An image authentication technology based on depth residual network

Jiafa Mao, Danhong Zhong, Yahong Hu, Weiguo Sheng, Gang Xiao and Zhiguo Qu

College of Computer Science and Technology, Zhejiang University of Technology, Hang Zhou, People’s Republic of China; Department of Computer Science, Hangzhou Normal University, Hang Zhou, People’s Republic of China; Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, People’s Republic of China

ABSTRACT
The traditional image authentication technique generally determines the image attribution by extracting specific features and combining the similarity calculation algorithm. Because of the selected features dimensions, characterization and other factors, the accuracy and speed of image authentication have been restricted. In this paper, Recog-Net, an end-to-end image authentication model based on convolution neural network has been proposed. Deep residual network is chosen as the features extractor. Mahalanobis distance and threshold method are used to complete the image authentication. Experiments show that the performance of the extractor’s features, compared with the traditional features and the features of other convolution neural network architectures, is more excellent, with a high degree of generality, recognition rate and robustness, still having these advantages even after a substantial compression. The Recog-Net for image authentication is able to accurately authenticate the images tampered with certain range.

ARTICLE HISTORY
Received 22 November 2017
Accepted 23 February 2018

KEYWORDS
Convolution neural network; residual network; image authentication; image features

1. Introduction
The rapid development of new media has become a new engine of economic development. The development of new media raises higher requirements for digital copyright protection technology. For the increasingly serious copyright issues, there is an urgent need for a technology to determine the image attribution problems and tampering problems.

‘A picture is worth a thousand words’. When seeing pictures, human beings can make a complete explanation of a story according to their existing background knowledge. Just like the human beings, computers can also obtain the concept and meaning from a picture. To understand the content of a picture, the computer needs to extract the effective features of it, then use these features to build the data model. This is quite different from human beings that achieve the same goal using their existing background knowledge (Tian, 2013). The key to establish the image features model is to extract the image features, and the higher capability of features description, the more suitable for image authentication. How to extract the primary features of the image and what type of features is extracted play an important role in the process of image processing.

The most common image features including colour, texture and shape features. Most of the established image annotations and image extraction systems are based on these features, and most of their performance depends on the representation of the extracted features. In general, there are three kinds of features representation methods, i.e. global features, the features based on image blocks and local features (Tian, 2013). It is still a challenging problem of computer vision to extract the ideal features that can reflect the intrinsic content of image as much as possible.

1.1. The motivations
In the past, low-level image features are widely used, such as colour features (Huo, Jiao, Wang, & Yang, 2016; Zhang, Zhang, & Guo, 2016) and contrast (Adu, Xie, & Gan, 2016; Qi, Yu J, Ma, Li, & Tian, 2015), and there are some popular features extracting algorithms like HOG (Dalal & Triggs, 2005; Kittipanya-Ngam & Lung, 2011; Wei, Guo, Wang, & Wan, 2015), LBP (Banerjee, Moelker, Niessen, & van Walsum, 2012; Ojala & Harwood, 1996), HSV (Jasmine & Kumar, 2014; Wang, Li, Wang, & Liu, 2012), SIFT and SURF (Lowe, 2004; Zhong & Chen, 2014), GIST (Oliva &
Torralba, 2001). HOG features describe the image by calculating the gradient of the image’s pixels and counting the gradient histograms of the image. LBP features obtain the texture information about the image by local binary calculation of the image. HSV features describe the colour features of the image by counting the colour histograms of the image. SIFT and SURF construct the image pyramid to find the scale invariant points in the image to extract the key points in the image. GIST features describe the image features in a macroscopic view. Normally colour distribution and contrast are used to detect salient object, but in some night scenes there is often a low contrast, low brightness, low signal noise ratio, almost no colour information can be obtained (Mu, Xu, Zhang, & Zhang, 2017). Therefore, it can’t perform well in complex data sets if only one single feature is used, even though we hope to improve the quality of features by combining the various features. It is only mechanical superposition and can’t integrate the advantages of various features.

Image authentication technology needs to process a large number of complex data and it can’t accurately identify its attribution without excellent image features. So what is needed is not a simple feature, but the semantic description of the image. We are eager to find a way to extract the features of an image that can differentiate most of the images.

In recent years, deep learning has gradually replaced the traditional machine learning algorithms and become the most popular research subject. When studying image features, instead of a simple extraction of a certain type of features, deep learning starts from the bottom, then enhances the features step by step, and finally abstracts the advanced features. It can make up for the traditional features’ defects of oneness and break the limitations of traditional features of global features semantic description of the image. Convolution neural network, as an important algorithm of deep learning, is an effective method to extract image’s complex features and high-level abstract features (Najafabadi et al., 2015). It implicitly learns features of the training data, avoiding extracting features intuitively. Convolution neural network can output the result after inputting the original image. It has been successfully applied to different visual projects due to its strong features learning ability, such as image classification (Jie & Yan, 2014; Krizhevsky, Sutskever, & Hinton, 2012), scene annotation (Farabet, Couprie, Najman, & LeCun, 2013), target detection (He, Zhang, Ren, & Sun, 2014; Wen, Shao, Xue, & Fang, 2015), video analysis (Cai et al., 2016) and so on. So we decide to do some research on the deep learning, it is more potential to find the method we need in this direction. We use convolution neural network and transfer learning to extract image’s features and get the global features semantic description of the image so as to authenticate the image.

