Research on target object recognition based on transfer-learning convolutional SAE in intelligent urban construction

BING XIE¹, ZHEMIN DUAN¹, BING ZHENG², (Member, IEEE), LIPING LIU²
¹School of Electronics and Information, Northwestern Polytechnical University, Xi’an 710072, China
²School of Electrical and Control Engineering, North University of China, Taiyuan, 030051)

Corresponding author: BING XIE (e-mail: yinyunhua@nuc.edu.cn)

This work was supported by the Science and Technology Innovation Team of Henan Province Supports Project (Grant: 17IRTSTHN014), and Equipment Pre-research Key Laboratory Fund Project (Grant: 61420010404004).

ABSTRACT With the development of society, intelligent urban construction is drawing attention globally. It embeds sensors and equipment into various environmental monitoring target objects to achieve environmental management and decision-making wisdom in a more granular and dynamic manner. However, how to achieve target objects’ recognition in the dynamic environment is of essential importance for intelligent urban construction. Due to the shape, color and other characteristics for target objects are more similar, which makes it difficult to identify the target types based on the low-level features such as color, shape, etc. In this paper, we attempt to apply the deep neural network composed of sparse autoencoders based unsupervised feature learning to identify the various types of target objects. On the other hand, due to the fact that the quantities of target objects which are obtained by environmental monitoring may be not sufficient, cannot get more high-level visual features through feature training, which affects the accuracy of the subsequent target recognition. A cross-domain feature learning scheme for target objects recognition using convolutional sparse auto-encoder has been presented. In order to improve the recognition speed, feature weights selection method based on a correlation analysis is further proposed for the purpose of reducing the amount of global features which are taken from target-domain images. Experimental results show that compared with non-transfer feature learning algorithm and underlying visual feature recognition algorithm, the new algorithm proposed in this paper has higher accuracy and robustness. Feature selection can reduce the computational time of global feature extraction and recognition by about 30% while improving recognition performance.

INDEX TERMS Target objects recognition, Sparse auto-encoder, Transfer learning, Intelligent urban

I. INTRODUCTION
As the Internet of Things and big data has been developed rapidly, the pace of intelligent urban construction has been promoted [1-3]. It can utilize Internet of Things technology to embed sensors and equipment into various environmental monitoring target objects to achieve environmental management and decision-making wisdom in a more granular and dynamic manner [4-7]. If the target objects in the dynamic environment cannot be effectively identified and classified, which will lead to more valuable information cannot be provided for the construction of smart cities, and inevitably lead to the fact that smart cities cannot truly realize informationization and intelligence. Therefore, it is of great importance and essential significance for intelligent urban construction to realize target objects recognition in an dynamic environment. Due to the shape, color and other characteristics for target objects are more similar, it is difficult to identify the target types effectively. However, the key to target recognition is the feature extraction of target object. At present, feature extraction algorithm which has been commonly used is based on statistical feature extraction such as: Local Binary Patterns (LBP) [8] based on time domain, Generalized Search Trees (GIST) [9] based on frequency domain, Fourier transform, wavelet transform and other feature extraction methods. However, these feature extraction methods need to establish complex feature models to extract the underlying visual features such as image color, texture, shape, etc. Only the images’ local surface information can be expressed by means of using these features, and it has been confronted with some difficulties in
fully expressing the global detail feature of the image. Therefore, the target recognition based on traditional underlying feature extraction method cannot accurately identify target objects in complex and dynamic environment. There are also special manual features to identify the target object based on the salient features of the target object and the scene. However, these dedicated manual features relying on manual experience and prior knowledge selection not only need spend a lot of energy and time, and also deal with computationally intensive and complex image preprocessing. Therefore, this recognition model is much more solid, the parameters are difficult to adjust, and it is not universal. When facing new complex scenes, it is often difficult to maintain a high recognition rate. A single feature often fails to effectively represent the target object, and combining multiple features [10] between the targets can obtain more information of the target object and thus better describe the characteristics of the target object. However, getting the best combination of features is much difficult, which not only require the analysis and selection of the characteristics of specific target objects and scenes with the help of prior knowledge, but also takes a lot of labor cost to carry out feature analysis, combination and selection. Therefore, this method does not necessarily achieve the desired results.

In recent years, deep learning technology has become a research hotspot because of its outstanding performance in the field of artificial intelligence [11-14]. By simulating the hierarchical structure of the human brain, Convolutional Neural Network model (CNN) [15-18] has been a monitored learning method. It first extracts the inputting data features and then abstracts them in the higher layer network. In this way, the higher the layer of the neural network, the more likely to express the target objects' high features. Compared with the underlying visual features, artificial design features and multi-feature combination methods, this method can better capture the most essential features of the target object, which not only brings new opportunities for artificial intelligence related research, but also has brought new breakthroughs in the yield of target recognition. However, if its performance is good or not is decided by the number of training samples marked. Therefore, the supervised deep neural network model is not suitable for the recognition of small number of target images. At present, unsupervised feature learning has received extensive attention from researchers at home and abroad [19-20]. The sparse auto-encoder [21-22] method for extracting high-level features of target objects by self-recovery training using a large number of unmarked sample data is effectively extended to the limited target samples, which does not require labeled samples and define features in advance. By using the hierarchical structure inherent in the neural network, the number of hidden layer units is set, and under the premise of ensuring that the input and output are "equivalent" in a certain sense, the hierarchical structure parameters are continuously adjusted [23-24]. However, this model still requires a large amount of unmarked data for the feature learning, dataset can not provide a lot of unmarked samples for training, which makes it difficult to carry out unsupervised feature learning. Theoretically, the training data of the unsupervised learning phase and the training data of the supervised recognition phase are not required to have the same feature distribution. Therefore, it is possible to solve the problem that unsupervised feature learning cannot be trained due to the limited number of target samples by means of transfer learning method. Because the transfer learning data set contains a large number of similar images, which will lead to local feature weights SAE learned from transfer learning dataset may be redundant or irrelevant [25-26], and increase the time for subsequent global feature extraction of target object based on the convolutional network, and affect the speed of recognition. We attempt to use the correlation analysis to select the features learned by the sparse autoencoder before the convolution operation in order to reduce the feature weights dimension and avoid over-fitting.

