DYNAMIC TRACKING ATTENTION MODEL FOR ACTION RECOGNITION

Chien-Yao Wang[1]*, Chin-Chin Chiang*[1]*, Jian-Jiun Ding*[2]*, and Jia-Ching Wang*[1]

[1]Department of Computer Science and Information Engineering,
National Central University, Taiwan, R.O.C.
[2]Graduate 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.

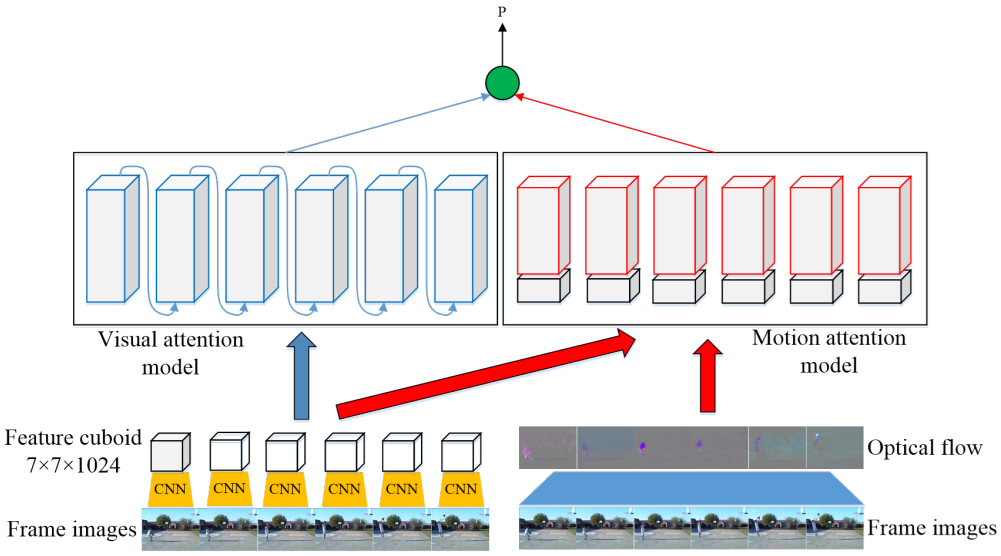


Fig. 1. Overview of proposed action recognition system.

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.

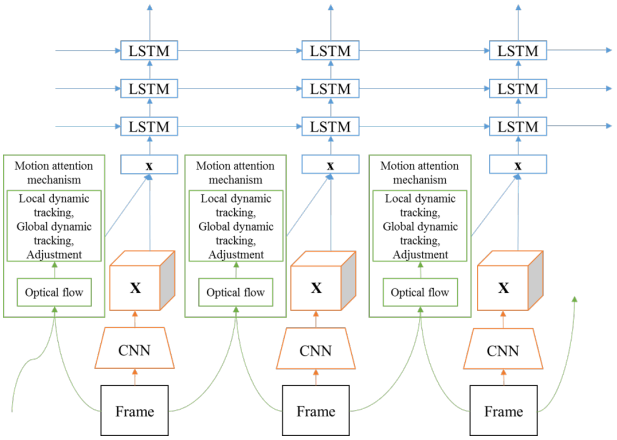
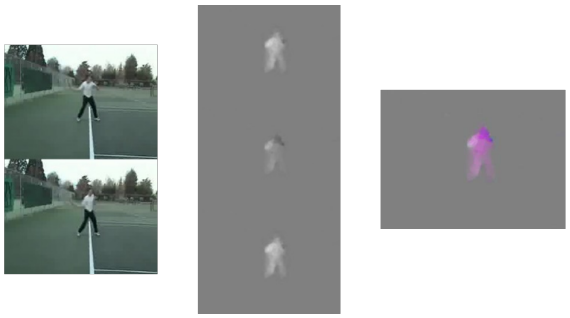


Fig. 2. Architecture of dynamic tracking attention model.

Fig. 3. Optical flow images.



(a) Sequential RGB images. (b) Optical flow images. (c) Pseudo colorized optical flow image.

Fig. 4. Optical flow extraction.

(a) RGB image. (b) Optical flow with only local dynamic tracking. (c) Optical flow with local and global dynamic tracking.

