A NOVEL MICRO-EXPRESSION RECOGNITION APPROACH USING ATTENTION-BASED MAGNIFICATION-ADAPTIVE NETWORKS

Mengting Wei, Wenming Zheng, Yuan Zong, Xingxun Jiang, Cheng Lu, Jiateng Liu

Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China.

ABSTRACT

Micro-Expression recognition (MER) is a challenging task due to the short duration and low intensity of Micro-Expressions. A popular method to tackle this is magnifying MEs so as to enlarge the expression intensity to make recognition easier. However, the single fixed magnification strategy, widely used in existing works of MER, is not appropriate for different subjects, because each subject has specific expression intensity corresponding to different MEs. To cope with this issue, we propose a novel Attention-based Magnification-Adaptive Network (AMAN) to learn adaptive magnification levels for the ME representation. The network consists of two modules: magnification attention (MA module) to adaptively focus on appropriate magnification levels of different MEs, and frame attention (FA module) to focus on discriminative aggregated frames in a ME video. Extensive experiments on three widely used databases manifest that our method yields state-of-art results compared with other methods.

Index Terms— Facial expression magnification, Attention mechanism, Micro-expression recognition, Transfer learning

1. INTRODUCTION

Micro-Expression (ME) is a temporary facial expression that appears unconsciously when people try to hide their true feelings [1]. For this characteristic, it is more likely to reflect people’s real emotions and thus contribute to potential applications in terms of clinical diagnosis and intelligence. Whereas, compared with conscious expressions, MEs are characterized by lower intensity and shorter duration, which makes accurate Micro-Expression Recognition (MER) a challenging task.

Targeting at the low intensity of MEs, some works attempt to enhance the intensity of facial movements to improve recognition performance. Lei et al. [2] adopted learning-based video motion magnification network to magnify facial movements of the apex frame. In this way, relations of different facial area would be more apparent so as to promote the performance of the posterior recognition task. Li et al. [3] used Eulerian Motion Magnification (EMM) [4] to magnify the apex frame from original MEs which could enlarge the difference among different ME categories so that the network is easier to learn discriminative features. In these works, when choosing the level of motion magnification, they just set one fixed level for all subjects and MEs. Nevertheless, since different subjects have different facial anatomical structure, even for the same stimulus, different subjects would present facial muscle movements at different intensities. In this case, a fixed magnification level is apparently not appropriate for every subject, and such problem would potentially decrease the performance of subsequent classification tasks, as shown in Fig. 1.

To address the issue aforementioned, in this paper, we propose a novel Attention-based Magnification Adaptive Network (AMAN) to learn appropriate magnification levels for MEs in a more flexible way. To alleviate the influence of individual differences and different pattern of MEs, we try to design a magnification-adaptive method. Instead of setting a fixed magnification level for all ME sequences, we use a set of different amplification factors (AFs) for the same image. Our network consists of two attention modules:

1. One module is for weighting different magnification levels, namely magnification attention (MA module). In MA module, we employ attention mechanism to assign different weights among a set of AFs so the network is able to adaptively learn discriminative magnification levels.

2. The other module is designed to weight different frames in a video, namely frame attention (FA module). In FA module, we introduce attention mechanism to assign different weights among a set of selected frames so the network is able to focus on more discriminative aggregated frames.

2. PROPOSED METHOD

2.1. Basic Idea

Videos in different ME datasets have various temporal length [5–7]. Considering this, for a given ME sequence, we firstly use a Time Interpolation Model (TIM) to reconstruct the video so that all the videos are consistent in length. Then we successively sample \( n \) number of frames from the beginning at a certain interval. Each selected frame is then magnified with a set of amplification factors (AFs). After that, for a
set of AFs corresponding to a single frame, we use a pre-trained CNN to extract their feature representations. Next, we use our magnification attention module (MA module) to aggregate these representations into a frame representation. Subsequently, for these frame representations in a video, we employ our frame attention module (FA module) to aggregate them into a final representation of the video. This final representation is eventually sent into a fully-connected layer for classification. The whole framework is shown in Fig. 2.

