Video Event Understanding using Natural Language Descriptions

Vignesh Ramanathan[*] Percy Liang[†] Li Fei-Fei[†]
[*]Department of Electrical Engineering, Stanford University
[†]Computer Science Department, Stanford University
{vigneshr, pliang, feifeili}@cs.stanford.edu

Abstract

Human action and role recognition play an important part in complex event understanding. State-of-the-art methods learn action and role models from detailed spatio temporal annotations, which requires extensive human effort. In this work, we propose a method to learn such models based on natural language descriptions of the training videos, which are easier to collect and scale with the number of actions and roles. There are two challenges with using this form of weak supervision: First, these descriptions only provide a high-level summary and often do not directly mention the actions and roles occurring in a video. Second, natural language descriptions do not provide spatio temporal annotations of actions and roles. To tackle these challenges, we introduce a topic-based semantic relatedness (SR) measure between a video description and an action and role label, and incorporate it into a posterior regularization objective. Our event recognition system based on these action and role models matches the state-of-the-art method on the TRECVID-MED11 event kit, despite weaker supervision.

1. Introduction

The ability to differentiate complex events is a key step towards video understanding and has spurred significant research in recent years [17, 8, 23]. Complex events can be thought of as compositions of atomic actions performed by people holding different roles. In this work, we provide a method to learn these action and role models based on easily obtainable natural language descriptions of event videos (see Fig. 1). We rely entirely on these descriptions and do not require separate ground truth annotations of roles and actions.

The use of action and/or role models trained with extensive spatio temporal annotations has shown to boost event recognition performance in videos [8, 12]. Such detailed annotations require expensive human effort and severely restrict the scalability with the inclusion of more actions and roles. On the other hand, complex event datasets like TRECVID-MED11 [1] event kit and *MPII* Cooking [22] are accompanied by natural language descriptions, which are easy to obtain and incur only a one-time annotation cost during the collection of a dataset. Internet repositories such as YouTube already have accompanying descriptions, and require no annotation at all.

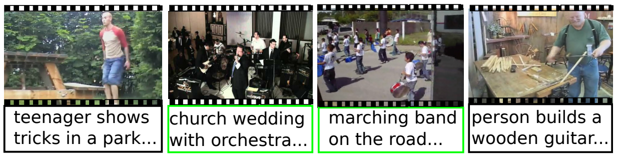

teenager shows tricks in a park...
church wedding with orchestra...
marching band on the road...
person builds a wooden guitar...

Figure 1. Our method relies on natural language video descriptions to train action and role models. Sample videos along with their descriptions are shown. The descriptions of videos containing the action "play instrument" are bounded in green, but we do not use action/role labels during training.

Unfortunately, natural language descriptions only provide a *coarse* high-level summary of the events occurring in videos. This coarseness leads to two challenges. First, the canonical action labels (e.g., "play instrument") might not appear in the description (e.g., "church wedding with orchestra")—see Fig.1. Bridging the gap between high-level natural language descriptions and low-level action/role labels is a challenging problem in natural language semantics. To tackle this, we define a new semantic relatedness (SR) measure between an action/role label and a natural language description. The measure is based on Latent Dirichlet Allocation [3] trained on YouTube descriptions.

The second challenge is that natural language descriptions do not specify the spatiotemporal extents of actions and roles. To cope with this missing data, we use the Posterior Regularization (PR) framework [7]. Specifically, we represent a video as a bag of spatiotemporally-localized human tracklets, and define an action and role assignment variable for each tracklet. The natural language supervision then imposes a soft constraint that at least one of tracklets in the video is assigned to a semantically-related action/role label.

