Double Compression Detection Based on Markov Model of the First Digits of DCT Coefficients

Lisha Dong, Xiangwei Kong, Bo Wang
School of Information and Communication Engineering
Dalian University of Technology
Dalian, P.R. China
deliadls@163.com, {kongxw, bowang}@dlut.edu.cn

Xin'gang You
Beijing Institute of Electronic Technology and Application
Beijing, P.R. China
youxg@dlut.edu.cn

***Abstract*—Double compression usually occurs after the image has gone through some kinds of tampering, so double compression detection is a basic mean to assess the authenticity of a given image. In this paper, we propose to model the distribution of the mode based first digits of DCT (Discrete Cosine Transform) coefficients using Markov transition probability matrix and utilize its stationary distribution as features for double compression detection. Experiment results show the effectiveness of the proposed method and comparison has been made to show the improvement by using this second order statistical model.**

Keywords-double compression detection;Markov model; first digits of DCT coefficients

I. INTRODUCTION

With the development of the image processing software, tampering a digital image becomes an easy job for non-specialist and the tampered image is hardly detected by the naked eyes, thus to ensure the authenticity and integrity of a given image, an effective method to detect image tampering is needed. To tamper an image, the typical steps usually include cutting and pasting, post-processing and resaving. For most of cases, the tampered image is resaved in a JPEG format in order to save storage space while maintaining the image quality. If the image for tampering is originally stored in a JPEG format, then double compression occurs when the second compression quality factor is different with the original one. Thus, if an image is detected to have gone through double compression operation, a basic conclusion can be drawn that the image might be tampered.

Until now, several methods have been proposed to detect double compression. In [1], A.C.Popescu pointed out that there were periodical artifacts on the histogram of DCT (Discrete Cosine Transform) coefficients for the double compressed image. Fourier transform of the zero-mean histogram of the DCT coefficients was used as a quantitative measure to discriminate between single compressed and double compressed images. This method was improved by B. Mahdian et al. [2] by considering only the magnitudes of DCT coefficients corresponding to AC components. D.Fu et al. [3] proposed a novel statistical model based on Benford's law for double compression detection on the consideration that the probability distribution of DCT coefficients of a single compressed image follows a generalized Benford's law but the law is violated by double compression. A more detailed result was given in [4], and the detection method was improved by utilizing the probability distributions of the DCT coefficients from individual AC modes. C.Chen et al. [5] proposed to use Markov random process to model the JPEG coefficient 2-D arrays and the transition probability matrix was used as features to detect double compression. Y. Chen et al. [6] proposed a novel quantization noise model to characterize single and double compressed images, and the uncompressed ground truth image could be approximated using image restoration techniques. X. Feng et al. [7] proposed three features based on the periodic artifacts and discontinuities in the signal histogram to detect JPEG re-compression.

Inspired by the above works, in this paper, in order to detect double compression operation, we propose to model the distribution of the mode based first digits of DCT coefficients using Markov transition probability matrix and extract features by calculating the stationary distribution of the modeled Markov chain. Experiment results show that the detection accuracy can be improved by considering the second order statistical model.

The rest of the paper is organized as follows. In section 2, the distribution of the first digits of DCT coefficients is explored and double compression effects on it are analyzed. In section 3, Markov transition probability matrix and the stationary distribution of the Markov chain are introduced first, and then a block diagram of the feature generation algorithm is given, followed by a detailed description of the propose method. Experimental results and discussion are shown in section 4, and the final conclusion is drawn in section 5.

II. DISTRIBUTION OF THE FIRST DIGITS OF DCT COEFFICIENTS

As is proposed by D. Fu et al. [3], the probability distribution of the first digits of the DCT coefficients before quantization follows Benford's law, which is first proposed by Benford [8] in 1938. The Benford's law states that for a table of physical constants or statistical data, the distribution of the first digits of these data is logarithmic, as is shown in (1):

$$p(x) = \log_{10}(1+\frac{1}{x}), x = 1,2,...,9. \quad (1)$$

978-0-7695-4541-7/11 $26.00 © 2011 IEEE
DOI 10.1109/ICIG.2011.100

Although for the singly compressed image, the quantized DCT coefficients don't follow Benford's law quite well, it still follows a parametric logarithmic function, as is shown in (2):

$$p(x) = N \log_{10}(1 + \frac{1}{s + x^q}), x = 1, 2, ..., 9. \quad (2)$$

Where N is a normalization factor and s and q are model parameters varying according to different images with different quality factors.

