Attentive Human Action Recognition

Minh Hoai Nguyen
minhhoai@cs.stonybrook.edu
Human Behavior Analysis

Action/Interaction/Activity
Human Behavior Analysis

1. Recognition (Past)
2. Prediction (Future)
3. Early Recognition (Present)
Why should we care about human behavior?
Applications of Human Behavior Analysis

Security
UK has > 1.8M CCTV cameras

Healthcare
US: 3.4% suffer major depression

Human-robot interaction
Is she initiating a handshake?

Network optimization
Human Behavior Analysis

Main problems:
- Recognition
- Early Recognition
- Prediction
- Human action
- Face behavior
- Visual Attention

Supporting problems:
- Human/Hand Detection
- Crowd counting

Machine Learning problems:
- Unlabeled samples
- Weak/noisy annotated samples
- Complementary samples

Interdisciplinary problems:
- Fundamental questions:
  - Human behavior
  - Human perception
  - Human learning
Human Behavior Analysis

Main problems:
- Recognition
- Early Recognition
- Prediction
- Human action
- Face behavior
- Visual Attention

Machine Learning problems:
- Unlabeled samples
- Weak/noisy annotated samples
- Complementary samples

Supporting problems
- Human/Hand Detection
- Crowd counting

Interdisciplinary problems
- Fundamental questions:
  - Human behavior
  - Human perception
  - Human learning

512 people
Talk Outline

• Pulling Actions of out Context

• Active Vision for Early Recognition
Goal: Recognizing Human Actions

• Human action recognition is challenging due to:
  • The subtlety of human actions
  • The complexity of video space

• It’s difficult to identify and attend to the relevant information
A Common Approach

• Collect positive and negative data and train a classifier
• For example, consider training a KISS classifier
A Common Approach

• Collect positive and negative data and train a classifier.
• For example, consider training a KISS classifier.

Positive Examples

SVM

Negative Examples

Problem
Context

• Various types of context
  – Background context, e.g., living room
  – Camera context, e.g., slow panning camera
  – Situation context, e.g., a dance

• Context often co-occurs with the actions
  – The learned classifier focuses on the contextual cues instead of the action content
  – The learned classifier does not generalize well
Our Goal

Positive Examples

Negative Examples

SVM

kiss + Context

kiss + Context

kiss + Context

kiss + Context
Technical Challenges

Positive Examples

Kiss + Context

Co-occurring

SVM

Kiss + Context

Negative Examples

Sit Up

Drive

Handshake

Co-occurring

Co-occurring

Co-occurring

Positive Examples

Kiss + Context
One Approach: Provide Detailed Annotation

- But it is difficult to obtain detailed annotation:
  - Laborious process => unscalable
  - Ambiguous
Our Idea

Action Context

Action sample

Difference

Context

Conjugate sample

Action
Our Idea
Our Idea

Action sample

Context

Conjugate sample

Difference

Similarity

Action

Context
How to obtain Conjugate Samples?

Pre-Action

Action Sample

Post-Action

Conjugate Sample

Action Sample
How to use conjugate samples **effectively**?

Possible approaches

1. **Subtract** conjugate sample from action sample
   - No direct pixel to pixel correspondence

2. Use conjugate samples as **negative** examples
   - Ignore the importance of context for recognition

3. Use conjugate samples as **positive** examples
   - Noisy labels!
Proposed Framework

• Separate Action & Context factors

- **Action is different**
  - $f: \text{action extractor}$
  - $g: \text{context extractor}$
  - $f(a)$
  - $g(a)$
  - $L_s$

- **Context is similar**
  - $f: \text{action extractor}$
  - $g: \text{context extractor}$
  - $f(c)$
  - $g(c)$
  - $L_d$

- Similarity measure
- Difference measure
Proposed Framework

• Separate Action & Context factors
• Combine Action & Context factors for classification
Proposed Framework

• Joint learning of:
  • Action extractor $f$
  • Context extractor $g$
  • Classifier $h$
For Prediction (Testing)
Our Framework Enables Detailed Analysis

