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

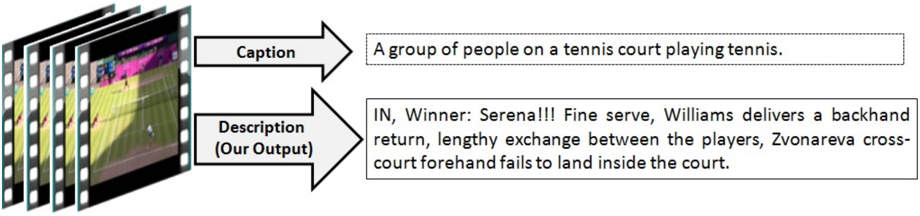
Abstract

Automatically describing videos has ever been fascinating. In this work, we attempt to describe videos from a specific domain – broadcast videos of lawn tennis matches. Given a video shot from a tennis match, we intend to generate a textual commentary similar to what a human expert would write on a sports website. Unlike many recent works that focus on generating short captions, we are interested in generating semantically richer descriptions. This demands a detailed low-level analysis of the video content, specially the actions and interactions among subjects. We address this by limiting our domain to the game of lawn tennis. 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. Extensive analysis demonstrate that our approach addresses both semantic correctness as well as readability aspects involved in the task.

1 Introduction

Annotating visual content with text has attracted significant attention in recent years [13, 14, 18, 19, 23, 26, 30]. While the focus has been mostly on images [14, 19, 23, 26, 30], of late few methods have also been proposed for describing videos [13, 18]. 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. This becomes challenging in a domain independent scenario due to almost innumerable possibilities. We make an attempt towards this goal by focusing on a domain specific setting – lawn tennis videos. For such videos, we aim to predict detailed (commentary-like) descriptions rather than small captions. Figure 1 depicts the problem of interest and an example result of our method. It even depicts the difference between a caption and a description. Rich description generation demands deep understanding of visual content and their associations with natural text. This makes our problem challenging.

For the game of tennis, which has a pair of players hitting the ball, actions play the central role. Here, actions are not just simple verbs like 'running', 'walking', 'jumping' etc. as in the early days of action recognition but complex and compound *phrases* like 'hits a forehand volley', 'delivers a backhand return' etc. Although learning such activities add to the complexities of the task, yet they make our descriptions diverse and vivid. To further



A group of people on a tennis court playing tennis.

IN, Winner: Serena!!! Fine serve, Williams delivers a backhand return, lengthy exchange between the players, Zvonareva cross-court forehand fails to land inside the court.

Figure 1: For a test video, the *caption* generated using an approximation of a state-of- the-art method [14], and the *description* predicted using our approach. Unlike recent methods that focus on short and generic captions, we aim at detailed and specific descriptions.

integrate finer details into descriptions, we consider constructs that modify the effectiveness of nouns and verbs. Though phrases like 'hits a nice serve' ,'hits a good serve' and 'sizzling serve' describe similar action, '*nice*', '*good*' and '*sizzling*' add to the intensity of that action. We develop a model that learns the effectiveness of such phrases, and builds upon these to predict florid descriptions. Empirical evidences demonstrate that our approach predicts descriptions that match the videos.

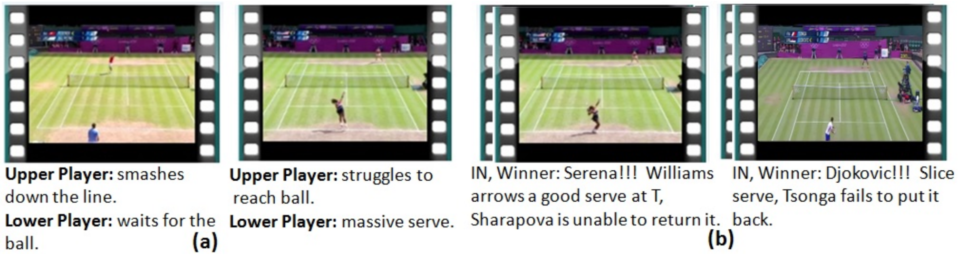
2 Related Work

Here we briefly discuss some of the works related to action recognition, image and video description generation and tennis video analysis.

Action recognition: This has been studied in a variety of settings [6, 12]. Initial attempts used laboratory videos of few well defined actions. In recent years the interest has been shifted to the actions captured in the natural settings like movies [11, 20]. In all these cases, the actions of interest are visually and semantically very different. At times when there is low inter-action and high intra-action variability in human actions, people have used fine grained action classifications to distinguish between subtle variations [27].

