Recent advances in media generation techniques have made it easier for attackers to create forged images and videos. State-of-the-art methods enable the real-time creation of a forged version of a single video obtained from a social network. Although numerous methods have been developed for detecting forged images and videos, they are generally targeted at certain domains and quickly become obsolete as new kinds of attacks appear. The method introduced in this paper uses a capsule network to detect various kinds of spoofs, from replay attacks using printed images or recorded videos to computer-generated videos using deep convolutional neural networks. It extends the application of capsule networks beyond their original intention to the solving of inverse graphics problems.

Index Terms— computer-generated video, replay attack, forgery detection, capsule network

2. RELATED WORK

In this section, we group forgery detection approaches into replay attack detection and computer-generated image/video detection on the basis of the features used and their target. Note that some approaches are two-fold while others are applicable only to certain types of attacks. We also provide some basic information about capsule networks and the dynamic routing algorithm that made this kind of network practical.
2.1. Replay Attack Detection

Prior to the current deep learning era, LBP methods were the primary defense against replay attacks [7, 8]. The method introduced by Kim et al. [17], which is based on local patterns of the diffusion speed (local speed patterns), achieves higher accuracy than that of LBP-based methods. Now, with the introduction of deep learning, the ability to detect replay attacks has been greatly improved. The method of Yang et al. [18] uses a support vector machine to classify features extracted by a pre-trained convolutional neural network (CNN). That of Menotti et al. [19] uses a similar procedure but optimizes the filters in an available high-performance CNN architecture. The method of Alotaibi and Mahmood [20] uses nonlinear diffusion based on an additive operator splitting scheme in their own CNN. The recently introduced method of Ito et al. [21] leverages a pre-trained CNN and utilizes the whole image instead of only the extracted face region.

2.2. Computer-Generated Image/Video Detection

There are several state-of-the-art methods for detecting images or videos generated by computer using, for example, a deepfake technique for face swapping [6], the Face2Face method for facial reenactment [1], or the deep video portraits technique [2] for the purpose of forgery. Fridrich and Kodovsky [10] proposed a hand-crafted-feature noise-based approach for steganalysis that can also be used for forgery detection. Cozzolino et al. [22] implemented a CNN version of this approach. Raghavendra et al. [23] described the special case of fine-tuning two available CNNs while Rossler et al. [11] used only one CNN. Bayar and Stamm [24], Rahmouni et al. [25], Afchar et al. [13], Quan et al. [26], and Li et al. [9] proposed their own networks. Li et al.’s network [9], for example, is video based and uses temporal information to detect eye blinking. We used a hybrid approach [12] incorporating part of a pre-trained VGG (Visual Geometry Group)-19 network [27] and a proposed CNN. Zhou et al. [28] proposed a two-stream network.

2.3. Capsule Networks

Hinton et al. [14] addressed the limitations of CNNs applied to inverse graphics tasks and laid the foundation for a more robust “capsule” architecture in 2011. However, this complex architecture could not be effectively implemented at the time due to the lack of an efficient algorithm and the limitations of computer hardware. Instead, easy-to-design easy-to-train CNNs became widely used. Now, with the introduction of the dynamic routing algorithm [15] and the expectation-maximization routing algorithm [16], capsule networks have been implemented with remarkable initial results. Two recent studies demonstrated that, with the agreement between capsules calculated by the dynamic routing algorithm, the hierarchical pose relationships between object parts can be well described. This has improved the accuracy of vision tasks. Application of a capsule network to the forensics task, the focus of this paper, is a challenging problem. However, the agreement between capsules achieved by using the dynamic routing algorithm could boost detection performance on complex and nearly flawless forged images and videos.

3. CAPSULE-FORENSICS

3.1. Overview

The proposed method (Fig. 1) works for both images and videos. For video input, the video is split into frames in the pre-processing phase. The classification results (posterior probabilities) are then acquired from the frames. The probabilities are averaged in the post-processing phase to get the final result. The remaining parts are constructed the same way as when the input is an image.

In the pre-processing phase, faces are detected and scaled to 128 × 128. Like we did in our previous work [12], we use part of the VGG-19 network [27] to extract the latent features, which are the inputs to the capsule network. Unlike we did in our previous work, we take the output of the third maxpooling layer instead of three outputs before the ReLU layers. We do this because we need to reduce the size of the inputs to the capsule network.

3.2. Capsule Design

The proposed network consists of three primary capsules and two output capsules, one for real and one for fake images (Fig. 2). The latent features extracted by part of the VGG-19 network [27] are the inputs, which are distributed to the three primary capsules (Fig. 3). As in our previous work [12], statistical pooling, which is important for forgery detection, is used. The outputs of the three capsules (u_{ij}) are dynamically
culated using equation 3, in which the output capsule  \( v \) is the dimension of the output capsule \( v_j \).}

\[ v_j = \text{squash}(s_j) = \frac{||s_j||^2}{1+||s_j||^2} s_j \]  
\hspace{1cm} (1)

Unlike Sabour et al.'s work [15], we use the cross-entropy loss function:

\[ L = - (y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \]  
\hspace{1cm} (2)

where \( y \) is the ground truth label and \( \hat{y} \) is the predicted label calculated using equation 3, in which \( m \) is the dimension of the output capsule \( v_j \).

\[ \hat{y} = \frac{1}{m} \sum_i \text{softmax} \left( \frac{v_i}{v_j} \right) \]  
\hspace{1cm} (3)

The use of equation 3 instead of simply using the length of the output capsules [15] promotes separation between the two output capsules on each dimension.