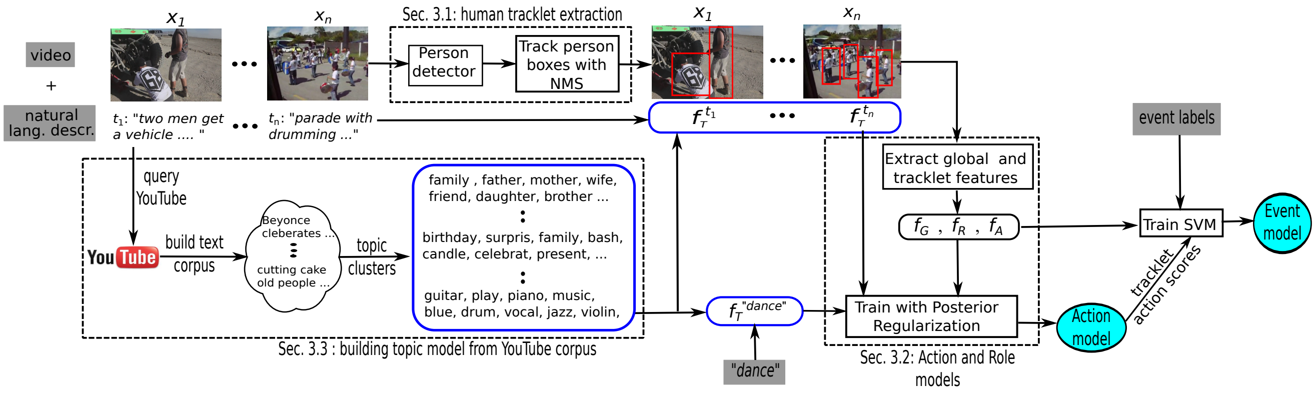

video
+
natural lang. descr.
x_1
x_n
t_1: "two men get a vehicle .... "
t_n: "parade with drumming ..."
Sec. 3.1: human tracklet extraction
Person detector
Track person boxes with NMS
x_1
x_n
$f_T^{t_1}$
$f_T^{t_n}$
query YouTube
YouTube
build text corpus
Beyonce cleberates ...
cutting cake old people ...
topic clusters
family , father, mother, wife, friend, daughter, brother ...
birthday, surpris, family, bash, candle, celebrat, present, ...
guitar, play, piano, music, blue, drum, vocal, jazz, violin,
Sec. 3.3 : building topic model from YouTube corpus
$f_T^{\text{"dance"}}$
"dance"
Extract global and tracklet features
f_G , f_R , f_A
Train with Posterior Regularization
Sec. 3.2: Action and Role models
Action model
tracklet action scores
event labels
Train SVM
Event model

Figure 2. An overview of the system. Inputs to the system are shaded in grey.

We first evaluated our approach on action and role classification, showing that our SR measure improves accuracy over existing measures. We also considered event recognition, where the state-of-the-art method [8] requires detailed spatiotemporal annotations of atomic actions. We incorporated our action/role models, which are trained only on natural language descriptions, into our event recognition model. On the TRECVID-MED11 event kit, our model matches [8] despite using weaker supervision.

2. Related Work

Natural language processing for vision Recent works attempting to leverage the vast amount of textual data available with Internet images have developed vision-specific semantic relatedness measures [25, 24] to identify the link between part-based object attributes and image classes. However, such measures are derived from general text like Wikipedia, and are therefore, less suited for language pertaining to human actions/roles specific to a set of events. Other attempts to use textual descriptions in conjunction with attribute recognition were presented in [2, 19]. While, [2] is restricted to simple part-based attributes directly mentioned in the image description, [19] involves humans in the loop to actively describe a group of images through visual attributes. Another line of work [9, 29] jointly considers multiple modalities including text descriptions to perform image annotation, retrieval or segmentation. [23] transfers composite action videos to an attribute space enabling comparison with textual corpus.

Previous works like [13, 16, 4] use time-synchronized movie scripts or closed captions to identify video segments corresponding to specific actions. Again, these methods rely on presence of the action label in the script or use a pretrained classifier [13] to identify the action-text in a script and require temporal annotations. [31] performs tag prediction by using meta data provided along with YouTube videos. [18] processes descriptions of action segments to automatically discover a set of action classes.

In contrast to the above methods, we learn models based on natural language descriptions which may not contain the action and role label. In particular, we construct a topic model based measure specific to our task.

Action, role and event recognition [8, 12] showed significant improvement in event recognition by using atomic action and role detectors as a part of their event recognition model. Both methods required spatio temporal annotation of action and roles in the training videos to learn the models. Other works which have investigated the use of social roles in video understanding include [30, 5]. [14] uses attributes to perform action recognition in videos.

Weakly supervised action models Discriminative spatio temporal regions in videos or images to localize the actions in [27, 21, 28]. Similar in spirit to these works, we try to localize the human actions and roles. However, we develop a model with latent action and role assignments to different human tracklets in a video.

3. Our Approach

An overview of our system is shown in Fig. 2. We first use natural language video descriptions to train action and role models. The prediction scores from the model are then used to train event recognition models.

In our setup, each training video is accompanied by a natural language description, which might or might not contain the action label present in the video. Formally, we denote our training dataset by $(\langle x_1, t_1\rangle, \ldots, \langle x_n, t_n\rangle)$, where x_i is a video belonging to different event classes and t_i is the corresponding textual description. No textual descriptions are present in the test data.