Image and video description: Template based approaches [13, 15, 17, 18, 19, 21] have been very popular for describing the visual content since the beginning. In case of videos, recent works [13, 18] have focused on complex task of recognizing *actor-action-object* relationships rather than simple nouns, verbs etc. E.g., Niveda *et al.* [18] generate descriptions for short videos by identifying the best SVO (*subject-verb-object*) triplet. Sergio *et al.* [13] extend this for out-of domain actions (not in training set) by generating brief sentences that sum up the main activity in a video with broader verb and object coverage. Barbu *et al.* [3] produce rich sequential descriptions of videos using detection based tracking and body-posture codebook. Template based approaches preserve generality but often take away the human aspect of the text. To overcome this, data driven methods [23, 26, 30] have been used to predict text for query images. Rather than generating a description, these methods retrieve the best matching description from a text corpus. These methods often have an advantage of producing syntactically better results, as retrieved descriptions are human-written and thus well formed.

Tennis video analysis: Early attempt [22] for tennis video analysis and annotation focused on detecting relative positions of both players with respect to court lines and net to determine action types. Use of special set-ups and other audio-video cues to detect players, ball and court lines with utmost precision were soon followed [7, 24]. While a few attempts have been



Upper Player: smashes down the line.
Lower Player: waits for the ball.

Upper Player: struggles to reach ball.
Lower Player: massive serve.

(a)

IN, Winner: Serena!!! Williams arrows a good serve at T, Sharapova is unable to return it.

IN, Winner: Djokovic!!! Slice serve, Tsonga fails to put it back.

(b)

Figure 2: Dataset contents: (a) Annotated-action dataset: short videos aligned with verb phrases. (b) Video commentary dataset: game videos aligned with commentary.

made to automate commentary generation in simulated settings (*e.g.*, for soccer [29, 31]), there has not been much success for games in natural settings. Our approach is designed to be effective in real-life game play environments and does not assume any special/simulated set-up.

3 Background, Motivation and Dataset Description

Lawn tennis is a racquet sport played either individually against a single opponent (*singles*), or between two teams of two players each (*doubles*). We restrict our attention to singles. Videos of such matches have two players – one in the upper half and the other in the lower half of a video frame. A complete tennis match is an amalgamation of sequence of 'tennis-sets', each comprising of a sequence of 'tennis-games' played with service alternating between consecutive sets. A 'tennis-point' is the smallest sub-division of match that begins with the start of the service and ends when a scoring criteria is met. We work at this granularity.

We use broadcast video recordings for five matches from *London Olympics* 2012 for our experiments. The videos used are of resolution 640 × 360 at 30 fps. Each video is manually segmented into shots corresponding to 'tennis-points', and is described with a textual commentary obtained from [1]. This gives a collection of video segments aligned with corresponding commentaries. In total, there are 710 'tennis-points' of average frame length 155. We refer to this collection as 'Video-commentary' dataset. This serves as our test dataset and is used for the final evaluation. In addition to this, we create an independent 'Annotated-action' data set comprising 250 short videos (average length of 30 frames) describing player actions with verb phrases. Examples of the verb phrases include *'serves in the middle'*, *'punches a volley'*, *'rushes towards net'*, etc. In total, we have 76 action phrases. We use this collection to train our action classifiers. Figure 2 shows samples from our dataset.

As an additional linguistic resource for creating human readable descriptions, we crawl tennis commentaries (with no corresponding videos). This text corpus is built using (human-

Name	**Contents**	**Role**
Annotated-action	250 action videos and phrases	Classification and Training
Video-commentary	710 game videos and commentaries.	Testing
Tennis Text	435K commentary lines	Dictionary Learning, Evaluation and Retrieval

Table 1: Dataset statistics: Our dataset is a culmination of three standalone datasets. Table describes them in detail, along with the roles they play in the experiments.