Therefore, this paper proposes target recognition algorithm based on transfer-learning convolutional sparse automatic encoders with feature weights selection. The algorithm is designed as follows: Firstly, SAE is used to perform unsupervised feature learning on a large number of unmarked graphs in cross-domain databases to obtain local features; Secondly, a correlation analysis method based feature weights selection is used to reduce redundancy of the local feature vectors learned by SAE. Thirdly, local features are convoluted on target object images by using CNN to obtain their global feature responses. Finally, the global feature response obtained is sent to logistic regression or the softmax regression model to achieve the recognition of the two or three types of target objects.

The innovations of this paper are as follows: 1) The deep neural network model composed of sparse autoencoder based unsupervised feature learning, convolutional neural network and classifier is used to identify the various types of target objects in intelligent urban; 2) A transfer-learning feature learning scheme using SAE model has been presented to overcome the limitation of small amount of training data; 3) Since feature weights learned by SAE are irrelevant or redundant, feature weights selection based on correlation analysis method is proposed to reduce the computational complexity of global feature extraction based on CNN, and the time required for target objects recognition is reduced while the recognition performance has not be decreased.

The remainder of this paper is organized as follows: The proposed algorithm for target objects recognition will be introduced in Section 2. Experimental results will be presented in Section 3 and overall discussions will be given in Section 4. Conclusions will be drawn in the last section.

2. PROPOSED ALGORITHM

The algorithm put forward in this paper has been consisted by four parts: 1) unsupervised feature learning used by SAE
in cross-domain database, 2) selection of feature weights, 3) extraction of global feature in the target domain and 4) target objects recognition. The proposed algorithm’s overall framework of target recognition is presented in FIGURE 1.

![Image](image_url)

**FIGURE 1. The overall framework of targets recognition algorithm**

As shown in FIGURE 1, the whole framework has been mainly made up of four parts: feature learning in cross-domain database which is unsupervised, selection of feature weights, extraction of global feature in the target domain and target recognition.

1) Unsupervised feature learning in the cross-domain dataset.

First and foremost, image patches are sampled from the STL-10 image dataset at random, which was set up for promoting unsupervised feature learning algorithms. In order to learn local features, image patches are sent into SAE after the preprocessing stage.

2) Selection of feature weights.

By filtering the local feature weights iteratively through correlation analysis, the feature weights which are closely correlated in each pair are eliminated.

3) Extraction of global feature in the target domain.

Then by means of employing convolution operation, global features are extracted successfully from object images in the target domain. After convolution, in order to further decrease the feature maps’ resolution, the pooling operation is then conducted.

4) Object images recognition.

Finally, object image samples’ global feature vectors in the target domain are sent into regression model for two and three types of target images recognition.

### 2.1 UNSUPERVISED FEATURE LEARNING BASED SAE

SAE is an improved form, which is able to handle the sparse constraint for the hidden layer unit response in the AE model [27]. Firstly, the majority of neurons in the network have been under an inhibition state; thus with the help of the back propagation (BP) training method, the least cost function is applied to learn the target object’s key feature response [28]. Since the sampled adjacent image blocks have a high degree of correlation, which result in a large amount of input redundancy. Whitening transform is a effective method that can reduce the redundancy of the input image block. In this paper, a zero-phase component analysis (ZCA) method [29] is availed so as to whiten multiple image sub-blocks. The architecture of an auto-encoder neural network with whitening transformation is presented in FIGURE 2.

![Image](image_url)

**FIGURE 2. The architecture of an SAE neural network including whitening transformation**

As it is shown in the above table, it is assumed that the size of the i-th image block acquired from the target image would be \( n \times n \), sorting by RGB components can obtain \( m = n \times n \times 3 \) dimensional vectors. The input vector after whitening is \( \hat{x}^{(i)} = W_{\text{white}} x^{(i)} \), \( W_{\text{white}} \) represents the \( m \times m \) dimensional whitening transform coefficient matrix. The s-dimensional hidden layer response vector of SAE is as shown in Eq. (1):

\[
d_{\text{SAE}}^{(i)} = \sigma(W_{\text{SAE}}^{(i)} \hat{x}^{(i)} + b_{\text{SAE}}) = \sigma(W_{\text{SAE}} W_{\text{white}} x^{(i)} + b_{\text{SAE}})
\]

where, \( W_{\text{SAE}} \) is the input weight coefficient for connecting the SAE hidden layer and each image block after whitening transform, \( b_{\text{SAE}} \) is the input offset, and \( \sigma(\cdot) \) represents the function of activation. After the whitening stage, \( W_{\text{ZCA}} \) white has been weight coefficient overall, which gives the representation of the mapping relationship between the original data and the hidden layer. After whitening, since the range of \([0, 1]\) will be exceeded by the input value, it is not necessary to use the activation function \( \sigma(\cdot) \) to map the SAE output when reconstructing the data.