4. EVALUATION

To evaluate the advantage of using random noise, we tested the proposed method with and without using random noise (Capsule-Forensics-Noise and Capsule-Forensics, respectively). The random noise was generated from a normal distribution \( N(0, 0.01) \) and was used in the training phase only. Two iterations \( r = 2 \) were used in the dynamic routing algorithm. We used the half total error rate (HTER) \( \frac{FRR + FAR}{2} \) and accuracy \( \frac{TP + TN}{TP + TN + FP + FN} \) as metrics.

4.1. Replay Attack Detection

To determine the ability of the proposed method to detect replay attacks, we compared its performance with that of eight state-of-the-art detection methods on the well-known Idiap REPLAY-ATTACK dataset [7]. As shown in Table 1, the proposed method with random noise (Capsule-Forensics-Noise), as well as our previous method [12], had an HTER of zero.

<table>
<thead>
<tr>
<th>Method</th>
<th>HTER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chigovska et al. [7]</td>
<td>17.17</td>
</tr>
<tr>
<td>Pereira et al. [8]</td>
<td>08.51</td>
</tr>
<tr>
<td>Kim et al. [17]</td>
<td>12.50</td>
</tr>
<tr>
<td>Yang et al. [18]</td>
<td>02.30</td>
</tr>
<tr>
<td>Menotti et al. [19]</td>
<td>00.75</td>
</tr>
<tr>
<td>Alotabib et al. [20]</td>
<td>10.00</td>
</tr>
<tr>
<td>Ito et al. [21]</td>
<td>00.43</td>
</tr>
<tr>
<td>Nguyen et al. [12]</td>
<td><strong>00.00</strong></td>
</tr>
<tr>
<td>Capsule-Forensics-Noise</td>
<td><strong>00.00</strong></td>
</tr>
</tbody>
</table>

Fig. 4. Average results calculated by primary capsules and output capsules from real and fake images generated with Face2Face method [1]. Three primary capsules have significantly different reactions between real and fake inputs. Although their weights are also different, there is strong agreement in the output capsules.
4.2. Face Swapping Detection

We determined the ability of our proposed method to detect face swapping using a deepfake technique on the deepfake dataset proposed by Afchar et al. [13] at both the frame and video levels. As shown in Tables 2 and 3, our proposed method with random noise (Capsule-Forensics-Noise) had the highest accuracy in both cases.

Table 2. Accuracy of face swapping detection at frame level on deepfake dataset [13].

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meso-4 [13]</td>
<td>89.10</td>
</tr>
<tr>
<td>MesoInception-4 [13]</td>
<td>91.70</td>
</tr>
<tr>
<td>Nguyen et al. [12]</td>
<td>92.36</td>
</tr>
<tr>
<td>Capsule-Forensics</td>
<td>94.47</td>
</tr>
<tr>
<td>Capsule-Forensics-Noise</td>
<td>95.93</td>
</tr>
</tbody>
</table>

Table 3. Accuracy of face swapping detection at video level on deepfake dataset [13].

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meso-4 [13]</td>
<td>96.90</td>
</tr>
<tr>
<td>MesoInception-4 [13]</td>
<td>98.40</td>
</tr>
<tr>
<td>Capsule-Forensics</td>
<td>97.69</td>
</tr>
<tr>
<td>Capsule-Forensics-Noise</td>
<td>99.23</td>
</tr>
</tbody>
</table>

4.3. Facial Reenactment Detection

We determined the ability of our proposed method to detect facial reenactment on the FaceForensics dataset [11], which was created using the Face2Face method [1]. We strictly followed the authors’ guidelines for processing the data. As shown in Table 4, on average, the proposed method (with and without noise) had performance comparable to that of the best-performing state-of-the-art methods.

We also tested our method at the video level and compared its performance with that of Afchar et al.’s MesoNet facial video forgery detection network [13]. For our method, we used only the first ten frames instead of the entire video. As shown in Table 5, our method outperformed Afchar et al.’s network.

4.4. Fully Computer-Generated Image Detection

Finally, we compared the performance of our proposed method with that of state-of-the-art methods on computer-generated images (CGIs) and photographic images (PIs) on the dataset proposed by Rahmouni et al. [25]. Once again, as shown in Table 6, our method had the best performance and had perfect accuracy on full-size test images.

5. CONCLUSION

Our comprehensive experiments demonstrated the feasibility of building a general detection method that is effective for a wide range of forged image and video attacks. They also demonstrated that capsule networks can be used in domains other than computer vision. The proposed use of random noise in the training phase proved beneficial in most cases. Future work will mainly focus on evaluating the ability of the proposed method to resist adversarial machine attacks, especially on the proposed random noise at test time, and enhancing its ability. It will also focus on making the proposed method robust against mixed attacks, on detecting anomalies, and on raising this critical issue in the research community.

6. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051) and by JST CREST Grant Number JPMJCR18A6, Japan.
7. PREFERENCES