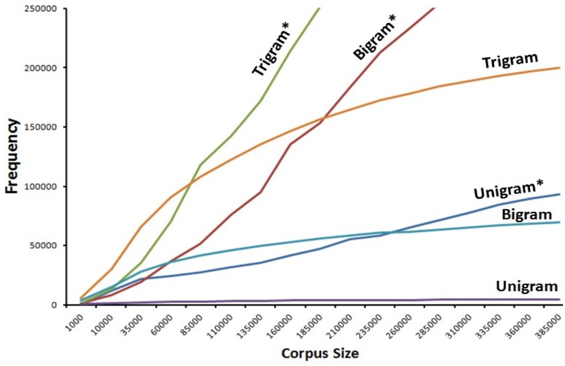


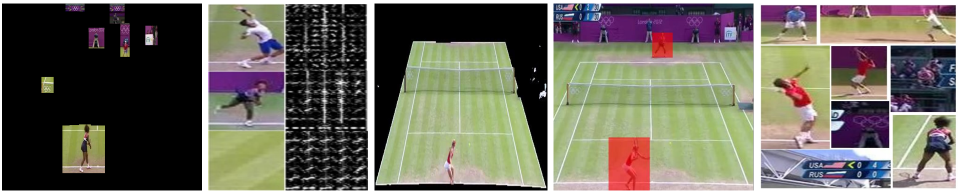
Figure 3: Trend illustrating saturation of word-level trigram, bigram and unigram counts in domain specific settings. Emphasized(*) labels indicate corpus of unrestricted tennis text (blogs/news) and remaining labels indicate tennis commentary corpus. Bigram* and Trigram* frequencies are scaled down by a factor of 10 to show the comparison.

written) commentary of 2689 lawn tennis matches played between 2009-14 from [1]. A typical commentary describes the players names, prominent shots and the winner of the game. We refer this collection as ‘Tennis-text’. Table 1 summarizes the three datasets. Note that all the datasets are independent, with no overlap among them.

We seek to analyse how focusing on a specific domain confines the output space. We compute the count of unique (word-level) unigrams, bigrams and trigrams in tennis commentaries. Each commentary sentence in the Tennis-text corpus is processed individually using standard Natural Language ToolKit (NLTK) library, and word-level n-gram frequencies of corpus are computed. The ‘frequency’ (count) trends of unigrams, bigrams and trigrams plotted over ‘corpus size’ (number of lines in corpus) are depicted in Figure 3. We compare these with the corresponding frequencies in unrestricted tennis text mined from on-line tennis news, blogs, etc. (denoted by ‘*’ in the figure). It can be observed that in case of tennis commentary, the frequency of each n-gram saturates well within a small corpus as compared to corresponding frequencies of unrestricted text. The frequency plots reveal that the vocabulary specific to tennis *commentary* is indeed small, and sentences are often very similar. Hence, in a domain specific environment, we can create rich descriptions even with a limited corpus size.

4 Approach

Our goal is to automatically describe video segments of *tennis points* in the Video-commentary dataset. We begin by learning phrase classifiers using Annotated-action dataset. Given a test video, we predict a set of action/verb phrases individually for each frame using the features computed from its neighbourhood. Since this sequence of labels could be noisy, these are smoothed by introducing priors in an energy minimization framework. The identified phrases along with additional meta-data (such as player details) are used to find the best matching description from the Tennis-text dataset. Major activities during any tennis game take place on each side of the net, we analyse videos by dividing them across the net in the subsequent sections (referred as ‘upper’ and ‘lower’ video/frame). In almost all tennis broadcast videos, this net is around the center of a frame, and thus can be easily approximated.



(a) (b) (c) (d) (e)

Figure 4: Retrieving player details: (a) Extracted foreground regions. (b) Visualization of HOG features for players and non-player regions. (c) Court detection. (d) Player Detection. (d) Some examples of successful and failed player detections.

4.1 Court Detection, Player Detection and Player Recognition

We begin by identifying/tracking both players on the tennis court. The playing court in lawn tennis has a set of prominent straight white lines, each of which adds meaning to the game in a unique way. We detect these lines using the Hough transform and consider only the most prominent lines (by keeping a threshold on length). A bounding-box is created encompassing the set of identified lines, which is then considered as foreground seed for the GrabCut algorithm [28]. This in turn returns the playing field as shown in Figure 4(c).

