Spatio-temporal attention mechanisms for Activities of Daily Living

SRIJAN DAS
RESEARCH SCHOLAR
STARS TEAM, INRIA
Outline

- Introduction
- Related Work
- Contributions
- Experimental Evaluation
- Conclusion
Introduction
What does activity recognition involve?
Detection: are there people?
Action recognition: what are they doing?
This is a nursing home. One nurse is crouching to comfort a fallen patient while another runs to get help.

"AI-complete": full semantic understanding necessary for success.
Human Activity Recognition

There are various types/levels of activities

- The ultimate goal is to make computers recognize all of them reliably.
Activity Classification

Categorization of segmented videos
- Input = a video segment containing only one activity

- Punching
- Hugging
- Waving
- Shaking hands
Why is activity recognition important?

User videos

- YouTube
- Twitch
- Facebook

~300 hours of videos per minute
- Video indexing and retrieval

Monitoring cameras

- Streaming videos 24/7
- Surveillance
- Patient/elderly monitoring

Media

- Content analysis, experience enrichment
- Recommendation systems
- Advertising
- Sports analytics

Wearables/robots

- Streaming videos to be analyzed in real-time
- Lifelogging
- Robot operations and actions
Why is activity recognition important?

User videos
- YouTube
- Twitch
- Facebook

Monitoring cameras
- Streaming videos 24/7
  - Surveillance
  - Patient/elderly monitoring

Media
- Content analysis, experience enrichment
  - Recommendation systems
  - Advertising
  - Sports analytics

Wearables/robots
- Streaming videos to be analyzed in real-time
  - Lifelogging
  - Robot operations and actions

~300 hours of videos per minute
- Video indexing and retrieval
Web videos vs Activities of Daily Living (ADL)
Challenges in ADL

- Same background
- High intra-class variation
Challenges in ADL

- Typing a keyboard
- Reading

- Same background
- Actions with subtle motion
Challenges in ADL

- Wear a shoe
- Taking off a shoe

- Same background
- Actions with similar appearance
Objectives

• Modelling Actions and dealing with the temporal domain.

• Focus on Activities of Daily Living (ADL), in particular
  
  • Fine-grained actions (similar appearance & subtle motion actions, temporally opposite actions)
    
    • In real-world settings (different camera views, low subject resolution, presence of occlusions)
Related Work
Traditional video classification

Video data

 Representation

Semantic labels

n*m-D data (e.g., n = 320*240, m = 1000 frames)

k-D vector (e.g., 1000)

s labels (e.g., s = 6)
Video classification: Deep learning

Video data
Convolutional Neural Networks (CNN)
Semantic labels

Learn millions of internal parameters from Data
• Representations optimized for the training data
Video classification with deep learning

Input: a fixed number of frames, Output: a class label

**Two-stream CNNs**
- 1 frame RGB + 10 frames of optical flow

[Carreira and Zisserman, 2017]

**Sequential models RNNs**
- model ‘sequences’ of per-frame CNN representations (RGB/3D Poses)

[J. Ng et al., 2015]

**3-D XYT CNNs**
- 15~99 frames (RGB + Flow)
- Facebook C3D, Google I3D
Video Classification with 3D CNNs

Facebook C3D [Tran et al., 2015]
- Spatio-temporal filters for short video segments (e.g., 15 frames) – coupling space and time

Google I3D [Careira et al., 2017]
- Extended by inflation from Spatial domain
Limitations of 3D CNNs

- Rigid spatio-temporal kernels limiting them to capture subtle motion
- No specific operations to help disambiguate similarity in actions.
- 3D (XYT) CNNs are not view-adaptive.
Do we need them all?

- The girl is drinking water from a bottle

- Do you really need the whole video to infer that?
Do we need them all?

- Isn’t this enough for an inference?
Do we need them all?

- Can you recognize this action?
Do we need them all?

- Now probably you can answer!!!

- So, temporal relationship is important.