- Understand the contribution of the action factors

- Understand the contribution of the context factors
Summary of Our Approach

- Separately model action & context
- Combine them for action recognition
- Enable fine-grained analysis (action/context channel)
- Conjugate samples are easy to collect

Positive Examples
- Sit Up
- Drive
- Drive

Negative Examples
- Sit Up
- Drive

Automatically

kiss + Context
Experiment 1

• ActionThread dataset [Hoai & Zisserman, ACCV14]
  • 3035 video clips
  • 13 action classes

• C3D Network and features [Tran et al., ICCV15]
  • 3D convolution network
  • Input: block of 16 RGB frames
  • Feature dimension: 4096
### Experiment 1 Results

<table>
<thead>
<tr>
<th>Action</th>
<th>NotUsed</th>
<th>AsNegative</th>
<th>AsPositive</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswerPhone</td>
<td>27.3</td>
<td>27.3</td>
<td>28.7</td>
<td>43.0</td>
</tr>
<tr>
<td>DriveCar</td>
<td>51.2</td>
<td>51.6</td>
<td>48.9</td>
<td>53.1</td>
</tr>
<tr>
<td>Eat</td>
<td>36.9</td>
<td>37.7</td>
<td>35.7</td>
<td>42.1</td>
</tr>
<tr>
<td>Fight</td>
<td>45.7</td>
<td>48.6</td>
<td>41.2</td>
<td>61.1</td>
</tr>
<tr>
<td>GetOutCar</td>
<td>29.7</td>
<td>31.4</td>
<td>30.1</td>
<td>27.2</td>
</tr>
<tr>
<td>ShakeHand</td>
<td>26.9</td>
<td>26.0</td>
<td>26.6</td>
<td>35.2</td>
</tr>
<tr>
<td>Hug</td>
<td>43.9</td>
<td>44.1</td>
<td>45.5</td>
<td>54.7</td>
</tr>
<tr>
<td>Kiss</td>
<td>67.4</td>
<td>66.4</td>
<td>67.0</td>
<td>72.6</td>
</tr>
<tr>
<td>Run</td>
<td>82.0</td>
<td>80.6</td>
<td>79.9</td>
<td>85.6</td>
</tr>
<tr>
<td>SitDown</td>
<td>36.3</td>
<td>35.7</td>
<td>36.0</td>
<td>45.2</td>
</tr>
<tr>
<td>SitUp</td>
<td>17.1</td>
<td>13.4</td>
<td>15.2</td>
<td>15.7</td>
</tr>
<tr>
<td>StandUp</td>
<td>31.9</td>
<td>31.0</td>
<td>31.4</td>
<td>28.5</td>
</tr>
<tr>
<td>HighFive</td>
<td>49.3</td>
<td>40.6</td>
<td>48.8</td>
<td>58.5</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>42.0</td>
<td>41.1</td>
<td>41.2</td>
<td>47.9</td>
</tr>
</tbody>
</table>

