2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Frame Level Annotations for Tennis Videos

Mohak Sukhwani
IBM Research
Bangalore, India
mosukhwa@in.ibm.com

C.V. Jawahar
IIIT Hyderabad
India
jawahar@iiit.ac.in

***Abstract*—Content based indexing is critical to the effective access of the multimedia data. To this end, visual data is often annotated with textual content for bridging the semantic gap. In this paper, we present a method to generate frame level fine grained annotations for a given video clip. Access to the frame level fine grained annotations lead to rich, dense and meaningful semantic associations between the text and video. This in turn makes the video retrieval systems more accurate. We demonstrate the use of probabilistic label consistent sparse coding and dictionary learning with a K-SVD algorithm to generate 'fine grained' annotations for a class of videos – lawn tennis. The algorithm simultaneously learns a classifier and a dictionary to generate the frame level annotations for the tennis videos using available textual descriptions. The utility of the proposed algorithm is demonstrated on a publicly available tennis dataset comprising of tennis match videos from Olympics games.**

I. Introduction

Multimedia sharing sites have gained substantial popularity in recent years. Both video and image contents are changing the ways we interact with the data on web. With the surge in the availibilty of online video content, it is increasingly becoming complex to index and annotate them with appropriate tags. These tags imbibe the contextual and semantic information in the video (or media, generally). They are used to facilitate the media content search and access. Considerable work has been done in the past for the multimedia annotations both for images [7], [17], [22] and videos [19], [5], [8], [3]. The annotation tags produced by such methods are substantially detailed and could be used for retrieval settings. However, in the case of videos we require more *fine grained* annotations (tags gets finer and specific as the interval gets smaller on the temporal axis) to be useful in real-life applications. Few have attempted the problem of tagging and localization in videos [2], [3], [12], [14]. Video tagging and temporal localization based on social knowledge [3], Youtube videos annotation using Flickr images [2] using zero-shot visual similarity of keyframes and images, video shot modeling using multiple instance learning [12], semantic tagging of personal video content using user generated tags in an image folksonomy [14] are some of the examples of previous video annotation attempts. In our present work, we propose an approach to generate fine grained annotations for tennis videos.

Using current retrieval methods, a video search query like 'Serena hits a forehand shot' might (depending on the details captured by annotations) return videos where Serena plays a forehand shot. We propose a method to generate video annotations at 'frame level' granularity (see Fig. 1). In such a case, we even can identify 'time-stamps' where Serena hits a forehand shot. This becomes extremely challenging in a domain independent setting due to innumerable possibilities. Focus on a specific domain confines the output space [19]. We therefore restrict ourself to lawn tennis videos and take a step towards fine grained video annotations. The input in our case is a video and a corresponding *caption*. We associate each frame with a corresponding tag and the annotation which is a collection of similar *action phrases*.

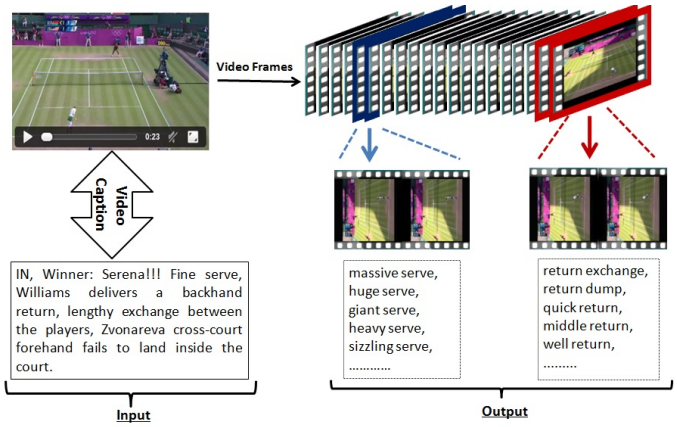


Fig. 1. Given a collection of tennis videos along with the linked captions, our approach generates annotations for constituent frames of the input video. The approach aligns video frames with the corresponding 'action phrases'. Window size (shown in red and blue colors) of 'two' assigns similar labels to the adjacent two frames.