\[
\hat{x}^{(i)} = W_{\text{white}}^{T} a_{\text{SAE}}^{(i)} + b_{2}
\]

where, \( \hat{x}^{(i)} \) is the i-th restored sample, the output weight is represented by \( b_{2} \) as the output offset. For the purpose of preventing over-fitting, actions are taken to maintain the hidden layer response sparsity, the sparsity penalty term and the weight attenuation term are added to the cost function which is shown in Eq. (3).

\[
J_{\text{SAE}}(W_{\text{SAE}}, b) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \| \hat{x}^{(i)} - W_{\text{white}} x^{(i)} \|^2 + \beta \sum_{j=1}^{\rho} \log \frac{\hat{\rho}_{j}}{\rho_{j}} + (1 - \rho) \log \frac{1 - \hat{\rho}_{j}}{1 - \rho_{j}} + \lambda \|W_{\text{SAE}}\|^{2}
\]

where, \( N \) denotes the amount of unmarked training samples and \( \lambda \) is regarded as the weight attenuation coefficient. Sparse penalty weight is represented by \( \beta \) whiles is the quantity of hidden layer units, and \( \rho \) is the target sparse value. \( \hat{\rho} \) is the average response of all training samples on the \( j-th \) hidden layer unit. After training the SAE, the \( W_{\text{SAE}} \) can be obtained, and the \( W_{\text{SAE}} \) can obtain the local feature response on each image block, as shown in Eq. (4):

\[
a_T = \sigma(W_T x_T + b_T)
\]
where, $W_j$ denotes the overall weights of the local feature after the whitening transform, and $h_{1S}$ denotes the input offset, $x_T$ denotes the sampled image block of the target object.

### 2.2 Feature Selection Based on Correlation Analysis

The feature weights learned by the sparse autoencoder can also be assumed to be a set of complete-bases which effectively represents the most essential features [30-34]. Due to the similarity between the sample data in the cross-domain database [35]. Therefore, local features obtained by SAE will have a large amount of redundancy, which will increase the time for subsequent global feature extraction of the target sample based on the convolutional network. In light of the hypothesis which is to the effect that a good feature subset contains features which maintains no correlation with each other [36-37], steps are taken to conduct the selection of feature after SAE through eliminating one of each pair of feature weights which highly correlated with each other. Therefore, before the convolution operation, this paper uses the correlation analysis to select the features learned by the sparse autoencoder in order to reduce the feature dimension and prevent from being over-fitting. Assumed that the $q$-th and $p$-th row eigenvectors of the global weight matrix $W$ for feature extraction are $w_p$ and $w_q$, and the correlation coefficients can be as shown in Eq. (5):

$$\psi_{pq} = \frac{\text{cov}(w_p, w_q)}{\sigma_{w_p} \sigma_{w_q}}$$

where, $\text{cov}(w_p, w_q)$ denotes the covariance between $w_p$ and $w_q$:

$$\text{cov}(w_p, w_q) = \sum_{u=1}^{m} (w_p^u - \mu_{w_p})(w_q^u - \mu_{w_q})$$

where, $W$ denotes a weight vector's dimension, $w_p^u$ represents the $w_p$'s $u$-th element, $w_q^u$'s $u$-th element is denoted by $w_q^u$, $\mu_{w_p}$ refers to the mean value of $w_p$, and $\mu_{w_q}$ denotes the mean value of $w_q$. Accordingly, the standard deviation of $w_p$ is denoted in Eq. (7):

$$\sigma_{w_p} = \sqrt{\sum_{u=1}^{m} (w_p^u - \mu_{w_p})^2}$$

For instance, a correlation coefficient with a magnitude which is ranging from 0.7 to 1 demonstrates that there is a close relationship. The correlation coefficient with the magnitude which is higher than 0.3 but lower than 0.7 denotes a modest relationship. A weak relationship is indicated if the magnitude of a correlation coefficient is between 0 and 0.3. On the other hand, if magnitude is 0, then two weight vectors are irrelevant to each other. Hence, in this paper we use the absolute value of the correlation coefficient between $w_p$ and $w_q$ to evaluate the correlation between $w_p$ and $w_q$. As shown in Eq.(8):

$$r_{pq} = |\phi_{pq}| = \frac{\sum_{u=1}^{m} (w_p^u - \mu_{w_p})(w_q^u - \mu_{w_q})}{\sqrt{(w_p^u - \mu_{w_p})^2} \sqrt{(w_q^u - \mu_{w_q})^2}}$$

Through the above formula, a matrix of absolute correlation coefficients between different weight vectors can be obtained, as shown in Eq.(9):

$$r = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} & \cdots & r_{1q} & \cdots & r_{1s} \\ r_{21} & r_{22} & \cdots & r_{2p} & \cdots & r_{2q} & \cdots & r_{2s} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} & \cdots & r_{pq} & \cdots & r_{ps} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{q1} & r_{q2} & \cdots & r_{qp} & \cdots & r_{qq} & \cdots & r_{qs} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_{s1} & r_{s2} & \cdots & r_{sp} & \cdots & r_{sq} & \cdots & r_{ss} \end{bmatrix}$$

where, each element in the matrix represents the correlation coefficient between each pair of weight vectors, which takes the range [0, 1], since there is a mutual correlation between the two vectors ($r_{pq} = r_{qp}$), the correlation of any random vector with itself is always equal to 1, therefore, the absolute correlation coefficient matrix is a symmetric matrix. The specific steps of this chapter based on correlation analysis for local feature selection obtained by SAE are as follows:

1) Using unsupervised feature learning in the cross-domain to obtain the weight coefficient matrix $W \in \mathbb{R}^{s \times m}$ containing the s-dimensional weight vectors ($w_j \in \mathbb{R}^1 \times m$). And use the correlation analysis method to obtain the matrix of absolute correlation coefficients between each feature weight vector $r \in \mathbb{R}^{s \times s}$.