Significant improvement on average
### Experiment 1 Results

#### How conjugate samples are used

<table>
<thead>
<tr>
<th>Action</th>
<th>NotUsed</th>
<th>AsNegative</th>
<th>AsPositive</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswerPhone</td>
<td>27.3</td>
<td>27.3</td>
<td>28.7</td>
<td><strong>43.0</strong></td>
</tr>
<tr>
<td>DriveCar</td>
<td>51.2</td>
<td>51.6</td>
<td>48.9</td>
<td><strong>53.1</strong></td>
</tr>
<tr>
<td>Eat</td>
<td>36.9</td>
<td>37.7</td>
<td>35.7</td>
<td><strong>42.1</strong></td>
</tr>
<tr>
<td>Fight</td>
<td>45.7</td>
<td>48.6</td>
<td>41.2</td>
<td><strong>61.1</strong></td>
</tr>
<tr>
<td>GetOutCar</td>
<td>29.7</td>
<td><strong>31.4</strong></td>
<td>30.1</td>
<td>27.2</td>
</tr>
<tr>
<td>ShakeHand</td>
<td>26.9</td>
<td>26.0</td>
<td>26.6</td>
<td><strong>35.2</strong></td>
</tr>
<tr>
<td>Hug</td>
<td>43.9</td>
<td>44.1</td>
<td>45.5</td>
<td><strong>54.7</strong></td>
</tr>
<tr>
<td>Kiss</td>
<td>67.4</td>
<td>66.4</td>
<td>67.0</td>
<td><strong>72.6</strong></td>
</tr>
<tr>
<td>Run</td>
<td>82.0</td>
<td>80.6</td>
<td>79.9</td>
<td><strong>85.6</strong></td>
</tr>
<tr>
<td>SitDown</td>
<td>36.3</td>
<td>35.7</td>
<td>36.0</td>
<td><strong>45.2</strong></td>
</tr>
<tr>
<td>SitUp</td>
<td><strong>17.1</strong></td>
<td>13.4</td>
<td>15.2</td>
<td>15.7</td>
</tr>
<tr>
<td>StandUp</td>
<td><strong>31.9</strong></td>
<td>31.0</td>
<td>31.4</td>
<td>28.5</td>
</tr>
<tr>
<td>HighFive</td>
<td>49.3</td>
<td>40.6</td>
<td>48.8</td>
<td><strong>58.5</strong></td>
</tr>
<tr>
<td>Mean</td>
<td>42.0</td>
<td>41.1</td>
<td>41.2</td>
<td><strong>47.9</strong></td>
</tr>
</tbody>
</table>

Table 1. Action recognition results on the ActionThread dataset. The table shows average precision values; a higher number indicates a better performance. All four settings use the same C3D architecture \[x\]. NotUsed is the baseline method that does not use conjugate samples. AsNegative and AsPositive are the methods that use conjugate samples as negative and positive training examples respectively. Our method achieves significantly better performance than the other methods.
# Experiment 1 Results

<table>
<thead>
<tr>
<th>Action</th>
<th>NotUsed</th>
<th>AsNegative</th>
<th>AsPositive</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswerPhone</td>
<td>27.3</td>
<td>27.3</td>
<td>28.7</td>
<td>43.0</td>
</tr>
<tr>
<td>DriveCar</td>
<td>51.2</td>
<td>51.6</td>
<td>48.9</td>
<td>53.1</td>
</tr>
<tr>
<td>Eat</td>
<td>36.9</td>
<td>37.7</td>
<td>35.7</td>
<td>42.1</td>
</tr>
<tr>
<td>Fight</td>
<td>45.7</td>
<td>48.6</td>
<td>41.2</td>
<td>61.1</td>
</tr>
<tr>
<td>GetOutCar</td>
<td>29.7</td>
<td><strong>31.4</strong></td>
<td>30.1</td>
<td>27.2</td>
</tr>
<tr>
<td>ShakeHand</td>
<td>26.9</td>
<td>26.0</td>
<td>26.6</td>
<td><strong>35.2</strong></td>
</tr>
<tr>
<td>Hug</td>
<td>43.9</td>
<td>44.1</td>
<td>45.5</td>
<td><strong>54.7</strong></td>
</tr>
<tr>
<td>Kiss</td>
<td>67.4</td>
<td>66.4</td>
<td>67.0</td>
<td><strong>72.6</strong></td>
</tr>
<tr>
<td>Run</td>
<td>82.0</td>
<td>80.6</td>
<td>79.9</td>
<td><strong>85.6</strong></td>
</tr>
<tr>
<td>SitDown</td>
<td>36.3</td>
<td>35.7</td>
<td>36.0</td>
<td><strong>45.2</strong></td>
</tr>
<tr>
<td>SitUp</td>
<td><strong>17.1</strong></td>
<td>13.4</td>
<td>15.2</td>
<td>15.7</td>
</tr>
<tr>
<td>StandUp</td>
<td><strong>31.9</strong></td>
<td>31.0</td>
<td>31.4</td>
<td>28.5</td>
</tr>
<tr>
<td>HighFive</td>
<td>49.3</td>
<td>40.6</td>
<td>48.8</td>
<td><strong>58.5</strong></td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>42.0</strong></td>
<td><strong>41.1</strong></td>
<td><strong>41.2</strong></td>
<td><strong>47.9</strong></td>
</tr>
</tbody>
</table>

