Detection of double compression in MPEG-4 videos using refined features-based CNN

S.-H. Nam, W. Ahn, M.-J. Kwon, and I.-J. Yu

Double compression is accompanied by various types of video manipulation and its traces can be exploited to determine whether a video is a forgery. This Letter presents a convolutional neural network for detection of double compression in MPEG-4 videos. Through analysis of the intra-coding process, we utilize two refined features for capturing the subtle artifacts caused by double compression. The discrete cosine transform (DCT) histogram feature effectively detects the change of statistical characteristics in DCT coefficients and the parameter-based feature is used as auxiliary information to help the network learn double compression artifacts. When compared with state-of-the-art networks and forensic method, the results show that the proposed approach achieves a higher performance.

Introduction: MPEG-4 is a typical lossy compression standard and has been widely adopted in surveillance systems and various devices. Due to storage issues, raw videos are single-compressed before distribution and when videos have been tampered with by an editing tool, a second compression occurs during the storage process. Therefore, video tampering accompanies double compression and artifacts caused by double compression can be utilized to trace the occurrence of such tampering [1]. Forensics is a technique that verifies the integrity of the content, and researchers presented handcrafted feature-based and convolutional neural network (CNN)-based approaches for capturing low-level features. Jiang et al. [2] presented a feature-based method using the Markov feature in discrete cosine transform (DCT) domain. Bayer and Stamm proposed a constrained convolutional neural network (CNN) that could successfully capture subtle artifacts of manipulations. Boroumand et al. [3] proposed a deep residual network for steganalysis that could suppress the content and extract noise-like stego signals.

However, the related work [2] is unsuitable for practical applications because it requires a classifier for each pair of compression parameters used in first and second compression. The CNNs in [3][4] show high performance for learning forensic features but show a slight reduction in performance in the DCT domain where MPEG-4 compression traces remain. To overcome the aforementioned drawbacks, this Letter proposes the refined features suitable for the given forensic task and CNN architecture that can effectively learn the features.

Analysis of intra-coding: Intra-coding is used to reduce spatial redundancy. A frame to which intra-coding has been applied is called an intra-coded frame (I-frame), and is placed at the first position for all group of pictures (GOP). As depicted in [2][3], intra-coding in MPEG-4 part 2 divides a frame into 8 x 8 blocks and then quantizes the DCT coefficients of each block with the default quantizer scale value for all group of pictures (GOP). As depicted in [2][5], intra-coding is used to reduce spatial redundancy because it requires a classifier for each pair of compression parameters.

Proposed approach: The whole procedure of the proposed approach consists of four steps as illustrated in Fig. 1.

Step 1) Input acquisition: Let video v with resolution W x H be represented by v = [v0, v1, ..., vL], where vt denotes a t-th decompressed frame in RGB format. The proposed approach uses the decompressed I-frame indicated by i = (vt) as input. Here, t denotes the index of the I-frame. The pre-processed input is defined as I = \{I(j) \} j \in \{1, 2, ..., L\}, where Y(j) denotes the Y-channel of the input after RGB-to-YUV conversion.

Step 2) DCT histogram feature generation: The DCT coefficients of each 8 x 8 block of I are computed. For frequency-specific analysis, the obtained coefficients are reshaped into \(D = \sum_{i=0}^{L-1} \sum_{j=0}^{W-1} D(i,j)\) with the same frequency for each channel. \(D_i\), which indicates each channel of \(D\), is the output of a 2D convolution operation with stride 8 between \(I(i,j)\) and \(H_c\), where \(H_c\) is a 2D basis for frequency c, \(c \in \{0, 1, 2\}\). For each channel, a cumulative histogram is calculated based on \(D_i\) and the boundary value \(b\) of the bin, where \(b \in \{0, 1, 2\}\). \(B_{c,b}\) is the b-th bin in a cumulative histogram for \(c\) and it means the average number of values in \(D_i\) that are greater than \(b\). \(B_{c,b}\) is computed as follows: \(B_{c,b} = \frac{1}{W \times H} \sum_{i=0}^{W-1} \sum_{j=0}^{H-1} T(D(i,j), -b)\), where \(T(\cdot)\) indicates the threshold function that turns positive and negative numbers into 1 and 0, respectively. Computing \(B_{c,b}\) for all \(c\) and \(b\) completes the cumulative histogram and DCT histogram feature \(F_i\), can be generated by calculating the difference between adjacent bins in the cumulative histogram: \(F_i = \{f_{i,c,b} = B_{c,b+1} - B_{c,b}\},\) where the size of \(F_i\) is \(2W \times H \times 1\).

Step 3) Parameter-based feature generation: \(M_q\) is a quantization matrix of size 8 x 8 and \(Q_t\) denotes the quantizer scale value of the last compression. Since \(M_q\) and \(Q_t\) are essential for quantization and de-quantization on intra-coding, they include useful information when CNN learns double compression artifacts. Hence, we generate the parameter-based feature \(F_3\) based on \(M_q\) and \(Q_t\), and \(F_3\) can be computed as: \(F_3 = V(M_q) \times Q_t\), where \(V(\cdot)\) denotes a vectorization function and the size of \(F_3\) is \(64 \times 1\).

