TennisVid2Text: Fine-grained Descriptions for Domain Specific Videos

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

Annotating visual content with text has attracted significant attention in recent years [1, 2, 3, 4]. While the focus has been mostly on images [2, 3, 4], of late few methods have also been proposed for describing videos [1]. The descriptions produced by such methods capture the video content at certain level of semantics. However, richer and more meaningful descriptions may be required for such techniques to be useful in real-life applications. We make an attempt towards this goal by focusing on a domain specific setting – lawn tennis videos. Given a video shot from a tennis match, we intend to predict detailed (commentary-like) descriptions rather than small captions. Figure 1 depicts the problem of interest and steps involved in our method. Rich descriptions are generated by leveraging a large corpus of human created descriptions harvested from Internet. We evaluate our method on a newly created tennis video data set comprising of broadcast video recordings of matches from London Olympics 2012. Extensive analysis demonstrate that our approach addresses both semantic correctness as well as readability aspects involved in the task.

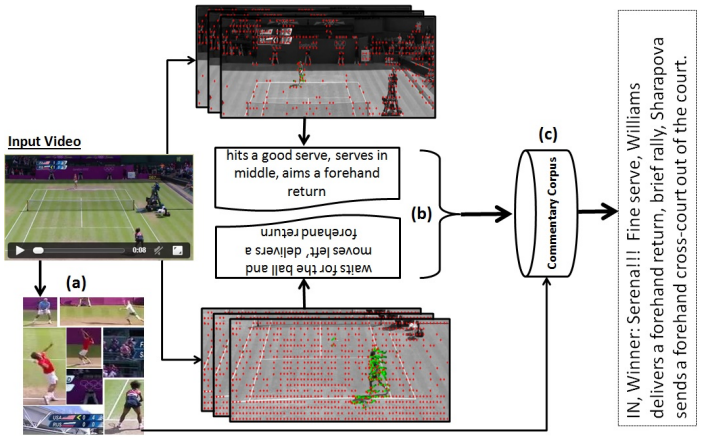

Figure 1: For a test video, the *description* predicted using our approach is dense and human-like. Above figure demonstrates our approach – a) Player Identification b) Verb Phrase Prediction c) Description Generation

Given a test video, we predict a set of action/verb phrases individually for each frame using the features computed from its neighbourhood. The identified phrases along with additional meta-data are used to find the best matching description from the commentary corpus. We begin by identifying two players on the tennis court, Figure 1(a). Regions obtained after isolating playing court regions assist us in segmenting out the candidate player regions through background subtraction using thresholding and connected component analysis. Each candidate foreground region thus obtained is represented using HOG descriptors over which a SVM classifier is trained to discard non-player foreground regions. The candidate player regions thus obtained are used to recognize players using using CEDD descriptors and Tanimoto distance.

Verb phrases are recognized, Figure 1(b), by extracting features from each frame of input video using sliding window. Since this typically results into multiple firings, non-maximal suppression (NMS) is applied. This removes low-scored responses that are in the neighbourhood of responses with locally maximal confidence scores. Once we get potential phrases for all windows along with their scores, we remove the independence assumption and smooth the predictions using an energy minimization framework. For this, a Markov Random Field (MRF) based model is used which captures dependencies among nearby phrases. We add one node for each window sequentially from left to right and connect these by edges. Each node takes a label from the set of action phrases. The energy function for nodes $\mathcal{V}$, neighbourhood $\mathcal{N}$ and labels $\mathcal{L}$ is:

$$E = \sum_{p \in \mathcal{V}} D_p(f_p) + \sum_{p,q \in \mathcal{N}} V_{pq}(f_p, f_q) \quad (1)$$

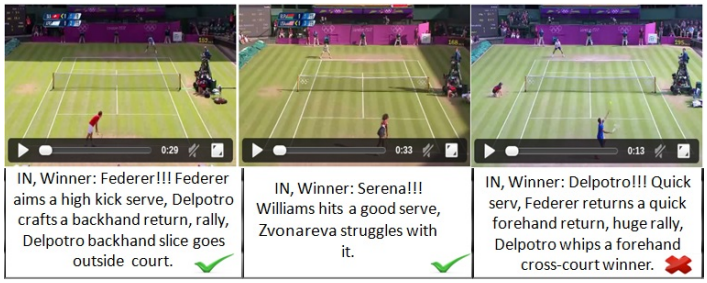

Figure 2: Sample outputs: Example videos along with their descriptions. The 'ticked' descriptions match with the ground truth, while the 'crossed' ones do not.

Here, D_p denotes *unary phrase selection cost* and V_{pq} denotes *pairwise phrase cohesion cost* associated with two neighbouring nodes. The solution of this energy minimization returns a set of actions phrases that describe the input video. We formulate the task of predicting the final description,Figure 1(c), as an optimization problem of selecting the best sentence among the set of commentary sentences in corpus which covers most number of unique words in obtained phrase set. We even employ Latent Semantic Indexing (LSI) technique while matching predicted phrases with descriptions and demonstrate its effectiveness over naïve lexical matching.

The proposed system is benchmarked against state-of-the-art methods. We compare our performance with recent methods in Table 1(right). Caption generation based approaches [1, 2] achieve significantly low score owing to their generic nature. Compared to all the competing methods, our approach consistently provides better performance. Table 1(left) demonstrates the effect of variations in corpus size on BLEU scores. It can be observed that the scores saturate soon, which validates that in domain specific settings, rich descriptions can be produced even with small corpus size.

Corp#	Voc#	B1	B2	B3	B4
100	85	0.379	0.235	0.154	0.095
500	118	0.428	0.251	0.168	0.107
5K	128	0.458	0.265	0.178	0.111
30K	140	0.460	0.277	0.182	0.113
50K	144	0.461	0.276	0.183	0.114

Method	B1	B2	B3	B4
Guadarrama [1]	0.119	0.021	0.009	0.002
Karpathy [2]	0.135	0.009	0.001	0.001
Rasiwasia [3]	0.409	0.222	0.132	0.070
Verma [4]	0.422	0.233	0.142	0.075
This work	0.461	0.276	0.183	0.114

Table 1: **Left:** Variation in BLEU score with corpus size. 'Corp#' refers to number of commentary lines, 'Voc#' is vocab dimensionality. **Right:** Performance comparison with present state-of-the-art methods. .

Our approach demonstrates the utility of the simultaneous use of vision, language and machine learning techniques in a domain specific environment to produce semantically rich and human-like descriptions. The proposed method can be well adopted to situations where activities are in a limited context and the linguistic diversity is confined.

[1] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-vemantic alignments for generating image descriptions. In *CVPR*, 2015.

[3] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.

[4] Yashaswi Verma and C. V. Jawahar. Im2text and text2im: Associating images and texts for cross-modal retrieval. In *BMVC*, 2014.