2) Since the matrix of absolute correlation coefficients is symmetrical, the correlation between different weight vectors can be represented by the upper triangular matrix, it is only necessary to select features for each upper triangular element in the matrix of absolute correlation coefficients. Given a threshold $r_T$, except for the diagonal elements, other elements of the upper triangular matrix will be preserved if it is equal to or greater than threshold $r_T$. On the other hand, the weight vector $w_p$ or the weight vector $w_q$ is removed when the threshold $r_T$ is exceeded by the absolute correlation coefficient $r_{pq}$ of a pair of weight vectors (i.e. $w_p$ and $w_q$).

3) Since the weight attenuation term is added to the
objective function of SAE, the weight coefficient obtained by the sparse auto-encoder after local feature learning for each image block is relatively low. Therefore, it is possible to determine which one of a pair of highly correlated weight vectors is deleted using the magnitude of the weighting coefficient of the automatic encoder (AE). Assume that the p-th row weight vector in the auto-encoder weight matrix W is \( \mathbf{W}_{AEp} \), the q-th row weight vector in \( \mathbf{W}_{AE} \) is \( \mathbf{W}_{AEq} \), and if \( \| \mathbf{W}_{AEp} \|_2^2 \) is equal to or greater than \( \| \mathbf{W}_{AEq} \|_2^2 \), then The p-row weight vector \( \mathbf{w}_p \) for the feature selection will be deleted.

4) After determining the weight vector which is about to be deleted for each pair of feature weight vectors, the row element corresponding to the weight vector discarded is removed form the overall weight matrix \( \mathbf{W}_{SAEwhite} \). Thus, a small amount (ie, \( t < s \)) of local weight vectors is eventually retained for subsequent extraction of global feature of the target object.

2.3 GLOBAL FEATURE EXTRACTION FOR TARGET IMAGES BASED ON CNN

After learning the local feature weights from the image sub-blocks using the sparse autoencoder, these weights will be used to extract the entire image’s features. The image recognition system based on convolution auto-encoder uses the convolution operation of CNN network to scan on the image, and the local feature weight is used as the detector to obtain the global feature response. Global features are extracted from all the large-size target object images in the target domain by means of neural network which includes a pooling layer and a convolutional layer. CNN model structure frame [38-39] is shown in FIGURE 3.

![Global feature extraction structure Diagram of Target object based on CNN](image)

FIGURE 3. Global feature extraction structure Diagram of Target object based on CNN

As shown in Fig.3, it is assumed that SAE has learned K local features on the \( n^2 \times n \) image blocks (K is the number of hidden layer units in the SAE network), Having obtained N feature weight vectors \( \mathbf{W} \in \mathbb{R}^{t \times m} \) after the selection of features, they are convolved on the whole image in the target domain so as to acquire an \((d - n + 1) \times (d - n + 1)\) array of global feature activations. With the aim to promote computational efficiency of global extraction, a 2D convolution in each color channel is carried out respectively and the three color channels’ convolution results are added up at last. In light of the fact that an \( l \times l \) image including three color channels and \( N \) weight vectors obtained after feature selection, it becomes able to obtain an \((l - n + 1) \times (l - n + 1)\) array of feature activations with \( N \times 3 \) after convolution (the convolutional stride is 1). The outputs of the convolutional layer is then combined by a pooling layer aggregating the feature activation in spatial regions. The pooling operation is capable of aggregating the convolved feature activations to reduce the resolution of the feature pattern obtained by the convolutional neural network. This not only reduces the feature dimension to avoid overfitting, but also achieves spatial scaling invariance. Mainly speaking, average pooling, summing pooling and maximum pooling are included in the pooling methods. In the present study, the whitening pretreatment is adopted in the convolutional SAE structure, therefore, the average pooling is most suitable for our application.

2.4 TARGET OBJECTS RECOGNITION BASED ON CLASSIFIER MODEL

The global feature activations extracted based on the convolutional network is finally sent to the classifier for supervised training. In present study, the regularized Logistic regression model [40] and Softmax regression model [41] are added to the CNN network respectively to realize both the classification of the two types and three types of target objects.