We assume a fixed set of actions $\mathcal{A}$ and roles $\mathcal{R}$ and define additional variables $\langle y_i, z_i\rangle$ for each x_i. Here, $y_i^a \in \{-1, 0, 1\}$ indicates whether the label of the video corresponding to the action a is negative, unknown or positive. We define $z_i^r \in \{-1, 0, 1\}$ similarly for the role r. These variables are not observed in the training data.

3.1. Human tracklet extraction

Complex event videos are composed of many atomic actions and roles, confined to spatio temporal regions. We attempt to incorporate this locality by representing a video x_i as a bag of human tracklets $\mathcal{H}_i$. The action or role occurring in a video would then correspond to one or more of these tracklets $h \in \mathcal{H}_i$. As illustrated in the corresponding section of Fig. 2, we obtain tracklets by running a human detector [6] across different segments in a video and tracking the resulting bounding boxes within a temporal window of 100 frames. In our experiments, we uniformly partition a video into 20 different segments and obtain 5 tracklets in each segment based on non-maximal suppression. We choose the top 50 tracklets based on their detection scores.

3.2. Action and role model

We define a conditional random field (CRF) to model the actions and roles of different tracklets in a video, similar in spirit to [12]. However, we neither assume perfect human tracking nor complete person-wise action and role labels for training.

We assume that each video x_i has a set of human tracklets given by $\mathcal{H}_i$. The potential $\Phi(x, h, a, r)$ of making action assignment $a \in \mathcal{A}$ and role assignment $r \in \mathcal{R}$ to the tracklet h in video x is given in Eq. 1.

$$\begin{aligned}\Phi(x, h, a, r) \quad &= \quad w_g(a, r) \cdot f_g^x + w_{in}(a, r) \qquad (1)\\ &\quad\quad + w_{ac}(a) \cdot f_{ac}^h + w_{ro}(r) \cdot f_{ro}^h,\end{aligned}$$

where $f_g^x \in \mathbb{R}^{d_g}$ is the global video feature of x. The features $f_{ac}^h \in \mathbb{R}^{d_{ac}}$, and $f_{ro}^h \in \mathbb{R}^{d_{ro}}$ are the action and role features for the human tracklet h respectively. The global weight is denoted by $w_g \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{R}| \times d_g}$, where $w_g(a, r) \in \mathbb{R}^{d_g}$ gives the global weight for action a and role r. Similarly, $w_{in} \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{R}|}$ is the weight for joint action and role assignment to a track, with $w_{in}(a, r) \in \mathbb{R}$ corresponding to action a and role r. The action-weight corresponding to a is given by $w_{ac}(a) \in \mathbb{R}^{d_{ac}}$ and the role-weight for role r is given by $w_{ro}(r) \in \mathbb{R}^{d_{ro}}$. The total set of weights to be learned are then represented by $w = (w_g, w_{in}, w_{ac}, w_{ro})$.

With a slight abuse of notation, we let $a_i \in \mathcal{A}^{|\mathcal{H}_i|}$ be the action-labels assigned to the tracklets in video x_i, and $a_i^h \in \mathcal{A}$ denote the action-label of the tracklet h in the video. Similarly, we let $r_i \in \mathcal{R}^{|\mathcal{H}_i|}$ be the role-labels associated with x_i and $r_i^h \in \mathcal{R}$ be the role-label of tracklet h. The probability $p(a_i, r_i; w)$ of this assignment to video x_i is given by Eq. 2.

$$p(a_i, r_i; w) = \frac{1}{Z_i} \exp\left(\sum_{h \in \mathcal{H}_i} \Phi(x_i, h, a_i^h, r_i^h) \right), \qquad (2)$$

where Z_i is the partition function for the video x_i.

The log-likelihood of making action and role assignments $\mathbf{a} = (a_1, \ldots, a_n), \mathbf{r} = (r_1, \ldots, r_n)$ respectively across n videos is given by Eq. 3

$$L(\mathbf{a}, \mathbf{r}; w) = \sum_{i=1}^{n} \log p(a_i, r_i; w) \qquad (3)$$

Features: The global video feature uses multiple channels through HOG3D [10], ASR, OCR, MFCC [20] and SIFT [15] features. The features f_{ac} and f_{ro} are bag of words HOG3D features extracted from the tracklet h.

