DYNAMIC TRACKING ATTENTION MODEL FOR ACTION RECOGNITION

Chien-Yao Wang1, Chin-Chin Chiang1, Jian-Jiun Ding2, and Jia-Ching Wang1

1Department of Computer Science and Information Engineering, National Central University, Taiwan, R.O.C.
2Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

ABSTRACT

This paper proposes a dynamic tracking attention model (DTAM), which mainly comprises a motion attention mechanism, a convolutional neural network (CNN) and long short-term memory (LSTM), to recognize human action in a video sequence. In the motion attention mechanism, the local dynamic tracking is used to track moving objects in feature domain and global dynamic tracking corrects the motion in the spectral domain. The CNN is utilized to perform feature extraction, while the LSTM is applied to handle sequential information about actions that is extracted from videos. It effectively fetches information between consecutive frames in a video sequence and has an even higher recognition rate than does the CNN-LSTM. Combining the DTAM with the visual attention model, the proposed algorithm has a recognition rate that is 3.6% and 4.5% higher than that of the CNN-LSTMs with and without the visual attention model, respectively.

Index Terms— Action recognition, long short-term memory (LSTM), deep learning, attention model, convolutional neural network

1. INTRODUCTION

Image recognition is important in daily life. With the aid of feature extraction and machine learning techniques, many advanced algorithms about face recognition, hand gesture recognition, and iris recognition have been developed. Compared to other sub-topics about image recognition, action recognition is rather challenging since to recognize an action from a video, each frame must be analyzed and features should be extracted from multiple images.

In recent years, several neural networks with deeper and more complicated architectures have been developed. They are called deep neural networks (DNNs) [1–9]. Some DNNs are derived from neural networks, like the CNN [1, 2] and the LSTM [8, 9]. The CNN is frequently used to perform feature extraction and as a classifier in the final layer. The LSTM is a recurrent neural network (RNN) based model. It is effective in many pattern recognition problems.

The CNN is composed of mostly convolutional layers and pooling layers. It has been extensively used in image related technologies. For example, multiple CNN models have been used in face recognition [1] and hand gesture recognition [2]. Additionally, it can also be used in audio recognition [3]. For example, Abdel-Hamid et al. [4] performed a convolution operation on an audio spectrogram. He et al. [5] collected feature maps using multiscale filters. In our paper, the GoogLeNet [6] was adopted to generate features of images.

The LSTM has many applications, including scene labeling and image description. Johnson et al. [8] used the CNN-LSTM for video analysis. In [9], the LSTM was applied to gesture recognition. In [7], the GoogLeNet and the LSTM was adopted to generate image descriptions.

In action recognition, various methods are used to extract sequential information. Scovanner et al. [10] developed the 3D scale-invariant feature transform (SIFT). Moreover, the 3D histogram of the oriented gradient (HOG) [11], the speed up robust features (SURF) [12], and the local binary patterns (LBP) [13] were also applied to action recognition. Unlike the above hand-crafted low-level features, the attention model is used to extract information at times and places on which human attention is focused. Bottom-up and top-down saliency detection, which is based on the distribution of low-level features and semantics in images or video, plays an important role in the attention model. Trainable visual attention models that use RNNs were developed in [14]. V. Mnih et al. [15] presented a visual attention model using the CNN-LSTM.

Although the visual attention model is very effective in elucidating the meaning of images or videos, it often only considers the information in single frame. Moving objects capture human attention and should play a more important role in action recognition. This paper develops an attention model that is based on the information about motion extracted from videos. The proposed motion attention model, called the dynamic tracking attention model (DTAM), not only considers the information about motion but also perform dynamic tracking of objects in videos. Moreover, in addition to the DTAM, a visual attention model is adopted in the proposed system for action recognition.
2. SYSTEM OVERVIEW

The proposed system combines two models—the baseline visual attention [15] and the proposed DTAM. Figure 1 presents an overview of the proposed system. First, a CNN is applied to perform feature cuboid extraction on each frame of an action video. Then, the proposed DTAM uses the information about feature cuboids and the changes of object locations between two consecutive frames to generate a motion-based attention model. Thereafter, LSTMs generate a visual attention model and learn the DTAM. Finally, the results obtained using the visual attention model and the proposed DTAM are combined to yield action recognition results from the video of interest.