1.2. The contribution

After comparing the traditional features with the features of convolution neural networks, it is found that the features of convolution neural networks are much better than the traditional features and what we thought proves right. But we also found that the existing models don’t work well. We refer to a variety of existing models and build a composite model. One part of the model is pre-trained by the Image-net dataset, which is the most often used dataset in deep learning. Other parts are fine-tuned by self-made dataset. We change the structure of the convolution neural network and tune the super parameter many times to obtain a better one because it’s easy to be over fitting. It can be proved by experiments that the features of the improved convolution neural network model in our paper has good generality and the image authentication technology based on convolution neural network is unique and has a high accuracy even if in the complex data set. The main contributions to this paper are as follows:

(1) A new image authentication model called Recog-Net is established. The new model can solve the problem that traditional features extraction has a low accuracy in the complex data sets.

(2) Our paper improves the existing convolution neural network structure to make it perform well in learning image’s features.

2. Related work

According to the colour features, texture features and shape features, there are the following traditional features extraction algorithms widely used. Combining these features extraction algorithms with the dimension-reduction, the establishment of pyramid, construction of bag of words model and other methods can generate efficient features extractors.

2.1. HOG Features (Dalal & Triggs, 2005; Kittipanya-Ngam & Lung, 2011; Wei et al., 2015)

In 2005, Navneet Dalal and Bill Triggs proposed histogram of Oriented Gradient (Dalal & Triggs, 2005), which describes a local image property by counting edge gradient information. HOG calculation includes image correction, detecting window unit segmentation, unit pixel gradient histogram statistics, block histogram synthesis and histogram concatenation features generation and other
steps. The calculation of the unit pixel gradient can be expressed as:

\[ G_x(x, y) = H(x + 1, y) - H(x - 1, y) \]  
\[ G_y(x, y) = H(x, y + 1) - H(x, y - 1) \]  

In formulas, \( G_x(x, y), G_y(x, y), H(x, y) \) represent the horizontal gradient, vertical gradient and pixel values of the pixels \((x, y)\) in the input image respectively. The gradient values and directions at the pixel \((x, y)\) are:

\[ G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \]  
\[ \alpha(x, y) = \tan^{-1}\left( \frac{G_y(x, y)}{G_x(x, y)} \right) \]

Pre-processing Gamma transform, regional unit block statistic, histogram intensity normalization guarantees HOG descriptor having robust anti-illumination and anti-shading performance. Gamma transform can achieve the normalization of the image, reduce the local shadow and illumination changes of the image, Gamma transform compression formula:

\[ I(x, y) = I(x, y)^{\text{Gamma}} \]

Gamma’s value can be adjusted according to requirements, such as 1/2.

Directional gradient histogram features are widely used in the field of computer vision such as object detection, recognition. Its application includes pedestrian detection, face detection (Guo & Chen, 2013) and scene classification (Kobayashi, 2013). The gradient histogram preserves the edge information about the image, and has a good ability to represent the appearance of the image. However, for the problem of flexible target detection, the variable shape of the objects, limits the performance of HOG in some extent.

### 2.2. LBP Features (Banerjee et al., 2012; Ojala & Harwood, 1996)

In 1996, Ojala proposed a local binary pattern algorithm (Ojala & Harwood, 1996), and it is normally used for texture classification (Dalal & Triggs, 2005). It can describe the local texture features of the image concisely and efficiently. The LBP features describe an image processing technique within a grayscale range, for the grayscale image with an 8-bit or 16-bit input source. The LBP features compare the centre point of the window with the neighbourhood point, and then recode to form a new feature to eliminate the impact on the external scene of the image. To a certain extent, it solves the complex scene (lighting transformation) features description problem. The above calculation process can be expressed by the following formula:

\[ \text{LBP}(x_c, y_c) = \sum_{p=0}^{P-1} 2^p s(i_p - i_c) \]  

\((x_c, y_c)\) is centre pixel, \(i_p\) is the gray value of neighbourhood pixels, \(i_c\) is the gray value of centre pixel, \(s(x)\) is sign function, defined as:

\[ s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \]

LBP includes texture classification and segmentation, and has gray-scale invariance. It can be combined with some simple comparing method to make it better. The shortcomings of LBP are that it doesn’t have the rotation invariance, so it is ineffective for some applications, and it is necessary to improve the rotation invariance of LBP (Ojala & Harwood, 1996).

Most of the early texture classification methods use a single value to quantify the texture. With a simple texture information distribution, we can get very good results, which means that we should use the distribution of eigenvalues rather than through a single value. The gray-scale difference method achieves the best performance by distinguishing most of the texture information.

The calculation methods of LBP features include image gray scale, detection window unit segmentation, unit pixel binary computation, gray histogram of unit pixel statistics, block histogram synthesis and histogram concatenation features generation etc. steps.

### 2.3. HSV Colour Histogram Features (Jasmine & Kumar, 2014; Wang et al., 2012)

The colour space contains many types, such as RGB, HSV, YCbCr, CMY, etc. (Wang et al., 2012). By default, the general image use RGB colour space, but this colour space is not suitable for rendering to human vision. Although HSV describes the colour, from the human visual system in the tone, saturation, value, it is very conducive to image processing and identification. Hue refers to the different colours, saturation refers to a variety of colours, and the value is the degree of light and darkness. The conversion to RGB to HSV space can be done with the following formula:

\[ H = \begin{cases} \theta & G \geq B \\ 2\pi - \theta & G \leq B \end{cases} \]

\[ S = \frac{3}{(R + G + B)} [\min(R + G + B)] \]

\[ V = \max(R, G, B) \]
In (8), \( \theta = \begin{cases} \frac{(R-G) + (R+B) \times 0.5}{(R-G)^2 + (R-B)(G-B)^2} \end{cases} \), R, G, B represent the different colour components of the image pixels, \( H, S, V \) correspond to the conversion of the image saturation and the response value of brightness components.