Training with posterior regularization: We present a method to learn the model by minimizing L from Eq. 3, assuming the labels $\langle y_i, z_i \rangle$ are given. We will later use natural language annotations to derive these labels in Sec. 3.3. We wish to learn model weights while making latent action and role assignments to each tracklet in the video. The setup is close to the Multi Instance Multi Label framework of [32]. However, to facilitate learning of a model with action-role relations, we adopt the more general posterior regularization framework [7]. This enables us to optimize the likelihood subject to soft constraints on the predicted action and role distribution. Formally, let $Q(\mathbf{a}, \mathbf{r})$ be a distribution of action and role assignments to the training videos. We wish to ensure that, in a video tagged as positive for a specific action, the number of tracklets corresponding to the action is at least one. Similarly, in negative videos, the number of tracklets corresponding to an action should be zero. The same follows for roles. We use these constraints to learn a model by solving the optimization problem in Eq. 4.

$$\begin{aligned}\min_{\substack{w, Q,\\ \delta \geq 0, \eta \geq 0}} \quad & \frac{1}{2}\|w\|^2 - \mathbb{E}_Q[L] - H_Q + \sum_{i,a} \delta_i^a + \sum_{i,r} \eta_i^r \\ \text{subject to} \quad & \mathbb{E}_Q[N_i(a)] \geq 1 - \delta_i^a, \quad \forall y_i^a = +1 \\ & \mathbb{E}_Q[N_i(a)] \leq \delta_i^a, \quad \forall y_i^a = -1 \qquad (4) \\ & \mathbb{E}_Q[M_i(r)] \geq 1 - \eta_i^r, \quad \forall z_i^r = +1 \\ & \mathbb{E}_Q[M_i(r)] \leq \eta_i^r, \quad \forall z_i^r = -1,\end{aligned}$$

where $N_i(a) = \sum_{h \in \mathcal{H}_i} \mathbf{1}(a_i^h = a)$, $M_i(r) = \sum_{h \in \mathcal{H}_i} \mathbf{1}(r_i^h = r)$ and H_Q is the entropy of distribution Q.

We optimize Eq. 4 using a modified Expectation Maximization algorithm shown in Sec. A of the supplementary document.

3.3. Using natural language video descriptions

The natural language description of a video contains rich information about the event context and can help infer the presence of specific actions and roles in the video. For instance, Fig. 1 provides examples of descriptions which do not contain the action label "play instrument".

Three measures based on WordNet, World Wide Web (WWW) and Wikipedia were introduced in [25] to determine the semantic relatedness between class names and attributes. The WordNet metric is a poor indicator of similarity between concepts not linked by a hypernym hierarchy. For instance, in Fig. 1 it would be unable to recognize the relation between "marching band" and "play instrument" which do not fall under the same subtree. The resulting poor performance of this measure was also noted in [25], making it less useful for our purpose. The WWW metric is tailored to measure the similarity of only a pair of terms based on co-occurrence in Internet repositories but does not offer a concept based similarity. While the Wikipedia SR measure uses Wikipedia concepts, it relies on a generic knowledge base, provides no dimensionality reduction and is not task-specific. We address this issue by building a task-specific language topic model and using it to define the SR measure.

Topic model based SR: A natural source for video descriptions is the vast collection of user-provided descriptions of YouTube videos. Hence, as shown in Fig. 2, we build a text corpus by querying YouTube for frequent terms from the training descriptions. We generate a topic model from this corpus with 200 topics. Since the text corpus was obtained based on training descriptions, the generated topic clusters often capture frequent actions and roles in the data. Sample topic clusters are shown in Fig. 2.

All video descriptions t_i can now be represented by a 200 dimensional vector $f_d^{t_i}$ specifying the distribution of the topics in the description. An action a can be represented by f_d^a giving the topic distribution over the action label (f_d^r is defined similarly for a role r). The cosine similarity $sim\left(f_d^a, f_d^{t_i}\right)$ provides the proximity of a video x_i to an action a. We will refer to this measure as the *topic model SR*. Each training video can now be assigned a training label $y_i^a \in \{-1, 0, 1\}$ based on a threshold τ as shown in Eq. 5 (z_i^r is defined similarly).

$$y_i^a = \begin{cases} 1 & \text{if } sim(f_d^a, f_d^{t_i}) \geq (1-\tau) \text{ or} \\ & \; t_i \text{ contains action label } a \\ -1 & \text{if } sim(f_d^a, f_d^{t_i}) \leq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Handling outliers Discovering semantic relatedness in textual space is challenging and these measures can be unreliable. While y_i^a, z_i^r from Eq. 5 can be used in PR, it would result in a significant number of outliers. We handle this problem by defining a pool of potential positive and negative examples according to Eq. 5 as shown in the first step of Fig. 3 and letting the model gradually choose more examples from this pool in successive iterations as illustrated in the third and fourth step of Fig. 3. This is achieved through a self-paced learning scheme introduced in [11].

