



# Structure preserving dimensionality reduction for visual object recognition

Jinjoo Song<sup>1</sup> · Gangjoon Yoon<sup>2</sup> · Heeryon Cho<sup>1</sup> · Sang Min Yoon<sup>1</sup>

Received: 12 July 2017 / Revised: 13 December 2017 / Accepted: 18 January 2018/

Published online: 31 January 2018

© Springer Science+Business Media, LLC, part of Springer Nature 2018

**Abstract** Robust object recognition has drawn increasing attention in the field of computer vision and machine learning with fast development in feature extraction and classification techniques, and release of public datasets, such as Caltech datasets, Pascal Visual Object Classes, and ImageNet. Recently, deep learning based object recognition systems have shown significant performance improvements in visual object recognition tasks using innovative learning methodology. However, high dimensional space searching and recognition is time consuming, so performing point and range queries in high dimension is reconsidered for object recognition. This paper proposes optimized dimensionality reduction using structured sparse principle component analysis. The proposed method retains high dimensional feature structures, removes redundant features that do not contribute to similarity, and classifies the query image in a large database. The qualitative and quantitative experimental results, including a comparison with the current state-of-the-art visual object recognition algorithms, verify that the proposed recognition algorithm performs favorably in reducing the query image dimension and number of training images.

**Keywords** Dimensionality reduction · Object recognition · Structured sparse PCA

---

✉ Sang Min Yoon  
smyoon@kookmin.ac.kr

Jinjoo Song  
decpearl@kookmin.ac.kr

Gangjoon Yoon  
gangjoon@gmail.com

Heeryon Cho  
heeryon@kookmin.ac.kr

<sup>1</sup> HCI Laboratory, College of Computer Science, Kookmin University, Seoul, Republic of Korea

<sup>2</sup> National Institute for Mathematical Sciences, Daejeon, Republic of Korea

## 1 Introduction

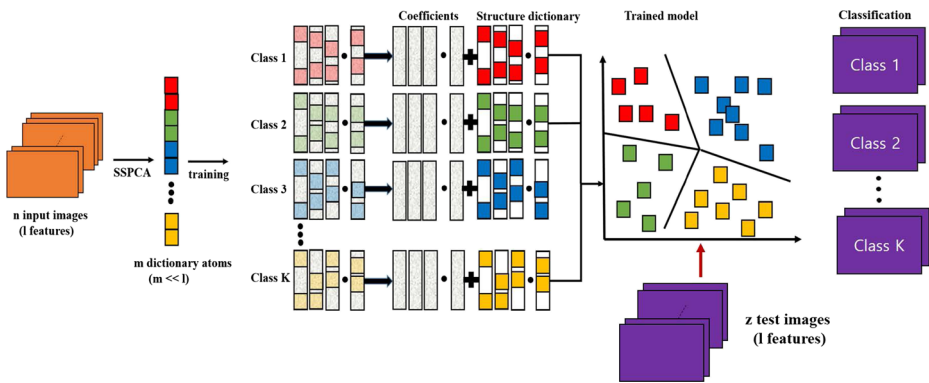
Representation of high dimensional data has been one of the most important research topics across a wide variety of information processing fields, including pattern recognition, machine learning, signal processing, data compression, and database navigation, because good data representation usually results in better performance in solving the problems related to those fields. Generally, good representation means that the representation preserves as much essential and meaningful information about the data as possible, while expressing it in a simpler way. Two common methods have been adopted to achieve this goal: dimensionality reduction and redundant feature removal that retains the structure. Dimensionality reduction attempts to transform high dimensional data into a lower dimensional form, simplifying and cleaning data for pattern recognition and machine learning tasks. Redundant feature removal retaining the main structure tries to find the optimized set of key features.

The data representation problem has become more challenging as very large high dimensional datasets have become available. This is because the number of possible distinct configurations of a set of variables increases as the number of variables increases. Many algorithms have been proposed and developed to tackle this problem, such as principle component analysis (PCA) [24], factor analysis (FA) [2], projection pursuit (PP) [20], independent component analysis (ICA) [29], etc. These algorithms come from different fields and have different approaches, but can be considered to have the same goal: finding a better representation. Hence, these algorithms focus on how to reduce dimensions or build a proper dictionary and show the resultant benefits.

PCA is one of the most popularly adopted methods in dimensionality reduction, which finds linear combinations of orthogonal factors that effectively represent the data. However, the PCA factors are still affected by all original data variables, which means they retain unwanted features when representing the data. In addition, the factors' meanings are often difficult to interpret. Several alternatives have been proposed to overcome the limitation of PCA, such as non-negative matrix factorization (NMF) [30] and sparse PCA (SPCA) [50]. In particular, the underlying motivation for sparse representation-based dimension reduction is that even though the signal is in high-dimensional space, it can actually be obtained in some lower-dimensional subspace due to its sparsity. However, the main problem of typical dimension reduction of sparse representation is that it does not consider the data structure, and can result in structural information loss. Structured sparsity [19] was proposed to solve this problem by optimal selection over structures like groups of input features. In this paper, we present a novel visual object recognition methodology by applying structured sparse PCA (SSPCA) to effectively reduce the dimension of the high-dimensional input features and retain the meaningful features and structures. The benefit of exploiting the structure of data was proved by Jenatton et al. [21]. According to them, given any intersection-closed groups of patterns of variables, we can build some quasi-regularization norms ( $\Omega$ ) enforcing that the support set regularized by  $\Omega$  belongs to the group patterns when solving the classification and regression tasks.

Figure 1 shows the proposed approach for object recognition. Finding the optimal dimensionality from very high dimensional data using SSPCA means it trains the dataset of the recognition task, while retaining recognition performance comparable to the previous state of the art approaches. The main contribution of this paper is summarized as below:

- The method to find the optimal dimensionality reduction for high dimensional data when representing features.



**Fig. 1** Our structure preserving dimensionality reduction system for robust visual object recognition. This system is separated into three parts: SSPCA-based dimension reduction, decomposition of the coefficients and the dictionaries, and similarity measure and classification

- Analysis of the relationships among the optimized size of the reduced dimension, number of training images, and non-optimized high dimensional features.
- Structure-based similarity measure and classification instead of traditional energy-based measure and classification.