Fig. 5. Effect of dynamic tracking on optical flow.

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 word 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.

(a) Optical flow. (b) Dynamic tracking attention without adjustment. (c) Dynamic tracking attention with adjustment.

Fig. 6. Comparison between dynamic tracking attention with and without adjustment.

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 224×224 and feature maps were chosen from the last convolutional layer with size 7×7 of GoogLeNet. The output motion attention maps of the proposed DTAM were pooled to size 7×7.

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 1. Evaluation of motion attention mechanisms.

Motion attention	Recognition rate
Optical flow	83.83%
Ours	**90.12%**

4.2.2. Comparison between visual attention model and DTAM

Table 2 presents the results that were obtained for each class of actions using different attention models. The visual attention model used herein was proposed by Sharma *et al.* [25]. The combination of the proposed DTAM and the visual attention model is called the Visual+DTAM model.

Based on the results that were obtained using the visual attention model and the DTAM, the visual attention model provides a good recognition rate for riding a bike and playing tennis, which are the actions about the interaction between a human being and an object. For the actions of diving, high jumping, and swinging, the DTAM performs better. The visual attention model classified basketball shooting with only 57.6% accuracy, as it wrongly classified many videos as high jumping perhaps because the basketball is out of sight in some frames, and the movements of the players in these wrongly classified video include several jumps.

4.2.3. Overall comparison

Table 3 presents the recognition rates achieved in experiments using different attention models and the combinations of the visual and the proposed DTAM models (denoted by Visual+DTAM) with different weights. The single three-layered LSTM has a recognition rate of 86.52%, the visual attention model has a recognition rate of 87.52%, and the proposed DTAM achieves a recognition rate of 90.12%. The different ratios (2:1, 1:1, and 1:2) of visual attention model to the DTAM are examined, and the recognition rates are 88.92%, 90.12%, and 91.02%, respectively. It shows the proposed DTAM can obviously improve the performance of action recognition.

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

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

Table 3. Overall performance comparison.

Approach	Recognition rate
LSTM	86.52%
Visual attention model [25]	87.72%
DTAM	**90.12%**
Visual+DTAM (2:1)	**88.92%**
Visual+DTAM (1:1)	**90.12%**
Visual+DTAM (1:2)	**91.02%**

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 optical 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

[1] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, 2015.

[2] L. Pigou, S. Dieleman, P. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," *Workshop at the European Conference on Computer Vision*, 2014.

[3] S. Sukittanon, A. Surendran, J. Platt, and C. Burges, "Convolutional networks for speech detection," *Interspeec*, 2004.

[4] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language pprocessing*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *European Conference on Computer Vision*, 2014.

[6] C. Szegedy, Y. Jia W. Liu, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[8] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," *arXiv:1511.07571*, 2015.

[9] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[10] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *ACM International Conference on Multimedia*, 2007.

[11] A. Klaser, M. Marcin, and S. Cordelia, "A spatio-temporal descriptor based on 3d-gradients," *British Machine Vision Conference*, 2008.

[12] G. Willems, T. Tinne, and V. Luc, "An efficient dense and scale-invariant spatio-temporal interest point detector," *European Conference on Computer Vision*, 2008.

[13] B. Nair and V. Asari, "Regression based learning of human actions from video using HOF-LBP flow patterns," *IEEE International Conference on Systems*, 2013.

[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R.Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *arXiv:1502.03044*, 2015.

[15] V. Mnih, H. Nicolas, and G. Alex, "Recurrent models of visual attention," *Advances in Neural Information Processing Systems*, 2014.

[16] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[17] H. Wang and C. Schmid, "Action recognition with improved trajectories," *IEEE International Conference on Computer Vision*, 2013.

[18] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," *IEEE International Conference on Computer Vision*, 2007.

[19] P. Wang, Y. Cao, C. Shen, L. Liu, and H. Shen, "Temporal pyramid pooling based convolutional neural networks for action recognition," *arXiv:1503.01224*, 2015.

[20] J. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[21] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," *European Conference on Computer Vision*, 2004.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[23] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F. F. Li, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, 2015.

[25] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv: 1511.04119*, 2015.