Using conjugate samples as positive or negative examples might hurt the performance.
**Experiment 1: Combining with DTD**

- Combine with Dense Trajectory Descriptors

<table>
<thead>
<tr>
<th>Method</th>
<th>Average precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretrained C3D [29]</td>
<td>35.4</td>
</tr>
<tr>
<td>Finetuned C3D [29]</td>
<td>42.0</td>
</tr>
<tr>
<td>Factor-C3D [Proposed]</td>
<td><strong>47.9</strong></td>
</tr>
<tr>
<td>DTD [32]</td>
<td>45.3</td>
</tr>
<tr>
<td>Non-Action [36]</td>
<td>48.0</td>
</tr>
<tr>
<td>DTD + C3D</td>
<td>52.0</td>
</tr>
<tr>
<td>DTD + Factor-C3D [Proposed]</td>
<td><strong>55.3</strong></td>
</tr>
</tbody>
</table>
Video clips that excite the action and context channels the most

Action Components

Context Components

<table>
<thead>
<tr>
<th>AnswerPhone</th>
</tr>
</thead>
</table>

Figure 5. Representative patterns with highest action and context activation values.

Figure 6. Pairwise context difference between GetOutCar (left), Kiss (middle), and Run (right) with other action classes. Smaller context difference indicates higher context similarity. In terms of contextual similarity, DriveCar is close to GetOutCar, Hug is close to Kiss, while Fight is close to Run.

4.2. Imperfect conjugate samples

So far, we have assumed that we can retrieve perfect conjugate samples that do not contain the target human actions in consideration. This assumption is true in general, unless we are forced to exclusively work with a dataset that have no corresponding pre- or post-action sequences such as UCF101 \[27\] and Hollywood2 \[21\]. We question whether it is necessary to have a black-and-white separation between conjugate and action samples with respect to the action content. We want to study the benefits and drawbacks of the proposed framework when the action samples may not contain the entire human action and the conjugate samples cannot be guaranteed to exclude all of the action. In this section, we describe the experiments to study this scenario, using UCF101 \[27\] and Hollywood2 \[21\] datasets.

Given a training video in either UCF101 or Hollywood2 dataset, we extract a sequence of 16 video frames and use it as the action sample. From the same video, we extract another sequence of 16 frames before or after the action sample to create the corresponding conjugate sample. These two video samples have the same context, and what distinguish them is the difference between the two dynamics stages of a human action.

Using the generated conjugate samples, the proposed framework achieves a mean average precision of 54.7% on the Hollywood2 dataset, outperforming the baseline C3D network (49.8%) where conjugate samples are not used. On the three splits of the UCF101 dataset, the proposed framework achieves an average accuracy of 84.5%, outperforming the baseline C3D network (82.3%) that does not use the conjugate samples. On both Hollywood2 and UCF101, the performance gains for using conjugate samples are significant, even though the approach of generating conjugate samples is not ideal. The performance gains can be attributed to the ability of the framework to force the action extractor to focus on the dynamics of the action, rather than the scene context. On the other hand, the performance gains on Hollywood2 and UCF101 are not as high as the performance gain obtained on the ActionThread dataset, where we could extract proper conjugate samples.

The results reported in previous paragraph should not be compared directly to the highest reported numbers in previous publications, which have been obtained by combining multiple features and methods \[6\], \[35\]. The proposed method provides complementary benefits to other methods, and the state-of-the-art results can be achieved by combining them, as shown in Table 4.