The remainder of this paper is organized as follows. Section 2 reviews related work in the view of dimensional reduction, sparse representation and learning, and visual object recognition, and Section 3 provides the technical detail of how to adopt SSPCA for solving the object recognition problem. Section 4 presents quantitative and qualitative evaluation of our approach, and analyses of the experimental results. Section 5 summarizes and concludes this paper.

## 2 Related works

### 2.1 Dimensionality reduction

As data dimensionality grows, the size of dimensional space increases exponentially. This is called the curse of dimensionality [4], which must be tackled particularly for many tasks in pattern recognition and machine learning fields. If dimensionality can be reduced, this enables lower computational cost and improved performance by removing noise and less informative features, and finding more general regions or rules applicable to new data for a variety of tasks. Several dimensionality reduction techniques have been introduced: PCA [24], multidimensional scaling (MDS) [10], ICA [29], Kernel PCA [17], semidefinite embedding (SDE) [44], Isomap [41], and locally linear embedding (LLE) [38].

PCA [24] is one of the most classical and popular methods for dimension reduction. It finds a linear reduced subspace having lower dimensionality that preserves as much variability as possible of the original data. However, its computational cost increases according to the complexity of the source data, and it is limited by linearity. MDS [10] is another classical dimensionality reduction method, and maps the original dimensional space to a lower dimensional space based on the proximities, indicating the similarities between variables. In MDS, the variables are represented as points in a lower dimensional space, and

the distance between the points corresponds to the similarities. However, similar to PCA, MDS has the limitation of linearity. ICA [29] is a method to process non-Gaussian data, which PCA cannot deal with. ICA is a statistical method using the data distribution, but it still finds a linear representation. Kernel PCA [17] is a nonlinear dimensionality reduction method that introduces a kernel to PCA. Kernel PCA is applicable to many high dimensional nonlinear datasets, but reconstruction of training data and test samples is difficult. A variation of kernel PCA is SDE [44] or maximum variance unfolding (MVU) [43]. SDE differs from Kernel PCA in that when choosing a kernel function it does not use a pre-defined kernel function, but learns a kernel matrix with the assistance of semidefinite programming. Isomap [41] is a nonlinear dimensionality reduction algorithm based on MDS. This algorithm finds the neighbors of each data point with the assumption that only neighboring points know pair-wise distances, then computes the geodesic pair-wise distances between all other data points, and performs low dimensional embedding via MDS based on these geodesic distances. It is not difficult to determine the proper value of  $k$  for neighborhood graph construction for Isomap. LLE [38] was proposed to address nonlinear dimensionality reduction, and identifies the neighbors of each data point, computes weights that best describes the data point as a linear combination of its neighbors, and finds the low dimensional embedding of points, which are described with the previously determined weights. LLE achieves faster optimization than Isomap using an eigenvector based optimization technique. Marjan et al. represented high dimensional data as a chaos theory-based appearance model for object tracking task [1]. Some approaches try to solve the high dimensional data problem from the perspective of subspace clustering [14, 49].

## 2.2 Sparse representation and learning

Sparse representation and learning has been widely used in various research fields subsequent to the theoretical proof that sparse representation in a general dictionary is unique and can be found using  $l_1$  minimization [48]. Solving sparse representation and learning involves seeking the sparsest linear combination of basis functions from an overcomplete dictionary. There are many approaches to obtain the sparse solutions, but they can be grouped into three broad categories: convex relaxation [3], greedy pursuits [8], and combinatorial algorithms. Convex relaxation approaches include basis pursuit (BP) [12], interior point approaches, projected gradient methods, and iterative thresholding. To emphasize the original data structure while retaining sparseness, Huang et al. [18] proposed an efficient dynamic group sparsity concept that adaptively learns the dynamic group structure in practical applications. They also provided theoretical proofs for less measurement requirement and lower computation complexity in group sparsity. Inspired by conventional group sparsity, Zhang et al. [47] proposed leveraging feature sparsity and clustering properties to regularize feature selection. Yang et al. [23] extended the robust encoding ability of group sparse coding with spatial correlations among training regions to better depict and index image content. Group sparse representation usually focuses on more specified structures, whereas structured sparse representation works on general coding structures. De Pierrefeu et al. extended the PCA framework by adding a structural constraint, TV-elastic net penalty, to sparse PCA [11].

## 2.3 Object recognition

Breakthroughs in deep convolutional neural networks have dramatically improved object recognition and detection performances in recent years [9, 28, 31, 37]. However, other

approaches that exploit the sparseness of image features have also been promising in realizing the state of the art object recognition performance. Initially, to imitate the human vision system, biologically motivated approaches emulated quantitative models of the visual cortex to recognize objects within images. Serre et al. presented a model that mimicked visual cortical processing to realize position- and scale-tolerant edge detection [39]. The features used in Serre et al.'s system were more flexible than template based approaches, and more selective than histogram based descriptors.

Other biologically inspired systems have attempted to model the three properties of mammalian primary visual cortex, i.e., spatially localized, oriented, and selective to structure at different scales [36]. Field proposed a coding strategy that maximized sparseness of the statistical structure of natural images [13] to mathematically (algorithmically) model these three properties. Olshausen and Field subsequently proposed a learning algorithm that searched for a sparse code by introducing two global objectives; one that preserved information and the other represented sparseness of the features [36]. They demonstrated that the three properties emerge when only the two global objectives are placed on a linear coding of natural images. Mutch and Lowe modified Serre et al.'s model by eliminating weaker responses that disagreed with the locally dominant responses [33], and matched only the dominant orientation within a feature, rather than comparing all orientations. They found increasing the sparsity of the features to be helpful in improving generalization performance.

As the usefulness of the sparse image features were tested and confirmed, more studies exploiting sparse image representation followed. For example, Kavukcuoglu et al. proposed a predictive sparse decomposition (PSD) algorithm that simultaneously learned an overcomplete linear basis set, and produced a smooth and easily computed approximator to predict the optimal sparse representation [25]. Naikal et al. proposed a Sparse PCA algorithm that selected the subset of informative features based on sparse coefficients in the first few principal vectors [34]. They also introduced an algorithm to speed up Sparse PCA using the augmented Lagrangian method. Sohn et al. studied an efficient training method for sparse and conventional restricted Boltzmann machines (RBMs) through the connections between mixture models and RBMs, and proposed a mid-level feature extraction method using convolutional RBMs [40]. Oliveira et al. proposed a sparse spatial coding (SSC) algorithm that combined a sparse coding dictionary learning, spatial constraint coding stage, and online classification to improve object recognition [35].