We further modify the self-paced method to treat $f_d^{t_i}$ as an additional global feature for x_i in the initial iterations but gradually reduce it to zero across the iterations. Intuitively, we are leveraging the textual information present along with videos to choose good examples in the initial phase of the training. However, as the model grows confident with more iterations, the textual features are ignored resulting in a model which only uses video features. The complete details are shown in Sec. B of the supplementary.

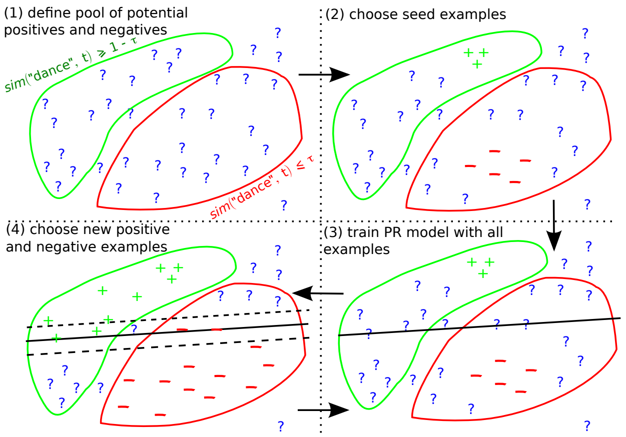

(1) define pool of potential positives and negatives
sim("dance", t) ⩾ 1 - τ
sim("dance", t) ≤ τ
(2) choose seed examples
(3) train PR model with all examples
(4) choose new positive and negative examples

Figure 3. An overview of our self-paced approach shown for one action or role. The green and red boundaries indicate the positive and negative pool of samples chosen using the topic model SR.

3.4. Training the event model

We use the action and role detection scores to perform video event classification. The expected number of tracklets corresponding to different actions and roles are used as additional features along with the global video features to train a linear SVM. We use the same set of global video features from Sec. 3.2. Similar to [8], we first train separate event classifiers for each individual feature mentioned in Sec. 3.2 and finally treat the event classification score from these classifiers as global video features. Since only a small set of actions and roles are usually related to an event, we add an additional L_1 regularization term for the action and role feature weights to encourage only the relevant action and/or role scores to be selected.

4. Experiments

We test our event, action and role classification models on the TRECVID-MED11 event kit. The dataset contains videos belonging to 15 complex event classes. Each video is accompanied by a *synopsis* describing the events in the video, and only a few of them mention the atomic actions and objects present in the video. We use the same training and testing splits as [8].

4.1. Implementation details

We define *crude* action labels $\tilde{y}_i$ and role labels $\tilde{z}_i$ for each video x_i based on simple text processing. We set $\tilde{y}_i^a =$

1 if t_i contains the action label a, $\tilde{y}_i^a = -1$ if none of the natural language descriptions in the event class of x_i contain the action label a; otherwise, we set $\tilde{y}_i^a = 0$. We define $\tilde{z}_i^r \in \{-1, 0, 1\}$ similarly for video x_i and role r. These crude labels are used to train baseline models that does not use the complete video description as well as to initialize the self-paced scheme in Sec. 3.3. The value of τ is set to consider the top 300 (30) videos closest to the action (role) description as potential positives.

In our experiments, we train separate one-vs-all models for each action and role. While training an action (role) model, we consider the relation of the action (role) to all the roles (actions) including a null role (action). In practice, this makes the learning more tractable and also performs better than training a single model considering all actions and roles together.

4.2. Action and role classification

A set of 62 “atomic events” were used in [8]. Some of these events were non-human actions like vehicle movement. We select a subset of 46 classes which involve one or more humans. We choose only the action classes which are directly mentioned at least once in the training data descriptions. We consider a set of 13 roles appearing in different events, as listed in Tab. 2. Each video in the test set is annotated with the actions and roles present in it for evaluation.