2.2. AMAN Model

2.2.1. MA module

Motion magnification techniques provide us convenience to capture small motions imperceptible to the naked eyes. In some cases, the motion is so small that the magnification results are inclined to noise, so MER methods using hand-designed filters to magnify micro-expressions may be optimal. Considering this, we employ deep convolutional networks to learn the filters directly and use transfer learning strategy to apply magnification characteristics learned in other databases [8] to micro-expression databases. As Wu et al. and Wadhwa et al. [4, 9] have defined on motion magnification, a single image in a consecutive video can be described as:

$$I(x, t) = f(x + \delta(x, t))$$  

(1)

where $\delta(x, t)$ is the motion field at position $x$ and time $t$, our goal is to magnify the motion so the magnified image $I_{mag}$ becomes

$$I_{mag}(x, t) = f(x + (1 + amp)\delta(x, t))$$  

(2)

where $amp$ is the magnification factor.

In our practice, for a single frame selected in a ME video, we firstly use the pre-trained network [8] to magnify it with a set of magnification factors ranging from 1 to $k$, denoted as $amp_1$ to $amp_k$. The network has three inputs: a magnification factor and two images between which there is a slight displacement (in different phases of the same video). This is formulated as:

$$I^i_{mag} = f(I_{on}, I_{sel}, amp_i)$$  

(3)

where $I_{on}$ is the onset frame, $I_{sel}$ is another frame in the same video we select for magnification, $I^i_{mag}$ is a generated image after magnification, $amp_i$ represents amplification factor. $f(\cdot)$ is the network.

For a magnified image with a specific amplification factor $amp_i$, we use the backbone network to extract its feature representation, denoted as:

$$F^T_i = g(I^i_{mag})$$  

(4)

where $g(\cdot)$ is the backbone network, and $F^T_i$ denotes the feature representation of a single magnified image.

Since we expect that magnification levels contributing more to recognition are weighted more importance, we assign different attention scores on these feature representations of magnified MEs from the same frame. Our MA module is composed of a linear fully-connected (FC) layer and a sigmoid activation function. Attention weights among feature representations are calculated by:

$$\alpha_i = \sigma(F^T_i q^0)$$  

(5)

where $q^0$ is the parameter of FC layer, and $\sigma(\cdot)$ denotes sigmoid function. The new weighted feature representation of a single magnified frame can be formulated as:

$$F^attn_i = \alpha_i F^T_i$$  

(6)

With those weighted representations of all magnified frames, feature representation of the original frame is aggregated by:

$$F_m = \frac{\sum_{i=1}^{k} F^attn_i}{\sum_{i=1}^{k} \alpha_i}$$  

(7)

2.2.2. FA module

Many methods have demonstrated that the apex frame (frame with largest movement intensity) contributes most to Micro-Expression recognition (MER) [10, 11]. Nevertheless, in a ME sequence, the intensity of the expression actually changes in a consecutive way, which means that more than one frames can be thought as ‘apex’. In view of this, we use multiple frames instead of the single apex frame in the video. Each frame is managed with the MA module to get its feature representation. We assign a set of attention scores on these frame feature representations once again so the network could adaptively focus on more discriminative frames. The FA module has the same structure as MA module. In this module, attention weights among these feature representations are formulated as:

$$\beta_j = \sigma(F^T_{m,j} p^0)$$  

(8)
where \( p^0 \) is the parameter of another FC layer, and \( F_{m_j} \) denotes the aggregated feature representation of the \( j \)-th image in a ME sequence. Then, the new weighted feature representation of original representation is denoted as:

\[
F_{att}^{m_j} = \beta_j F_{m_j}^T
\]  

(9)

We aggregate these new frame representations with their attached attention scores by:

\[
F_v = \frac{\sum_{j=1}^{n} F_{att}^{m_j}}{\sum_{j=1}^{n} \beta_j} \quad (10)
\]

\( F_v \) is the global representation of a ME video. The network takes it as the final representation for the following classification.

### 3. EXPERIMENTS

#### 3.1. Experimental Setup

##### 3.1.1. Datasets Preprocessing and Experiments Detail.

Three ME databases are used to evaluate our method: CASME II [5], SAMM [6] and SMIC-HS [7]. On the CASME II and the SAMM, we select categories with more than 10 samples, following the rules most methods [12, 13] adopt for five classification. On the SMIC-HS, we use all samples in this database. In a ME sequence, we first calculate 68 landmarks of the whole face utilizing [13] in the onset frame and then align the face area in line with these landmarks. All the images are resized to 224×224.