The colour histogram can be obtained by counting the number of occurrences of each colour in the image matrix. The histogram is invariant to the translation and rotation of the image plane, and the change is very slow at the change of the viewing angle. The colour histogram \( H \) of a given image is defined as a vector:

\[
H = \{H[0], H[1], \ldots, H[l], \ldots, H[N]\}
\]  
(9)

where \( i \) is the number of occurrences of this colour in the colour histogram, \( H[i] \) represents the number of pixels in the HSV colour in the image, and \( N \) is the dimension in the colour histogram. To compare histograms with different sizes, the colour histogram needs to be normalized. The formula for colour histogram normalization is as follows (Jasmine & Kumar, 2014):

\[
H' = \frac{H}{P}
\]
(10)

where \( P \) represents the total number of pixels in the image. After the colour histogram is normalized, the colour difference between the images can be compared by the method of histogram intersections, distance and centre distance.

### 2.4. SIFT/SURF Features and BOW, FV, SPM encoding (Lowe, 2004; Zhong & Chen, 2014)

In 1999, Lowe proposed scale invariant features (SIFT) algorithm, and improved it in 2004 (Lowe, 2004). SIFT extracts the stable points in the scale image, and matches an object of the image by comparing these points. SIFT algorithm has been proved to be a very powerful algorithm in computer vision application. The SIFT descriptor has good robustness to deformations (such as translation, rotation, and affine). SIFT extracts the potential extreme points in the two images by constructing the Gaussian pyramids and searching for local extreme position and scaling them in a series of Gaussian difference images. The key to construct the pyramid is to build a scale space. In (Lowe, 2004), the Gaussian convolution kernel is the only transform kernel to implement the scale transformation and it is the only linear kernel. Gaussian convolution kernel can be expressed as:

\[
G(x, y, \sigma) = \frac{1}{2\pi \sigma^2} e^{-\left(\frac{x^2 + y^2}{2\sigma^2}\right)}
\]  
(11)

Let \( I(x, y) \) represent an image, \( L(x, y, \sigma) \) represent a scale image at a corresponding scale, \( * \) represent a convolution operation, then the scale transformation can be expressed as:

\[
L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)
\]
(12)

Then, each extreme point is described by an eigenvector, including the position, scale, and direction. The next step is to match these key points. The best matching candidate of each key point in the relevant image is obtained by sensing the nearest neighbour region of the image. The nearest neighbour region is represented with the key points as the minimum Euclidean distance between the eigenvector (Zhong & Chen, 2014).

The process of extracting sift features points is slow, and the dimension of each features point is 128. Because the number of features points in each graph is different, the eigenvector dimension of each graph is inconsistent, so we first extract the sift features, combine it with the bag of words model (BOW) (Banerji, Sinha, & Liu, 2013) and the spatial pyramid matching (SPM) algorithm (Lazebnik, Schmid, & Ponce, 2006), or combine it with the fisher vector coding (FV) (Perronnin & Dance, 2007) to represent the eigenvector.

The meaning of the bag of words model is, after the extraction of all the features, similar parts are gathered together to form a visual vocabulary, and the vocabulary is used as a dimension to create a histogram for image representation. Each dimension represents a complete visual vocabulary to describe the image.

SPM algorithm considers the spatial information and the image is divided into a number of blocks. The features of each sub block are counted respectively, and later all the features of the block are spliced to form complete features. In the details of the block, a multi-scale block method is chosen, i.e. the image is divided into many blocks of different scale to show a hierarchical structure of the pyramid.

The basic idea of Fisher vector coding is to construct a visual dictionary with GMM, which is to express an image by the gradient of the likelihood function. After Fisher vector coding, the image features’ dimension is increased and the image is described better.

### 2.5. GIST Features (Oliva & Torralba, 2001)

It is a kind of scene features description. Normally features descriptors describe the local features of a picture, but if we want to describe the scene of the picture, such as: ‘There are some pedestrians on the street’ we must identify whether the image has streets, pedestrians and other objects through the local features, and then determine whether this scene meets the condition. But undoubtedly, this amount of calculation is huge, and the eigenvector may be too large to be stored and processed in memory. This forces us to use a more ‘macroscopic’ method to...
describe the whole picture rather than the local features of the picture.

The features of GIST are a macroscopic feature to describe the image by describing the space. The following five aspects can be used to describe the spatial features, and they are: degree of naturalness, degree of openness, degree of roughness, degree of expansion, degree of ruggedness (Oliva & Torralba, 2001).

The calculation method of GIST features includes extraction of image colour features and texture features, and then these features are concatenated to form the gist features vector, that is the gist features of the image.

After combining with the dimension reduction, the bag of words model, the spatial pyramid matching and other methods, the ability in learning features of the traditional features extraction algorithm can be improved. However, according to a large number of experiments, the improved traditional features extraction is still very limited because of its single features. The features also perform badly in the complex data set.