Studies have utilized spatial pyramid matching (SPM), with some incorporating sparse coding, have also contributed to improving the object recognition performance. For example, Lazebnik et al. adapted the pyramid matching scheme of Grauman and Darrell [15], which found an approximate correspondence between two sets of vectors, and pioneered the SPM method [26]. Their SPM method repeatedly subdivided the image using a hierarchy of rectangular windows and computed the histograms of local features at increasingly fine resolutions. Bosch et al., rather than focusing on the entire scene of each image, automatically learned the region of interest, and applied random forests to the spatial pyramid representation [6]. Yang et al. extended the SPM method by computing a spatial pyramid image representation based on sparse codes for SIFT [32] features, rather than the K-means vector quantization in the traditional SPM [23]. Boureau et al. investigated performance for different pairings of coding (vector quantization vs. sparse coding) and pooling (average vs. max) modules within the spatial pyramid framework and showed that sparse coding systematically outperformed vector quantization irrespective of the pooling modules, and that max pooling improved linear classification performance over average pooling [7]. He et al. [16] added a spatial pyramid pooling layer on top of the last convolutional layer of CNNs [27]

to eliminate the fixed input image size requirement posed by current CNNs. This allowed training of variable-size images, possibly increasing scale invariance and reducing overfitting. Other object recognition methods include fast approximated locality constrained linear coding that utilizes the locality constraint to project each descriptor into its local coordinate system [42].

### 3 Structure preserving dimensional reduction for effective object recognition

We consider a method for recognizing objects from data by extracting appearance features enabling us to represent them in lower dimension while retaining the structure of the high dimensional features to efficiently represent and measure the similarity and classification. To effectively and efficiently recognize desired objects, we propose an object recognition method by classifying the coefficients of the image's appearance model.

To use a smaller number of dictionary vectors than the descriptor dimensions, the number of dictionary vectors need to be sufficient to reasonably reflect the similarity the objects of a class share with each other, and to distinguish the mutually distinct structures in different classes as well. Although PCA is one of the most popularly adopted methods in dimensionality reduction solution, the PCA factors [Robert9] are still affected by all original data variables, and there are insufficient PCA aspects to express each individual characteristic that objects of the same type have in common. Thus, it is difficult for PCA to deliver the common features of objects of the same type. However, SSPCA [22] complements the PCA constraints, and can be an excellent candidate to represent the data because it allows a dictionary to be learned by exploiting a priori structural constraints as well as sparsity while reducing dimensionality [21].

The current paper applies SSPCA to generate dictionary vectors from appearance features. The proposed object recognition procedure comprises two steps: training the dictionary using SSPCA, and recognition by measuring similarity.

#### 3.1 Symbols and notations

Before proceeding to the technical details, we introduce the notation and symbols to help understanding. To denote real numbers, we use lower case letters for indices, parameters and dimensions, and upper case letters for the constants used for the recognition process. We use boldfaces to denote multi-dimensional terms such as images, vectors, and matrices. Vectors are considered to be columns. Table 1 shows the detailed symbols and notations.

#### 3.2 Dictionary learning and dimensionality reduction

We collected a large training dataset  $\mathcal{B}$  of 2D images and categorized the images into  $M$  groups  $\{\mathbb{B}_i\}_{i=1}^M$  where each  $\mathbb{B}_i$  consists of images containing the same (relevant) type of objects.

To create feature descriptors, we randomly choose  $K$  images  $\{\mathbf{b}_{i,j}\}_{j=1}^K$  from each group  $\mathbb{B}_i$ . By resizing, we may assume that all the images have the same size divided into  $S$  subimages of the same size. From each subimage of  $\mathbf{b}_{i,j}$ , we created a feature descriptor [32] of dimension  $P$ , and arranged them in a column vector,  $\mathbf{x}_{i,j} \in \mathbb{R}^{P \cdot S}$ . Thus, we created the descriptor  $\mathbf{x}_{i,j}$  corresponding to images  $\mathbf{b}_{i,j}$  for  $i = 1, \dots, M, j = 1, \dots, K$ .

**Table 1** Symbols and notations used in this paper

Symbol	Description
$\mathcal{B}$	dataset consisting of categorized image groups
$\mathbb{B}_i, i = 1, \dots, M$	categorized groups in the database
$\mathbf{b}_{i,j}, j = 1, \dots, K$	randomly selected images from $\mathbb{B}_i, i = 1, \dots, M$
$\mathbf{x}_{i,j} \in \mathbb{R}^{P \cdot S}$	descriptors obtained from $\mathbf{b}_{i,j}$
$\mathbf{b}$	input image
$\mathbf{x} \in \mathbb{R}^{P \cdot S}$	descriptor obtained from $\mathbf{b}$
$\Omega$	mixed $\ell_1/\ell_2$ norm on $\mathbb{R}^{128s}$
$\mathbf{X} \in \mathbb{R}^{P \cdot S \times K \cdot M}$	training descriptor matrix
$\mathbf{U} \in \mathbb{R}^{P \cdot S \times r}$	dictionary matrix
$\mathbf{C} \in \mathbb{R}^{r \times K \cdot M}$	coefficient matrix
$\mathbb{G}, \mathbb{G}_i$	sets of positive integers
$\mathbf{z}, \mathbf{y}, \mathbf{w}_k$	vectors in $\mathbb{R}^n$
$\{\mathbf{e}_i\}_{i=1}^n$	orthonormal basis for $\mathbb{R}^n$
$\mathbb{W}_k$	the vector space spanned by $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k\}$
$\gamma > 0$	regularization parameter
$M$	number of categorized object types
$K$	number of randomly chosen images from each group $\mathbb{B}_i$
$S$	number of subimages forming each image
$r$	number of dictionaries
$n_k$	dimension of the vector space $\mathbb{W}_k$
$P$	dimension of the feature vectors obtained from each subimage
$i, j, k$	indices
$\langle \cdot, \cdot \rangle$	Euclidean inner product
$\ \cdot\ _2$	Euclidean $\ell_2$ -norm $\ \mathbf{v}\ _2^2 = \langle \mathbf{v}, \mathbf{v} \rangle$
$\ \cdot\ _0$	the $\ell_0$ pseudo norm counting the nonzero components of a vector
$\cdot^T$	transpose operator
$\lfloor x \rfloor$	the integer greater than or equal to $x$

Normally,  $P \cdot S$  is large, e.g.  $P = 128$  and  $S \geq 100$ , and we propose to reduce the dimensionality. That is, we need to find dictionaries such that every vector  $\mathbf{x}_{i,j}$  has a well fitted sparse representation. Before finding such dictionaries, we note that for each  $i = 1, \dots, M$ , feature vectors  $\{\mathbf{x}_{i,j}\}_{j=1}^K$  come from the images containing the same kind of objects and they are likely to share similar structures.