<table>
<thead>
<tr>
<th>Method</th>
<th>UCF101 Average Accuracy</th>
<th>Hollywood2 Mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3D</td>
<td>82.3</td>
<td>49.8</td>
</tr>
<tr>
<td>Factor-C3D</td>
<td>84.5</td>
<td>54.7</td>
</tr>
<tr>
<td>DTD</td>
<td>85.9</td>
<td>67.5</td>
</tr>
<tr>
<td>DTD + C3D</td>
<td>90.8</td>
<td>69.5</td>
</tr>
<tr>
<td>DTD + Factor-C3D</td>
<td>91.3</td>
<td>71.3</td>
</tr>
<tr>
<td>EigenTSN</td>
<td>95.3</td>
<td>75.5</td>
</tr>
<tr>
<td>EigenTSN + C3D</td>
<td>95.8</td>
<td>76.1</td>
</tr>
<tr>
<td>EigenTSN + Factor-C3D</td>
<td>95.8</td>
<td>76.7</td>
</tr>
</tbody>
</table>

Table 4. Action recognition results on UCF101 and Hollywood2.

The comparison between Factor-C3D and C3D is the direct measurement for the advantage of the proposed framework with conjugate samples. State-of-the-art performance can be achieved when combining the proposed method with others. We report accuracy for UCF101 and mean AP for Hollywood2.
Video clips that excite the action and context channels the most

<table>
<thead>
<tr>
<th>Action Components</th>
<th>Context Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswerPhone</td>
<td></td>
</tr>
<tr>
<td>Hug</td>
<td></td>
</tr>
<tr>
<td>ShakeHand</td>
<td></td>
</tr>
<tr>
<td>SitUp</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Representative patterns with highest action and context activation values.

Figure 6. Pairwise context difference between GetOutCar (left), Kiss (middle), and Run (right) with other action classes. Smaller context difference indicates higher context similarity. In terms of contextual similarity, DriveCar is close to GetOutCar, Hug is close to Kiss, while Fight is close to Run.

4.2. Imperfect conjugate samples

So far, we have assumed that we can retrieve perfect conjugate samples that do not contain the target human actions in consideration. This assumption is true in general, unless we are forced to exclusively work with a dataset that have no corresponding pre- or post-action sequences such as UCF101 \[27\] and Hollywood2 \[21\]. We question whether it is necessary to have a black-and-white separation between conjugate and action samples with respect to the action content. We want to study the benefits and drawbacks of the proposed framework when the action samples may not contain the entire human action and the conjugate samples cannot be guaranteed to exclude all of the action. In this section, we describe the experiments to study this scenario, using UCF101 \[27\] and Hollywood2 \[21\] datasets.

Given a training video in either UCF101 or Hollywood2 dataset, we extract a sequence of 16 video frames and use it as the action sample. From the same video, we extract another sequence of 16 frames before or after the action sample to create the corresponding conjugate sample. These two video samples have the same context, and what distinguish them is the difference between the two dynamics stages of a human action.

Using the generated conjugate samples, the proposed framework achieves a mean average precision of 54.7% on the Hollywood2 dataset, outperforming the baseline C3D network (49.8%) where conjugate samples are not used. On the three splits of the UCF101 dataset, the proposed framework achieves an average accuracy of 84.5%, outperforming the baseline C3D network (82.3%) that does not use the conjugate samples. On both Hollywood2 and UCF101, the performance gains for using conjugate samples are significant, even though the approach of generating conjugate samples is not ideal. The performance gains can be attributed to the ability of the framework to force the action extractor to focus on the dynamics of the action, rather than the scene context. On the other hand, the performance gains on Hollywood2 and UCF101 are not as high as the performance gain obtained on the ActionThread dataset, where we could extract proper conjugate samples.

The results reported in previous paragraph should not be compared directly to the highest reported numbers in previous publications, which have been obtained by combining multiple features and methods \[6, 35\]. The proposed method provides complementary benefits to other methods, and the state-of-the-art results can be achieved by combining them, as shown in Table 4.