3. Features extraction based on convolution neural network

The traditional features extraction is only for certain type of feature, such as LBP can perform very well in texture features, while the disadvantages of these algorithms are also obvious. They are only able to extract one type of features and it can’t give a global description of the image to have an excellent performance in image authentication. Convolution neural network can realize the representation of the global image by the features decomposition and reconstruction in multilevel. Therefore, this paper tries to extract the features of image with convolution neural network, and proves that it is more powerful than traditional features.

3.1. Convolution neural network

3.1.1. Basic framework of convolution neural network

Convolution neural network is the first truly successful algorithm to use multi-level network architecture, which is a multi-layer non-full-connected neural network. Each layer is composed of a plurality of two-dimensional planes, and each plane consists of multiple independent neurons. Convolution neural network contains two special network architectures, i.e. convolution layer and down sampling layer. The convolution layer and the down sampling layer can have multiple layers, which reflects the depth of the convolution neural network. The convolution layer contains multiple features planes to complete the features extraction task. The features plane is composed of neurons. Each features plane represents the features of the upper layer. All the neurons of the same features plane share a connection weight. The down-sampling layer can’t exist independently, and it must follow a convolution layer. The most commonly used sampling method is the pooling, i.e. the input features image is divided into rectangular areas, and calculation is made to these areas. The operation of taking the maximum of each rectangle is called the maximum pooling, and taking the mean value is called mean pool (Jia, Yang, & Liu, 2014).

Convolution neural network learns the features of an original image by the cooperation of convolution layer and the down-sampling layer. Classical BP algorithm is combined to adjust the parameters and to update weight. Weight updating formula used in BP network is shown in Equation (13) (Jia et al., 2014):

\[ \omega(t + 1) = \omega(t) + \eta \delta(t)x(t) \]  

where \( x(t) \) represents the output of the neuron, \( \delta(t) \) represents the error term of the neuron, and \( \eta \) represents the learning rate. In the convolution neural network, the network architecture of the convolution layer adopts the discrete form of convolution, and it is expressed as:

\[ X_{\beta}^\gamma = f \left( \sum_{\alpha \in M_{\beta}} x_{\alpha}^{\gamma-1} k_{\alpha \beta}^\gamma + b_{\beta}^\gamma \right) \]  

where \( M_{\beta} \) represents the collection of the input features maps, \( k \) represents the convolution kernel, \( \gamma \) represents the number of layers of the network, \( b \) represents the added bias for the input features maps. For the specific output map, the input mapping features can be obtained by different convolution kernel. \( f \) represents the activation function used for convolution neurons, and the most commonly used activation functions include the sigmoid function and the hyperbolic tangent function. The difference between these two is that the sigmoid function maps \([ -\infty, +\infty ]\) to \([ 0, 1 ]\), the hyperbolic tangent maps \([ -\infty, +\infty ]\) to \([-1, 1]\).

The function of the down-sampling layer is to sample the features in previous convolution layer to reduce the dimension of the features data, the number of output features maps is unchanged, but the size of output features maps will be significantly reduced compared with the input features maps, the down-sampling layer can be expressed as follow:

\[ x_{\beta}^\gamma = x_{\beta}^{\gamma-1} \times k_{\beta}^\gamma \]  

where \( x_{\beta}^{\gamma-1} \) represents the input features maps, \( k_{\beta}^\gamma \) represents the convolution kernel.

In general, convolution neural network is composed of input layer, convolution layer, down-sampling layer, the
whole connection layer and the output layer. The down-sampling layer must follow a convolution layer and they can be repeated alternately. Convolution layers usually use smaller convolution kernel, such as 3*3 or 5*5. The down-sampling layer reduces the dimension of the result from convolution layer. For example, we can use a down-sampling region which size is 2*2 to sample a feature map getting from the previous convolution layer, then the size of the feature map will become the half. Figure 1 shows a simple architecture of convolution neural network. The size of the input image is $32 \times 32$ and the size of the convolution kernel $k$ is $5 \times 5$. Convolution layer C1 containing 6 features maps in the size of $28 \times 28$ is the result of convoluting the input image. Different features map denotes the different features of the input image and each pixel in the feature map is convoluted by the $5 \times 5$ pixels of the previous layer and a convolution kernel in size of $5 \times 5$. S2 is a down-sampling layer. It sums the pixels' value of a $2 \times 2$ region in C1 layer and then adds a bias, and finally it uses a sigmoid function to get mapped. S2 layer reduces the dimension of the feature map from C1. Similarly, we can get C3 layer which contains 16 features maps in size of $5 \times 5$ using the output of S2 as input. Then we sample the C3 layer to get the S4 layer in size of $1 \times 1$. Finally, all the pixels of all the features maps of S4 are connected and drawn into one-dimensional vector to get F5. F5 is the feature learning from convolution neural network and then we will get the output through softmax function.

### 3.1.2. Convolution neural network architecture contrast

Convolution neural network only containing 2 convolution layers is rarely used in practical applications. Usually multi-layer convolution layers are chosen because the features only from one convolution layer are very limited. The more layers it has, the more global the features are. In recent years, many new structures of convolution neural network emerge. The general convolution neural network architecture includes a pre-convolution, features extractor of pooling layer and a classifier of fully connected layer, so the improvement of CNN is mostly to strengthen the features extractor. The common methods include reducing convolution kernel, decomposing and reducing parameters, increasing depth and width of the network, and modifying the data stream connection. Now there are several popular network architectures of convolution neural network, as illustrated in Table 1. From the analysis of Table 1, the improvement of the network makes the classification error rate gradually decrease, and most of the network models increase the width of the network. VGG introduces $3 \times 3$ small convolution kernels to replace large-scale kernel, reducing the parameters and deepening the network. GoogLeNet introduces the inception modules which use the $1 \times 1$ convolution kernel to reduce the dimension to deepen the network, and the auxiliary classifier solves the problem that low-layer convolution layer gradient diffuses. On the basis of GoogLeNet, Inception v2/v3 furthers to decompose and reduce the parameters on the square convolution kernel, increasing the ability of the features describing. ResNet uses Residual connection to solve the problem that when the depth of network increases and the gradient diffuses, then the performance of ResNet will get blocked.