A complementary but essential task to multimedia retrieval is to associate the multimedia content with semantically meaningful captions. Several approaches have been proposed to achieve such associations between text and multimedia content. Amongst all these approaches, the two most popular practices have been either to generate captions using template based NLG techniques [7], [5] or to retrieve from a collection of available descriptions [19], [17]. In retrieval based approach, description is generated by using either image-to-image similarity [17] or video-to-text similarity [19]. Our work is a precursor to methods using text-based similarities over available descriptions for multimedia retrieval. In a domain

*Mohak Sukhwani is presently at IBM Research, India. He did this work as a IIIT Hyderabad student.

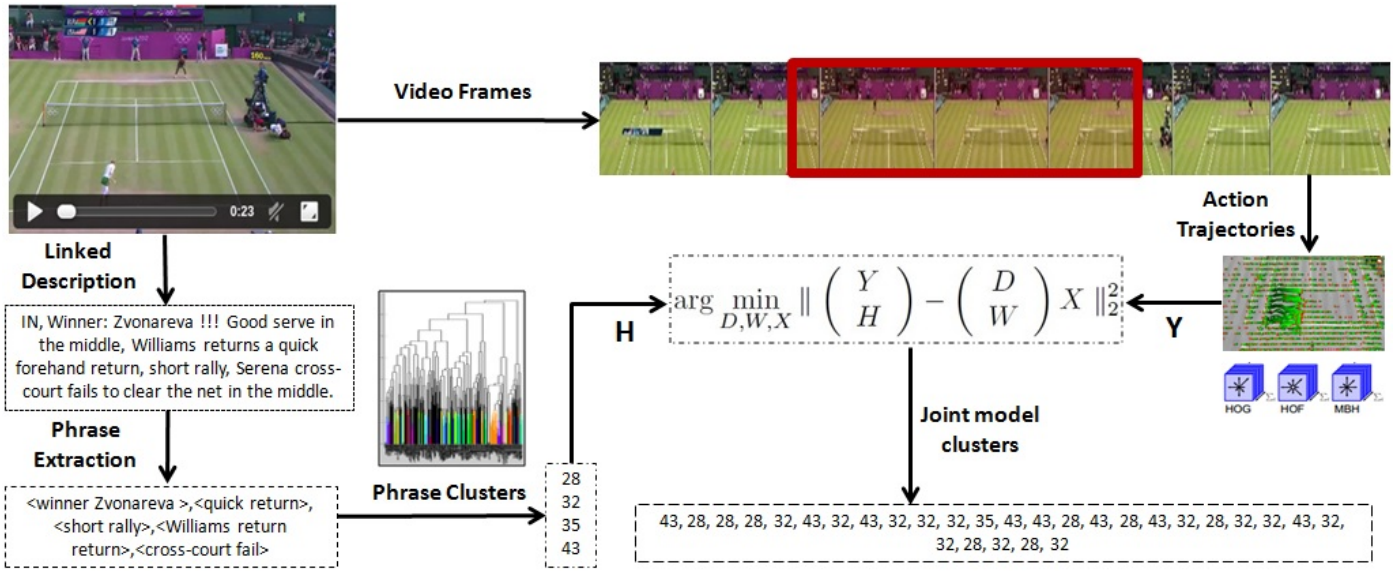


Fig. 2. The proposed approach: Every frame is associated with a corresponding tag which is a collection of similar *action phrases*. Both trajectory matrix (Y) and phrase-cluster matrix (H) are stacked together to compute the dictionary. The dictionary is used to generate frame level annotations for the input videos. We represent each group of 'phrases' by a cluster number.

specific video retrieval setting, our method would assist in better retrieval results as it encodes local temporal structures in tennis videos and aligns the frames with appropriate fine grained annotations. We address the problem of automatically identifying and aligning 'annotations' that best describe the constituent video frames.