Table 4. Action recognition results on UCF101 and Hollywood2.

<table>
<thead>
<tr>
<th>Method</th>
<th>UCF101 Accuracy</th>
<th>Hollywood2 AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3D</td>
<td>82.3</td>
<td>49.8</td>
</tr>
<tr>
<td>Factor-C3D [Proposed]</td>
<td>84.5</td>
<td>54.7</td>
</tr>
<tr>
<td>DTD</td>
<td>85.9</td>
<td>67.5</td>
</tr>
<tr>
<td>DTD + C3D</td>
<td>90.8</td>
<td>69.5</td>
</tr>
<tr>
<td>DTD + Factor-C3D [Proposed]</td>
<td>91.3</td>
<td>71.3</td>
</tr>
<tr>
<td>EigenTSN</td>
<td>95.3</td>
<td>75.5</td>
</tr>
<tr>
<td>EigenTSN + C3D</td>
<td>95.8</td>
<td>76.1</td>
</tr>
<tr>
<td>EigenTSN + Factor-C3D [Proposed]</td>
<td>95.8</td>
<td>76.7</td>
</tr>
</tbody>
</table>

4.3. Separating action and context in still images

The proposed framework for separating action and context is not exclusive to video data. In this section, we perform some controlled experiments on still images. Action
Pairwise Context Differences

- Smaller difference indicates higher context similarity
- DriveCar vs. GetOutCar
### Pairwise Context Differences

- **GetOutCar Classifier**
  - AnswerPhone
  - DriveCar
  - Eat
  - Fight
  - GetOutCar
  - ShakeHand
  - Hug
  - Kiss
  - Run
  - SitDown
  - SitUp
  - StandUp
  - HighFive

- **Kiss Classifier**
  - AnswerPhone
  - DriveCar
  - Eat
  - Fight
  - GetOutCar
  - ShakeHand
  - Hug
  - Kiss
  - Run
  - SitDown
  - SitUp
  - StandUp
  - HighFive

- Smaller difference indicates higher context similarity
  - DriveCar vs. GetOutCar
  - Kiss vs. Hug
Role of Context Features

• Being able to separate action components $f(a)$ from context components $g(a)$ brings flexibility
• We can tune the contribution of the context component for classification: $f(a) + \gamma g(a)$
Experiments on Hollywood2 & UCF101 Datasets

- Hollywood2 dataset
  - 12 action classes
  - Around 1600 clips from various movies

- UCF101 dataset
  - 101 action classes
  - Around 13K clips collected from YouTube
  - But the video clips are trimmed to the action parts only
    - No preceding/following context sequences

- We consider *imperfect* conjugate samples
Using Imperfect Conjugate Samples

<table>
<thead>
<tr>
<th>Method</th>
<th>UCF101</th>
<th>Hollywood2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3D</td>
<td>82.3</td>
<td>49.8</td>
</tr>
<tr>
<td>Factor-C3D [Proposed]</td>
<td>84.5</td>
<td>54.7</td>
</tr>
<tr>
<td>DTD</td>
<td>85.9</td>
<td>67.5</td>
</tr>
<tr>
<td>DTD + C3D</td>
<td>90.8</td>
<td>69.5</td>
</tr>
<tr>
<td>DTD + Factor-C3D [Proposed]</td>
<td>91.3</td>
<td>71.3</td>
</tr>
<tr>
<td>EigenTSN</td>
<td>95.3</td>
<td>75.5</td>
</tr>
<tr>
<td>EigenTSN + C3D</td>
<td>95.8</td>
<td>76.1</td>
</tr>
<tr>
<td>EigenTSN + Factor-C3D [Proposed]</td>
<td>95.8</td>
<td>76.7</td>
</tr>
</tbody>
</table>
Summary

• First part
  • Task: recognition
  • A method to separate action from context

• Second part
  • Task: early recognition
  • Consider multiple cameras and learn a selection policy

• Learn to attend to the relevant information

Thank You!