3. PROPOSED METHOD

In this section, the proposed motion attention model, DTAM, and its adjustment are discussed. Figure 2 presents the architecture of the proposed motion attention model, which is composed of the motion attention mechanism, CNN, and LSTM units. The CNN is applied to the video frames and the output feature cuboid is obtained. Then, the motion attention generated from optical flow images is used as weights. Finally, the LSTM is used to determine the action recognition result.

3.1. Dynamic Tracking Attention Model (DTAM)

3.1.1. Extraction of information about motion

There are numerous works about using optical flow maps in action recognition [16–18]. Wang et al. [19] studied the relationship between RGB images and optical flow images to elucidate the effectiveness of information extraction. Ng et al. [20] combined RGB and optical flow images in recognition. The proposed DTAM adopts optical flow [21] to extract information about motion. Figure 3 presents an example of optical flow images of basketball shooting.
3.1.2. Dynamic tracking of motion

The CNN yields state-of-the-art results in semantic image segmentation [22]. Feature cuboids $x$-$y$-$c$ are extracted by a CNN where $x$ is the horizontal coordinate, $y$ is the vertical coordinate, and $c$ is the category. They contain spatial and semantic information. The activation of feature maps in the $x$-$y$ plane can find the location of objects and the activation of feature maps in the $c$ domain can identify the objects in the image. The activation states of feature cuboids and the information of motion can be used to track different moving objects dynamically. In the proposed DTAM, the two dynamic tracking techniques are local dynamic tracking and global dynamic tracking.

Since the optical flow can extract information about motion at any location in an image, the proposed DTAM is able to find out the motion of objects. Accordingly, local dynamic tracking samples the optical flow along the trajectory of each feature map in the feature cuboid. Optical flow images [21] have three dimensions, which are the information along horizontal axis and the vertical axis and the magnitude of difference, respectively. Figure 4 presents the optical flow with local dynamic tracking for the video of the human category. The left-hand side of the figure gives an example of two sequential RGB images, the middle presents the three-dimensional optical flow images obtained from the two sequential RGB images, and the right-hand side presents the pseudo colorized optical flow image.

While local dynamic tracking can extract the true motion of objects in a video, it may not be able to determine the actual motion in the real world if the camera was moving when shooting a video. Global dynamic tracking can estimate the motion of the camera and correct the weights of the motion attention model. After local dynamic tracking is applied to optical flow images, global dynamic tracking is used to remove the motion of the camera. Figure 5 presents the optical flow with and without global dynamic tracking when a human is in the center of a video.

3.2. Adjustment of motion attention weight map

In the proposed DTAM, objects that move faster are given greater weights. Figure 6 compares the images that are obtained by dynamic tracking attention with and without adjustment. The middle of the figure presents the dot product of the weight map without adjustment and the original image. The right hand side of the figure presents the map with adjustment.

The dynamic-tracking generated optical flow image has values around 128 where the range of brightness is between 0 and 255. If the weights are used directly, then most pixels in the optical flow image will have weights around 128. Therefore, the difference from 128, which is the middle value, is used as the new weight for each pixel.