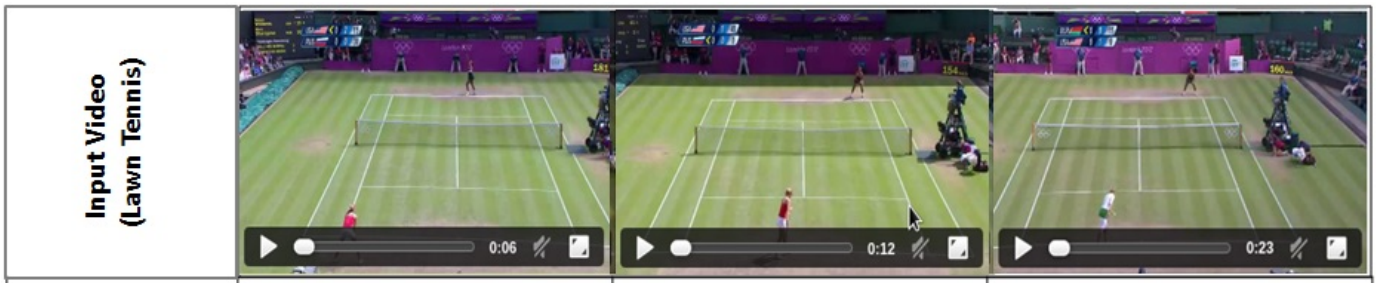
Unlike current video tagging and event detection methods [6], [13] we extract action phrases from the linked captions and simultaneously localize the spatio-temporal actions. Deep Event Network, [6], detects high-level events and localizes the key evidences based on CNNs. Deep networks are utilized in multimedia event detection and recounting [6]. Convolutional Neural Networks (CNNs) have also shown promising results for action classification and recognition [11], [18]. Related to our task is the problem of video classification by identifying key segments and their transitions [20], [21], [16]. These methods focus on high-level event classification and generate recounting for video understanding and automatic tagging. We intend to be more fine grained in our present approach and temporally localize multiple actions in domain restricted setting of tennis videos. Such frame level annotations are useful for computer vision and machine learning based applications. The better the association of the visual data and the textual description, better is the generated model.

Earlier attempts for tennis video analysis and annotations focused on player and action detections [15]. Recent work on tennis commentary generation [19] suggests a more holistic approach of tennis scene understanding. It simultaneously uses vision, language and machine learning techniques to produce semantically rich and human-like descriptions for lawn tennis videos. Our approach bears a resemblance to the task of learning main steps from input narrated instruction videos [1]. We take a step towards weak labelled unsupervised 'action phrase' recognition for tennis videos and suggest a unified objective function to identify prominent action phrases from verbal descriptions. The identified phrases are then aligned to appropriate video frames.

We obtain the alignment by optimising an appropriate objective function. This is done using KSVD which yields similar action phrases for local temporal structures of videos composed of similar actions. We use probabilistic label constrained KSVD for learning sparse dictionary for recognition. The suggested approach (Section II) utilizes the video descriptions to learn varied constructs and associates them with action features extracted from videos. It extracts various linguistic phrases [7] from available sentences and clusters them into groups of semantically similar [9] action phrases, for example phrases like 'huge rally' and 'contested rally' are part of one cluster and 'backhand catches net' and 'Serena catches net' belong to the same cluster. The phrase clusters are thereafter used for classifying action features – similar looking actions are labelled with similar phrase clusters. A group of semantically similar phrases belong to the same phrase clusters and the set of frames encoding similar actions are assigned similar phrase clusters as corresponding labels. Our main contributions in this direction are:

1) a joint model to assign video frames into appropriate phrase bins under weak label supervision, and
2) use of probabilistic label consistent dictionary for fine grained classification.