Therefore, we apply the SSPCA proposed by Jenatton et al. [22] to find such dictionaries. Let  $\mathbf{X}$  be the  $P \cdot S \times K \cdot M$  matrix having  $\mathbf{x}_{i,j}$  as its column vectors. To apply the SSPCA, we first divide the set  $\mathbb{G} = \{1, 2, \dots, P \cdot S\}$  into  $t = \lfloor \frac{P \cdot S}{M} \rfloor$  subsets  $\mathbb{G}_1, \mathbb{G}_2, \dots, \mathbb{G}_M$ , where  $\mathbb{G}_i = \{(i - 1)t + 1, (i - 1)t + 2, \dots, i \cdot t\}$  for  $i = 1, \dots, M - 1$ ;  $\mathbb{G}_M = \{(M - 1)t + 1, (M - 1)t + 2, \dots, P \cdot S\}$ ; and  $\lfloor x \rfloor$  denotes the largest integer smaller

than or equal to  $x$ . Then we define the sparsity-inducing mixed  $\ell_1/\ell_2$  norm  $\Omega$  over a vector  $\mathbf{u} \in \mathbb{R}^{P \cdot S}$ ,

$$\Omega(\mathbf{u}) = \sum_{j=1}^M \left( \sum_{k \in G_j} u_k^2 \right)^{\frac{1}{2}} \quad \text{for } \mathbf{u} = (u_1, u_2, \dots, u_{P \cdot S}). \tag{1}$$

With this setting, we apply the SSPCA dictionary search algorithm to the descriptors  $\{\mathbf{x}_{i,j}\}_{i=1,j=1}^{M,K}$  to find dictionaries  $\{\mathbf{u}_i\}_{i=1}^r \in \mathbb{R}^{P \cdot S}$ ,

$$\begin{aligned} (\mathbf{U}, \mathbf{C}) = & \underset{\substack{\mathbf{U} \in \mathbb{R}^{P \cdot S \times r} \\ \mathbf{C} \in \mathbb{R}^{r \times K \cdot M}}}{\text{argmin}} \frac{1}{2(P \cdot S) \cdot (K \cdot M)} \|\mathbf{X} - \mathbf{UC}\|_F^2 \\ & + \lambda \sum_{j=1}^r \Omega(\mathbf{u}_j) \quad \text{subject to } \|\mathbf{c}_j\|_2 \leq 1, \quad j = 1, \dots, K \cdot M \end{aligned} \tag{2}$$

where  $\mathbf{U} = (\mathbf{u}_k)_{k=1}^r$  is the  $P \cdot S \times r$  dictionary matrix;  $\mathbf{C} = (\mathbf{c}_j)_{j=1}^{K \cdot M} \in \mathbb{R}^{r \times K \cdot M}$  is the coefficient matrix;  $\|\cdot\|_F$  is the Frobenius matrix norm,  $\|\mathbf{A}\|_F^2 = \text{trace}(\mathbf{AA}^T) = \sum_{i=1}^n \sum_{j=1}^m a_{ij}^2$ ;  $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times m}$  and  $\|\cdot\|_2$  is the Euclidean norm; and  $\Omega$  is the mixed norm defined in (1) which controls the sparsity and the structure of the support of  $\mathbf{d}_j$  (for details on the norm (see [22] and references therein).

Then, for  $\ell = (i - 1) \cdot K + j$ , ( $i = 1, \dots, M, j = 1, \dots, K$ ), the  $\ell$ -th column vector  $\mathbf{x}_{i,j}$  of  $\mathbf{X}$  is (approximately or exactly) expressed by a linear combination of the  $r$  dictionaries  $\mathbf{u}_k$ 's with coefficients  $(c_{k\ell})_{k=1}^r$ ,

$$\mathbf{x}_{i,j} \approx \sum_{k=1}^r c_{k\ell} \mathbf{u}_k = \mathbf{Uc}_\ell, \quad \mathbf{c}_\ell = (c_{1\ell}, \dots, c_{r\ell})^T. \tag{3}$$

Thus, each descriptor vector  $\mathbf{x}_{i,j} \in \mathbb{R}^{P \cdot S}$  is represented (approximately) in terms of quite a few dictionaries, i.e., we reduce the dimensionality from  $P \cdot S$  to  $r$ . In the following section, we generate support vector machines using the  $\mathbf{u}_k$ 's dictionaries.

We end this section with the following observation. Theorem 1 shows that we don't expect that the larger dictionary set delivers better recognition performance. Empirically, there exists an optimal choice for  $r$  that depends on the number of samples ( $K \cdot M$ ) and is shown to be  $\lfloor \frac{K \cdot M}{2} \rfloor$ .

**Theorem 1** For vector  $\mathbf{z}$  and finite set  $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$  in  $\mathbb{R}^n$ , let  $\{\mathbf{y}_k\}_{k=1}^m$  be the orthogonal projections in  $\mathbb{R}^n$  of  $\mathbf{z}$  into the vector space  $\mathbb{W}_k$  spanned by  $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k\}$ , i.e.,

$$\mathbf{y}_k = \underset{\mathbf{y} \in \mathbb{W}_k}{\text{argmin}} \|\mathbf{z} - \mathbf{y}\|_2, \quad \text{for } k = 1, 2, \dots, m. \tag{4}$$

Then,

$$\|\mathbf{z} - \mathbf{y}_1\|_2 \geq \|\mathbf{z} - \mathbf{y}_2\|_2 \geq \dots \geq \|\mathbf{z} - \mathbf{y}_k\|_2 \tag{5}$$

and

$$\|\mathbf{y}_1\|_0 \leq \|\mathbf{y}_2\|_0 \leq \dots \leq \|\mathbf{y}_k\|_0, \tag{6}$$

where  $\|\mathbf{z}\|_0$  is the  $\ell_0$  pseudo norm of a vector  $\mathbf{z}$  defined to be the number of non-zero entries in  $\mathbf{z}$ .