1) The logistic regression model is used to classify the two types of target objects. Add a logistic regression prediction function based on linear regression \( h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \) to realize the generalized linear model design of logistic regression for the classification of \( y = \{0,1\} \), As shown in Eq.(10).

\[
\begin{align*}
\mathbf{p}(y = 1 | x, \theta) &= \frac{1}{1 + e^{-\theta^T x}} = h_\theta(x) \\
\mathbf{p}(y = 0 | x, \theta) &= 1 - h_\theta(x) 
\end{align*}
\]  

(10)

Where, \( \theta \) represents the weight of the observation value of each feature in the whole prediction function. To determine which the class a sample belongs to, the value of the sample \( \theta^T x \) need be calculated and substituted it into the Eq.(10) to obtained \( h_\theta(x) \). If \( h_\theta(x) \) is greater than 0.5, the sample belongs to class of \( y = 1 \), whereas \( h_\theta(x) \) belongs to the class of \( y = 0 \). \( \theta \) is a parameter estimation problem, which can be used the Maximum Likelihood method to estimate the weight value. Assuming that the number of observed samples is \( n \), the training sample feature set is \( x = \{x_0, x_1, \ldots, x_n\} \), and the observed values are \( y = \{y_0, y_1, \ldots, y_n\} \). Substituting them into the Eq.(11), joint distribution function is shown in Eq.(12).

\[
\mathbf{p}(y | x, \theta) = (h_\theta(x))^{y} (1 - h_\theta(x))^{1-y} 
\]  

(11)
\begin{equation}
L(\theta \mid x, y) = \prod_{i=1}^{n} p(y_i \mid x_i, \theta) \\
= \prod_{i=1}^{n} (h_\theta(x_i)^{y_i}(1-h_\theta(x_i))^{1-y_i})
\end{equation}

(12)

The parameter \( \theta_0, \theta_1 \ldots \theta_n \) can be obtained by finding the maximum value of \( L(\theta \mid x, y) \). Performing a logarithmic operation on Eq.(12) can obtain the following results:

\[ l(\theta) = \log(L(\theta \mid x, y)) \]
\[ = \sum_{i=1}^{n} (y_i \log h_\theta(x_i) + (1-y_i) \log(1-h_\theta(x_i))) \]

(13)

When \( l(\theta) \) takes the minimum value, the value of \( \theta \) is the best weight parameter, so that the estimation problem in logistic regression is transformed into solving a set of \( \theta^* \) as shown in Eq.(14).

\[ \theta^* = \arg \min_{\theta} (l(\theta)) \]

(14)

\( \theta^* \) can be obtained through using the gradient descent method. As shown in Eq.(15).

\[ \frac{\partial}{\partial \theta_j} (l(\theta)) = \sum_{i=1}^{m} \left[ y_i \frac{h_\theta(x_i)}{h_\theta(x_i) - (1-y_i)} - \frac{1}{1-h_\theta(x_i)} \right] \frac{\partial}{\partial \theta_j} g(\theta^T x_i) \]

\[ = \sum_{i=1}^{m} \left[ y_i (1-g(\theta^T x_i)) - (1-y_i) g(\theta^T x_i) \right] x_{ij} \]

\[ = \sum_{i=1}^{m} (y_i - g(\theta^T x_i)) x_{ij} = \sum_{i=1}^{m} (h_\theta(x_i) - y_i) x_{ij} \]

(15)

After each iteration, \( \theta^* \) can be updated with Eq.(16):

\[ \theta_j := \theta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} (h_\theta(x_i) - y_i) x_{ij} \]

(16)

where, \( j \) represents the \( j \) types of characteristics of the sample, and \( \alpha \) represents the size of the iteration step. Substituting the obtained set of feature weights \( \theta_0, \theta_1 \ldots \theta_j \) into Eq.(16), the generalized linear model of logistic regression can be obtained, and using it to classify the target objects into two types.

The softmax regression model is used to sort the target objects as three types. Assuming that there are \( Q \) kinds of recognition results, the marker value at this time can be expressed as \( y(i) \in \{0,1,2\ldots Q\} \). Therefore, the Softmax model uses a \( Q \)-dimensional vector to represent the probability that a sample corresponds to each marker value. The model parameter at this time is a matrix of \( Q \) columns, each column of the matrix \( \theta(1), \theta(2), \ldots \), is a connection parameter corresponding to each tag value. Assuming that the eigenvector \( \theta(Q) \) corresponding to the i-th sample is

\[ V(i) \], the probability that the sample corresponds to each tag value in the Softmax model can be expressed as:

\[ h_\theta(x(i)) = \begin{bmatrix} p(y(i)=1 \mid x(i), \theta) \\
\vdots \\
\vdots \\
\vdots \\
p(y(i)=k \mid x(i), \theta) 
\end{bmatrix} \]

(17)

where, \( \theta \) denotes the model parameter, \( \alpha \) corresponds to the serial number of the \( Q \) tag values. The cost function of the Softmax model can be expressed as:

\[ J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} [y(i)=j] \log \frac{e^{\theta_j^T x(i)}}{\sum_{l=1}^{k} e^{\theta_l^T x(i)}} \]

(18)

where, \( 1\{\bullet\} \) denotes the indication function, the value is 1 when the condition in the parentheses is true, otherwise the value is 0.