The next section describes the framework of the fine grained tennis annotations generation. Localized frame level annotations binding the temporal information are the output of the proposed approach. We evaluate our method on a public tennis



Linked Commentary	IN, Winner: Serena!!! Huge serve. Ace !!!	IN, Winner: Sharapova!!! Quick serve, Sharapova crafts a forehand return, Serena goes for a forehand down the line but catches the net	IN, Winner: Zvonareva !!! Good serve in the middle, Williams returns a quick forehand return, short rally, Serena cross-court fails to clear the net in the middle.
Phrases Generated	<winner Serena>, <huge serve> <ace>	<winner Sharapova>,<quick serve>, <Sharapova craft return>, <Serena catch net>,<Serena go>	<winner Zvonareva >,<quick return>, <short rally>,<Williams return return>,<cross-court fail>
Phrase Clusters	4, 23	23, 32, 36, 42, 46	28, 32, 35, 43
Joint Model (output)	4, 23, 23, 4	42, 23, 32, 42, 42,42	43, 28, 28, 28, 32, 43, 32, 43, 32, 32, 32, 35, 43, 43, 28, 43, 28, 43, 32, 28, 32, 32, 43, 32, 32, 28, 32, 28, 32

Fig. 3. Illustration of the approach: First two rows correspond to input videos and the linked descriptions. The extracted phrases and assigned phrase clusters are shown in the next two rows. The last row demonstrates output phrase clusters obtained using the proposed approach (Number of such clusters depends on the duration of an input video). Every cluster index represent the group of similar action phrases for the local temporal structures of the input video.

video dataset [19] in the penultimate section.

II. Fine grained annotations

The video segments of tennis game are input to the proposed approach. We begin by text based phrase clustering which is a precursor to learn dictionary parameters using KSVD. The computed dictionary is then used to generate frame level annotations for the input videos. Fig. 2 summarises the steps involved in our method.

A. Phrase Extraction and Clustering

We automate the process of phrase extraction from a given video descriptions using CoreNLP toolkit [1]. We use 'collapsed-coprocessed-dependencies' [4], i.e. dependencies involving prepositions and conjuncts are collapsed to reflect direct relation between content words. Each sentence is mapped to the list of nine distinct phrase encodings – (subject), (object), (subject;verb), (object;verb), (subject;prep;object), (object;prep;object), (attribute;subject), (attribute;object) and (verb;prep;object). We believe these nine encodings assimilate all possible information in a linked description and input sentence, Fig. 4. Since the generated phrases are targeted to increase the overall retrieval efficiency we keep player names intact in the extracted phrases.

The extracted phrases are clustered using hierarchical agglomerative clustering with Semantic Textual Similarity (STS) measure described in [9]. The similarity scores describe the degree of equivalence in the underlying semantics of paired snippets of text (phrases). The distance metric uses a word similarity feature combining both LSA word similarity and WordNet knowledge. The text clusters computed are visualized using a dendrogram as shown in the Fig. 2. In an agglomerative setting, each phrase starts with its own cluster and subsequently, the pairs of text clusters are merged as one climbs up the hierarchy.

[1] http://nlp.stanford.edu/software/corenlp.shtml

B. Action Phrase Alignment

For action representation, we extract dense trajectory features [23] over space-time volumes (using parameters as in [19]). We use trajectory, HOG, HOF and MBH descriptors to describe the actions and process them similar to [19]. Given a set of tennis videos, we extract action features from both upper and lower part of the frames (capturing actions of both upper and lower player respectively). Action features are computed from neighbouring frames using a sliding window (neighbourhood of size 30 frames with no overlap). The computed action features (from both the halves) are stacked over each other and represent the overall feature vector of the frames embodied in each sliding window. We aim to classify the stacked action feature vector into the phrase bins computed in Section II-A.

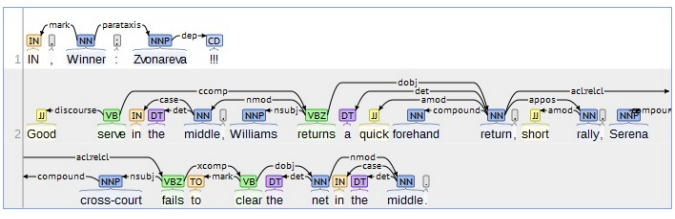


Fig. 4. Parsed dependencies (collapsed and propagated) for an input commentary text. We only keep nine selected encodings and discard others to assimilate all possible phrase information in the linked sentence.