### Table 1. Convolution neural network model comparison.

| Network model         | Network depth | Additional technology                  |
|-----------------------|---------------|----------------------------------------|
| AlexNet               | 5             | –                                      |
| ZF-Net (Zeiler & Fergus, 2014) | 5             | widen network                         |
| VGG_CNN_M/S/F (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014) | 5             | widen network                         |
| VGG                   | 16–19         | $3 \times 3$ small convolution kernel parameter reduced |
| GoogLeNet/Inception v1 (Szegedy et al., 2015) | 22           | Inception module, auxiliary classifier added |
| Inception v2/v3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) | 46           | Improved Inception module, convolution kernel decomposition |
| Residual-Net (He et al., 2016) | 50/101/152   | Residual connection                    |
| Inception-ResNet/Inception v4 (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017) | 175          | Combined with Residual connection and Inception module |
Plentiful features of the image can be obtained by extracting the output of the intermediate convolution layer and the postposition fully connected layer in CNN. It is pointed out in (Srivastava, Greff, & Schmidhuber, 2015) that the output of the neural network layer which has been trained in a large database called ImageNet has a shocking baseline capability and using a small database to fine tune the pre-training model can get better classification performance. Transfer learning based on the pre-training and fine tuning model has become the important way to improve the performance of convolution neural network and accelerate the convergence of the network.

However, according to the experiments, our paper holds that the existing network structure is not perfect while training the general data set and still has space for improvement. We can strengthen the features extractor and improve the ability of learning features.

3.2. Improved convolution neural network architecture – Res-ExtNet

3.2.1. Network architecture

In this paper, an image features extraction network, i.e. Res-ExtNet, based on deep residual network is proposed to improve the existing convolution neural network architecture. As shown in Figure 2, Res-ExtNet network is composed of features extractor, features adjustment layer and classifier. Residual-Net is used as the features extraction model and the network depth is deepen through residual connection and block overlay. All these help to improve the quality of the features and ensure that the features are rich and varied. The pre-residual network consists of four convolution stacks of 48 layers. The input images will be resized to 224 × 224 × 3, and the parameters of convolution layer in the block are all reduced by 3 × 3 small convolution kernel. The features adjustment layer consists of a convolution layer P1 (512 output) and a fully connected layer P2 (512 outputs). Layer P1 uses the 1 × 1 convolution kernel to enhance the general features in ResNet through the ImageNet dataset and generates the target dataset features. Layer P2 enhances the features of P1, and performs the end-to-end features dimension reduction to reduce the storage space. The number of nodes in P2 can be set according to the actual storage requirements, and we will do some experiments to determine whether dimension has influence on performance of the features in P2 layer.

3.2.2. Training details

The network training includes pre-training and features adjustment. The first stage is training residual network in the ILSVRC2012 data set with the features adjustment layer disabled and classifier is modified to 1000 class classification, and training hyper parameters refers to the MSRA ResNet (He, Zhang, Ren, & Sun, 2016). The second stage is adjusting features in the specific data set and the dimension reduction processing, and the output layer will be set to corresponding data set category. In this stage, the features extractor parameter is loaded by the pre-training model, the features adjustment layer and the classifier weight parameter are initialized by the mode ‘Xavier’.

In our experimental environment, the process of ‘Xavier’ initializing is that: the input dimension of the layer of the parameters is defined as n, and the output dimension is defined as m, then the parameters will be initialized through the way of uniform distribution in the range of \[-\sqrt{\frac{6}{m+n}}, +\sqrt{\frac{6}{m+n}}\]. Hence, the variance of the input weight of neurons (while back propagating, it can be token as output weight) can be 1/n, and information distributes uniformly across the network. The detailed derivation is described in (Glorot & Bengio, 2010). In order to ensure the stability of the features when training, the convolution kernels of the first 3 convolution blocks of ResNet do not accept training, the fourth convolution block allows adjustment.

Before the training, the input image will be resized to 224 × 224. The width and height of the image are scaled in the same proportion to make the short edge meet 224 pixels, then we will get the standard image after centre cutting rather than stretching image directly. While
training, we will use horizontal flip to increase the number of images and take the pretreatment, i.e. the input images subtract mean value of the each channel of input images.

3.2.3. Generality of the model
The network architecture proposed in this paper has good generality. The existing pre-training model can replace the features extractor under the situation of limited computing resources, such as CaffeNet, VGG_CNN, Inception and so on. Simulation of using multiple pre-training model are carried out and the results are compared to confirm that the model proposed has good generality. More details please refer to the experiment part. The features adjustment layer is derived from Ren ShaoQing NOC network (He et al., 2014), and it proves that increasing the convolution layers can improve the network detecting effect. The features adjustment layer can modify the network’s size according to actual conditions, improve the quality of features by increasing the depth of network, and can accelerate the convergence by reducing the parameters to reduce the width of network.