\section*{3 EXPERIMENTAL VERIFICATION}

Under the condition without selection of feature weights, having learned local feature weights on image patches sampled from the cross-domain STL-10 dataset using SAE model. Then, we get global layer visualizing feature weights of all the target images using CNN model in the target domain. Finally, in order to classify the target object images into different types, the cross-domain higher layer local visualizing feature weights are sent into the Softmax regression model. We also obtain local feature weights on the target domain dataset without using transfer-learning and conduct three types of target object recognition experiments. We further use the traditional lower layer visualizing feature extraction algorithm including color histogram, LBP, and GIST to carry out recognition experiments. The recognition performance of unsupervised feature learning methods with and without transfer-learning is thoroughly compared with that of conventional lower layer visualizing feature using color, texture, edge. The selection of feature weights is finally adopted after local feature weights acquired on the cross-domain STL-10 dataset using SAE and feature selection’s effect upon the performance of transfer-learning based three types of target recognition has been evaluated. There are 100,000 vehicle and animal images which are unlabeled included in the STL-10 dataset applied for cross-domain unsupervised feature learning. These sample images are not directly related to subsequent target images to be identified. Partial samples of the STL-10 dataset is shown in Fig. 4. The target
dataset used for recognition experiment is 200 two types of target objects, which including cars and trucks on the road. The number of the target dataset used for recognition experiment of three types of target objects is 300, which including tower cranes, chimney and street light. Each type includes 100 number target images whose size is of $64 \times 64$. Partial two types of target object images for recognition experiment are shown in FIGURE 5. Partial three types of target object images for recognition experiment are shown in FIGURE 6.

3.1 LOCAL FEATURE EXTRACTION IN TRANSFER AND NON-TRANSFER LEARNING BASED ON SAE

Transfer learning based the SAE model is used to obtain local features. By sampling 100,000 $8 \times 8$ image patches at random from the cross-domain STL-10 dataset, SAE with 300 hidden units ($s=300$) is adopted to study local features in an unsupervised learning manner. During whitening transformation, the regularization term $\epsilon$ is set to 0.1. The parameters of the object function are $\lambda = 3e - 3$, $\beta = 5$ and $\rho = 0.035$. On the other hand, image patches are also gathered from three types of target images while local features are learned from these image patches without non-transfer learning. All the parameter settings applied for non-transfer learning on three types of target images are the same as that on the cross-domain STL-10 dataset. The visualization of local features learned from the STL-10 dataset and three types of target datasets are presented in FIGURE 7(a) respectively. The edge strength visualization of local features using SAE learned from the object dataset and that from the STL-10 dataset are shown in FIGURE 7(b).

As shown in FIGURE 7, it is able to infer that edges of the local feature learned from the STL-10 database are much more clear. Nonetheless, the local feature weights learned on the three types of target dataset are relatively fuzzy, and the edge performance of each local feature weights are obscure, which demonstrates that the result of unsupervised feature learned from a small sample dataset is not perfect enough in spite that a huge quantity of training examples are gathered. The reason is that there has been a high degree of similarity among the massive the tremendous image patches gathered from a limited number of three target images.

3.2 TARGET OBJECTS RECOGNITION WITHOUT FEATURE SELECTION
Based on small number of target images recognition without feature selection, the performance of unsupervised feature has been firstly evaluated. During the process of global feature extraction using CNN, 1 pixel is fixed as the stride in the process of convolution operation and the size of the pooling region is set to 40×40. Besides the performance of target recognition based on transfer-learning, non-transfer learning method that extracting features from the target domain (two or three types of target images) has also been conducted for testing the performance of target recognition.

For conventional low-level visual feature extraction methods (color histogram, LBP, and GIST), experiments have also been carried out to measure the performance of targets recognition using multiple low-level visual features, such as color histogram features, texture features, and edge features. Since the maximum number of iterations (MaxIter) of the Softmax regression model in the optimization procedure will excrete influence upon the results of recognition. MaxIter’s various values are applied to train logistic regression and the softmax regression models. With an interval of 10, the maximum number of iterations ranges from 10 to 100. In the experiments, a 5-fold cross-validation scheme has been availed and the average recognition accuracy, precision, and recall indicator are applied to evaluate all the methods. The calculation formula of the respective indicators are as shown in the formula (19), the formula (20), and the formula (21).

$$ \text{precision} = \frac{TP}{TP + FP} \times 100\% \quad (19) $$

$$ \text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (20) $$

$$ \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (21) $$

where, the quantity of positive samples that are judged to be positive is denoted by TP with FP representing the amount of negative samples which are regarded to be positive, FN represents the number of negative samples that are determined to be negative, and NP represents the number of positive samples that are judged to be negative class.

In this paper, conventional low-level visual feature extraction methods, the transfer learning method and non-transfer learning feature extraction methods based on SAE model are used to acquire low-level visual feature and high-level visual feature for identification of target objects. The results of two and three types of target objects recognition performance using different feature extraction methods are presented respectively in TABLE 1 and TABLE 2.

<table>
<thead>
<tr>
<th>Different feature extraction methods</th>
<th>Recognition Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision (%)</td>
</tr>
<tr>
<td>Transfer-learning</td>
<td>90.63</td>
</tr>
<tr>
<td>Non-transfering learning</td>
<td>87.43</td>
</tr>
<tr>
<td>Color histogram</td>
<td>73.23</td>
</tr>
<tr>
<td>GIST</td>
<td>75.23</td>
</tr>
<tr>
<td>LBP</td>
<td>76.13</td>
</tr>
<tr>
<td>Color histogram+GIST+LBP</td>
<td>76.32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 1 RECOGNITION PERFORMANCE OF TWO TYPES OF TARGET OBJECTS USING DIFFERENT FEATURE EXTRACTION METHODS</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Different feature extraction methods</th>
<th>Recognition Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision (%)</td>
</tr>
<tr>
<td>Transfer-learning</td>
<td>88.63</td>
</tr>
<tr>
<td>Non-transfering learning</td>
<td>82.23</td>
</tr>
<tr>
<td>Color histogram</td>
<td>73.23</td>
</tr>
<tr>
<td>GIST</td>
<td>78.23</td>
</tr>
<tr>
<td>LBP</td>
<td>76.23</td>
</tr>
<tr>
<td>Color histogram+GIST+LBP</td>
<td>75.32</td>
</tr>
</tbody>
</table>