C. Dictionary Learning

A compact and a discriminative dictionary is learnt for sparse coding which is later used for classification. Let, Y be a matrix of stacked video features (from the upper and the lower halves) of the training videos (collection of N sliding windows). A single training video is represented by the consecutive columns of Y and the number of such columns equate to the count of sliding windows encompassing the whole video. Every column represents feature vector of frames in one sliding window, each of n dimensions, i.e., $Y = [y_1, y_2, \ldots, y_N] \in R^{n \times N}$. We learn a single re-constructive dictionary, D, with K items for sparse representation of Y:

$$< D, X >= \arg\min_{D,X} \| Y - DX \|_2^2, s.t. \forall i, \| x_i \|_0 \leq T \quad (1)$$

Here, $D = [d_1, d_2 \ldots, d_K] \in R^{n \times K}$ is the learnt dictionary, $X = [x_1, x_2 \ldots, x_N] \in R^{K \times N}$ are sparse codes of Y and T is sparsity constraint factor (each sparse code has at-most T items). We leverage the supervised information (i.e. phrase labels obtained after text-based hierarchical agglomerative clustering) of the feature vectors to learn an efficient dictionary. In order to make dictionary optimal for the classification task, we include the classification error term in the objective function. Similar to [10], we use a linear predictive classifier $f(x; W) = Wx$ and learn the weights, W. The following objective function for learning a dictionary, encompasses both reconstruction and classification errors:

$$< D, W, X >= \arg\min_{D,W,X} \| Y - DX \|_2^2 + \| H - WX \|_2^2, \quad s.t. \forall i, \| x_i \|_0 \leq T \quad (2)$$

The above function encodes an explicit correspondence between the dictionary items and the phrase labels. The term $\| H - WX \|_2^2$ represents the classification error, with W being the classifier (weights) parameters. $H = [h_1, h_2 \ldots h_N] \in R^{m \times N}$ are the class labels of Y. A video (represented by consecutive columns of Y) can have multiple (extracted) phrases, so we assign equal probabilities to the corresponding phrases in each column of H. Number of columns in H for each video depend on the number of features columns in Y for each video, Fig. 5.

A video with 90 frames will constitute $(90/30) = 3$ columns (considering sliding window size of 30 frames) in matrix Y and H; y_i corresponds to stacked dense trajectory feature of 30 frames and h_i represents equal probabilities of phrase labels (all three corresponding columns in H will be identical). $h_i = [0, 0.33, 0.33, 0, 0.33, \ldots, 0, 0]$ would mean that phrases identified from linked description belong to phrase clusters $2, 3, 5$ (non-zero entries in h_i). We assign each phrase cluster an equal probability $(1/3 = 0.33)$. Intuitively this would mean the input video is comprised of these three cluster labels and hence classification error should be minimized with respect to these three labels. We use KSVD to find the optimal solution for all the parameters simultaneously. The objective function can be re-written as:

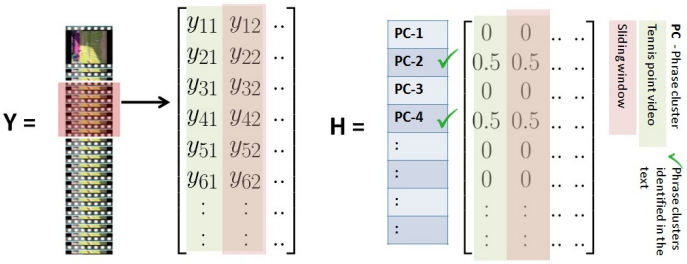


Fig. 5. Matrix structure: Considering the sliding window size of 'n' frames, a video with m frames ($n < m$) will constitute (m/n) columns in matrix Y and H; Each column of Y corresponds to the dense trajectory feature of frames in sliding window and each column of H represents equal probabilities of phrase cluster labels detected in the input linked text (m/n adjacent columns in H are identical).