In a tennis broadcast video background is (nearly) static after the serve begins, and remains the same till a player wins a point. Based on this, the candidate regions for players are segmented through background subtraction using thresholding and connected component analysis. Each candidate foreground region thus obtained is represented using HOG descriptor [8]. To prune away false detections (i.e., non-player regions), multi-class SVM with RBF-kernel is employed distinguishing between 'upper-player', 'lower-player' and 'no player' regions. Figure 4(b) visualizes HOG features for few examples. The detected windows thus obtained are used to recognize players. In any particular tournament (and in general) players often wear similar colored jerseys and depict unique stance during game play, we use these cues to recognize them. We perceive both color and stance information using CEDD [4] descriptor. This descriptor captures both colour and edge (for stance) information of each detected candidate player region. We use Tanimoto distance [4] to build our classifier. Classifier scores averaged over initial ten frames are used to recognize both the players. Figure 4(d) highlights the players detected in a frame, and Figure 4(e) shows some true and false detections.

4.2 Learning Action Phrases

We learn phrase classifiers using 'Annotated-action' dataset. For representation, we use descriptors as described in [32], and extract dense trajectory features over space-time volumes (using default parameters). For each trajectory, we use Trajetory, HOG, HOF (histograms of optical flow) [5] and MBH (motion boundary histogram) [9] descriptors. While HOG captures static appearance information, HOF and MBH measure motion information based on optical flow. The dimensions of each of these descriptors are: 30 for Trajectory, 96 for HOG, 108 for HOF and 192 for MBH. For each descriptor, bag-of-words (BOW) representation is adopted (with vocabulary size 2000). We take square root of each element in a feature vector before computing the codebook (similar to RootSIFT [2]). The final representation is concatenation of BOW histograms of all the descriptors. Using this, a 1-vs-rest SVM classifier (with RBF kernel and χ^2 distance) is learned for each phrase. In all, we have 76 verb phrases - 39 for

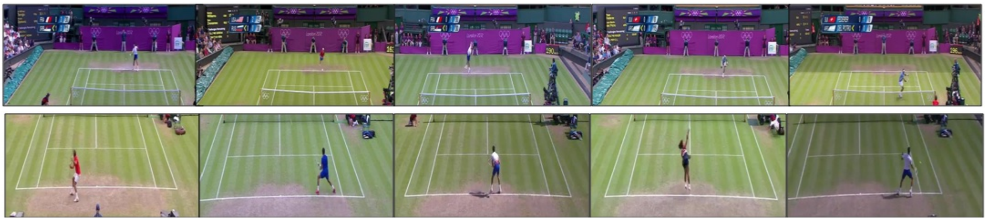


Figure 5: Example frames depicting varied actions. Upper frames are shown at the top, and lower frames at the bottom. Here, upper and lower frames do not correspond to same video.

upper and 37 for lower player. Figure 5 illustrates some examples of player actions.

4.3 Verb Phrase Prediction and Temporal Smoothing

Given a (test) video, we recognize verb phrase for each frame by extracting features from neighbouring frames using sliding window (neighbourhood of size 30 frames). 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 ν, neighbourhood $\mathcal{N}$ and labels $\mathcal{L}$ is:

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

Here, $D_p(f_p)$ denotes *unary phrase selection cost*. This is set to $1 - p(l_j|x_p)$, where $p(l_j|x_p)$ is the SVM score of the phrase label l_j for node x_p, normalized using the Platt's method [25]. The term $V_{pq}(f_p, f_q)$ denotes *pairwise phrase cohesion cost* associated with two neighbouring nodes x_i and x_j taking some phrase label. For each pair of phrases, this is determined by their probability of occurring together in the game play, and is computed using frame-wise transition probability. In our case, since there are two players, and each player's shot depends on the other player's shot and his own previous shot, we consider four probability scores: $p(l_{iP_1}, l_{jP_1})$, $p(l_{iP_1}, l_{jP_2})$, $p(l_{iP_2}, l_{jP_2})$ and $p(l_{iP_2}, l_{jP_1})$. Here, $p(l_{iP_1}, l_{jP_2})$ refers to the probability of phrase l_i of *player*1 and l_j of *player*2 occurring together during game play. We compute pairwise cost, $1 - p$, for each of the four probabilities and solve the minimization problem using a loopy belief propagation (BP) algorithm [16].