The answer is yes but we need to have an attention mechanism to provide weightage to them!
Attention mechanism

Hard attention

- Hard decisions while choosing parts of the input data.
- Cannot be learned easily through gradient decent (no global optimization).

Soft attention

- Weighs the RoI dynamically, taking the entire input into account.
- Can be trained end-to-end (global optimization).
Example videos of soft-attention in the state-of-the-art

Sharma et al., (ICLRW 2015)
Example videos of soft-attention in the state-of-the-art

Xiaong et al., (AAAI 2018)
Disadvantages of the existing attention mechanisms

- Existing attention mechanisms are based on RNN classification models.

- Performance is lower due to lack of spatio-temporal coupling.

- Lacks the use of highly informative 3D pose information. These poses are robust to illumination, view and describes the human dynamics.
Contributions

• SPATIAL ATTENTION
• SPATIO-TEMPORAL ATTENTION
• EXTRA LAYER OF TEMPORAL ATTENTION FOR COMPLEX ACTIVITIES
Proposed Attention Mechanism

1. **Spatial Attention (WACV 2019)**
   - **Objective**: To focus on the pertinent human body parts involved in an action
   - **Method**: 3D ConvNet (RGB input) + RNN (to weight the body parts from the evolution of skeleton sequences).

2. **Spatio-temporal Attention (ICCV 2019)**
   - **Objective**: To incorporate spatial and temporal attention in the same model
   - **Input**: RGB + 3D skeleton; **Method**: 3D ConvNet (RGB input) + 1 RNN (to compute spatial attention mask and temporal attention mask separately)

3. **Extra layer of Temporal Attention for Complex Activities (WACV 2020)**
   - **Objective**: To focus on the pertinent temporal segments in a video
   - **Method**: 3D ConvNet (RGB input) + G RNNs + (G+1) RNNs (to weight the temporal segments from the corresponding poses at a granularity G).

**Input**: RGB -> classification Network
3D skeleton -> attention network
Spatial attention model (WACV 2019)
Spatial attention model

An end-to-end Spatial attention network for human action recognition.

- A method to classify actions from RGB-D videos based on spatio-temporal representation of human body parts.

- A novel RNN attention model. The attention model uses articulated poses to compute the importance of human body parts.

- A joint strategy to tightly couple 3D ConvNet classification networks and the RNN attention model using a regularized cross-entropy loss.
Demo of Spatial attention model

Raw Video - RGB

Action Label - drinking
Spatio-temporal attention mechanism (ICCV 2019)
Spatio-temporal attention model

An end-to-end Spatial & temporal attention network for human action recognition.

- A method to classify actions from RGB-D videos based on spatio-temporal representation of video.

- Dissociate spatial and temporal attention mechanism (instead of coupling them) [architecture is based on the study of retinal ganglion cells in the primate visual system]
Extra layer of temporal attention for complex activities in ADL (WACV 2020)
Extra layer of temporal attention for complex activities in ADL

An end-to-end temporal Model for temporally complex human action recognition. This is done by

- splitting a video into several temporal segments at different levels of temporal granularity

- employing a two-level pose driven attention mechanism. First to manage the relative importance of the temporal segments within a video for a given granularity. Second to manage the relative importance of the various temporal granularities.
What is temporal granularity?

A video of person drinking is represented with coarse to fine granularity ($G_{max} = 4$)
What are temporal segments?

The video with temporal granularity $G = 3$ has 3 temporal segments and so on.