$$< D, W, X >= \arg\min_{D,W,X} \| \begin{pmatrix} Y \\ H \end{pmatrix} - \begin{pmatrix} D \\ W \end{pmatrix} X \|_2^2, \quad s.t. \forall i, \| x_i \|_0 \leq T \quad (3)$$

$$< D^*, X >= \arg\min_{D^*,X} \| Y^* - D^* X \|_2^2, s.t. \forall i, \| x_i \|_0 \leq T \quad (4)$$

Here, D^* represents the final dictionary with optimal reconstruction and classification error – the features from same class have similar sparse codes and those from different classes have dissimilar sparse codes. The dictionary is subsequently used for action classification.

Initialization: We initialize the parameters D_0 and W_0 similar to the way suggested in [10]. We use the multivariate ridge regression, with quadratic loss and L_2 norm regularization.

Classification: We compute D and W using the KSVD algorithm. Both D and W are transformed and normalized before classification. For a given video, we compute sparse representation (x_i) of dense trajectories action features (y_i) using modified dictionary, $\hat{D}$:

$$x_i^* = \arg\min_{x_i} \| y_i - \hat{D} x_i \|_2^2, s.t. \forall i, \| x_i \|_0 \leq T \quad (5)$$

Thereafter, we use the linear predictive classifier $\hat{W}$ to estimate the labels of frames:

$$\arg\max_j (l_j) \ , \ l_j = \hat{W} x_i^* \quad (6)$$

These labels correspond to the phrase cluster ids to which given group of frames belong. Every frame of the input video would have a tag associated with it. These tagged associations add a textual meaning to the video frames. Both video frames

and the available textual descriptions are thus co-clustered to generate finer annotations.

III. Experiment and Results

We restrict our attention to the singles game of lawn tennis and use 314 'tennis-points' videos from 'Video-commentary' dataset of lawn tennis dataset introduced in [19]. A 'tennis-point' begins with the start of the service and ends when a scoring criteria is met. Each video in this dataset has a corresponding commentary text aligned to it. In all there are 45K video frames with an average length of each video being 147 frames. Videos of lawn-tennis matches have two players – one on each side of the net (approximately at the center of every frame). Similar to [19], we analyse the videos by dividing them across the center net and compute features for upper and lower parts separately.

For the experiments, we partition the dataset into 60% train and 40% test data. The video dataset is collection of broadcast video recordings of matches from London Olympics 2012. Each video has a linked textual description describing its content. The videos used are all of resolution 640×360. The training set is used to learn the dictionary and other model parameters. At the time of text based phrase clustering, we use all available linked descriptions. While extracting the phrases from the available descriptions, we replace all possible words with respective synonyms determined using WordNet synsets [7]. Overall, we have 50 phrase clusters during the text based phrase clustering – these clusters determine the groups and the bin to which a description would belong. Number of clusters (i.e. phrase bins) were empirically selected on the basis of the most relevant qualitative results with experiments being performed over $25, 30, 40$ and 50 bins.

We compute the action phrases and the clusters as described in Section II-A. Dense trajectory features [23] are used as action features – trajectory, HOG, HOF and MBH descriptors