*Proof* For  $k = 1, \dots, m$ , let  $n_k$  be the dimension of the vector space  $\mathbb{W}_k$  spanned by  $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k\}$ . Since the subspaces  $\mathbb{W}_1 \subseteq \mathbb{W}_2 \subseteq \dots \subseteq \mathbb{W}_k$  are nested,  $n_1 \leq n_2 \leq \dots \leq n_k$ . Gram-Schmidt orthogonalization implies that there exists an orthonormal basis  $\{\mathbf{e}_i\}_{i=1}^n$  for  $\mathbb{R}^n$ , such that for  $k = 1, \dots, m$ ,  $\{\mathbf{e}_i\}_{i=1}^{n_k}$  is a basis for  $\mathbb{W}_k$ . Thus, every vector  $\mathbf{w} \in \mathbb{W}_k$  is uniquely represented as

$$\mathbf{w} = \sum_{i=1}^{n_k} \langle \mathbf{w}, \mathbf{e}_i \rangle \mathbf{e}_i \quad \text{and} \quad \|\mathbf{w}\|_2^2 = \sum_{i=1}^{n_k} \langle \mathbf{w}, \mathbf{e}_i \rangle^2, \tag{7}$$

where  $\langle \cdot, \cdot \rangle$  denotes the Euclidean inner product in  $\mathbb{R}^n$ .

Let  $\mathbf{z} \in \mathbb{R}^n$  be given arbitrarily. Then for  $k = 1, \dots, m$ , the vector  $\mathbf{y}_k$  in (4) is the orthogonal projection of  $\mathbf{z}$  onto  $\mathbb{W}_k$ , i.e.,

$$\mathbf{y}_k = \operatorname{argmin}_{\mathbf{y} \in \mathbb{W}_k} \|\mathbf{z} - \mathbf{y}\|_2 = \sum_{i=1}^{n_k} \langle \mathbf{z}, \mathbf{e}_i \rangle \mathbf{e}_i$$

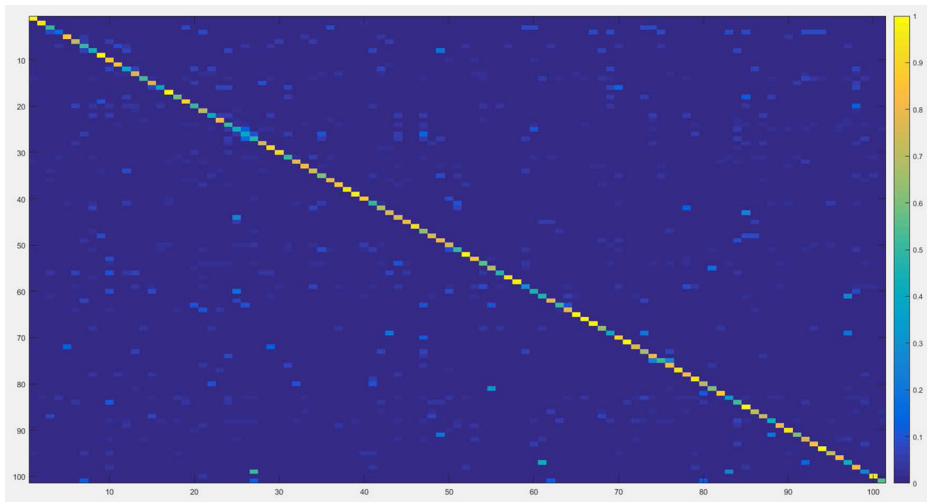
and

$$\|\mathbf{z} - \mathbf{y}_k\|_2^2 = \sum_{i=n_k+1}^n |\langle \mathbf{z}, \mathbf{e}_i \rangle|^2.$$

Thus, the inequalities (5) and (6) hold, which completes the proof. □

### 4 Experimental result

We prove the existence of an optimal number of dictionary elements and evaluate the performance of our method. The schematic experimental setup and evaluation for SSPCA based object recognition is illustrated and we report the results based on quantitative and qualitative evaluations in two parts: Section 4.1 covers the experimental setup, and Section 4.2



**Fig. 2** Confusion matrix of our SSPCA-based visual object recognition using Caltech 101 dataset. The performance is evaluated using the classifiers learned with 30 training images per class. Each row indicates the true class and each column represents the assigned class

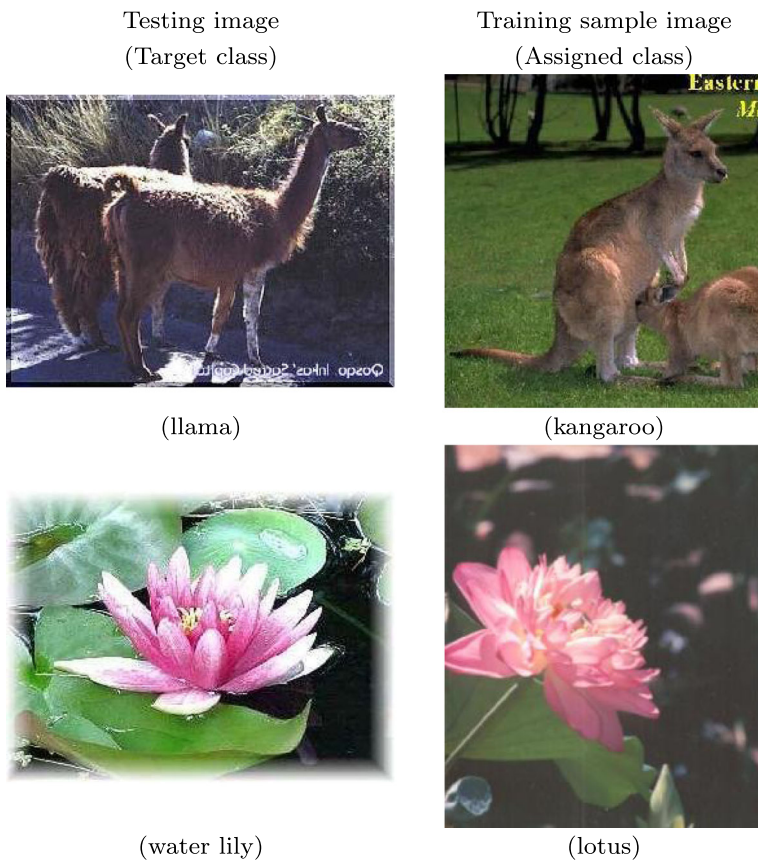
presents object recognition results from a public dataset and analysis of the proposed approach based on the experimental results and previous techniques.

