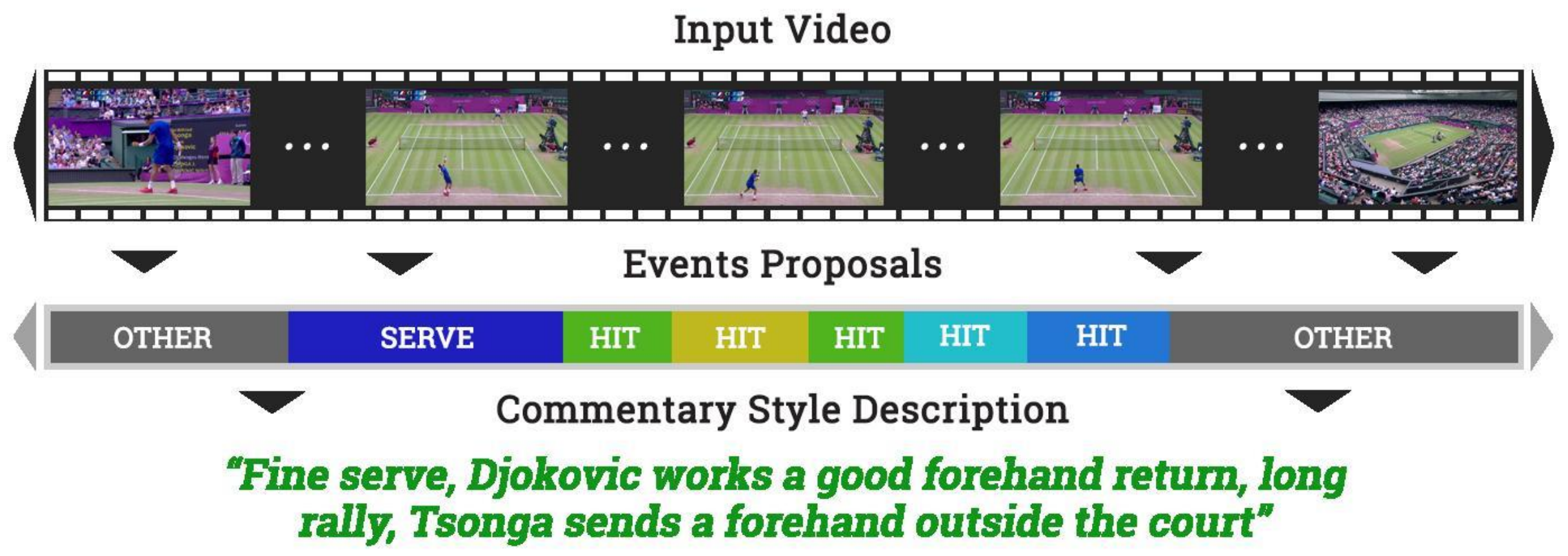
TenniSet: A Dataset for Dense Fine-Grained Event Recognition, Localisation and Description

Hayden Faulkner, Anthony Dick - ***The University of Adelaide***

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.


Input Video
...
Events Proposals
OTHER
SERVE
HIT
HIT
HIT
HIT
HIT
OTHER
Commentary Style Description
"Fine serve, Djokovic works a good forehand return, long rally, Tsonga sends a forehand outside the court"

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.

786455 Frames

525 Minutes

746 Descriptions

223 Vocabulary

5 Matches

11 Sets

118 Games

746 Points

1017 Serves

2551 Hits

Serve Far (SF)	Serve Near (SN)
Hit Far Left (HFL)	Hit Far Right (HFR)
Hit Near Left (HNL)	Hit Near Right (HNR)
Other (O)	

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

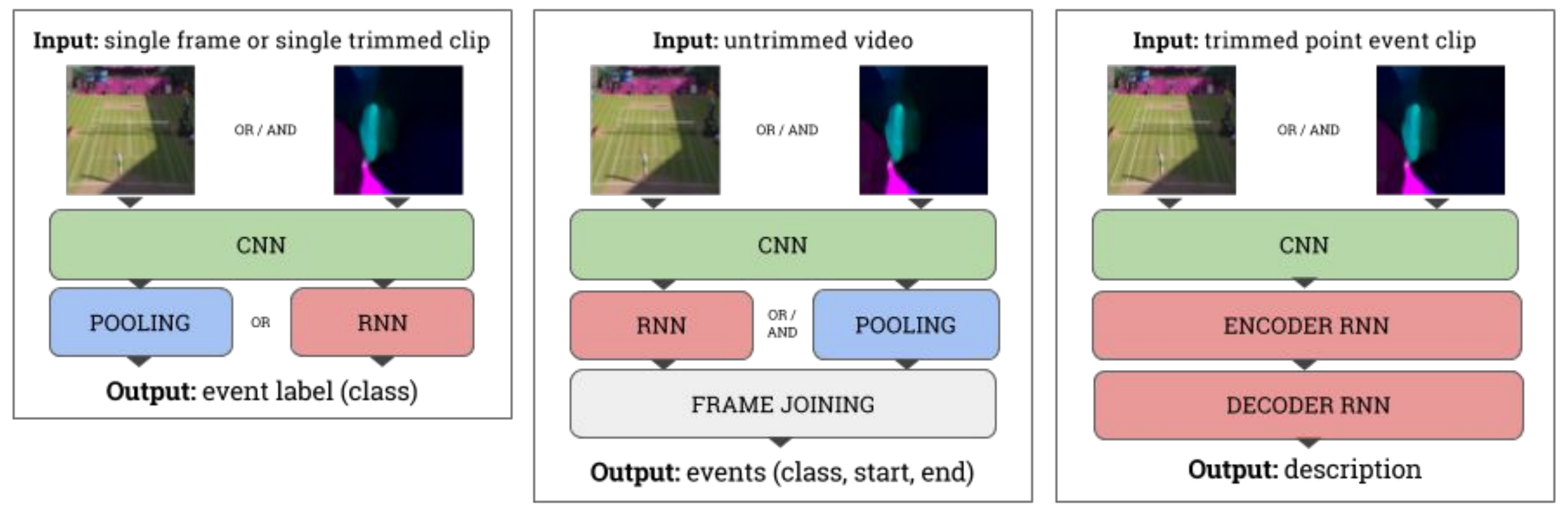
- Points have a caption attribute which contains a single sentence which describes the point
- Descriptions are similarly parsed to remove individual names and forehand/backhand mentions converted to far player (fp), near player (np), right shot (rs), and left shot (ls)