4. EXPERIMENTS

4.1. Experimental Setup

The UCF-11 dataset [23] contains 1599 videos with 11 classes of actions, which are bike-riding, diving, golfing, football-playing, high jumping, horse-riding, basketball-shooting, volleyball-playing, swinging, tennis-playing, and dog-walking. Training data were generated from 80% of the videos of each class, and the remaining videos provided the test data. GoogLeNet [6], which is trained using the ILSVRC14 dataset [24], was adopted in the CNN feature cuboid extraction stage of the attention models. In the experiments, frames in videos were resized to $224 \times 224$ and feature maps were chosen from the last convolutional layer with size $7 \times 7$ of GoogLeNet. The output motion attention maps of the proposed DTAM were pooled to size $7 \times 7$. 
4.2. Experimental Results

4.2.1. Comparison of motion attention mechanisms

The motion attention mechanism that is used in the proposed DTAM exhibits local dynamic tracking, global dynamic tracking, and weight adjustment. This subsection compares the action recognition rates of the motion attention mechanisms using optical flow and our method. Table 1 presents the relevant experimental results. The motion attention model using the original optical flow [21] achieves an recognition rate of 83.83%. By contrast, our motion attention mechanism achieves an recognition rate of 90.12%. The results reveal that our method improves the performance of the motion attention model which is based on the optical flow.

<table>
<thead>
<tr>
<th>Motion attention mechanism</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical flow</td>
<td>83.83%</td>
</tr>
<tr>
<td>Ours</td>
<td>90.12%</td>
</tr>
</tbody>
</table>

Table 2. Detailed evaluation of attention models in action recognition rate.

<table>
<thead>
<tr>
<th>Attention model</th>
<th>Visual</th>
<th>DTAM</th>
<th>Visual+DTAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Riding bike</td>
<td>100%</td>
<td>81.8%</td>
<td>95.5%</td>
</tr>
<tr>
<td>Diving</td>
<td>94.3%</td>
<td>97.1%</td>
<td>94.3%</td>
</tr>
<tr>
<td>Golfing</td>
<td>97%</td>
<td>97%</td>
<td>97%</td>
</tr>
<tr>
<td>Playing football</td>
<td>96.7%</td>
<td>96.7%</td>
<td>96.7%</td>
</tr>
<tr>
<td>High jumping</td>
<td>82.4%</td>
<td>97.1%</td>
<td>94.1%</td>
</tr>
<tr>
<td>Riding horse</td>
<td>96%</td>
<td>96%</td>
<td>98%</td>
</tr>
<tr>
<td>Basketball shooting</td>
<td>57.6%</td>
<td>72.7%</td>
<td>75.8%</td>
</tr>
<tr>
<td>Playing volleyball</td>
<td>96%</td>
<td>96%</td>
<td>96%</td>
</tr>
<tr>
<td>Swing</td>
<td>73.3%</td>
<td>83.3%</td>
<td>80%</td>
</tr>
<tr>
<td>Playing tennis</td>
<td>81.8%</td>
<td>72.7%</td>
<td>77.3%</td>
</tr>
<tr>
<td>Walking dog</td>
<td>90%</td>
<td>90%</td>
<td>90%</td>
</tr>
</tbody>
</table>

Table 3. Overall performance comparison.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>86.52%</td>
</tr>
<tr>
<td>Visual attention model</td>
<td>87.72%</td>
</tr>
<tr>
<td>DTAM</td>
<td>90.12%</td>
</tr>
<tr>
<td>Visual+DTAM (2:1)</td>
<td>88.92%</td>
</tr>
<tr>
<td>Visual+DTAM (1:1)</td>
<td>90.12%</td>
</tr>
<tr>
<td>Visual+DTAM (1:2)</td>
<td>91.02%</td>
</tr>
</tbody>
</table>

The proposed DTAM can obviously improve the performance of action recognition.

5. CONCLUSIONS

This paper proposed a deep-learning action recognition system that is based on a new motion attention mechanism, a CNN, and an LSTM. This system combines the visual attention model with the proposed DTAM. Our motion attention mechanism dynamically tracks moving objects based on information about motion that is extracted from the optical flow. In the experiments, the proposed DTAM is compared with the original flow motion attention model, the visual attention model, and a system without an attention model. The proposed DTAM improves the recognition rates by 6.29%, 2.4%, and 3.6%, respectively. Additionally, the combination of the proposed DTAM and the visual attention model has a recognition rate of 91.02%, which is 1% even higher than that of using only the DTAM.
6. REFERENCES