#### 4.1 Experimental setup

The experiments were conducted on the Caltech 101 dataset, which consists of 101 classes and one background class. Each class contains 31 to 800 images having the size of approximately  $300 \times 200$  pixels. The images of each class vary in color, pose, and illumination, making the classification task challenging. We use 101 classes, excluding the background class. For pre-processing, we extract features from each image using SIFT [32] on a standard 4 GHz machine with 2 GB RAM.



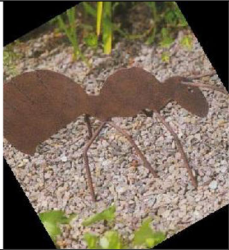
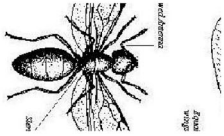


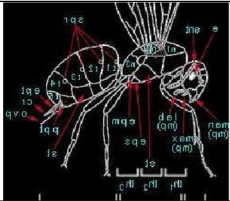
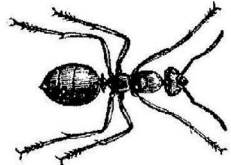


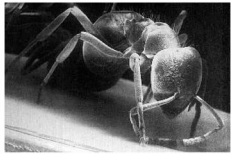

#### 4.2 Quantitative and qualitative evaluation of visual object detection

For fair performance evaluation and comparison, we start by focusing on the proposed approach's effectiveness and efficiency to reduce the dimensionality of the visual objects and number of training images.



**Fig. 3** Examples of misclassification. The testing images with the target class are on the left column; the assigned class of each testing image and the assigned class's training images are on the right column

Figure 2 shows the confusion matrix of the proposed algorithm learned with 30 training samples per class, where the row denotes true class and the column denotes the assigned class. Some of the highly misclassified classes are llama (29.2%) and water lily (25%) classes. Figure 3 displays some examples of misclassified results of these classes. The images of the llama and the water lily classes are assigned to the kangaroo and lotus classes respectively. The image features of the target classes and the assigned classes are considerably similar, and this is the main reason for the low recognition rate of the two classes. The class with the lowest accuracy (16.7%) is the ant class, which has twelve test images. As shown in Fig. 4, only two testing images are recognized as an ant; ten other images are classified into ten different classes. This is because the features and the structures of an ant itself are uncomplicated; i.e., an ant is a relatively simple object consisting of some lines and circle-like shape. As a result, the factors like illumination, texture, pose, and background have a stronger influence on the classification than in the case of the objects composed of more complicated features and structures.

		
ant	ant	trilobite
		
sunflower	motorbike	octopus
		
chandelier	tick	scorpion
		
anchor	platypus	snoopy

**Fig. 4** The object recognition result of the ant class. Only two test images are recognized as an ant. Ten other images are classified into ten different classes

**Table 2** Recognition rate of different dictionary element size ( $r$ ) and training sample combination

Training samples	15	20	25	30
$r=500$	73.79%	76.25%	79.93%	81.76%
$r=1000$	75.30%	78.07%	80.64%	82.73%
$r=1500$	76.17%	78.38%	81.71%	83.92%
$r=2000$	76.24%	79.83%	81.64%	83.51%

We then conducted the experiments to evaluate the robustness of the proposed approach by reducing the number of training images because traditional learning and recognition approaches are largely dependent on the number of training samples; we seek to prove the Theorem 1 experimentally. The number of training images was varied; 15, 20, 25, and 30 images were randomly chosen from each class for training and the remaining images were used for testing. The online dictionary pertaining to the chosen training images was learned using SSPCA, which enabled us to decide how many feature dimensions were used, i.e., the number of dictionary elements. Table 2 displays the proposed method's results applied to the task of object recognition using Caltech-101 dataset. We trained the classification model according to the different number of samples and dictionary elements. Generally, increasing the dictionary elements or the number of training samples increases performance. However, when using 25 or 30 training samples, the optimal number of the dictionary elements was 1500, not 2000. When the classification model was trained with 30 samples per class and 1500 dictionary elements, the best accuracy of 83.92% was achieved. This implies that overfitting occurred when more than the optimal number of dictionary elements were used. On the other hand, the combined use of 15 or 20 training samples and 1500 dictionary elements did not produce the best results. Using 2000 dictionary elements showed better performance than the use of 1500 dictionary elements, given 15 or 20 training samples. This means that, as the number of training samples decreases, more dictionary elements are needed to build a more accurate classification model. Thus, the number of dictionary elements and training samples influence the optimality of dimensions. Table 2s shows that, when using 30 training samples, the optimal number of dictionary elements is 1500, i.e., optimal dimensionality = 1500, rather than full dimension of the given features. As the number of feature dimensions increases, overfitting occurs; on the other hand, less dimensions cause underfitting. Increasing the feature dimensions do not guarantee better performance; hence, the optimal number of feature dimensions should be investigated for best performance.

