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

Segment-based Models for Event Detection and Recounting

Rama Kovvuri Ram Nevatia
University of Southern California
Los Angeles, USA
Email: {nkovvuri,nevatia}@usc.edu

Cees G.M.Snoek
University of Amsterdam
Amsterdam, The Netherlands
Email: cgmsnoek@uva.nl

***Abstract*—We present a novel approach towards web video classification and recounting that uses video segments to model an event. This approach overcomes the limitations faced by the classical video-level models such as modeling semantics, identifying informative segments in a video and background segment suppression. We posit that segment-based models are able to identify both the frequently-occurring and rarer patterns in an event effectively, despite being trained on only a fraction of the training data. Our framework employs a discriminative approach to optimize our models in distributed and data-driven fashion while maintaining semantic interpretability. We evaluate the effectiveness of our approach on the challenging TRECVID MEDTest 2014 dataset. We demonstrate improvements in recounting and classification, particularly in events characterized by inherent intra-class variations.**

I. Introduction

User generated videos have been growing at a rapid rate. These videos typically do not come with extensive annotations and metadata; even category level labels may be missing or noisy. For efficient retrieval and indexing of such videos, it would be useful to have automated methods that classify a video into one of the known categories but also to identify key segments and provide semantic labels for them to enable rapid perusal and other analyses. Given an input video, our framework provides a user-defined event label (detection) and positive evidence for the same with their locations and labels (recounting).

The tasks of detection and recounting are challenging due to large intra-class variances in structure, imaging conditions and possible presence of long segments not directly related to the event. As shown in Figure 1, a video with caption ”Marriage Proposal” can contain various backgrounds such as a ”restaurant”, ”basketball court” and ”outdoors”. However, effectively identifying instances such as ”Getting down on one knee”, ”Proposal speech” and ”Wearing a ring” can help in identifying the event despite the variations.

Popular approaches to model events can be divided into holistic or part-based. Holistic approaches (*e.g.*, [2] [3]), model an event using distributions of low-level features from various modalities such as appearance, scene, text, motion of its constituent videos. It is common to encode features using Fisher Vectors [4] which are aggregated using different type of pooling [5] such as max and average pooling. While these approaches achieve reasonable performance for detection, they do not identify positive segments or provide semantic interpretation of the results needed for tasks like recounting. Also, they work well for videos that are *trimmed*, *i.e.*, where almost the entire video corresponds to a single event category. Other methods [6] [7] have tried to use semantic features by computing concept scores using a dictionary of object [6] and/or action detectors [8] applied to each frame and aggregating the scores. While these methods can provide some semantic interpretation of the video, by emphasizing the high-scoring concepts [7] [9], their utility for localization and recounting is still limited. There are also difficulties like the concept dictionaries may not be well matched to the concepts in the video and concept detectors may not perform uniformly across datasets [10], which may be considered as problems of ”transfer learning”.

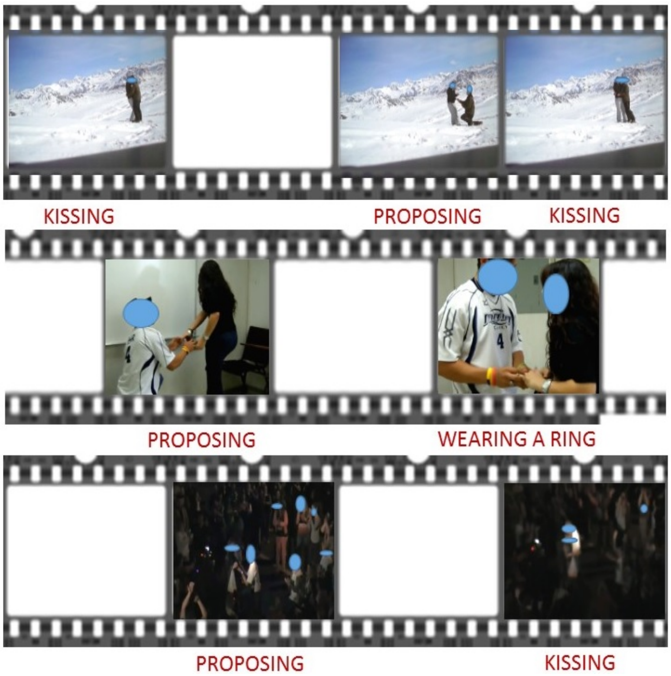

Fig. 1: Exemplars from the event ”Marriage Proposal” in the TRECVID MED Dataset [1] showcasing the variance in backgrounds; actions and their order of occurrence in unconstrained videos.