From the above TABLE 1 and TABLE 2, it could be inferred that the traditional low-level visualising feature extraction methods, such as Color histogram, GIST, LBP, and multiple low-level visual features combined is worse than the recognition performance by using either transfer learning or non-transfer learning feature extraction methods based on the convolutional SAE model, which further proves that convolutional SAE can be used in the recognition of small number of target objects. Since the numbers of the target objects are relatively enough, the satisfactory recognition results can also be gotten with the help of the lower-level visual feature extraction method. In order to compare with the recognition performance of conventional low-level visual feature extraction methods, the transfer learning and non-transfer learning feature extraction methods on fewer number of target images, we select 100 number target from 200 two types of target and 150 number target from 300 three types of target as the new target dataset for recognition experiments, and each type includes 50 number images. The performance of recognition is evaluated from
the perspective of average accuracy which is tested by 5-fold cross-validation. The recognition results obtained by means of different methods under iterative test are shown in Figure 8.

From Figure 8 above, it is clear that the traditional lower-level visualizing feature extraction, such as color histogram, LBP, and GIST is obviously worse than the performance of the recognition of both non-transfer and transfer learning feature extraction method based on SAE, and the recognition performance based on the transfer learning feature extraction method is slightly better than the non-transfer learning feature extraction methods. The experimental result demonstrates that regardless of whether or not transfer learning is used, the feature extraction method based on unsupervised feature learning can acquire much better recognition performance than the lower-level visual feature extraction method. Transfer learning not only effectively make up for the problem of limited training data, but also acquire slight better recognition performance than non-transfer learning feature extraction methods.

With the aim to further compare with the recognition performance of both the non-transfer learning and transfer learning feature extraction methods on fewer number of target images, we select 60 number target from 300 three types of target as a new target dataset for recognition experiments, and each type includes 20 images. The recognition results of the three types of target objects during one experiment are shown in Figure 9.

From Figure 9, it would be clear that recognition performance based on the non-transfer learning method fails to efficiently achieve a high recognition precision. Nonetheless, based on transfer learning method, the recognition performance can achieve results with higher precision. This is because the number of three types of samples used for identification is only 60, after local feature extraction using SAE, the similarity between local feature vectors is smaller, and over-fitting is prone to occur. Therefore, the global feature vector of the target sample image obtained by convolving the local feature vector with small similarity to each target image does not represent all the features of the three types of target objects well, which results in poor recognition effect. Therefore, the transfer learning extraction method can be availed to sort the target objects when there is no sufficient target sample images.

With the aim to verify the influence of different hidden layer node numbers in SAE on performance of the recognition, we further respectively select the different node numbers, such as 50, 100, 150, 200, 250, 300, 350, and 400. The recognition performance results of three types of target objects under the different node numbers of the hidden layer are presented in Figure 10.
3.3 RECOGNITION OF TARGET OBJECTS WITH FEATURE SELECTION

Having evaluated the performance of transfer learning convolutional SAE for target objects recognition, in this section, experiments are conducted to test the influence of feature selection on the recognition performance of transfer learning method. Weight vectors are chosen from the s=300 weight vectors learned in Section 2.2 using different threshold values (e.g. 0.9, 0.7, and 0.5). Accordingly, 163, 154, and 136 weight vectors are finally kept after the selection of features. We only conduct three types of target recognition experiments while availing of all the feature weights acquired either with or without feature selection. The performance of recognition is estimated from the perspective of average accuracy tested by 5-fold cross-validation. Because the transfer learning feature extraction method achieves better accuracy in the recognition experiments, we only select the local features from the STL-10 database using SAE learned by means of feature weights selection and completes the recognition experiment on the three types of target objects. Three types of target recognition results using different feature weights is presented in the TABLE 3.

![FIGURE 10](image)

**FIGURE 10. Recognition performance results under the different node numbers of the hidden layer**

From FIGURE 10, we can see that at the beginning, as the node numbers of the hidden layer increase, the recognition performance of the proposed algorithm is apparently improved; when the node numbers of the hidden layer exceeds 200, no obvious promotions in the recognition of the algorithms are measured.

It can be inferred from TABLE 3 that the recognition performance of the transfer learning method can be further promoted by appropriate selection of features (e.g. \( r_1 = 0.9 \) and \( r_7 = 0.7 \)) even when the feature weight dimension is decreased by about half. The experimental result indicates that using the correlation analysis method to select the local feature weights obtained by SAE can not only reduce the correlation between the local feature weights and the number of features, but also the accuracy of recognition is not reduced. Since the proposed feature selection method has already been carried out before convolution operation, not only is it able to decrease the computational cost of the convolutional neural network for global feature extraction, but also reduce the time required for recognition. On the other hand, experiments are further conducted to estimate how much time feature extraction would cost while using various numbers of features. An Intel Corei7-3.2GHz CPU with an 16GB PC is used to evaluate time consumption of feature selection and target objects recognition. In one experiment, the average time of the different feature selections number and target objects recognition is shown in TABLE 4.