Step 4) CNN for refined features learning: We designed a CNN to detect double compression (CNNDC) using the refined features \(F_3\) and \(F_2\) as input. The generated \(F_3\) is fed into the network, which consists of three convolution blocks and a fully connected (FC) layers block. Each convolution block consists of 3 × 3 or 1 × 1 convolutional (Conv) layers, each of which is followed by batch normalization (BN), and a rectified linear unit (ReLU). The 3 × 3 Conv layer learns the relationship between neighboring elements and the 1 × 1 Conv layer learns the association between sequentially placed feature maps. The last layer of each block was applied with 2 × 2 max pooling (MaxPool) with stride 2 to reduce the dimensionality of the feature maps. After flattening the outputs of the last convolution block, the flattened feature was fed into FC layers block, which consists of FC layers, ReLU, and softmax. To induce CNNDC to learn auxiliary information, we concatenate generated \(F_3\), with the flattened feature and the activations of two FC layers. The output of the FC-2 layer was fed into a two-way softmax to predict the class label. We modeled the loss function as follows: \(L = L_c + L_w\), where \(L_c\) and \(L_w\) represent the cross-entropy loss and regularization term, respectively. \(L_c\) is computed as: \(L_c = -(1 - y) \log \left(\frac{e^{y h}}{e^0 + e^{y h}}\right)\).
Table 1: Detection results of MISLnet, SRNet, and the proposed CNNs

| Metric          | MISLnet | SRNet | CNNDC-x | CNNDC |
|-----------------|---------|-------|---------|-------|
| Accuracy (%)    | 90.0    | 91.0  | 91.4    | 91.8  |
| TPR (%)         | 75.4    | 79.7  | 84.3    | 91.8  |
| TNR (%)         | 70.2    | 81.1  | 84.3    | 91.4  |

Table 2: Detection accuracy (%) of MFM and proposed CNNDC

| Method                  | Single compression | Double compression |
|-------------------------|--------------------|--------------------|
| (3, 5)                  | (5, 3)             |
| MFM                     | 90.9               | 89.9               |
| CNNDC                   | 99.9               | 99.9               |

Conclusion: In this Letter, we propose refined features- and a CNN-based approach for detecting double compression in MPEG-4 videos. The proposed CNNDC improves performance by utilizing DCT histogram features to capture statistical changes in the DCT domain and parameter-based feature for providing auxiliary information on intra-coding. These refined features were selected through analyzing the intra-coding and each layer of CNNDC was arranged to learn the features efficiently. The experimental results demonstrated that CNNDC is superior to the state-of-the-art CNNs MISLnet and SRNet, which specialize in forensics and steganalysis, respectively. Moreover, our approach outperforms the conventional non-CNN-based forensic method in terms of accuracy and detection efficiency.

E-mail: shnam1520@gmail.com

References
1 Nam, S.H., Park, J., Kim, D., Yu, I.J., Kim, T.Y., and Lee, H.K.: ‘Two-stream network for detecting double compression of h. 264’ videos’, IEEE International Conf. on Image Processing, September 2019, pp. 111-115
2 Jiang, X., Wang, W., Sun, T., Shi, Y.Q., and Wang, S.: ‘Detection of double compression in MPEG-4 videos based on Markov statistics’, IEEE Signal Process. Lett., 2013, 20 (5), pp. 447-450
3 Bayar, B. and Stamm, M.C.: ‘Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection’, IEEE Trans. Inf. Forensics Sec., 2018, 13 (11), pp. 2691-2706
4 Boroumand, M., Chen, M., and Fridrich, J.: ‘Deep residual network for steganalysis of digital images’, IEEE Trans. Inf. Forensics Sec., 2019, 14 (5), pp. 1181-1193
5 Xvid Codec, [Online]. Available: http://www.xvid.org/
6 Stamm, M. C., Wu, M., and Liu, K.R.: ’Image forensics: An overview of the first decade’, 2013, IEEE Access, I, pp. 167-200
7 Ebrahimi, M., Bondh, L., Bonnett, T., Costanzo, A., Maggini, M., Tondi, B., and Tubaro, S.: ’Aligned and non-aligned double JPEG detection using convolutional neural networks’, J. Vis. Commun. Image Represent., 2017, 49, pp. 153-163
8 Park, J., Cho, D., Ahn, W., and Lee, H.K.: ‘Double jpeg detection in mixed jpeg quality factors using deep convolutional neural network’, European Conf. Computer Vision, September 2018, pp. 636-652
9 Montgomery, C. Xiph. org Video Test Media (derf’s collection), the xiph open source community, 1994. Online, https://media.xiph.org/video/derf/
10 Liu, J.Y, Song, R., Wu, C.H., Liu, T., Wang, H., and Xiao, C.C.L.: ‘MCLV: A streaming video quality assessment database’, J. Vis. Commun. Image Represent., 2015, 30, pp. 1-9
11 Ba, W., Nam, S.H., Yu, I.J., Kwon, M.J., and Lee, H.K.: ‘Dual-path convolutional neural network for classifying fine-grained manipulations in H. 264 videos’, Multimedia Tools and Applications, 2021, pp. 1-28