Part-based approaches use video-segments instead of entire videos for event models. For example, [11] represents an

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

event using a set of models from various temporal scales for human activity classification. Temporal structure of classifiers is embedded as a template and the models are learnt by mining iteratively for positive segments of motion features. While this approach captures the intra-class variance, it cannot be applied to web videos due to lack of temporal structure unlike human activity. [12] splits a video into fixed-length temporal segments and employs a variable-duration HMM to model the state-variations in the segments. Latent models are used to infer the temporal composition of a video. This method performs well on action datasets but unlikely to handle the variations in web videos due to rigid temporal constraints. [13] proposed a joint framework for detection and recounting where the positive segment locations are treated as latent variables. Their method uses global video model and part-segment models based on concept dictionary in conjunction to optimize for event classification and recounting. While this approach gives significant improvement over methods using only semantic features, it is still limited by the concept dictionary.

In our approach, we employ video-segments instead of entire videos for training event models. While it is impractical to provide large numbers of positive video exemplars to model each event, each video exemplar provides tens to thousands of video segments whose positive instances can be utilized to model an event. If the positive segments are identified and clustered, they can be used to discard the significant amount of ”outliers” or ”non-informative” segments found in unconstrained videos and structurally highlight the semantically meaningful parts of video. If the positive segments are labeled, they can also be used for tasks such as recounting of the videos. We train ensembles of models to depict the sub-categories of an event using the positive segments from video exemplars. We allow for data-sharing while training our models, enabling them to use segments not just from training data but from background segments which helps to overcome limited data which is common in long-tail distributions.

We use knowledge transfer from detectors of external concept dictionary only for initialization and concepts (subcategories) of an event are trained by mining groups of positive segments from exemplar videos themselves in a weakly supervised fashion. Unlike [11], which uses augmented initial models from various scales, this form of initialization has more semantic interpretrability of the models and higher incidence on positive segments. We also do not attempt to assign label to each segment of the video or model temporal composition of the constituent events, unlike [11] [12], and it is more amenable to unstructured videos.

Our overall framework is represented in Figure 2. Given a set of exemplar videos for each event, we first divide each video into fixed-length, non-overlapping segments and use responses of concept detectors to sample possible positive segments. We use the sampled segments as initial seeds and use iterative positive segment mining to group similar segments. From the resulting groups of segments, we train an SVM (”candidate” segment-based models) for each group. The contributions of each of the ”candidate” segment-based models towards event are evaluated in the next step using a greedy strategy. The top contributing ”candidate” segment-based models (”representative segment-based models”) are chosen to represent the event. For testing, we score all the segments of test video using the ”representative segment-based models” of an event and aggregate the scores. Final video-level score is obtained by averaging the scores from segments with top responses. Following sections contain a detailed description of our method.

We show both qualitative and quantitative results on the challenging MEDTest 2014 [1] dataset provided by NIST for classification and recounting tasks respectively.