Next, we investigated the influence of the classifier type. The learned dictionary was then used to create a classification model utilizing SVM or the Softmax function. Many numerical experiments were conducted to identify the optimal number of dictionary elements and then compared with the proposed method and other significant previous approaches. We trained the model with two classifiers: SVM and Softmax. As shown in Table 3, SVM showed better performance than Softmax, but both classifiers showed increased

**Table 3** Recognition rate of different classifier and training sample combination

Training samples	15	20	25	30
Softmax	74.53%	77.33%	80.25%	82.59%
SVM	76.17%	78.38%	81.71%	83.92%

**Table 4** Comparison of SSPCA-object recognition approach to the previous approaches, which do not use pretrained models

Training samples	15	20	25	30
Yang et al. [45]	73.2%	80.1%	82.7%	84.3%
Bo et al. [5]	–	–	–	82.5%
Zeiler et al. [46] (non-pretrained)	22.8%	–	–	46.5%
Our approach	76.2%	78.4%	81.7%	83.9%

performance with increased number of training samples, and showed the best accuracy using 30 training samples. Thus, the classifier type does not affect dimensional optimality.

### 4.3 The comparison with the previous approaches

Lastly, we compared our SSPCA-based object recognition result to the remarkable previous approaches that use various learning techniques without exploiting pretrained models. Table 4 compares the recognition performance of our approach with the previous approaches using the Caltech 101 Dataset. When using fifteen training samples, Yang et al. [45] achieves recognition accuracy below 75%; in comparison, our approach shows relatively good result with fifteen training samples. Our approach successfully removes useless information via dimensionality reduction while preserving the important structures and features for the task. Zeiler and Fergus [46], an algorithm that uses deep learning method, reported the recognition accuracy of 22.8% and 46.5% with 15 and 30 training images without using the ImageNet-pretrained model.

Nowadays, approaches adopting deep learning methodologies have shown considerably good performance in object recognition task. Zeiler and Fergus [46] reported the recognition accuracy of 83.8% and 86.5% in the Caltech101 Dataset classification task when using 15 and 30 training samples respectively with the ImageNet-pretrained model. He et al. [16], which uses a convolutional neural network (CNN), also reported the experiment result with the recognition accuracy of 91.9% when training with 30 samples. However, both methods pre-trained the classification model using the ImageNet 2012 dataset, which contained about 10 million images, and the methods benefit significantly from their pre-trained models.

## 5 Conclusion

We proposed an object recognition algorithm using optimal dimensionality reduction and verified the effectiveness of the proposed technique experimentally. The proposed object recognition method employs SSPCA and is effective at preserving important and crucial features while removing noisy and unimportant features by leveraging the strengths of both sparse representation and PCA.

We also proved that an optimal number of feature dimensions exists, which simplifies feature representation and improves performance. The core concept of finding the optimal dimensionality reduction from high dimensional features can be applied to various applications, such as image abstraction, image manipulation, and image composition.

The future work includes the investigation of the following issues: improvement of the classification model to recognize the objects with simple features and structures such as ants and adaptive discovery of the optimal number of dimensions given the number of training

samples. We will also work on SSPCA-based deep learning by combining SSPCA, which considers structure and sparsity, with many different deep learning algorithms, which learns features effectively, in order to deal with various machine learning tasks including object detection, image annotation or scene parsing with large scale datasets like the MS COCO dataset or the ImageNet dataset.

**Acknowledgements** J. Song and S.M. Yoon were supported by the National Research Foundation of Korea grants funded (No.2015R1A5A7037615, No.2016R1D1A1B04932889) and IITP (#2014-0-00501) by the Korean Government. H. Cho was support by the National Research Foundation of Korea (No. 2017R1A2B4011015). G.J.Yoon was supported by National Institute for Mathematical Sciences (NIMS).

## References

1. Abdechiri M, Faez K, Amindavar H, Bilotta E (2017) Chaotic target representation for robust object tracking. *Signal Process Image Commun* 54:23–35
2. Akaike H (1987) Factor analysis and AIC. *Psychometrika* 52(3):317–332
3. Arias RS A convex optimization algorithm for sparse representation and applications in classification problems. Ph.D. thesis, DigitalCommons@UTEP. <http://digitalcommons.utep.edu/dissertations/AAI3565935>
4. Bellman R (1957) *Dynamic programming*. Princeton University Press
5. Bo L, Ren X, Fox D (2013) Multipath sparse coding using hierarchical matching pursuit. In: *IEEE Conference on computer vision and pattern recognition*
6. Bosch A, Zisserman A, Mu X, Munoz X (2007) Image classification using random forests and ferns. In: *IEEE 11th International conference on computer vision (ICCV)*, pp 1–8
7. Boureau YL, Bach F, LeCun Y, Ponce J (2010) Learning mid-level features for recognition. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, pp 2559–2566
8. Chen L, Chen J, Gu Y (2012) Greedy pursuits: stability of recovery performance against general perturbations. In: *ICNC. IEEE Computer Society*, pp 897–901
9. Ciresan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J (2011) High-performance neural networks for visual object classification. *CoRR arXiv:1102.0183*
10. Davison ML (1983) *Multidimensional scaling*. Wiley, New York
11. De Pierrefeu A, Löfstedt T, Hadj-Selem F, Dubois M, Ciucci P, Frouin V, Duchesnay E (2016) Structured sparse principal components analysis with the tv-elastic net penalty. *arXiv:1609.01423*
12. Donoho DL (2006) Compressed sensing. *IEEE Trans Inf Theory* 52(4):1289–1306
13. Field DJ (1994) What is the goal of sensory coding? *Neural Comput* 6(4):559–601
14. Gan G, Ng MKP (2015) Subspace clustering with automatic feature grouping. *Pattern Recogn* 48(11):3703–3713
15. Grauman K, Darrell T (2005) The pyramid match kernel: Discriminative classification with sets of image features. In: *Proceedings of the IEEE international conference on computer vision*, vol II, pp 1458–1465. <https://doi.org/10.1109/ICCV.2005.239>
16. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 37(9):1904–1916. <https://doi.org/10.1109/TPAMI.2015.2389824>
17. Hoffmann H (2007) Kernel PCA for novelty detection. *Pattern Recogn* 40(3):863–874
18. Huang J, Zhang T (2010) The benefit of group sparsity. *Ann Stat* 38(4):1978–2004
19. Huang J, Zhang T, Metaxas D (2009) Learning with structured sparsity. *J Mach Learn Res* 12:1–30. <https://doi.org/10.1145/1553374.1553429>
20. Huber PJ (1985) Projection pursuit. *Ann Statist* 13(2):435–475
21. Jenatton R, Audibert JY, Bach F (2011) Structured variable selection with sparsity-inducing norms. *J Mach Learn Res* 12:2777–2824
22. Jenatton R, Obozinski G, Bach F (2010) Structured sparse principal component analysis. In: *International conference on artificial intelligence and statistics*, pp 1–13
23. Jianchao Y, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: *2009 IEEE Conference on computer vision and pattern recognition* pp 1794–1801

