

TennisVid2Text: Fine-grained Descriptions for Domain Specific Videos (Supplementary)

Mohak Sukhwani

<http://researchweb.iiit.ac.in/~mohak.sukhwani/>

C.V. Jawahar

<http://www.iiit.ac.in/~jawahar/>

CVIT

IIIT Hyderabad,

India.

<http://cvit.iiit.ac.in>

We provide results of various modules in this supplementary document. We also showcase the contents of our ‘Annotated-action’ dataset to highlight the variability in action phrases across the dataset.

1 Performance Analysis

1.1 Player Identification

We use domain specific cues, such as court lines and net position to detect players in field. Court lines assist us in isolation playing areas which in turn help us in identifying players. Table 1 summarizes our court detections and player detections results.

1.2 Phrase Classifiers

Phrases like ‘*hits a forehand return*’, ‘*returns a forehand return*’, ‘*works a forehand return*’, etc. are significantly similar in our case (high correlation between the classes); as a result looking for exact matches could be a strict constraint. We relax these conditions by looking for label matches that are considerably near rather than exact. Phrase pairs having overlapping subtopics have high similarity values when measured by standard cosine similarity. For exact matches the cosine similarity score is 1.0 while for relaxed case we consider pairs with cosine similarity score of more than 0.5. We report our results in Table 1, averaged over top five retrievals for both exact and relaxed matches. In all we have 76 verb phrases - 39 for upper and 37 for lower player. We also report average class wise accuracy of frequent classes in Table 2.

Using Domain Specific cues: Compared to a naive use of phrase classifier, our verb phrase classifier performs better by dividing the frame into upper and lower halves. Table 3 summarizes the effects of domain specific cues, i.e. division across the net.

1.3 Generation of Descriptions

Given a (test) video, we recognize verb phrases for each frame by extracting features from neighbouring frames using sliding window. We use MRF based smoothing to identify the final set of phrases. Using the identified phrases we build retrieval system over our corpus

Approach	Upper	Lower		Correct #	Actual #
Relaxed	0.869	0.705	Court	621	710
Exact	0.770	0.613	Player	746	1420

(a) (b)

Table 1: Phrase and Player recognition: a) Verb phrase recognition accuracy averaged over top five retrieval. b) Detections on test videos each with two players on single court. In all we have 710 videos in our test dataset and hence $710 \times 2 = 1420$ players in all. The players thus detected are then identified using recognition module.

Class	Upper		Lower		Class	Upper		Player	
	#	Acc.	#	Acc.		#	Acc.	#	Acc.
waits for ball	64	0.87	44	0.64	serves a good one	5	0.5	7	0.33
player focused	30	0.93	30	0.91	crafts a forehand return	8	0.33	8	0.53
audience seems enjoying	13	0.33	13	0.25	hits a backhand return	3	0.60	7	0.14

Table 2: Class classification accuracy for frequent phrases. # refers to number of times these phrases occur in annotated-action dataset and Acc. refers to classification accuracy.

using both naive lexical and LSI approach. Figure 2 and Figure 1 depict some of the successful and failure cases; descriptions shown in both the figures correspond to top retrieval of LSI output.

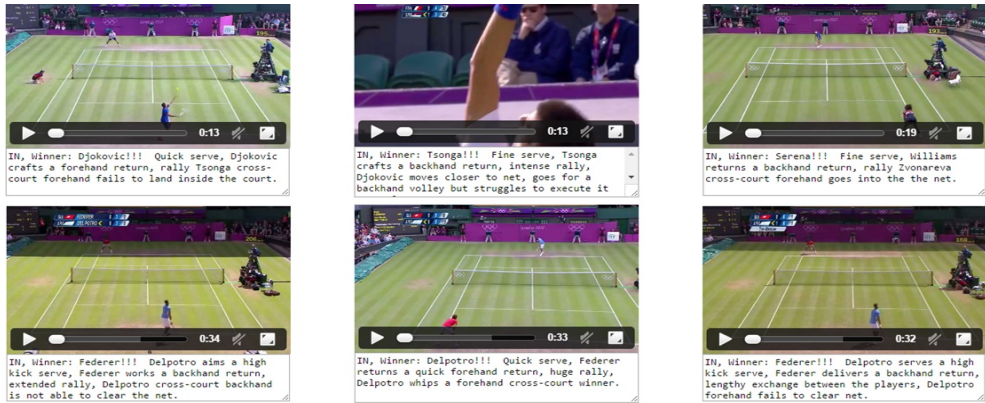


Figure 1: Success Cases: Example videos along with their generated descriptions that are similar to ground truth descriptions

2 Dataset contents

‘Annotated action’ dataset comprises of video shots linked with corresponding action phrases. Figure 3 showcases some of the contents of ‘Annotated action’ dataset.



Figure 2: Failure Cases: Example videos along with their generated descriptions which fail to match with ground truth descriptions. Descriptions in failed category involve wrong action sequences in descriptions.

Approach	Without Frame Division		With Frame Division	
	Upper	Lower	Upper	Lower
Relaxed	0.713	0.636	0.869	0.705
Exact	0.672	0.516	0.775	0.614

Table 3: Impact of harnessing domain relevant cues on verb phrase recognition accuracy.

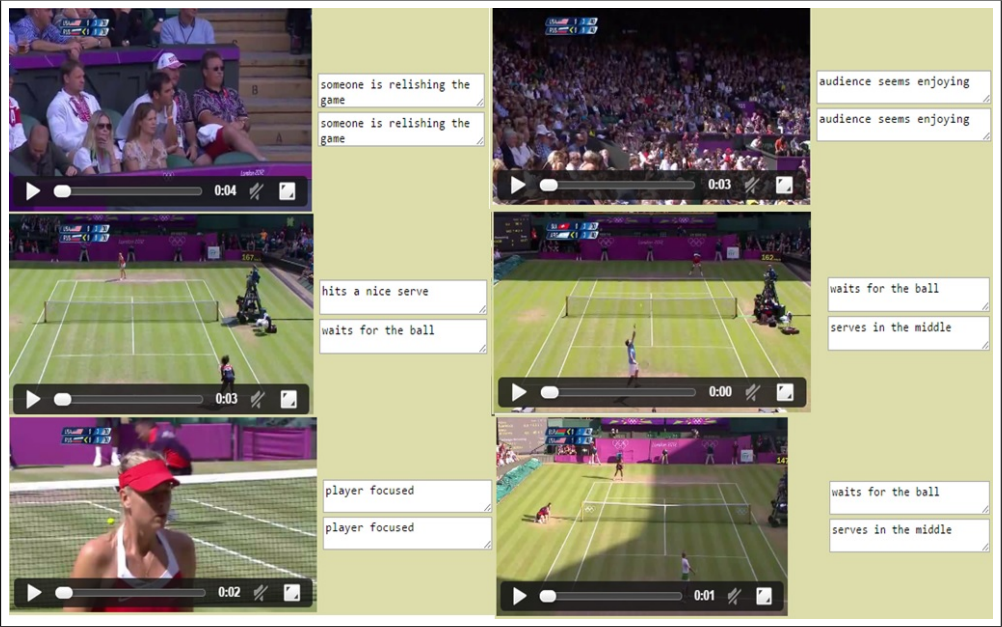


Figure 3: Annotated-action dataset: Video shots aligned with verb phrases. Every video shot is labelled with two phrases, the upper text correspond to actions of upper player and lower text to actions of lower player.