II. Segment-based Models

A. Seed Initialization

For training efficient segment-based models, we need an initialization scheme that can identify a subset of representative positive segments as seed segments. For this, we take advantage of the observation that highest responses of the top contributing concepts of an event are highly relevant to the event [6]. We use a mid-level feature representation to select concepts that are relevant to the event and then choose the segments that have high scores for these concepts. Note that some of the initial seed segments can be noisy and are pruned in the latter stages of training. Given a set of videos $\mathbb{V} = \{ V_1 \ldots V_N \}$ belonging to an event C, let each video V_i contain F_i frames. Given K concept detectors $\{D_1 \ldots D_K\}$, each concept detector is applied to $\mathbb{V}$. For each video segment $V_i(f_i)$, a K dimensional response vector $V_i^s(f_i)$ is obtained. We select top L relevant concepts per event C by computing the sum of l_1 normalized feature response vectors as follows : $\beta = \sum_{i f_i \in \mathbb{V}(*)} V_i^s(f_i)$. The top L concepts are then selected as $\underset{1<L<K}{max_L}\, \beta(k), k \in \{1, 2, ...K\}$. For each concept c_k, we choose positive segments $\{ V_{t_1}(f_{t_1}), \ldots V_{t_P}(f_{t_P})\} \in \mathbb{V}(*)$ that have the P highest responses for c_k in $\mathbb{V}(*)$. To obtain genuine maximal responses without redundancy, we apply non-maximal suppression.

We choose L and P values to be relatively high, to ensure that the initial seeds are oversampled. This way, the representation of an event can be exhaustive when the models are pruned in the latter stages. From the $L{*}P$ initial segment seeds chosen, we build candidate segment models $M_i, i \in C$. Each seed is used as a positive example and hard mined negative segments from background set are used as negative examples, to train exemplar SVMs [14].

B. Iterative positive segment mining

The candidate models trained in the previous step tend to over-fit to the single exemplar they are trained on. To generalize the models further, we need to retrain the models with positive segments similar to the exemplar. To avoid the problems faced by classical clustering models, we choose to mine the positive segments iteratively in a discriminative space. In each iteration, we group N_p positive examples that show high response to the current model. We then retrain