As is known that double compression cause periodical artifacts on the histogram of the DCT coefficients due to the different quantization steps utilized by the first and second JPEG compression. Experiments have also shown that the distribution of the first digits could be affected. Shown in Fig.1 are the mean distributions of the first digits of DCT coefficients for 500 different images for both single and double compression. From the figure we can see that for the single compression case, the first digits of DCT coefficients follow a generalized Benford's law, while for the double compression case, the distribution of the first digits show obvious violation to the logarithmic trend. The probability distribution of the first digits of DCT coefficients was used directly as features in [3] to classify double compressed images from the single compressed ones, and the result was improved in [4] by using mode based first digit features, i.e. using the probabilities of the first digits of DCT coefficients from individual AC modes. Both features used in [3] and [4] are first order statistics, which may cause information loss since the first order statistics can't reflect the correlations between adjacent pixels. So in this paper, we propose to use Markov model to characterize the mode based first digits of DCT coefficients, expecting that the second order statistical tool would give better performance.

III. Effective Features for Double Compression Detection

In this section, a set of effective features is proposed for double compression detection. Firstly, the Markov model is introduced as a tool to characterize the distribution of the mode based first digits of DCT coefficients, and the block diagram of the proposed feature generation algorithm is given afterwards.

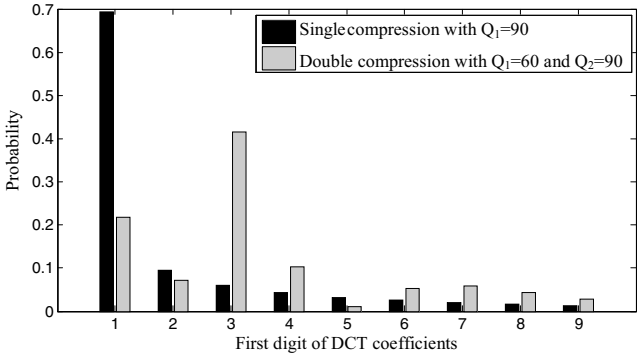

Figure 1. Distributions of the first digits of DCT coefficients for single compression with quality factor Q_1=90 and double compression with quality factor Q_1=60 followed by Q_2=90.

A. Markov Model

Since there are correlations between the DCT coefficients of JPEG images due to the image formation pipeline and the JPEG compression process, we expect that there are also some correlations between the first digits of DCT coefficients of individual modes. Markov process is proved to be a useful tool to characterize correlation, so in this paper, we model the mode based first digits of DCT coefficients as a Markov chain and use one-step transition probability matrix to characterize this process.

There are only ten states in this Markov model since the range of the first digits of DCT coefficients is from 0 to 9, so this is a finite-state Markov chain. Note that being different from [3] and [4], we include 0 in the range of the first digits to retain as much information as possible.

Let $F(i, j)$ to represent the elements of the first digits matrix of DCT coefficients in the position of ith row and jth column, then the transition probability matrix for F along the horizontal direction are given by:

$$p\{F(i,j+1) = v \mid F(i,j) = u\} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n-1}\delta(F(i,j) = u, F(i,j+1) = v)}{\sum_{i=1}^{m}\sum_{j=1}^{n-1}\delta(F(i,j) = u)}. \quad (3)$$

Where m and n are the number of rows and columns respectively, and

$$\delta(A = u) = \begin{cases} 1, if\ A = u \\ 0, otherwise \end{cases}, \quad (4)$$

$$\delta(A = u, B = v) = \begin{cases} 1, if\ A = u\ and\ B = v \\ 0, \qquad otherwise \end{cases}. \quad (5)$$

The transition probability matrix along other directions such as vertical, main diagonal and minor diagonal can be calculated in a similar way. For simplicity, we use only the horizontal direction in this paper. According to our experiment, if other three directions are also included, the dimension of the feature vector will be fourfold without an obvious improvement of the detection accuracy.