**TABLE 4**

<table>
<thead>
<tr>
<th>Number of Features</th>
<th>400</th>
<th>161</th>
<th>157</th>
<th>132</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time for recognition (ms)</td>
<td>(r_1=0)</td>
<td>(r_7=0.9)</td>
<td>(r_7=0.7)</td>
<td>(r_7=0.5)</td>
</tr>
<tr>
<td>Time consumption for feature selection (ms)</td>
<td>/</td>
<td>54</td>
<td>55</td>
<td>52</td>
</tr>
</tbody>
</table>

From TABLE 4, it can be inferred that due to the number of feature weights used for global feature extraction based CNN is reduced, recognition time can be decreased effectively with the assistance of feature reduction. In this sense, it is useful to conduct feature selection. By contrast, the selection of features is capable of improving the recognition accuracy to a slight extent. Moreover, the proposed feature selection method’s time consumption is marginally low (less than 60ms).

4 DISCUSSIONS

Accurately identify the target objects in the dynamic environment is of great significance for intelligent urban construction. Only in this way can we provide relatively more valuable information for the construction of intelligent urban. In this paper we have conducted extensive experiments on different types of target objects recognition using conventional low-level visual feature extraction methods, and convolutional SAE based methods with and without transfer learning. Feature weights selection method is further brought into the cross-domain convolutional sparse autoencoder model to achieve target objects recognition. In order to analyze it comprehensively, an overall discussion regarding the experimental results goes as follows.
1) Transfer learning convolutional SAE based methods.

From the experimental results mentioned above in Section 3.2, it can be inferred that the traditional lower layer visualizing feature extraction methods, such as Color histogram, GIST, LBP, and multiple low-level visual features extraction combined is worse than the recognition performance by using both non-transfer and transfer learning based on the convolutional SAE model, which proves that Convolutional SAE can be used for the recognition of target objects. In order to further compare with the recognition performance of both the transfer learning and non-transfer learning methods on fewer number of target objects, we select 150, 120, 60 number target objects from 300 three types of target sample dataset as a new target dataset for recognition experiments, respectively. Experiment results show that transfer learning method can achieve better recognition than non-transfer learning under the insufficient sample numbers. We also verify the impact of different hidden layer node numbers in SAE on recognition performance. It can be observed from Fig.10 that recognition performance does not change much when the number of hidden layer nodes exceeds 250, So we choose S=300 as the optimal value of the number of hidden layer nodes.

2) Feature weights selection based on correlation analysis.

From what has been found in experimental in Section 3.3, it has been much clear that feature selection which is based upon the correlation analysis can not only reduce the number used for global feature extraction and reduce the time required for target objects recognition, but also improve target objects recognition performance. For instance, it is able to reduce the global feature extraction’s time consumption while the recognition performance has not be decreased, which further demonstrates that a great deal of redundancy exists in the features learned by SAE from cross-domain dataset and it is necessary to carry out feature selection using analysis of correlation. It has been useful to perform feature selection in such manner from the viewpoint of recognition performance since it is a time-consuming to conduct CNN based global feature extraction.

5 CONCLUSIONS

In the construction of intelligent urban, accurately identifying the target objects in the dynamic environment can provide more valuable information, while the traditional recognition methods based on the underlying features such as color, shape and texture cannot effectively and accurately distinguish the target objects, which will inevitably lead to the fact that smart cities cannot truly realize informationization and intelligence. In the present study, a convolutional SAE based transfer learning method for small number of target objects recognition has been put forward. Additionally, for the purpose of reducing the computational cost of the convolutional neural network for global feature extraction, a feature selection method based on correlation analysis has been proposed for reduction the local feature weights learned by SAE. From the experimental results which are discussed above, a conclusion can be safely reached that a convolutional SAE method can be applied for target object recognition in the dynamic environment. Transfer learning can improve recognition performance of target objects when labeled training data are not sufficient. On the other hand, the experiments also demonstrate that the computational complexity can be reduced with the assistance of the proposed method of feature selection and the time required for recognition without the sacrifice in recognition performance. New ideas have been brought to other small number of target recognition problems in the construction of smart cities. Further, we will discuss high-level visual features acquired by unsupervised feature learning combining the low-level visualizing features, including edge, texture, and color to find the most suitable small number of target object recognition methods for the construction of smart cities.

ACKNOWLEDGEMENTS

This paper is supported by the Science and Technology Innovation Team of Henan Province Supports Project (Grant:17RTSTHN014), and Equipment Pre-research Key Laboratory Fund Project (Grant: 61420010404004).

REFERENCES


Bing Xie is a doctoral student in the School of Electronic Information, Northwestern Polytechnical University, Xi an, Shanxi, China. He received his BS degree in Measurement and Control Technology from Hefei Artillery Academy in 2010, and his MS degree in communication and information system from North University of China in 2013. His current research interest include image reconstruction, small target detection and recognition.

Zhemin Duan is a professor in the School of Electronic Information, Northwestern Polytechnical University, Xi an, Shanxi, China. He received his BS degrees from Northwestern Polytechnical University of Science & Technology in 1978 and his MS degree from Xi’An Jiaotong University in 1983. He is the author of more than 200 journal papers and has written four book chapters. His research interest include signal processing.

Liping Liu is a Senior engineer in the School of Electrical and Control Engineering, North University of China. Tai yuan, Shanxi, China. He received his BS and MS degrees from North China Institute of Technology of Information & Communition in 1982 and 1992, respectively, and his Ph.D. degree from Beijing Institute of Technology. His research interest include intelligent structure equipment test and test system design.