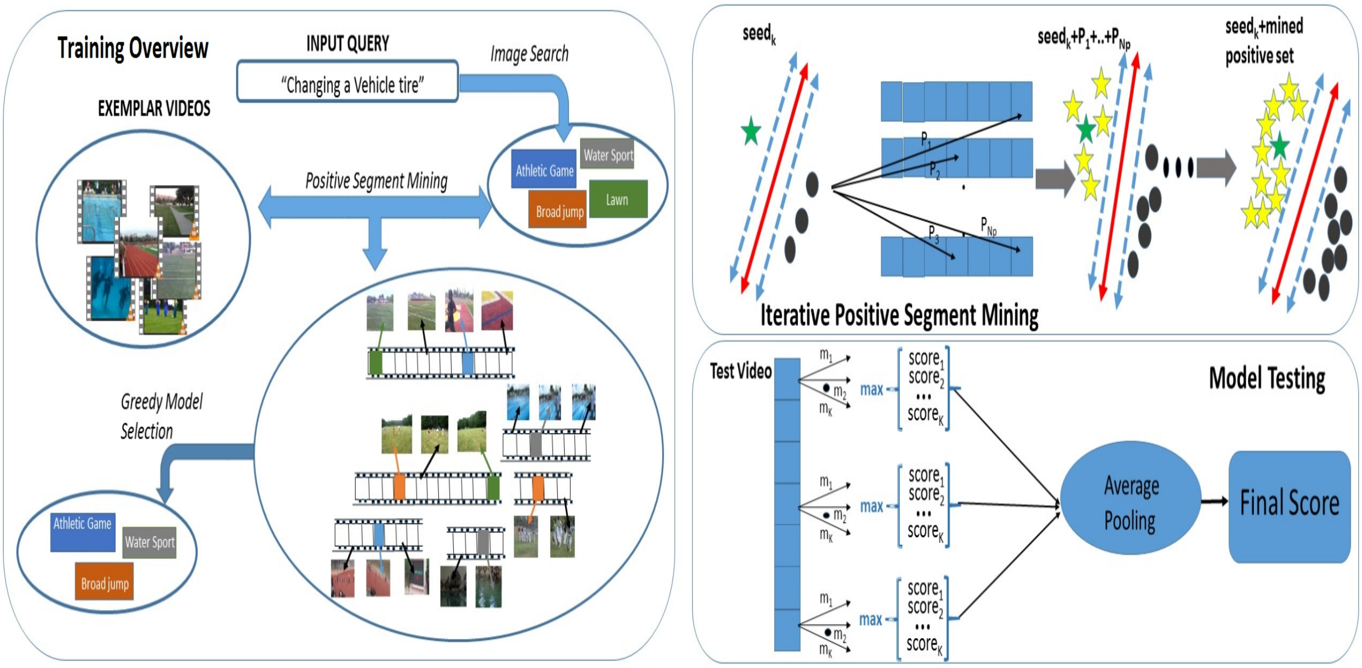

Fig. 2: Outline of the training, iterative positive mining and testing approaches. Training approach (left) selects relevant concepts for an input query and uses iterative positive mining to select segments and greedy selection to prune them. Iterative positive mining approach (right-up) trains an exemplar SVM [14] using the initial seed segment and iteratively mines for similar positive segments to generalize. Testing approach (right-down) generates scores for videos using max responses from the selected models and final score is generated by averaging the highest local responses.

the current model by including the mined positive examples in the exemplar set. This form of mining helps in learning more reliable templates by using the mined samples as a form of "regularization" to prevent overfitting and models long-tail distribution naturally [15]. It is also advantageous in discarding outliers, since there is no requirement for a sample to be bound to a cluster. Choice of each group is independent of the other groups' choices and can be trained in parallel for efficiency. The algorithm is iterative and alternates over the following two steps until it reaches convergence.

(i) Each candidate model M_i scores all the positive segments for an event and mines the top scoring N_p segments

(ii) Each candidate model re-trains to include the N_p mined positives with the existing positive set to improve the generalization of the candidate model.

Convergence of the algorithm is judged based on the Average Precision (AP) value of the candidate model, on a held-out validation set. The iteration is terminated when there is a marginal improvement in AP or when enough positive examples are mined, whichever happens earlier. Many of the candidate segment models, M_i, trained in this step are either noisy or redundant and need to be further pruned to build a representative set for each event.

C. Model Selection

From the pool of $|L| * |P|$ candidate models for each event, we need to select a subset $S \subset \{|L| * |P|\}$, that is representative of the event. Many of the candidate models are redundant due to over-sampling. So, the subset, S is chosen to maximize the mean Average Precision (mAP) on the training set excluding the positive segments used for training and their neighbors. Since the search over the entire subset space has high computational complexity, we opt for a greedy algorithm to choose the final representative models, $\widetilde{M_i}$, which works quite well in our experiments. We tried both the greedy model selection and greedy model elimination strategies to select the subset. We observe that, greedy selection gives similar performance as greedy elimination while being computationally faster. At each step, we add a model, $\widetilde{m_i^*}$ that maximizes the AP of the existing subset S. We use early stopping to prevent over-fitting.

D. Model Testing

Testing the segment-based models is different from video-level models primarily in two aspects. Firstly, since segment-based models are trained on the discriminative segments they are expected to have low responses for non-discriminative and outlier segments. This results in sparse high detection scores across the video segments. Averaging across the segments would result in a very low and noisy final score. Secondly, since each event C could be represented by more than one segment-based model,$\widetilde{M_i}$ from representative set, detection scores of various models for a segment have to be aggregated to obtain detection score for that segment. However, the detection scores of the models are not comparable and need to be calibrated across each other in probabilistic space.

To calibrate the detection scores of representative segment-based models, $\widetilde{M_i}$ of an event C, we use a held-out validation