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


Input: single frame or single trimmed clip
OR / AND
CNN
POOLING
OR
RNN
Output: event label (class)
Input: untrimmed video
OR / AND
CNN
RNN
OR / AND
POOLING
FRAME JOINING
Output: events (class, start, end)
Input: trimmed point event clip
OR / AND
CNN
ENCODER RNN
DECODER RNN
Output: 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)

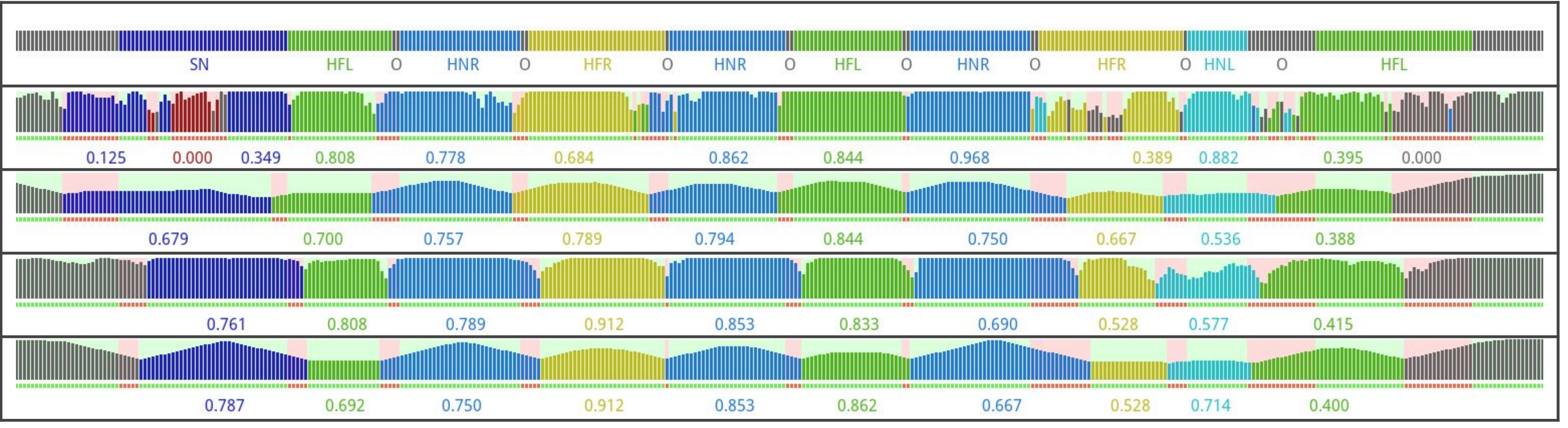
Caption Generation: Encoder-Decoder RNN (GRU)

Results

Framewise Classification and Event Detection

- Two stream CNNs (81%) have better framewise 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

Models - Two Stream	Framewise mAP	Event Det. mAP (.5 IOU)
Framewise	.816	.641
Mean Pooled (40)	.766	.809
BiRNN (40)	**.881**	**.814**
BiRNN (40) + Pool (40)	.726	.763



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

CNN Model	mAP	BLEU@4	METEOR	CIDEr	ROUGE-L
Rdm Retr'val	-	.0593	.1493	.2713	.3147
RGB Only	.6748	.1038	.2014	.5729	.4078
Flow Only	.7607	.0839	.1905	.4486	.4053
Two Stream	**.8157**	**.1284**	**.2223**	**.6777**	**.4518**

good serve aimed in the corner np only reaches to it	sharp angled slice serve is an ace over the net
good serve in the middle np crafts a rs return a brief rally fp rs cross-court is a winner	good serve in the middle np returns a rs return short rally np produces a rs cross-court winner the line
fp hits a flat bodyline serve np struggles to put it back	fp arrows a serve serve np return it over net
fp serves a good one np delivers a ls return good rally fp hits a ls cross-court drop-shot winner	fp serves a good one np delivers a ls return brief rally np hits to rs net the the net

Paper - hf.id.au/papers/DICTA17_Tennis.pdf

Code & Dataset - github.com/HaydenFaulkner/Tennis