4.4 Description Prediction

Let, $W = \{w_1, w_2, \ldots, w_n\}$ be set of unique words present in the group of phrases along with player names, and $S = \{s_1, s_2, \ldots, s_m\}$ be the set of all the sentences in the Tennis-text corpus. Here, each sentence is a separate and full commentary description. We formulate the task of predicting the final description as an optimization problem of selecting the best commentary from S that covers as many words as possible. Let, x_i be a variable which is 1 if sentence s_i is selected, and 0 otherwise. Similarly, let a_{ij} be a variable which is 1 if sentence s_i contains

Input with ground Truth	 IN, Winner: Serena!!! Williams arrows a good serve at T, Sharapova is unable to return it.	 IN, Winner: Federer!!! Good serve in the middle, Fedrer crafts a forehand return, short rally, Delpotro cross-court forehand fails to land inside the court.
Phrases	(waits for the ball), (waits for the ball), ... (prepares for serve) ,......, (hits a good serve),(sizzling serve), ... , ...	(prepares for serve),.....,(tosses ball for serve),........ , (hits a good serve),.... (waits for the ball),...,..... (returns a quick forehand return),..,(sprays a forehand).. ...
Descriptions (Top 2 retrievals)	1. IN, Winner: Serena!!! Williams hits a good serve, Sharapova struggles with it. 2. IN, Winner: Serena!!! Williams hits a good serve, Sharapova struggles with it.	1. IN, Winner: Delpotro!!! Fine serve, Delpotro works a forehand return, brief rally, Delpotro rushes to net and punches a forehand volley winner. 2. IN, Winner: Federer!!! Quick serve, Delpotro returns a quick forehand return, couple of shots exchanged, Delpotro nets a forehand down the line.

Figure 6: Illustration of our approach. Input sequence of videos is first translated into a set of phrases, which are then used to produce the final description.

word w_j, and 0 otherwise. A word w_j is said to be covered if it is present in the selected sentence ($\sum_i a_{ij}x_i = 1$). Hence, our objective is to find a sentence that covers as many words as possible:

$$\max_{i\in\{1,2,\ldots,m\},j\in\{1,2,\ldots,n\}} \sum a_{ij}x_i, \quad s.t. \sum_{i=1}^{m} x_i = 1, \;\; \forall a_{ij}, x_i \in \{0,1\} \tag{2}$$

In the above formulation, doing naïve lexical matching can be inaccurate as it would consider just the presence/absence of words, and fail to capture the overall semantics of the text. To address this, we adopt Latent Semantic Indexing (LSI) [10], and use statistically derived conceptual indices rather than individual words. LSI assumes an underlying structure in word usage that is partially obscured by variability in word choice. It projects derived phrases and corpus sentences into a lower dimensional subspace, and addresses the problems of synonymy (similar meaning words). Figure 6 illustrates the steps involved in our method by taking two examples. The verb phrase prediction and smoothing steps provide a set of relevant phrases. Number of such phrases depend on the size of the (test) video. This is evident from the second example (right), which is of longer duration and thus has more phrases predictions. These phrases are used to select the best matching commentary from the Tennis-text corpus. Since similar events are described by identical descriptions in text corpus, there could be instances where the retrieved descriptions are same – first example (left) in Figure 6.

5 Experiments and Results

5.1 Experimental Details

(1) Creating Textual Dictionary: We create textual dictionary using commentary descriptions from our Tennis-text corpus. The text is processed using standard NLTK modules. This involves tokenizing, filtering out common stopwords and stemming. The dictionary thus obtained is used to compute tf-idf based feature representation of the commentary text.

Corp#	Vocab#	B-1	B-2	B-3	B-4
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	B-1	B-2	B-3	B-4
Guadarrama [13]	0.119	0.021	0.009	0.002
Karpathy [14]	0.135	0.009	0.001	0.001
Rasiwasia [26]	0.409	0.222	0.132	0.070
Verma [30]	0.422	0.233	0.142	0.075
This work	0.461	0.276	0.183	0.114

Table 2: **Left:** Variation in BLEU score with corpus size. **Right:** Performance comparison with previous methods using the best performing dictionary. 'Corpus#' denotes the number of commentary lines, 'Vocab#' denotes the dimensionality of the textual vocabulary/dictionary and 'B-n' means n-gram BLEU score.