ID	Event Name	Video-level Models(VM) [6]	ELM [13]	Segment-based Models(SM)	VM+SM
21	Bike trick	0.0653	**0.0912**	0.0778	0.0696
22	Cleaning an appliance	0.0856	0.0910	0.1028	**0.1272**
23	Dog show	0.7729	0.6853	0.7840	**0.8194**
24	Giving direction	0.1093	0.1296	**0.1313**	0.1244
25	Marriage proposal	0.0208	**0.0459**	0.0266	0.0371
26	Renovating a home	0.0690	0.0673	0.0593	**0.0802**
27	Rock climbing	0.0812	**0.0889**	0.0850	0.0850
28	Town Hall meeting	0.3855	0.3674	0.4447	**0.4840**
29	Winning a race without a vehicle	0.2543	0.2978	0.2989	**0.3041**
30	Working on a metal crafts project	0.1032	**0.2186**	0.1238	0.1237
31	Beekeeping	0.7367	0.7532	0.73855	**0.7565**
32	Wedding shower	0.2545	0.2790	**0.2793**	0.2894
33	Non-motorized vehicle repair	0.2712	**0.3070**	0.2774	0.2841
34	Fixing musical instrument	0.4575	0.4067	0.4124	**0.4686**
35	Horse-riding competition	0.3534	0.3323	0.2782	**0.3842**
36	Felling a tree	0.1774	0.1952	0.2141	**0.2238**
37	Parking a vehicle	0.1719	0.1802	**0.2791**	0.2678
38	Playing fetch	0.0906	**0.0984**	0.0749	0.0842
39	Tailgating	0.2066	**0.2132**	0.1889	0.2001
40	Tuning a musical instrument	0.0781	0.1484	**0.2026**	0.1938
mAP		0.2373	0.2498	0.2540	**0.2704**

TABLE I: Comparison of MED performance (AP metric) on the NIST MEDTEST 2014 dataset using Video-level Models, Segment-based Models and Late Fusion.

set (VS) to mine non-redundant top P_s scores from positive segments V_j^s, $V_j \in C$ and a background set to mine N_s hard-negative scores. A learned sigmoid $(\alpha_{\widetilde{M_i}}, \beta_{\widetilde{M_i}})$ is then fit to each model, $\widetilde{M_i}$ and the detection scores $x_j = Sc(V_j^s, \widetilde{M_i})$ are rescaled to be comparable to each other as follows:

$$f(x_j \mid w(\widetilde{M_i}), \alpha_{\widetilde{M_i}}, \beta_{\widetilde{M_i}}) = \frac{1}{1 + e^{-\alpha_{\widetilde{M_i}}(w(\widetilde{M_i})^T x_j + \beta_{\widetilde{M_i}})}}$$

This calibration step also suppresses the responses of models that do not have high distinction in positive and negative scores by shifting the decision boundary towards the exemplars [14]. The final detection score, X_j, for each segment V_j^s is then obtained by max-pooling the calibrated scores,$f(x_j)$, of all the representative segment-based models, $\widetilde{M_i}$ of the event C.

$$X_j = max(f(x_j|\widetilde{M_i})), x_j = Sc(V_j^s, \widetilde{M_i})$$

Once the detection score for each segment is calculated, a video-level score is obtained by averaging the scores of local maxima of the video.

$$Sc(V_j) = avg(max_k(X_j)), X_j = Sc(V_j^s)$$

Non-redundancy of the scores is achieved through non-maximal suppression while averaging suppresses noisy responses.

E. Model Recounting

For identifying the positive evidence, we take the segments with local maxima scores.

$$Ev(V_j) = \{max_k(X_j)\}, X_j = Sc(V_j^s)$$

The corresponding labels of the positive evidence are identified by choosing the labels of the representative models that have the maximum scores for the segments with local maxima.

III. Experiments

In this section, we provide details about the dataset we used, various choices of parameters and evaluate the performance of our segment-based models.

A. Dataset

In our experiments, we use TRECVID MED14 [1] test video corpus and MED 14 event kit data for evaluation. The dataset contains unconstrained, Youtube-like web videos from the Internet consisting of high-level events. The MEDTest 14 has around 27,000 videos and the event kit consists of a 100Ex setting, providing approximately 100 exemplars per event. The "event kit" consists of 20 complex high-level events differing in various aspects such as background : outdoor (bike trick) vs indoor (town hall meeting); frequency : daily (parking a vehicle) vs uncommon (beekeeping); sedentary (tuning a musical instrument) vs mobile (horse-riding competition). A complete list of events is provided in table I.

B. Object Bank

For mid-level features, we choose an Object Bank [6] containing 15k categories of ImageNet. Each category is trained using a convolution network with eight layers and error back propagation. The responses for each category are obtained for each frame and the 15k dimensional vector is simply averaged across frames to obtain segment level and video level representations. The 15k objects are noun phrases that encapsulate a high diversity of concepts such as scenes, objects, people and activities.