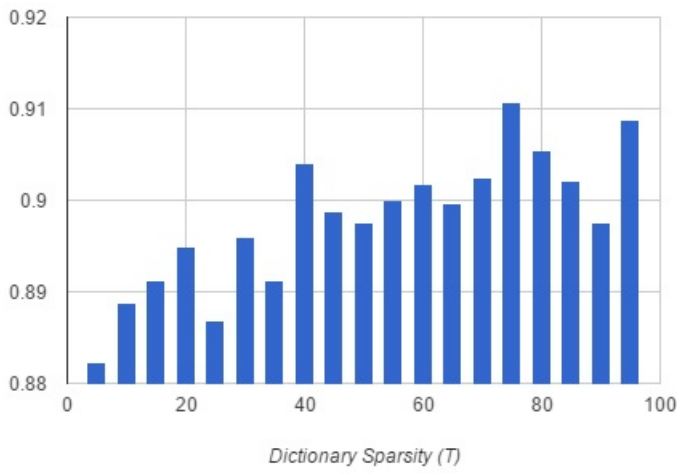


Fig. 6. Correct match accuracy: Initial phrase clusters (computed using only linked description) act as ground truth and are compared to clusters computed using the proposed approach.

Clusters (Only Description)	Aligned Clusters (Joint Model)
4,23	4,23,23,23
2,16,26,32,36,43	16,43
4,9,16,23,40	23,4,23,16,4,40,4,16,4,9
23,29,48	23,29
4,15,23,26,29,32	23,32,29,23,15
4,23	23,4,23,23,23,4,23,4,4,23,23,2,2,23, 18,23,18,49

TABLE I
Qualitative result for the assigned clusters: Clusters computed using the proposed solution bind temporal information and depend on the length of the video. The localized phrase clusters obtained using a joint model are richer and align themselves to respective frames of input tennis video.

describe the actions from both upper and lower part of the frames. Initial phrase clusters are used to evaluate the proposed approach. One should note that the count of phrases generated and number of the phrase clusters may differ. In column 3 (Fig. 3), '⟨ quick return ⟩' and '⟨ Williams return return⟩' belong to same clusters. We determine the final output of the proposed system by using classifiers described in section II-B. Number of such assigned phrase clusters depend upon the size and length of the (test) video and not on the size of the linked descriptions. This is evident from all the examples shown in Fig. 3.

To evaluate the temporal localization, we should have a one-to-one correspondence map between the identified phrase clusters in the video and the ground truth alignment. Owing to the lack of such data we demonstrate the effectiveness of the proposed system by comparing it with clusters obtained using only text-based model. Table I illustrates the qualitative improvement in phrase alignments obtained using a joint model. The clusters obtained using joint modelling bind the temporal information and depend on the length of the input video, which is not the case with text based modelling, Fig. 7. Standalone text based clusters depend only on the input descriptions. We consider the clusters computed using only descriptions as the ground truth and compare them to the clusters obtained using our proposed solution – an overlap of one cluster is considered as a match. Fig. 6 illustrates the match between ground truth and the computed clusters.

Discussions

Frame level fine grained annotations imply that the annotation interval gets smaller on a temporal axis and every frame has an annotation of its own. Frame-by-frame annotations are important for the tennis videos because such videos are composed of long shots, which in turn are a collection of many fine grained actions. These frame level annotations bind constituent frames of an input video to a textual tag. In one of our past works [19], we explicitly focused on integrating fine grained details into an overall descriptions for an input tennis video. The annotation descriptions generated in [19] were distinct for every input video but were significantly detailed