The action and role classification performance is evaluated by computing the average precision on the testing data as shown in Tab. 1, 2. We defube the expected number of tracklets performing an action in a videos as the corresponding action score for the video. Similarly, the expected number of tracklets holding a role in a video provides the role score.

“Full model” refers to the complete algorithm using video descriptions to train PR models in a self-paced setting. The different baselines are explained below. The first three baselines are trained only with crude labels $\langle \tilde{y}_i, \tilde{z}_i \rangle$.

- global only: use global video features to train a SVM.
- simple PR: train action or role models without considering joint action-role potential in Eq. 1.
- full PR: uses action-role relation in addition to tracklet features to train the PR model.
- wiki SR [25]: train full PR model by identifying positives and negative training examples based on the Wikipedia SR using a threshold as defined in Sec. 4.1, without self paced learning
- topic SR: Our full model without outlier handling through self paced learning.

Comparing the performance of global only and simple PR baselines in Tab. 1, 2, we observe that identifying human tracklets in the videos improves the overall action and role classification. The effect is even more prominent for roles, since roles are governed by the humans holding the roles, whereas most actions like reeling, cutting, jacking car, turning wrench can be determined by object manipulations in the scene. We notice that actions like kissing, walking, playing instrument, which can be determined by observing the complete or the upper human body, benefit more from the human tracklet representation compared to actions like cleaning, cutting, petting, turning wrench which are often shown as close-up shots of the hand. Similarly, human detectors fail to detect people when they are not upright, leading to a drop in performance.

Action	global only	simple PR	full PR	wiki SR [25]	topic SR	full model
bending	0.0604	**0.0708**	0.0689	0.0688	0.0586	0.0601
blowing candles	0.4616	0.4485	0.5088	**0.5222**	0.4934	0.5134
carving	0.2131	0.0229	0.0918	0.0794	0.0359	**0.2348**
casting	0.0046	0.0125	0.0118	0.0119	**0.0141**	0.0135
clapping	0.1433	0.1865	**0.2720**	0.2615	0.2236	0.2408
cleaning	**0.0262**	0.0047	0.0047	0.0048	0.0048	0.0240
cutting	**0.1928**	0.0794	0.0764	0.0776	0.0760	0.1906
cutting cake	0.0885	0.1361	0.1764	**0.2803**	0.1208	0.1764
cutting fabric	**0.1896**	0.0152	0.1541	0.1526	0.1557	0.1351
dancing	0.5941	0.5556	0.6189	0.6052	**0.6357**	0.6261
drilling	0.0570	0.0145	0.0142	0.0157	0.0661	**0.0910**
drinking	0.0258	0.0347	0.0445	**0.0556**	0.0421	0.0322
eating	0.0532	0.0522	**0.0613**	0.0558	0.0598	0.0569
falling	0.1081	**0.1697**	0.1523	0.1390	0.1513	0.1512
flipping	0.3995	0.4316	**0.4554**	0.2636	0.4364	0.4524
hammering	0.0794	0.0057	0.0056	0.0057	**0.2743**	0.2741
jacking car	**0.0734**	0.0185	0.0172	0.0164	0.0185	0.0373
jumping	0.5572	0.5184	0.5443	**0.5734**	0.5203	0.5586
kissing	0.1499	0.5232	0.4976	**0.5318**	0.4716	0.4976
laughing	0.0853	0.1508	0.1624	0.1605	**0.1753**	0.1611
lighting candle	0.0218	0.0437	**0.0805**	0.0772	0.0513	0.0631
open door	0.1276	0.0846	0.0692	0.0692	**0.1285**	0.0989
petting	**0.0253**	0.0103	0.0103	0.0103	0.0103	0.0115
planing	0.0525	0.0162	0.0140	0.0084	0.0449	**0.0555**
play instrument	0.1335	0.2424	0.2059	**0.2705**	0.2083	0.2059
pointing	0.0159	0.0437	**0.0466**	0.0398	0.0336	0.0238
polishing	0.0015	0.0015	0.0015	0.0015	0.0017	**0.0025**
pouring	0.0051	0.0061	**0.0103**	0.0038	0.0088	0.0026
pushing	0.2768	**0.2871**	0.1922	0.2865	0.1783	0.1824
reeling	0.4603	0.4675	0.4669	**0.4973**	0.4665	0.4788
rolling	**0.0533**	0.0074	0.0072	0.0065	0.0078	0.0091
sawing	0.0416	0.0305	0.0628	0.0667	0.0750	**0.2390**
sewing	0.3073	**0.4089**	0.2801	0.2839	0.2660	0.2588
shake	0.0067	0.0062	0.0062	0.0058	0.0064	**0.0101**
singing	0.0384	0.0900	0.0732	0.0721	**0.0901**	0.0742
sliding	0.0438	**0.0811**	0.0776	0.0750	0.0761	0.0806
stir	0.0398	0.2008	0.2037	**0.2071**	0.1975	0.1967
surfing	0.1039	0.1442	0.1494	**0.1510**	0.1302	0.1382
turning wrench	**0.0788**	0.0234	0.0233	0.0232	0.0232	0.0573
using knife	0.0015	0.0016	0.0017	**0.0023**	0.0016	0.0017
using tire tube	**0.0675**	0.0145	0.0141	0.0136	0.0138	0.0351
walking	0.1771	0.2562	0.2520	0.2110	**0.2697**	0.2557
washing	**0.1329**	0.0309	0.0307	0.0309	0.0307	0.0307
waving	0.0965	0.1302	0.1555	0.1413	0.1473	**0.1700**
wiping	**0.0428**	0.0253	0.0253	0.0253	0.0405	0.0369
writing	0.0400	0.1309	0.1080	0.1010	0.1413	**0.1856**
mean	0.1295	0.1356	0.1415	0.1427	0.1453	**0.1616**