C. Evaluation

1) Training parameters: For training the segment-based models, the first parameter choice is the number of initial seed models$(K * M)$. For the value of (K), a performance plateau was reached for $K = 50$. For M, lower values led to poor performance due to noisy estimates of the object bank, while higher values led to high redundancy in the initial seeds. $M = 5$ was chosen for our experiments. For discriminative clustering, $N_p = 10$ was used for collecting positives in each

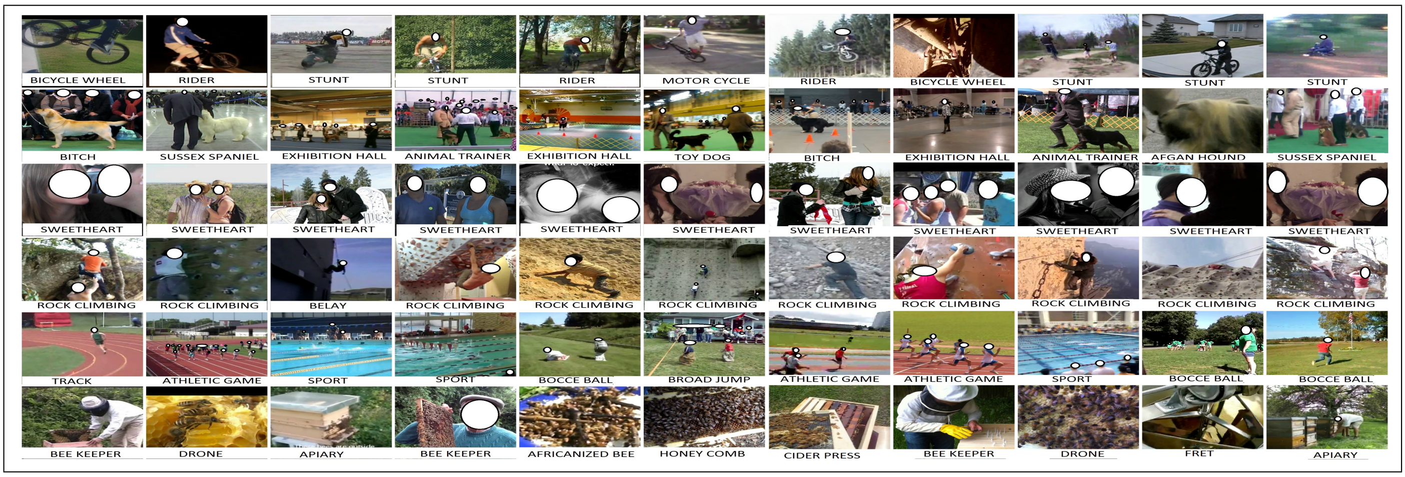

Fig. 3: Captions/labels generated by segment-based models for events Bike Trick, Dog Show, Marriage Proposal, Rock Climbing, Winning a race without a vehicle and Beekeeping (from top to bottom). The first ten out of the twenty positive test videos of the event are chosen and the middle frame of the segment is chosen for illustration. It can be seen that the captions are relevant to the segments.

ID	Event Name	Threshold 1 = 0.5		Threshold 2 = 0.7		Threshold 3 = 0.9		Average	
		VM	SM	VM	SM	VM	SM	VM	SM
23	Dog show	0.9612	0.9668	0.9082	0.8974	0.7619	0.7778	0.8771	**0.8806**
25	Marriage proposal	0.2801	0.3118	0.2726	0.2944	0.2686	0.2913	0.2737	**0.2991**
27	Rock climbing	0.7322	0.7506	0.7299	0.7480	0.7153	0.7351	0.7258	**0.7381**
29	Winning a race without a vehicle	0.6684	0.6841	0.6661	0.6818	0.6004	0.6580	0.6449	**0.6746**
mAP		0.6604	**0.6783**	0.6442	**0.6554**	0.5865	**0.6155**	0.6304	**0.6462**

TABLE II: Comparison of Average Precision(AP) of the ranked segments in test videos for Video-level Models(VM) and Segment-based Models(SM) for various thresholds.

iteration and at this rate most of the models stabilize in 3-4 iterations (30-40 exemplars). A maximum iteration limit of 20 (~200 exemplars) is set for the clustering, with most of the models reaching convergence far before except the highly noisy ones. For training and validation, we use a split of 67%-33% on the training videos.