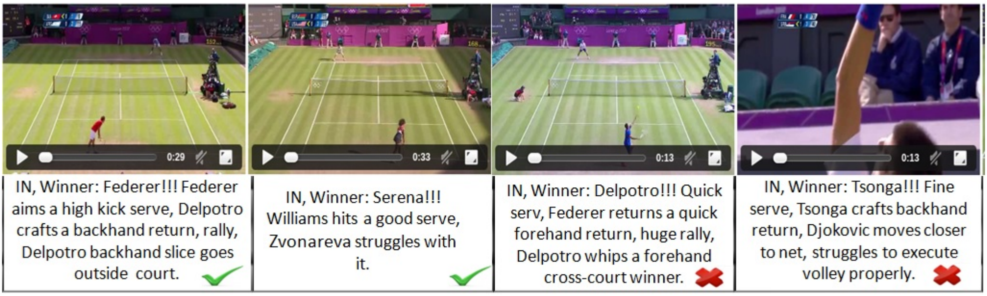
(2) Player Detection and Recognition: To learn multi-class SVM for differentiating between player and non-player regions, we use manually annotated examples. The bounding boxes (separate for 'upper-player' and 'lower-player') are identified from random video frames of 'Annotated action' dataset. In all we had 2432 'lower-player', 2421 'upper-player' and 13050 'no player' windows. Similarly for player recognition, we learn classifiers using CEDD features over 'lower-player' and 'upper-player' windows. Each window in this case is labelled as one of the eight unique players in our training set. The player recognition classifier is run over the candidate player-regions proposed by the previous module.

(3) Evaluation Criteria: We conduct both automatic as well as human evaluations to validate our approach. For automatic evaluation, we consider BLEU score that has been popularly used by several other relevant works such as [13, 14, 18, 19, 23, 30]. It measures n-gram agreement (precision) between test and reference text, with higher score signifying better performance. For each test video, the scores are averaged over the top five retrieved descriptions, by matching them with the ground truth commentary. As part of human evaluation, we collected judgements from twenty human evaluators with ample tennis exposure. Videos were presented in form 15 sets comprising of 6 videos each. Every evaluator was randomly presented (atleast) two sets and asked to rate both linguistic structure as well as semantics of a predicted description/commentary. The rating was done on a subjective scale of {'Perfect', 'Good', 'OK', 'Poor', 'Flawed'}. These were converted on a likert scale of 1-5, with 5 being 'Perfect' and 1 being 'Flawed'. We report average for both scores.

5.2 Comparison Baselines

The proposed system is benchmarked against state-of-the-art methods from two streams that are popular in predicting textual descriptions for visual data: description/caption generation and cross-modal description retrieval.

(1) Description generation: Since the approaches in this domain are either too generic [13, 14, 18], or designed for images [14], we evaluate by adapting them to our setting. In both [13, 18], a template-based caption is generated by considering a triplet of form (SVO). To compare with this setting, we align the best predicted verb phrase into a template 'player1 − verbPhrase1 , player2 − verbPhrase2'. Since a verb phrase is a combination of an action verb and an object, this template resembles the 'SVO' selection of [13, 18]. To compare with [14], we use the publicly available pretrained model and generate captions for key frames in a video. Since this is a generic approach, the captions generated are nearly similar, and the set of distinct captions is far less than the total number of key-frames. To associate a caption with a video, we pick the one with the highest frequency.



IN, Winner: Federer!!! Federer aims a high kick serve, Delpotro crafts a backhand return, rally, Delpotro backhand slice goes outside court. ✓

IN, Winner: Serena!!! Williams hits a good serve, Zvonareva struggles with it. ✓

IN, Winner: Delpotro!!! Quick serv, Federer returns a quick forehand return, huge rally, Delpotro whips a forehand cross-court winner. ✗

IN, Winner: Tsonga!!! Fine serve, Tsonga crafts backhand return, Djokovic moves closer to net, struggles to execute volley properly. ✗

Figure 7: Success and failure cases: Example videos along with their descriptions. The 'ticked' descriptions match with the ground truth, while the 'crossed' ones do not.

(2) Cross-modal description retrieval: Cross-modal retrieval approaches [26, 30] perform retrieval by matching samples from the input modality with those in the output modality (both of which are represented as feature vectors). While comparing with [30], we consider the best performing variant of their approach; i.e., the one that uses projected features with Euclidean distance as loss function. Note that our approach is also based on retrieving a description; however, it makes explicit use of low-level visual and textual cues unlike cross-modal retrieval approaches. This difference is also evident from the experimental results, where our approach is shown to retrieve better descriptions than [26, 30].

5.3 Results

To validate the utility of our design choices, we discuss performance of various modules in this section. More details on these are provided in the supplementary file.