Table 1. Action classification results. The highest score for a class is shown in bold font. The first three columns do not use the complete video descriptions, but train with labels $\langle \tilde{y}_i, \tilde{z}_i \rangle$.

From simple PR and full PR results in Tab. 1, 2, we notice that jointly modeling the action-role relations using posterior regularization increases the performance for ac-

Role	global only	simple PR	full PR	wiki SR [25]	topic SR	full model
bride	0.7115	0.7946	0.7880	0.7873	**0.8017**	0.7877
groom	0.6751	0.7755	0.7805	0.7857	0.7901	**0.7901**
priest	0.2686	0.4263	0.4094	**0.4468**	0.4002	0.4058
performer	0.1499	0.1212	0.1708	**0.1805**	0.1722	0.1737
musician	0.0990	**0.2933**	0.2468	0.2506	0.2334	0.2643
parent	0.2084	**0.2388**	0.2123	0.2028	0.2014	0.2245
birthday child	0.7442	0.8350	0.8212	0.8086	0.8359	**0.8359**
audience/guest	0.3163	0.4260	0.3623	0.4263	0.3807	**0.4429**
friends	0.2849	0.5619	**0.5671**	0.4979	0.5591	0.5641
fisherman	0.2677	0.0238	**0.2949**	0.2617	0.2925	0.2873
craftsman	**0.0957**	0.0284	0.0281	0.0285	0.0282	0.0265
mechanic	**0.0446**	0.0268	0.0267	0.0262	0.0266	0.0264
police/soldier	0.2267	0.2346	0.2122	**0.2398**	0.2384	0.2303
mean	0.3148	0.3682	0.3785	0.3802	0.3816	**0.3892**

Table 2. Role classification results. The highest score for a class is shown in bold font. The first three columns do not use the complete video descriptions, but train with labels $\langle \tilde{y}_i, \tilde{z}_i \rangle$.

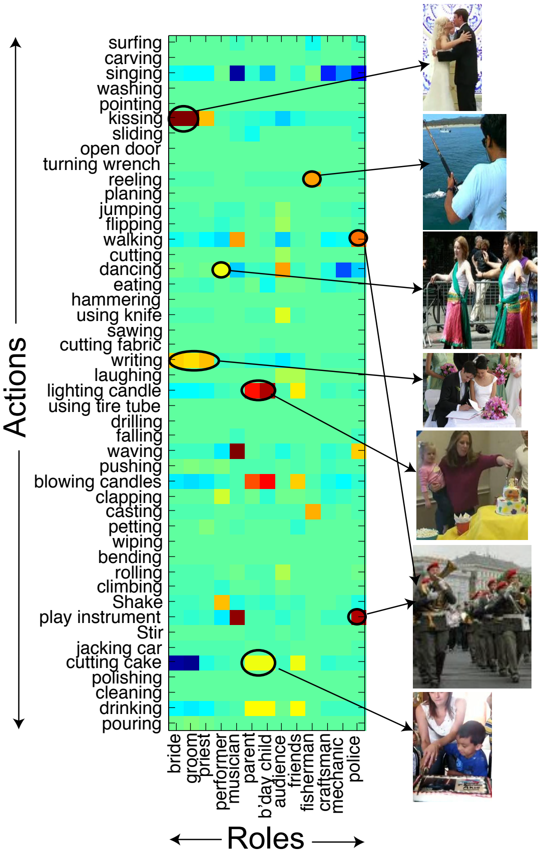


Figure 4. The weights corresponding to different action-role relation $w_I(a, r)$ are shown. Sample frames depicting the action-role relations corresponding to some high weights are also shown.