We use less than 1% of the available training segments to train all the events, showing the efficiency of our training procedure. Some events such as ”dog show” were efficiently represented with a single model. This indicates that if the events have low intra-class variance, representation is possible with very few models.

2) Multimedia Event Detection: We compare the performance of our segment-based models with a standard video-level model using the object bank features [6] and evidence localization model (ELM) [13]. For [6], we use a histogram intersection kernel SVM [16] to model the event and logistic regression based fusion when combining the two modalities. For [13], latent svm is used on the object bank that models both global and part-based models. A summary of the results per event is provided in Table I. For majority of events, AP of the segment-based models is better than the AP of the other methods, while late fusion with video-level models improves the performance significantly indicating some complementarity of ”modeling-segments” to ”modeling-context”. Note that AP of segment-based models is similar to that of ELM which uses both global and part-based models. Hence, a better comparison is with fusion results which are better than that of ELM. Also ELM is relatively slow as it uses a latent SVM for inference. Events such as ”Winning a race without vehicle”(running, swimming, potato race) and ”Tuning a musical instrument”(guitar, key board, snare drum) improve considerably, indicating that events that contain natural sub-categories are modeled more accurately using segment-based models. Sometimes, lack of sufficient data to model events leads to drop in performance as in the case of event ”horse-riding competition”, where the segment-based models produce high scores in the test videos that have strong incidence of horse, race track or jockey but they perform poorly when the race occurs in a grassy surface and horses appear in a very low resolution where incidence is on poorly trained ”paddock” model.

3) Multimedia Event Recounting: Multimedia Event Recounting (MER) generates summary of key evidence for event of a video, by providing when(event interval) and what(evidence label) and confidence of the evidence segments. To evaluate the performance of segment-based models for MER, we use the annotations provided by NIST for positive MEDTest videos of 4 events. The annotations provide the probability that a video segment belongs to the event. We use various thresholds to categorize the test segments into positive/negative for the event and report the Average Precision

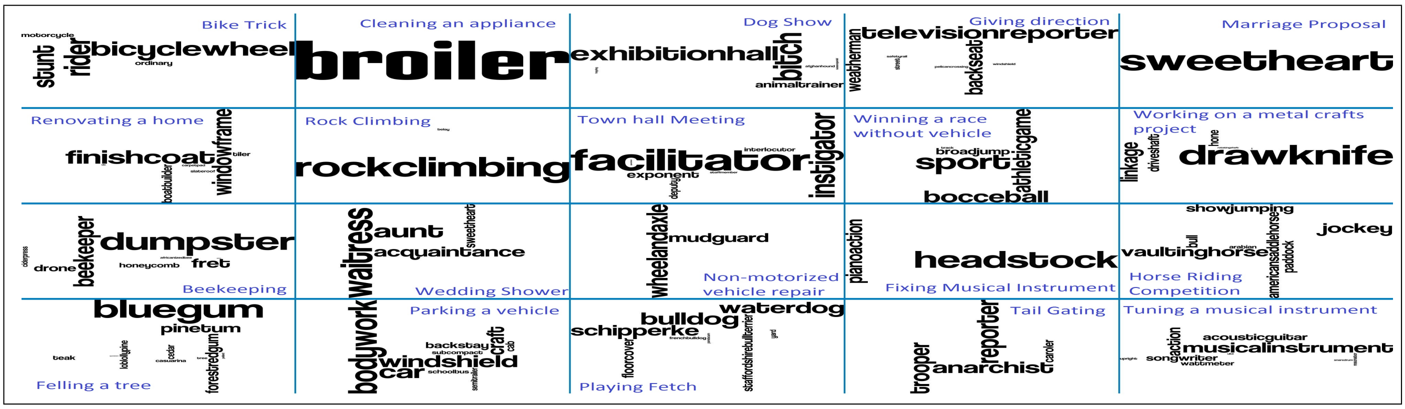

Fig. 4: Visualization of frequency of tags generated for events. (Tags are generated using labels of Segment-based Models.)

of the retrieved scores for the segments. We consider any overlap $> 50\%$ as positive. The average precision (AP) for each event at various thresholds, based on the rank of each segment are shown in Table II. The AP is consistently better for segment-based models, indicating that they are able to better discriminate the positive segments from the outliers.