5.3.1 Performance of Individual Modules

(1) Verb Phrase Recognition: Compared to a straightforward use of state-of-the-art technique [32] to recognize verbs, our verb phrase recognition module performs better by around 15%. This suggests the utility of harnessing relevant cues (dividing a frame into two halves) while working in a domain specific environment.
(2) Smoothing Vs. No Smoothing of Verb Phrase Predictions: In practice, using MRF for smoothing phrase predictions improved average BLEU score from 0.204 to 0.235.
(3) LSI-based Matching Vs. Lexical Matching: Employing LSI technique while matching predicted phrases with descriptions achieves an average BLEU score of 0.235, whereas naïve lexical matching achieves 0.198.

5.3.2 Video Description Performance

Table 2 (left) demonstrates the effect of variations in corpus size on BLEU scores. It can be observed that the scores saturate soon, which validates our initial premise that in domain specific settings, rich descriptions can be produced even with small corpus size. In Table 2 (right), we compare our performance with some of the recent methods. Here we observe that caption generation based approaches [13, 14] achieve very low BLEU score. [1] This attributes to their generic nature, and their current inability to produce detailed descriptions.

[1] Recall that [13, 14] work for generic videos and images, we approximate them (Section 5.2) for comparisons.

On the other hand, cross-modal retrieval approaches [26, 30] perform much better than [13, 14]. Compared to all the competing methods, our approach consistently provides better performance. The performance improvements increase as we move towards higher n-grams, with an improvement of around 50% over [30] for 4-gram. These results confirm the efficacy of our approach in retrieving descriptions that match the semantics of the data much better than cross-modal retrieval approaches. In human evaluation, we achieve an average score of 3.21 for semantics, and 3.9 for structure of the predicted descriptions. The scores reported are on scale of 1-5. Figure 7 depicts some success and failure examples (success means the topmost predicted description matches the ground-truth description).

6 Conclusion

We have introduced a novel method for predicting commentary-like descriptions for lawn tennis videos. 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 is fairly generic and can be adopted to similar situations where activities are in a limited context and the linguistic diversity is confined, however the output description can be semantically rich. Applications of our solution could range from content based retrieval to real life tennis coaching.

References

[1] Tennis Earth - webpage. http://www.tennisearth.com/.

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

[3] Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven J. Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell W. Waggoner, Song Wang, Jinlian Wei, Yifan Yin, and Zhiqi Zhang. Video in sentences out. *CoRR*, 2012.

[4] Savvas A. Chatzichristofis and Yiannis S. Boutalis. Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *ICVS*, 2008.

[5] Rizwan Chaudhry, Avinash Ravich, Gregory Hager, and Rene Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *CVPR*, 2009.

[6] Ivan Laptev Christian Schuldt and Barbara Caputo. Recognizing human actions: A local svm approach. In *CVPR*, 2004.

[7] W. J. Christmas, A. Kostin, F. Yan, I. Kolonias, and J. Kittler. A system for the automatic annotation of tennis matches. In *CBMI*, 2005.

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[9] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[10] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JASIST*, 1990.

[11] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, and Jean Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009.

[12] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *ICCV*, 2003.

[13] 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.

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

[15] Muhammad Usman Ghani Khan and Yoshihiko Gotoh. Describing video contents in natural language. In *Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, 2012.

[16] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In *CVPR*, 2009.

[17] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 2002.

[18] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond J. Mooney, Kate Saenko, and Sergio Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, 2013.

[19] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. In *CVPR*, 2011.

[20] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[21] Mun Wai Lee, A Hakeem, N. Haering, and Song-Chun Zhu. Save: A framework for semantic annotation of visual events. In *CVPR Workshop*, 2008.

[22] H. Miyamori and S. Iisaku. Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In *FG*, 2000.

[23] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[24] G. Pingali, A Opalach, and Y. Jean. Ball tracking and virtual replays for innovative tennis broadcasts. In *ICPR*, 2000.

[25] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 2000.

[26] 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.

[27] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[28] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH*, 2004.

[29] K. Tanaka, H. Nakashima, I Noda, K. Hasida, I Frank, and H. Matsubara. Mike: an automatic commentary system for soccer. In *ICMS*, 1998.

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

[31] Dirk Voelz, Elisabeth Andre, Gerd Herzog, and Thomas Rist. Rocco: A robocup soccer commentator system. In *Proceedings of RoboCup*, 1999.

[32] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action Recognition by Dense Trajectories. In *CVPR*, 2011.