For a finite-state Markov chain, there exists a stationary distribution π satisfying the following equations [9]:

$$\pi = \pi p. \quad (6)$$

Where p is the transition probability matrix and π is a vector with non-negative elements, which sum up to 1. What's more, if the Markov chain is irreducible and ergodic, then π is unique. According to the random process theory, if a Markov chain having state space S and transition probability matrix p, the stationary distribution π is unique when:

$$\lim_{n\to\infty} p^n(x, y) = \pi(y), \ y \in S. \quad (7)$$

In this case, the unique π can be calculated by solving the following equations:

$$\begin{cases} \pi = \pi p_r \\ \sum_j \pi_j = 1 \end{cases} \quad (8)$$

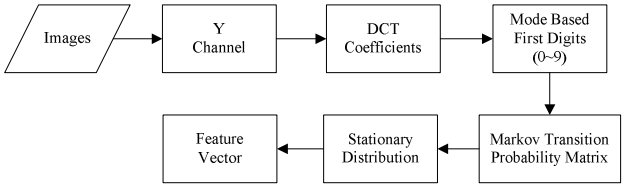

Figure 2. Block diagram of the proposed feature generation algorithm

Generally speaking, the Markov chain of the first digits of DCT coefficients is irreducible and ergodic, but as suggested in [9], to make the proposed method more general, the transition probability matrix can be refined by adding a positive perturbation and then scaled each row of the matrix to make its L1 norm to be one. Then in this manner, the stationary distribution can be ensured to be unique.

The dimension of stationary distribution π here is 10 as there are 10 finite-states in the proposed Markov model. After π is calculated, it will be used as features for double compression detection in the proposed method.

B. Propose Feature Generation Algorithm

The block diagram of the proposed feature generation algorithm is shown in Fig.2.

For a given image, the DCT coefficients of the Y Channel are extracted first. Then the first digits of DCT coefficients are calculated from individual AC mode as stated in work [4]. The digits of each individual mode are then used to generate the Markov transition probability matrix and the stationary distribution π of the Markov model is calculated which are finally fed to the classifier as features. Note that only Y Channel is used in our experiment. If Cb and Cr channels are also included in the experiment, the dimension of features will be tripled without greatly improving the performance, this is because down-sampling is usually applied to Cb and Cr channels which will reduce the correlations between adjacent DCT coefficients. What's more, quantization steps are usually larger for the Cb and Cr channels, which will also cause information loss.

As for the calculation of the first digits of DCT coefficients, small alternation has been made compared to [3-4]. To retain as much information as possible, those DCT coefficients whose value are zero are also included when calculating the Markov transition probability matrix for the reason that the number of zero coefficients varies from different quality factors and it can be treated as an intrinsic feature for double compression detection. To make a comparison with [4], only the first 20 AC modes are used exactly as suggested by [4], so in total, a feature vector of 10×20=200 elements is obtained for each given image.

IV. EXPERIMENT AND DISCUSSION

To testify the effectiveness of the proposed method, we select a database containing 500 TIF images which are taken by Kodak DC290 with the resolution of 720×480. These images are taken in different light conditions and have a variety of contents, including portraits, natural scenery, architecture, etc. Some of the image samples used in the experiment is shown in Fig.3.

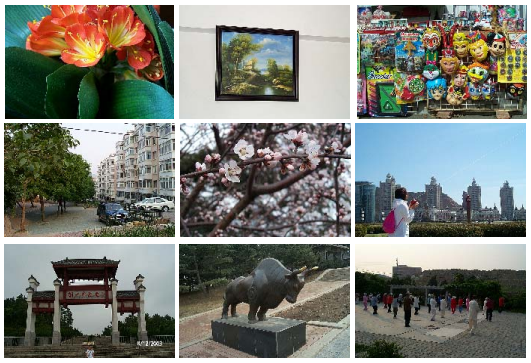

Figure 3. Some of image samples used in our experiment

First, these images are JPEG compressed with quality factor Q_1=50, 55, 60 … 95 respectively, generating 500×10=5000 single compressed images with different qualities. Then, for each set of the single compressed images, they are recompressed with quality factor Q_2=50, 55, 60…95 respectively to generate 5000 double compressed images. In this way, 100 groups of images are obtained, with each group including 500 single compressed images with quality factor Q_1 and 500 double compressed images with quality factor Q_1 followed by Q_2.

In the experiment, we use Support Vector Machine (SVM) [10] to classify the double compressed images from the single compressed ones. For each group of images, we randomly choose 3/5 of the images as ground truth for training and the remaining for testing. To ensure the effectiveness and stability of the proposed method, the experiments are repeated 20 times with the training sets and testing sets chosen randomly. The results shown in Table 1 are the average detection results of the 20 experiments, with the accuracy rounded to its nearest integers.

The detection result of work [4] is shown in Table 2 as a comparison with our proposed method. The same database and experimental environment are used and the detection accuracies are averaged over 20 random experiments.