4. Image authentication technology based on depth residual network

4.1. Image authentication model–Recog-net
Based on depth residual network, we present an end-to-end image authentication model called Recog-Net. This model consists of two modules, the image features extraction module and the image authentication module. Figure 3 shows the detailed authentication procedure using Recog-Net. Key points of image authentication is to find the features vector that can describe the image accurately with low time and space expense. It should also have a good performance in recognizing the image in most cases even though the image has been compressed or tampered with. Therefore, this paper focuses on how to extract the image features with strong characterization capacity.

4.2. Image features extraction module
Image features extraction includes a few steps. Firstly, we train convolution neural network – Res-ExtNet through the data set, adjust the parameters and get the label of image after classified. Then we remove the classifier and get the features of the adjustment layer, and these features can be regarded as the one learnt from the network with better generality, robustness and good characterization capability. All these features vectors and their labels are saved in a database. While training the image features extraction network, it’s hard to judge if we have got a good network because it’s easy to be over fitting. So we need to use a small learning rate and do more experiments.

4.3. Image authentication module
There are two kinds of output after image extraction, one is the features vector of the image, and the other is the image label after classification. The image authentication module puts the image to its corresponding category according to the image label, and then calculates the Mahalanobis (Mahalanobis & Vijaya Kumar, 2008) distance between the image and other images in this category. The calculating results is defined as $D_i(i = 1, 2, 3 \ldots)$, and $D_{\text{min}}$ is found out. The calculating result is used to judge whether the image to be authenticated is from this database or has the same source image as an image in this database. Threshold method is chosen to do the judgment, and the threshold, defined as $Th_m$, gets its appropriate value after multiple experiments.

If $D_{\text{min}} > Th_m$, the image has a quite large distance with all the images in this category, so it should not belong to this database. Otherwise, the image belongs to this database, and it must be the same image or has the same source image with the image to which the $D_{\text{min}}$ corresponds.

Mahalanobis distance is defined as: there are $M$ known sample vectors $X_1 \sim X_m$, $S$ is the covariance matrix, and the vector $\mu$ is the mean. The Mahalanobis distance of the sample vector $X_i$ to $\mu$ is expressed as:

$$D(X) = \sqrt{(X - \mu)^T S^{-1} (X - \mu)} \quad (16)$$

And the Mahalanobis distance between vector $X_i$ and vector $X_j$ is defined as:

$$D(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)} \quad (17)$$

If the covariance matrix is a unit matrix, then each sample vector is independent-identical-distribution. The formula is defined as:

$$D(X_i, X_j) = \sqrt{(X_i - X_j)^T (X_i - X_j)} \quad (18)$$

As can be seen, the formula (18) actually shows the Euclidean distance. If the covariance matrix is a diagonal matrix, the formula becomes normalized Euclidean distance. Therefore, Mahalanobis distance has similar calculation with Euclidean distance. Euclidean distance does not consider the correlation of each dimension, while Mahalanobis distance does and it also eliminates
the correlation between variables. In conclusion, Mahalanobis distance is more suitable for the distance calculation in this paper. But calculating Mahalanobis distance takes more time because the calculation formula is more complex. So we need to find a way to overcome the problem.

5. Experiment simulation

5.1. Experimental environment and background

The hardware configuration of the experiment is Intel core i7 6700 k, 8 core 16 lines, 3.4 GHZ dominant frequency, and NVIDIA GTX 1070 6 GB graphics memory. The experimental platform is ubuntu16.04 LTS with Python language. The experimental framework is Caffe and toolkit for Sklearn and Skimage are also included.

In order that the experimental results are effective and reliable, two data sets are used in the experiments. One is public data set Caltech-101 and the other is the self-made data set Scene-50. The Caltech-101, collected by Fei-Fei Li, Marco Andreetto and Marc’Aurelio Ranzato in September 2003 (Li, Fergus, & Perona, 2007), is an image database of California Institute of Technology. The Caltech-101 dataset consists of 9146 images and each image is about \(300 \times 200\) pixels. These images are divided into 101 different object categories, including the common categories, such as face, aircraft, motorcycles and other mixed categories. There are 40–800 images inside each category and most categories obtain 50 images.

As shown in Figure 4, the self-made data set Scene-50, consisting of 10,000 images, is made of images from Google and the video image frame from YouTube. There are altogether 50 categories, including hotels, museums, forests and other scenes, and each scene has about 200 images.

Both in the two datasets, training samples are about 70% and testing samples are about 30%.
5.2. Features performance simulation

This simulation is to verify the quality of the features extracted by our proposed algorithm, so that to test the accuracy of the authentication model. The experimental data set is Caltech-101 and our algorithm is compared with the low-level visual features, including HSV, SIFT+BOW, ScSPM, HOG, GIST and the convolution neural network model, including CaffeNet, VGG_CNN, GoogleNet. Brief introductions of these convolution neural network models are as follows.

CaffeNet: It is improved by BVLC (Berkeley Visual and Learning Center) based on AlexNet, and the network contains 5 convolution layers. The number of features image is between 96 and 384. There are 3 fully connected layers and 4096 neurons in per layer, and the last layer is the output layer.

VGG_CNN: This kind of convolution neural networks is similar to CaffeNe. It has a little more convolution layers. There are at most 512 features images in one layer and 3 fully connected layers. VGG_CNN has kinds of models such as VGG_CNN_M, VGG_CNN_F, VGG_CNN_S and so on. All of the networks are pre-trained by ILSVRC2012 data set and the error rate of Top-5 in testing phase is 13%–16%.