Segment-based models can also be used to provide labels to the informative segments without any post-processing due to the label assigned to each model. Figure 3 contains examples of labels produced by segment-based models for sample videos of some events. For events like ”marriage proposal” and ”rock climbing”, single models like ”sweetheart” and ”rockclimbing” are able to encapsulate majority of videos with precision. In the absence of specific labels from object bank, as in the case of ”swimming” and ”potato race” from event ”Winning a race without a vehicle”, it can be seen that semantically closer labels like ”sport” and ”broad jumping” have been assigned. This can be attributed to the inter-model dependencies in the object bank which are efficiently utilized by the discriminative clustering algorithm. Figure 4 shows the frequency distribution of tags that were generated using the labels for positive MEDTest videos of each category. It can be seen that the tags are highly relevant to the event categories.

IV. Conclusions and Future Work

In this paper, we formulated a novel approach using segment-based models that can be used to tackle event classification and recounting tasks simultaneously. Using the noisy pre-trained concepts, we trained discriminative models that can diversely represent an event with semantic interpretation which is useful for higher-level video tasks. The proposed method has been evaluated on the challenging TRECVID dataset, achieving promising results in both classification and recounting. The results are also significant given the small portion of the exemplar videos that was used to train the event models while achieving better performance.

In future, the models can be extended to enable data-sharing across different events or different datasets to overcome the limited data available for rare patterns of the events.

Acknowledgment

Supported by the Intelligence Advanced Research Projects Activity (IARPA)via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to re-produce and distribute reprints for Governmental purposes-not-withstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements,either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] P. Over, J. Fiscus, G. Sanders, D. Joy, M. Michel, G. Awad, A. Smeaton, W. Kraaij, and G. Quénot, “Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics,” 2014, in TRECVID.

[2] C. Sun and R. Nevatia, “Large-scale web video event classification by use of fisher vectors,” 2013, in WACV.

[3] D. Oneata, J. Verbeek, and C. Schmid, “Action and event recognition with fisher vectors on a compact feature set,” 2013, in ICCV.

[4] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” 2007, in CVPR.

[5] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos, “Dynamic pooling for complex event recognition,” 2013, in ICCV.

[6] M. Jain, J. van Gemert, and C. Snoek, “What do 15,000 object categories tell us about classifying and localizing actions?” 2015, in CVPR.

[7] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney, “Video event recognition using concept attributes,” 2013, in WaCV.

[8] C. Sun and R. Nevatia, “Active: Activity concept transitions in video event classification,” 2013, in ICCV.

[9] C. Sun, B. Burns, R. Nevatia, C. Snoek, B. Bolles, G. Myers, W. Wang, and E. Yeh, “Isomer: Informative segment observations for multimedia event recounting,” 2014, in ICMR.

[10] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, “Undoing the damage of dataset bias,” 2012, in ECCV.

[11] C.-W. C. Juan Carlos Niebles and L. Fei-Fei, “Modeling temporal structure of decomposable motion segments for activity classiffication,” 2010, in ECCV.

[12] L. F.-F. Kevin Tang and D. Koller, “Learning latent temporal structure for complex event detection,” 2012, in CVPR.

[13] C. Sun and R. Nevatia, “Discover: Discovering important segments for classification of video events and recounting,” 2014, in CVPR.

[14] T. Malisiewicz, A. Gupta, and A. A. Efros, “Ensemble of exemplar-svms for object detection and beyond,” 2011, in ICCV.

[15] X. Zhu, D. Anguelov, and D. Ramanan, “Capturing long-tail distributions of object subcategories,” 2014, in CVPR.

[16] S. Maji, A. C. Berg, and J. Malik, “Classification using intersection kernel support vector machines is efficient,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.