We use Resnet-18 as our backbone to extract shallow features of expressions. The dataset employed to pre-train is FER2017. When fine-tuning on MEs, we freeze the former residual blocks to extract some shallow features and release the last residual block for fine-tuning. To avoid subject dependence in the process, we adopt the leave-one-subject-out (LOSO) cross validation.

#### 3.1.2. Settings on Ceiling of Magnification Levels and Number of Frames.

Muscle movements in the facial area become more pronounced as the AFs increase during magnification. In order to ensure that most MEs are sufficiently enlarged, and meanwhile to avoid facial deformation caused by overlarge magnification levels, we conducted a series of comparison experiments on the CASME II. Related results are shown in Tab. 1.

As Tab. 1 presents, recognition accuracy firstly grows as the maximum level increases and then begins to decline when the level ceiling is over nine, indicating that from this level there may appear excessive deformation in the facial area of most MEs. Therefore, we set \( k = 9 \) as the ceiling of magnification level.

<table>
<thead>
<tr>
<th>k</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc(%)</td>
<td>69.85</td>
<td>72.34</td>
<td>75.40</td>
<td>69.23</td>
<td>66.41</td>
</tr>
</tbody>
</table>

To find appropriate number of frames selected in a video, in our experiments, we set \( n \) ranging from 1 to the maximum length(20) of the video on three datasets and plot the relation between recognition accuracy and number of frames, as shown in Fig. 3.

As can be seen from Fig. 3, recognition accuracy does vary with the change of \( n \) on three datasets, but not with large fluctuation, which further reflects the robustness of our method for selecting frames. We speculate that the robustness of this lies in our network’s ability to voluntarily weight more on the aggregated frames near the apex. We set \( n = 11 \) uniformly on three databases.
Table 2: The accuracy (ACC(%)) and F1-score (%) of different methods under the LOSO protocol on three datasets.

<table>
<thead>
<tr>
<th>MER method</th>
<th>CASME II</th>
<th>SAMM</th>
<th>SMIC-HS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>F1-score</td>
<td>Acc</td>
</tr>
<tr>
<td>LBP-SIP (2014) [14]</td>
<td>66.40</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DSSN (2019) [16]</td>
<td>70.78</td>
<td>72.97</td>
<td>57.35</td>
</tr>
<tr>
<td>TSCNN-I (2019) [17]</td>
<td>74.05</td>
<td>73.27</td>
<td>63.53</td>
</tr>
<tr>
<td>LGCconD (2020) [3]</td>
<td>62.14</td>
<td>60.00</td>
<td>35.29</td>
</tr>
<tr>
<td>RNMA (2020) [18]</td>
<td>65.90</td>
<td>53.90</td>
<td>48.50</td>
</tr>
<tr>
<td>GEME (2021) [19]</td>
<td>75.20</td>
<td>73.54</td>
<td>55.88</td>
</tr>
<tr>
<td>AMAN(ours) (2021)</td>
<td>75.40</td>
<td>71.25</td>
<td>68.85</td>
</tr>
</tbody>
</table>

*N/A - no results reported.

3.3. Experimental Results

Experimental results of our experiments and comparison to the state-of-the-art approaches are shown in Tab. 2.

From Tab. 2, we can see that our method exceeds most methods using handcraft features or high-level learned features. Specifically, AMAN yields comparable accuracy on the CASME II but gets further improvement on the other two datasets compared with GEME [19]. On the SAMM, our accuracy is much higher than TSCNN-I [17] by 5.32%. On the SMIC-HS, compared to TSCNN-I, our method improves the accuracy by 7.13%. AMAN achieves further improvements compared to those approaches using single-magnification techniques [2, 3] to magnify MEs, indicating that our method can automatically find the appropriate degree of magnification levels.

4. CONCLUSION

Our paper presents a novel attention-based network that can be adaptive to different magnification levels for Micro-Expression recognition (MER). The network consists of two modules: magnification attention module (MA module) for weighting magnification levels and frame attention module (FA module) for weighting aggregated frames in a Micro-Expression video. Extensive experimental results prove that our method yield superior performance compared to other state-of-the-art methods.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program under Grant 2018YFB1305203, and in part by the NSFC under grants U2003207 and 61902064.

6. REFERENCES

[1] W. J. Yan, Q. Wu, J. Liang, Y. H. Chen, and X. Fu, “How fast are the leaked facial expressions: The duration of