GoogLeNet: It is initially proposed by Google in 2014 and it was the champion in competition classification task of ILSVRC2014. It pioneers the use of Inception architecture, using $1 \times 1$ convolution kernel to reduce the dimension of the features image to deepen the network. Inception v1 has 22 convolution layers, not including the fully connected layer, and use auxiliary classifier to adjust the low-level layer weight. The error rate of top-5 is 6.67% in ILSVRC2014.

In the simulation, we use random forest classifier to do the image classification, and the number of decision tree is set to 25. From the analysis of Table 2, the traditional features extracting algorithm is superior to the convolution neural network in dimension. The reason is that the traditional features extracting algorithm uses PCA and other specific coding methods such as BOW, FV, SPM to reduce dimension. However, when considering the accuracy of image classification, the efficiency and performance of convolution neural network features extraction algorithm are much higher than the traditional features extraction engineering. It is because that the traditional extraction algorithm needs artificial design, relying on

| Features extraction algorithm               | Features dimension | Classification accuracy |
|---------------------------------------------|--------------------|------------------------|
| HSV                                         | 256                | 29.4                   |
| SIFT-PCA+BOW(1000) (Banerji et al., 2013)   | 1000               | 31.7                   |
| ScSPM(500)+PCA (Chatfield et al. 2014)      | 1024               | 36.9                   |
| HOG+PCA (Zhang et al., 2016)                | 128                | 35.1                   |
| GIST                                        | 960                | 41.8                   |
| CaffeNet(FC7)                               | 4096               | 89.4                   |
| VGG_CNN_F(FC7) (Szegedy et al., 2017)       | 4096               | 86.7                   |
| VGG_CNN_M(FC7) (Szegedy et al., 2017)       | 4096               | 89.0                   |
| VGG_CNN_S(FC7) (Szegedy et al., 2017)       | 4096               | 88.2                   |
| GoogLeNet(Pool5) (Mu et al., 2017)          | 1024               | 91.7                   |
| **Ours Res-ExtNet(FC_P2))**                 | **512**            | **94.1**               |
prior knowledge and the features from the traditional extraction algorithm has a bad generality. So it is difficult to catch the essential features of things in complex scenes. On the contrary, the convolution neural network based on the end-to-end learning, relying on big data and the space with high-dimensional parameters, gradually synthesizes the advanced features from the bottom to the top layer. The self-learning method with data driving ensures that the convolution neural network has excellent features extraction ability. In addition, comparing with various types of neural network model, the error rate of Res-ExtNet is 3%–8% lower. Through the experiment, we can prove that the quality of features from ResNet which is with deeper network and the residual connection, is higher than the features from the other shallow network.

5.3. Generality analysis of the image features extraction network

This experiment mainly tests the generality of the features from the features extractor in Res-ExtNet network. We use the self-made Scene-50 data set and use the CaffeNet, VGG_CNN_S and GoogLeNet to compare with Res-ExtNet. Here, we use the different features respectively from CaffeNet, VGG_CNN_S, GoogLeNet and Res-ExtNet to do the image classification and also analyze the result. Each network has two architectures with features adjustment layer or not, and the specific improvements are as follows.

Table 3. Network generality comparison.

| network model | Features adjustment layer | Features dimension | Classification accuracy |
|---------------|---------------------------|--------------------|------------------------|
| CaffeNet      | ×                         | 4096               | 64.9                   |
| CaffeNet      | √                         | 512                | 66.8                   |
| VGG_CNN_S     | ×                         | 4096               | 64.6                   |
| VGG_CNN_S     | √                         | 512                | 72.8                   |
| GoogLeNet     | ×                         | 1024               | 68.4                   |
| GoogLeNet     | √                         | 512                | 74.8                   |
| Ours Res-ExtNet | √                        | 512                | 77.1                   |

VGG_CNN_S: Remove the output layer. The neuron in the penultimate second layer -FC7 is set to 1024. Use the FC7 layer as the output layer and the features vectors from FC7 are the new image features. The architecture of features adjustment layer is FC_P1/512 and FC_P2/512, where FC_P1 and FC_P2 are layer name, and 512 is the dimension of the output in the layer.

GoogLeNet: Remove the auxiliary classifier. Replace the output layer of the main classifier with a convolution (size is $1 \times 1$, output is 512) as features extracting layer and use the features from the added convolution layer as output. The architecture of features adjustment layer is Conv_P1/512 and FC_P2/512, where Conv_P1, FC_P2 are the name of layers, and 512 is the dimension of the output in the layer.

It can be seen from Table 3 that the performance of features in the network using the features adjustment layer is more excellent than that in the original network, and the overall accuracy of the classification is increased by 3%–6%. According to the classifying accuracy comparison with other 4 kinds of networks with features adjustment layer, Res-ExtNet network performs better than the other networks. The 22 layer GoogLeNet with features adjustment layer also has a better performance than that of CaffeNet and VGG_CNN. The testing result is consistent with the result shown in the previous section. In summary, the neural network with features adjustment layer has good generality and supports multiple features extracting networks. The performance of features depends on the pre-training model architecture, and it is obviously improved compared with features from the original network.

5.4. Analysis of the effect of network dimension reduction

In this experiment, we mainly test the effect of network dimension reduction on the overall performance of features. We use self-made Scene-50 as data set, Res-ExtNet as the network model, and reduce the dimension of the features in FC_P2 layer.

