2009) 273–288.
- [25] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE Transactions on PAMI* 31 (2009) 210–227.
- [26] N. Shroff, P. Turaga, R. Chellappa, Video precis: highlighting diverse aspects of videos, *IEEE Transactions on MM* 12 (2010) 853–868.
- [27] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: *International Conference on Computer Vision*, 2003.
- [28] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: *European Conference on Computer Vision*, 2004.
- [29] L. Renninger, J. Malik, When is scene identification just texture recognition? *Vision Research* 44 (2004) 2301–2311.
- [30] S. Battiato, G. Farinella, G. Gallo, D. Ravi, Exploiting textons distributions on spatial hierarchy for scene classification, *Journal of Image and Video Processing*, 2010, pp. 7:1–7:13.
- [31] C. Yang, J. Shen, J. Fan, Effective summarization of large-scale web images, *ACM Multimedia*, 2011.
- [32] A. Castrodad, G. Sapiro, Sparse modeling of human actions from motion imagery, *IMA Preprint Series # 2378*, 2011.
- [33] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing* 13 (4(April)) (2004).
- [34] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, *International Journal of Computer Vision* 42 (3) (2001) 145–175.
- [35] K. Engan, S.O. Aase, J.H. Husoy, Frame based signal compression using method of optimal directions (MOD), in: *IEEE International Symposium on Circuits and Systems*, 1999.
- [36] M. Aharon, M. Elad, A. Bruckstein, The K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, in: *IEEE Transactions on Signal Processing*, 2006.
- [37] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Discriminative learned dictionaries for local image analysis, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [38] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: *International Conference on Machine Learning*, 2009.

Chunlei Yang received his B.S. degree in Electronic Engineering from Shanghai Jiaotong University, Shanghai, China, in 2007. He is currently pursuing the Ph.D. degree in Computer Science at the University of North Carolina at Charlotte, NC. His current research interests include multimedia analysis and large scale image summarization and classification.

Jialie Shen received the Ph.D. degree from the University of New South Wales, Australia. He is currently an Assistant Professor in School of Information Systems at Singapore Management University, Singapore. His research interests can be summarized as developing effective and efficient data analysis and retrieval techniques for novel data intensive applications. Particularly, he is currently interested in various techniques of multimedia data mining, multimedia information retrieval and database systems. Dr. Shen is a member of ACM SIGMOD and SIGI.

Jinye Peng received a M.S. degree in Radio Electronics from Northwest University, Xi'an, China, in 1996 and a Ph.D. degree in Signal and Information Processing from Northwestern Polytechnical University, Xi'an, China, in 2002. He was nominated as one of the "New Century Excellent Talents and sponsored by the" Program for New Century Excellent Talents in University by the Ministry of Education in 2007. He has been a professor since 2003. At present, he is a professor at the School of Electronics and Information, Northwestern Polytechnical University, and also acts as director of the Department of electronic science and technology. His current research interests include image/video analysis and retrieval, face recognition, and machine learning.

Jianping Fan received the M.S. degree in theory physics from Northwestern University, Xian, China, in 1994 and the Ph.D. degree in optical storage and computer science from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997. He was a Researcher at Fudan University, Shanghai, during 1998. From 1998 to 1999, he was a Researcher with the Japan Society of Promotion of Science (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From September 1999 to 2001, he was a Researcher in the Department of Computer Science, Purdue University, West Lafayette, IN. In 2001, he joined the Department of Computer Science, University of North Carolina at Charlotte, as an Assistant Professor, and then became an Associate Professor. His research interests include content-based image/video analysis, classification and retrieval, surveillance videos, and statistical machine learning.