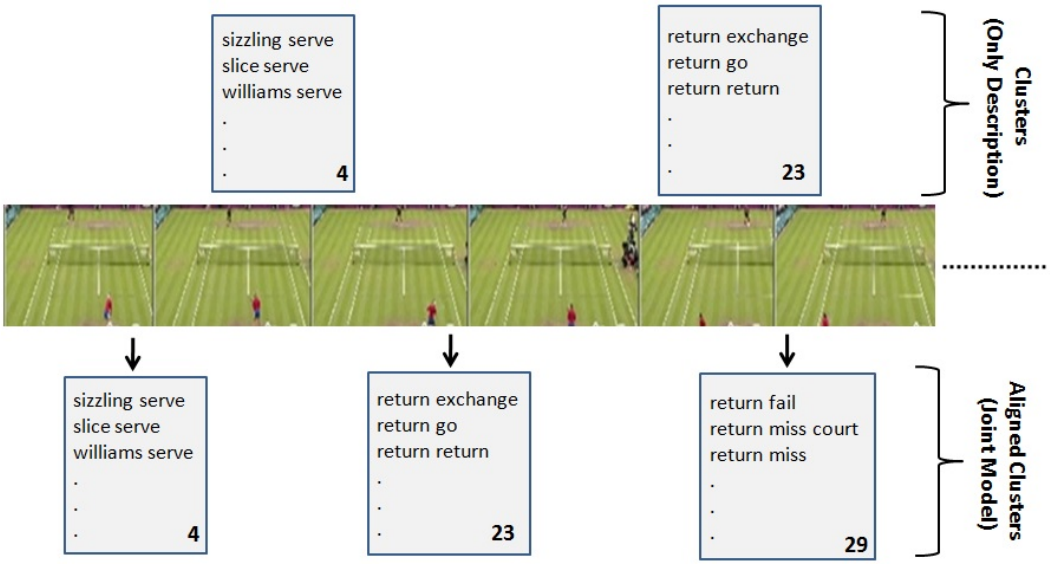


Fig. 7. Frame level annotations for the input video: The figure illustrates the difference between the clusters obtained using only the text descriptions and the clusters obtained using the proposed approach. The phrase clusters computed by the proposed approach are richer and bind temporal information within them. The arrows represent the binding of the phrase clusters with the frames. The numbers at the bottom-right corner depict cluster-ids of the phrase clusters generated.

and human like. In our present work the focus is to bind every frame with a corresponding tag.

IV. Conclusions

We describe an unsupervised approach to identify action phrases in domain specific tennis settings and temporally localize them in the given tennis videos. The complementary nature of both the video and the associated text is used to demonstrate a method that sequentially solves text and video clustering problems linked by joint constraints to obtain the frame level annotations. We co-cluster the video frames and the available textual descriptions in to generate 'fine grained' annotations. This could be useful for plethora of applications ranging from fine grained cross modal retrieval systems to modern day text based video summarization systems.

References

[1] J.-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien. Learning from narrated instruction videos. In *arxiv 1506.09215*, 2015.

[2] L. Ballan, M. Bertini, A. Del Bimbo, M. Meoni, and G. Serra. Tag suggestion and localization in user-generated videos based on social knowledge. In *SIGMM Workshop*, 2010.

[3] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Enriching and localizing semantic tags in internet videos. In *ACMMM*, 2011.

[4] M.-C. De Marneffe and C. D. Manning. The stanford typed dependencies representation. In *COLING Workshop*, 2008.

[5] D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. Hauptmann. Beyond audio and video retrieval: Towards multimedia summarization. In *ICMR*, 2012.

[6] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, 2015.

[7] A. Gupta, Y. Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.

[8] A. Habibian, T. Mensink, and C. G. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACMMM*, 2014.

[9] L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese. Umbc ebiquity-core: Semantic textual similarity systems. In *SEM*, 2013.

[10] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *CVPR*, 2011.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[12] G. Li, M. Wang, Y.-T. Zheng, H. Li, Z.-J. Zha, and T.-S. Chua. Shottagger: Tag location for internet videos. In *ICMR*, 2011.

[13] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney. Video event recognition using concept attributes. In *WACV*, 2013.

[14] H. Min, J. Choi, W. De Neve, Y. M. Ro, and K. N. Plataniotis. Semantic annotation of personal video content using an image folksonomy. In *ICIP*, 2009.

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

[16] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*. 2010.

[17] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2Text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[18] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[19] M. Sukhwani and C. V. Jawahar. Tennis Vid2Text : Fine-grained descriptions for domain specific videos. In *BMVC*, 2015.

[20] C. Sun and R. Nevatia. Discover: Discovering important segments for classification of video events and recounting. In *CVPR*, 2014.

[21] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[22] Y. Verma and C. Jawahar. Image annotation by propagating labels from semantic neighbourhoods. 2016.

[23] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.