$S_{31}$  $S_{32}$  $S_{33}$
Illustration of visual result of attention scores ($TS - att$ & $TG - att$) on the sample video

$G = 2$

High attention score 76.3%

$G = 3$

$G = 4$
Experimental Evaluation
• **NTU RGB-D** dataset, one of the largest available human activity dataset

- 58,000 videos
- 60 actions
- 40 subjects
- 80 views
Dataset Description


- 1194 videos
- 10 actions
- 10 subjects
- 3 views
Comparison with the state-of-the-art

Results on **NTU RGB-D** with cross-subject (CS) and cross-view settings (accuracies in %)

<table>
<thead>
<tr>
<th>Methods</th>
<th>CS</th>
<th>CV</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>VA-LSTM (ICCV 2017)</td>
<td>79.4</td>
<td>87.6</td>
<td>83.5</td>
</tr>
<tr>
<td>Glimpse Cloud (CVPR, 2018)</td>
<td>86.6</td>
<td>93.2</td>
<td>89.9</td>
</tr>
<tr>
<td>PEM (CVPR, 2018)</td>
<td>91.7</td>
<td>95.2</td>
<td>93.4</td>
</tr>
<tr>
<td>Spatial Attention (WACV 2019)</td>
<td>93</td>
<td>95.4</td>
<td>94.2</td>
</tr>
<tr>
<td>Spatio-temporal Attention (ICCV 2019)</td>
<td>92.2</td>
<td>94.6</td>
<td>93.4</td>
</tr>
<tr>
<td>Temporal Model (P-I3D base) (WACV 2020)</td>
<td>93.9</td>
<td>96.1</td>
<td>95</td>
</tr>
</tbody>
</table>

Results on **N-UCLA** with cross-view settings (accuracies in %)

<table>
<thead>
<tr>
<th>Methods</th>
<th>$V_{1,2}^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NKTM (CVPR, 2015)</td>
<td>85.6</td>
</tr>
<tr>
<td>Ensemble TS-LSTM (ICCV, 2017)</td>
<td>89.2</td>
</tr>
<tr>
<td>Glimpse Cloud (CVPR, 2018)</td>
<td>90.1</td>
</tr>
<tr>
<td>HPM+TM (CVPR, 2016)</td>
<td>91.9</td>
</tr>
<tr>
<td>Spatial Attention (WACV 2019)</td>
<td>93.1</td>
</tr>
<tr>
<td>Spatio-temporal Attention (ICCV 2019)</td>
<td>92.4</td>
</tr>
<tr>
<td>Temporal Model (P-I3D base) (WACV 2020)</td>
<td>93.5</td>
</tr>
</tbody>
</table>
Towards Real-world Action Recognition

Real-world challenges
- spontaneous acting
- low camera awareness
- high camera framing
- multi-view setting
- composite activities
- activities with different objects

18 subjects
31 activity classes
16.1k videos
7 camera views
Experimental evaluation on Toyota Smarthome dataset

Results on **Smarthome** with cross-subject (CS) and cross-view settings (accuracies in %)

<table>
<thead>
<tr>
<th>Methods</th>
<th>CS</th>
<th>CV₁</th>
<th>CV₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT (CVPR, 2011)</td>
<td>41.9</td>
<td>20.9</td>
<td>23.7</td>
</tr>
<tr>
<td>LSTM on 3D joints (CVPR, 2015)</td>
<td>42.5</td>
<td>13.4</td>
<td>17.2</td>
</tr>
<tr>
<td>I3D (CVPR, 2017)</td>
<td>53.4</td>
<td>34.9</td>
<td>45.1</td>
</tr>
<tr>
<td>I3D+NL (CVPR, 2018)</td>
<td>53.6</td>
<td>34.3</td>
<td>43.9</td>
</tr>
<tr>
<td>Spatial Attention (WACV 2019)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Spatio-temporal Attention (ICCV 2019)</td>
<td>54.2</td>
<td>35.2</td>
<td>50.3</td>
</tr>
<tr>
<td>Temporal Model (I3D base) (WACV 2020)</td>
<td>59.0</td>
<td>37.4</td>
<td>55.6</td>
</tr>
</tbody>
</table>
Conclusion

- Proposed end-to-end attention models (spatial and temporal) to focus on pertinent RoI and key frames in a video.

- Validation of the proposed methods on publicly available datasets and a real-world dataset outperforming the state-of-the-art results on them.

- Future perspectives include –
  - Domain adaptation for video understanding
  - Going towards weakly supervised action recognition
Thank You
Questions???