From the detection results we can see that both work [4] and our proposed method work well when $Q_2>Q_1$, with an accuracy of nearly 100%. But for the situation of $Q_2<Q_1$, our proposed method outperforms that of work [4] in most of the cases, especially when Q_1=95 and Q_2=50 and 55, in which cases work [4] fails to detect the double compression while our proposed method can still reach a relatively high detection accuracy. This is mainly because when $Q_2<Q_1$, the double compression artifacts are not as obvious as when $Q_2>Q_1$. But by calculating the Markov transition probability matrix of the mode base first digits of DCT coefficients, the artifacts can be magnified by considering the correlations between the neighbor DCT coefficients, thus resulting in higher detection accuracy.

V. CONCLUSION

In this paper, a set of effective features are proposed for double compression detection. By analyzing the distributions of the first digits of DCT coefficients for both single and double compression, we propose to utilize the Markov

transition probability matrix of the mode based first digits of DCT coefficients to characterize the double compression artifacts and use the stationary distribution of the Markov chain as features. Experiment results show that our proposed method outperforms that of work [4] in most of the cases especially when $Q_2<Q_1$. When Q_1 is as high as 95 and Q_2 equals to 50 and 55, in which cases most of the previous works have reported to fail the detection, the proposed method can still reach a relatively high classification accuracy, which shows that the second order statistical model can really improve the performance.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 60971095) and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] A. C. Popescu, "Statistical tools for digital image forensics," Ph.D. dissertation. Department of Computer Science, Dartmouth College, Hanover, NH, 2005.

[2] Babak Mahdian, Stanislav Saic, "Detecting double compressed JPEG images," Crime Detection and Prevention (ICDP 2009), 3rd International Conference on Digital Object Identifier, 2009, pp.1-6.

[3] D. Fu, Y.Q. Shi, and W. Su, "A generalized Benford's law for JPEG coefficients and its applications in image forensics," Proc. of SPIE Conference on Electronic Imaging, Security and Watermarking of Multimedia Contents, San Jose, USA, January 2007.

[4] Bin Li, Yun Q. Shi, Jiwu Huang, "Detecting doubly compressed JPEG images by using mode based first digit features," IEEE International Workshop on Multimedia Signal Processing, 2008, pp.730-735.

[5] Chunhua Chen, Yun Q. Shi, and Wei Su, "A machine learning based scheme for double jpeg compression detection," International Conference on Pattern Recognition, 2008, pp. 1–4.

[6] Yi-Lei Chen, Chiou-Ting Hsu, "Detecting doubly compressed images based on quantization noise model and image restoration," IEEE International Workshop on Multimedia Signal Processing, Rio De Janeiro, 2009, pp.1-6.

[7] X. Feng and G. Do-err, "JPEG re-compression detection," Proceedings of SPIE, 2010, vol. 7541, pp. 75410J-1--75410J-12.

[8] F. Benford, "The law of anomalous numbers," Proceedings of the American Philosophical Society, 1938, vol. 78, pp. 551–572.

[9] Wei Wang, Jing Dong and Tieniu Tan, "Image tampering detection based on stationary distribution of Markov chain," 17th IEEE International Conference on Image Processing, 2010, pp.2101-2104.

[10] C.C. Chang, C. J. Lin. LIBSVM: A Library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

TABLE I. DETECTION RESULTS OF THE PROPOSED METHOD (BY %)

Q_1 \ Q_2	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**
50	—	99	100	100	100	100	100	100	100	100
55	98	—	99	100	100	100	100	100	100	100
60	99	99	—	99	100	100	100	100	100	100
65	100	100	99	—	100	100	100	100	100	100
70	99	100	100	100	—	100	100	100	100	100
75	98	99	99	100	100	—	100	100	100	100
80	99	99	100	100	100	100	—	100	100	100
85	99	99	99	99	100	100	100	—	100	100
90	98	99	99	99	99	99	100	100	—	100
95	84	89	95	97	96	98	99	99	99	—

TABLE II. DETECTION RESULTS OF WORK [4] (BY %)

Q_1 \ Q_2	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**
50	—	99	100	100	100	100	100	100	100	100
55	98	—	99	100	100	100	100	100	100	100
60	99	99	—	100	100	100	100	100	100	100
65	99	99	99	—	100	100	100	100	100	100
70	99	99	99	99	—	100	100	100	100	100
75	93	98	99	99	99	—	100	100	100	100
80	98	98	99	99	99	99	—	100	100	100
85	96	96	98	99	99	99	99	—	100	100
90	90	95	97	96	98	99	99	99	—	100
95	50	50	74	83	78	88	93	98	99	—