![Figure 5. Comparison of the effect of network dimension reduction.](image-url)
As shown in Figure 5, we do dimension reduction on the features vectors of the features adjustment layer. The result has changed little and the change of the classification error rate is basically within 1%. In conclusion, while still having enough accuracy in a certain range, we can substantially compress the features to decrease the consumption of the storage and computation.

5.5. Analysis of image authentication network

This experiment mainly tests the accuracy of the image authentication module. We do a small range of rotation, affine, gray variation and other transformations on each picture of the Scene-50 data set to simulate the tampered images. These tampered images together with the original Scene-50 data set form a new data set named new Scene-50. New Scene-50 data set is used as training data. The accuracy of the image authentication module depends on the results of the previous image classification and the threshold selection, so the purpose of this experiment is to find the appropriate threshold. The initial value of the threshold is set by the statistical analysis after calculating the distance between the transformed image and the original image. According to a large number of experimental data (shown in Figure 6), the average value of Mahalanobis distance between the transformed image and the original image is about 1.2. However, the Mahalanobis distance between two completely different images is basically more than 3. Therefore, the threshold can be set roughly. To validate this we give a series of thresholds and do experiments. The result of the experiments is shown in Figure 7. Then we set the threshold to 2.1. The accuracy is 73.6% by using Caltech-101 public data set and it can reach 90.8% by using the Scene-50. In conclusion, the authentication model has quite good performance in image authentication.
6. Conclusion
Existing research and the experiments of this paper have proved that the convolution neural network plays an important role in features extraction. Our experimental results show that the generality and stability of traditional features are poor and the limitation is large, while the convolution neural network has a strong generality and stability.

Our experiments also show that Res-ExtNet architecture has a high accuracy and is superior to the existing convolution neural network in features learning. Res-ExtNet architecture is improved on the basis of the depth residual network model and it is composed of features extraction layer and features adjustment layer. The features adjustment layer is composed of convolution layer and fully connected layer. By introducing the residual network, it overcomes the constraints of the performance in the features extracting of the shallow network. It deepens the network and enhancing the quality of features by using the block structure and the residual connection. At the same time, the features in the adjustment layer of the Res-ExtNet architecture also has the function of dimension reduction. Accuracy can be kept even though the features vector’s dimension has been substantially compressed. The Res-ExtNet architecture has better performance than that of (Jie & Yan, 2014) although the depth of it is far less than that of (Jie & Yan, 2014). It is proved that the Res-ExtNet architecture proposed in this paper is a new breakthrough to the existing network.

The end-to-end image authentication model RecogNet based on the convolution neural network–Res-ExtNet also has a high accuracy. The image features based on Res-ExtNet has the ability of dimension reduction and has a good accuracy while the dimension of features is 32. Therefore, the image authentication model has the advantages of fast calculation and small storage space, and thus it has meaningful significance in image authentication. That means this image authentication technique can be widely used in the digital image copyright protection and authentication, video copyright protection, image similarity matching and other multimedia authentication. It can meet the requirements for digital copyright protection well.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This work is supported by the National Natural Science Foundation of China [grant number 61573316], [grant number 61373131], the Zhenjiang province Natural Science Foundation of China [grant number LY15F020032].

ORCID
DanHong Zhong http://orcid.org/0000-0002-8488-8574

References
Adu, J., Xie, S., & Gan, J. (2016). Image fusion based on visual salient features and the cross-contrast. Journal of Visual Communication and Image Representation, 40, 218–224.
Banerjee, J., Moelker, A., Niessen, W. J., & van Walsum, T. (2012). 3D LBP-Based rotationally invariant region description. Computer vision - ACCV 2012 workshops (pp. 26–37).
Banerji, S., Sinha, A., & Liu, C. (2013). A new bag of words LBP (BoWL) descriptor for scene image classification. International conference on computer analysis of images and patterns (pp. 490–497). Berlin: Springer.
Cai, Y., Zheng, W., Zhang, T., Li, Q., Cui, Z., & Ye, J. (2016). Video based emotion recognition using CNN and BRNN. Chinese conference on pattern recognition (pp. 679–691). Singapore: Springer.
Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets. BMVC 2014, Nottingham, The British Machine Vision Conference.
Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on (pp. 886–893). IEEE.
Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2013). Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis & Machine Intelligence, 35(8), 1915–1929.
Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research, 9, 249–256.
Guo, J. X., & Chen, W. (2013). Face recognition based on HOG multi-feature fusion and random forest. Computer Science, 40(10), 279–283.
He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), 1904–1916.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).
Huo, L., Jiao, L., Wang, S., & Yang, S. (2016). Object-level saliency detection with color attributes. Pattern Recognition, 49, 162–173.
Jasmine, K. P., & Kumar, P. R. (2014). Integration of HSV color histogram and LMEBP joint histogram for multimedia image retrieval. Intelligent computing, networking, and informatics (pp. 753–762). India: Springer.
Jia, S., Yang, D., & Liu, J. (2014). Fine classification of commodity images based on convolution neural network. Journal of Shandong University of Science and Technology (Natural Science Edition), 33(6), 91–96.
Jie, Z., & Yan, S. (2014). Robust scene classification with cross-level LLC coding on CNN features. Computer vision – ACCV 2014. (pp. 376–390), Springer International Publishing.
Kittipanya-Ngam, P., & Lung, E. H. (2011). HOG-Based Descriptors on rotation invariant human detection. Computer vision – ACCV 2010 workshops (pp. 143–152). Berlin: Springer.
