Motivation
In the past, video understanding models and datasets have been focused on separate tasks, however recent models are focusing on addressing multiple tasks at once. This leads to the growing need for cohesive datasets which can be utilised for different understanding tasks, we introduce one of the first of these datasets.

Dataset
The dataset is based on broadcast tennis footage, and is annotated for fine-grained action centric event recognition, temporal localisation and description. The annotations are more detailed, accurate, and structured, compared to other video understanding sets.

<table>
<thead>
<tr>
<th>786455 Frames</th>
<th>525 Minutes</th>
<th>746 Descriptions</th>
<th>223 Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Matches</td>
<td>11 Sets</td>
<td>118 Games</td>
<td>746 Points</td>
</tr>
<tr>
<td>1017 Serves</td>
<td>2551 Hits</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Serve Far (SF)</th>
<th>Serve Near (SN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit Far Left (HFL)</td>
<td>Hit Far Right (HFR)</td>
</tr>
<tr>
<td>Hit Near Left (HNL)</td>
<td>Hit Near Right (HNR)</td>
</tr>
<tr>
<td>Other (O)</td>
<td></td>
</tr>
</tbody>
</table>

Classes and Events
- Videos are annotated with temporal events which have domain related classes and attributes
- For tennis the events have an inherent hierarchy (Matches > Sets > Games > Points > Serves & Hits)
- From these events we look to classify 7 classes of events at the serve/hit hierarchy level
- These classes are based on court and racquet position, generalising across players and their ‘handedness’

Descriptions
- Points have a caption attribute which contains a single sentence which describes the point
- Descriptions are similarly parsed to remove individual names and forehand/backhand mentions converted to far player (fp), near player (np), right shot (rs), and left shot (ls)

Usage
- We built an event annotator GUI which can be used for annotating any video sequence with any types of events, and make it available in our code release
- Annotations are marked up in .json so are easily read and understood by both human and machine
- All data and models as well as code for processing, code for annotating, and code for training plus testing are freely available via the link in the footer

Tasks
1. Frame & Clip Classification into Events
2. Temporal Event Detection & Classification
3. Event (Point) Captioning / Description

Models
Visual Processing: Single and two stream VGG16 CNNs on RGB frames and / or Optical Flow frames
Temporal Processing: Mean Pooling or (Bi)RNN (GRU)
Sentence Generation: Encoder-Decoder RNN (GRU)

Results
Frame-wise Classification and Event Detection
- Two stream CNNs (81%) have better frame-wise mAP than single stream RGB (67%) and Flow (76%) CNNs
- Temporal mean pooling improves on accuracy compared to no temporal modeling however using a BiRNN outperforms mean pooling and even is hindered by it when used together
- Longer mean pooling windows have negative effects on shorter events

Sentence Generation
- Word matching metrics not a reliable indication of true performance, more for this set to object-centric sets, as individual words or small phrases have a great conceptual difference
- Empirically, sentences mostly correct however our model can suffer from repeated or misplaced words as well as the occasional concept error

<table>
<thead>
<tr>
<th>CNN model</th>
<th>mAP</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>0.649</td>
<td>0.368</td>
<td>0.471</td>
<td>0.372</td>
<td>0.518</td>
</tr>
<tr>
<td>Two Stream</td>
<td>0.768</td>
<td>0.550</td>
<td>0.691</td>
<td>0.572</td>
<td>0.766</td>
</tr>
</tbody>
</table>

Paper - hf.id.au/papers/DICTA17_Tennis.pdf
Code & Dataset - github.com/HaydenFaulkner/Tennis