24. Jolliffe I (1986) Principal component analysis. Springer, New York
25. Kavukcuoglu K, LeCun Y, Ranzato M (2010) Fast inference in sparse coding algorithms with applications to object recognition, pp 1–9. arXiv:1010.3467. <https://doi.org/10.1109/ICIP.2001.958968>
26. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 2169–2178. <https://doi.org/10.1109/CVPR.2006.68>
27. LeCun Y, Kavukcuoglu K, Farabet C (2010) Convolutional networks and applications in vision. In: ISCAS. IEEE, pp 253–256
28. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7533):436–444
29. Lee TW (1998) Independent component analysis, theory and applications. Kluwer Academic Publishers
30. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–91. <https://doi.org/10.1038/44565>
31. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC SSD: single shot multibox detector. arXiv:1512.02325
32. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
33. Mutch J, Lowe DG (2006) Multiclass object recognition using sparse, localized features. In: IEEE Conference on computer vision and pattern recognition, pp 11–18. <https://doi.org/10.1109/CVPR.2006.200>
34. Naikal N, Yang AY, Shankar S (2011) Informative feature selection for object recognition via Sparse PCA. In: Proceedings of the IEEE international conference on computer vision, pp 818–825. <https://doi.org/10.1109/ICCV.2011.6126321>
35. Oliveira GL, Nascimento ER, Vieira AW, Campos MFM (2012) Sparse spatial coding: a novel approach for efficient and accurate object recognition. In: Proceedings - IEEE international conference on robotics and automation, pp 2592–2598. <https://doi.org/10.1109/ICRA.2012.6224785>
36. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609. <https://doi.org/10.1038/381607a0>
37. Redmon J, Farhadi A YOLO9000: better, faster, stronger. arXiv:1612.08242
38. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326
39. Serre T, Wolf L, Poggio T (2005) Object recognition with features inspired by visual cortex. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 994–1000. <https://doi.org/10.1109/CVPR.2005.254>
40. Sohn K, Jung DY, Lee H, Hero AO (2011) Efficient learning of sparse, distributed, convolutional feature representations for object recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2643–2650. <https://doi.org/10.1109/ICCV.2011.6126554>
41. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323
42. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 3360–3367. <https://doi.org/10.1109/CVPR.2010.5540018>
43. Weinberger KQ, Saul LK (2006) An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In: AAAI. AAAI Press, pp 1683–1686
44. Weinberger KQ, Saul LK (2006) Unsupervised learning of image manifolds by semidefinite programming. *Int J Comput Vis* 70(1):77–90
45. Yang J, Li Y, Tian Y, Duan L, Gao W (2009) Group-sensitive multiple kernel learning for object categorization. In: IEEE International conference on computer vision
46. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision
47. Zhang S, Huang J, Li H, Metaxas DN (2012) Automatic image annotation and retrieval using group sparsity. *IEEE Trans Syst Man Cybern Part B* 42(3):838–849
48. Zhang Z, Xu Y, Yang J, Li X, Zhang D (2015) A survey of sparse representation: algorithms and applications. *IEEE Access* 3:490–530. <https://doi.org/10.1109/ACCESS.2015.2430359>
49. Zhu P, Zhu W, Hu Q, Zhang C, Zuo W (2017) Subspace clustering guided unsupervised feature selection. *Pattern Recogn* 66:364–374
50. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. *J Comput Graph Stat* 15(2):265–286. <https://doi.org/10.1198/106186006X113430>



**Jinjoo Song** received B.E. degree in Computer Science from Seogang University, Seoul, South Korea in 2002. She is currently working toward the Integrated Ph.D. degree in Computer Science at Kookmin University, Seoul, South Korea. Her current research interests include denoising and object recognition.

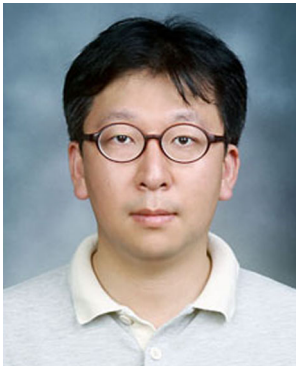


**Gangjoon Yoon** is a researcher at National Institute for Mathematical Sciences in Korea. He joined the National Agenda Project of Korea on Surveillance System. His research interests are in compressive sensing and computer vision based on mathematics and in functional analysis. He obtained his Master's degree and Ph.D. degree in Mathematics from Korea Advanced Institute of Science and Technology (KAIST) in 1995 and 1999 respectively.





**Heeryon Cho** received Master and Ph.D. degrees in Informatics from Kyoto University, Kyoto, Japan, in 2005 and 2009 respectively. In 2010, she was a postdoctoral researcher at the Center of Interaction Science, Sungkyunkwan University, Seoul, South Korea, and from 2011 to 2014, she was a senior researcher with the Yonsei Institute of Convergence Technology, Yonsei University, Incheon, South Korea. From 2014 to 2015, she was a research professor at the Department of MIS, Chungbuk National University, Cheongju, South Korea. She is currently a Research Professor at the College of Computer Science, Kookmin University, Seoul, South Korea. Her research interests include big data analysis, opinion mining and sentiment analysis.



**Sang Min Yoon** received B.E. and M.E. degrees in Electronics Engineering from Korea University, Seoul, Korea, in 2000 and 2002 respectively, and Dr.-Ing. degree in Informatik, TU Darmstadt, Darmstadt, Germany in 2010. Then, he joined the Digital Human Research Center, Advanced Industrial Science and Technology (AIST), Tokyo, Japan, as postdoctoral researcher. He was with Samsung Advanced Institute of Technology (SAIT) as a researcher in the area of face recognition and video surveillance from 2002 to 2005. Currently, he is working at the College of Computer Science at Kookmin University, South Korea, as an associate professor since 2012. His main interests are computer vision and pattern recognition including object detection and recognition, visual object tracking, 3D model retrieval, and human computer interface (HCI).