# Dynamic Tracking Attention Model for Action Recognition

*Chien-Yao Wang[1], Chin-Chin Chiang[1], Jia-Ching Wang[1], and Jian-Jiun Ding[2]*
[1]Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan, R.O.C.
[2]Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.
National Taiwan University

## 1. Major Contribution

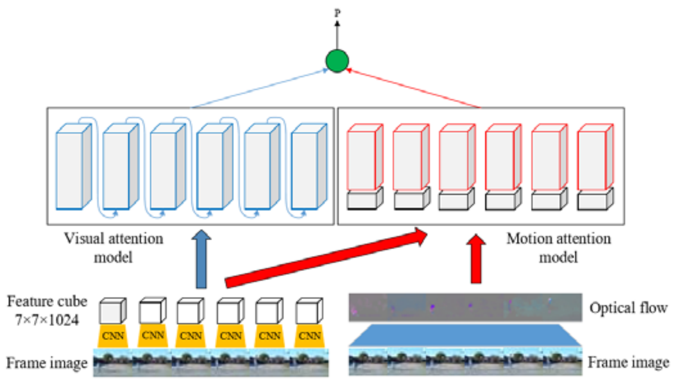
- ✓ The dynamic tracking attention model (DTAM), which comprises a convolutional neural network (CNN) and the long short-term memory (LSTM), is proposed.
- ✓ The proposed DTAM is to perform motion recognition from videos. It effectively fetches information between consecutive frames in a sequence .
- ✓ The recognition rate of the proposed algorithm is 3.6% and 4.5% higher than that of the CNN-LSTMs with and without the attention model, respectively

## 2. Existing Methods

- 3D scale-invariant feature transform (SIFT)
- 3D histogram of the oriented gradient (HOG )
- Speed up robust features (SURF)
- local binary patterns (LBP)
- RNN
- CNN-LSTM

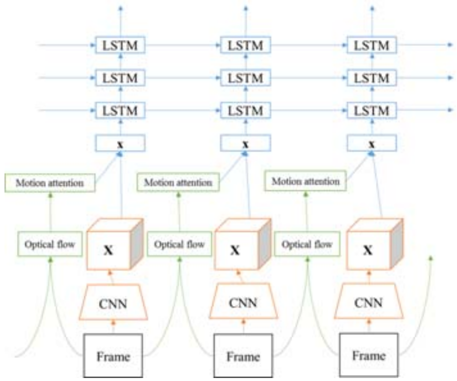
## 3. Proposed Method

Dynamic tracking attention model (DTAM),
It not only considers the information about motion but also perform dynamic tracking of objects in videos.




Overview of proposed action recognition system

## 4. Applied Techniques




Motion attention model

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

Optical flow image

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

Dynamic tracking of the optical flow

Global dynamic tracking can estimate the motion of the camera and correct the weights of the motion attention model.

**Adjustment of DTAM**

$$flow_{t,d,i} = \left|\mathbf{I}_{t,d,m}^{flow} - 128\right|, \quad d = 1,2,\ldots,D$$

$$a_{t,m}^{flow} = \sum_{d=1}^{D} \frac{flow_{t,d,m} - \min(flow_{t,d})}{D \times (\max(flow_{t,d}) - \min(flow_{t,d}))}$$

$$\mathbf{x}_t^{flow} = \sum_{m=1}^{K^2} a_{t,m}^{flow} \mathbf{X}_{t,m}$$

where $\mathbf{X}_{t,m}$ is the feature cuboid.

## 5. Simulation Results

The UCF-11 dataset 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.

Comparison of DTAM with/without adjustment

| Motion attention | Recognition rate |
|---|---|
| optical flow | 83.83% |
| DTAM | **90.12%** |

Classifications by the proposed attention model

| Class | Recognition result | | |
|---|---|---|---|
| | Visual | Motion | Overall |
| 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%** |

Results of action recognition obtained using the hybrid attention model with different weights.

| Architecture | Recognition rate |
|---|---|
| LSTM | 86.52% |
| Visual attention model [28] | 87.72% |
| **Proposed DTAM** | **90.12%** |
| **Overall (2:1)** | **88.92%** |
| **Overall (1:1)** | **90.12%** |
| **Overall (1:2)** | **91.02%** |

## 6. Conclusion

This paper proposed a deep-learning action recognition system that is based on the CNN and the LSTM. It dynamically tracks moving objects based on information about motion that is extracted from the optical flow.