tion classes like kissing, writing, lighting candle, cutting cake and the roles fisherman, birthday child, performer. In order to analyze the effect of action-role relations, we visualize the joint action-role weights in Fig. 4. As expected, we see strong correlation between certain action and role classes (highlighted by ovals). These correspondences result in improved classification accuracy for the respective classes. Sample frames pertaining to some high weights are shown besides the matrix in Fig. 4. We further demonstrate some qualitative results in Fig. 5, where the highest scoring tracklet in a video for a certain action is shown along with the corresponding role assignment.

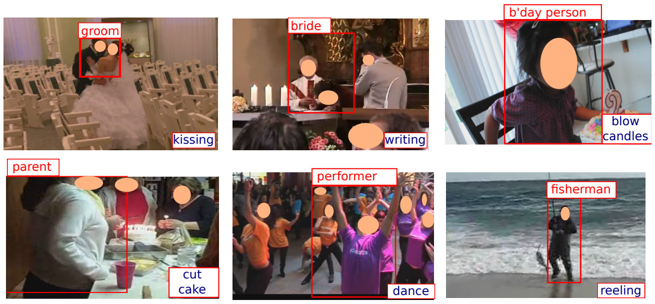


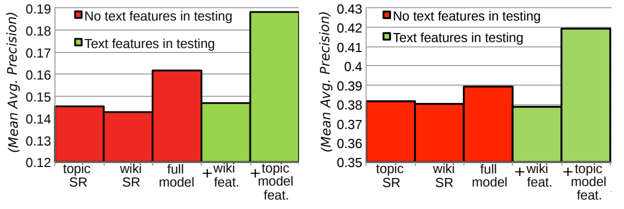
Figure 5. High scoring tracklets for few action models are shown for six test videos. The role assignments are also shown.

The Wikipedia SR and topic SR models trained by identifying positives and negatives based on a SR measure are seen to only marginally improve the performance. This can be accounted to the addition of false positive and negative training labels, demonstrating the difficulty in processing natural language descriptions. The results are seen to be worse in the case of Wikipedia SR measure, when using language at test time.

To analyze the utility of our topic model, we run experiments where natural language descriptions are assumed to be present both during training and testing. We train two separate PR models which use topic model based textual features and Wikipedia SR based textual features respectively as additional global features both during training and testing. The mean AP of these models correspond to the green bars in Fig. 6. For Wikipedia features, we concatenate the Wikipedia SR measure of a description with each of the action and role labels to form a feature. We treat $f_d^{t_i}$ as the topic model feature from description t_i. The considerable gain achieved by using topic model features during test time over other methods justifies the use of topic model SR for the current task. The lower performance of Wikipedia-based methods can be explained by the use of non task-specific corpus and lack of dimensionality reduction.

Further, note that in our adaptation of the self-paced approach, while natural language descriptions are not available during testing, we use features extracted from natural language descriptions in the initial iterations of training, and finally anneal their weights to zero. The effectiveness of our textual features as shown in Fig. 6 allows us to handle outliers introduced by the SR measure.

From Tab. 1, 2, we see that our full model, which handles outliers, is able to achieve significant improvement over the other methods. Our method is seen to be particularly ef-



(a) actions

(b) roles

Figure 6. The green bars correspond to the setting where natural language descriptions are used at test time. The red bars are from Tab. 1, 2.

fective for action classes like hammering, writing, planing, drilling, where the number of training video descriptions directly containing the action label were below 10. We show sample videos which were added as positives by our full method in Fig. 7, along with the corresponding descriptions. We notice the inclusion of videos whose descriptions do not contain the action directly.

4.3. Event classification

We compare the event classification performance of our model from Sec. 3.4 against baseline methods as well as state-of-the-art results from [8] in Tab. 3. Note that unlike our method, [8] used extensive spatio temporal annotation to learn completely supervised atomic action classification models. The prediction scores from these models were finally used to perform event classification. We report two sets of results from [8], one using action classification scores in linear ensemble SVM and the other using them in a joint CRF model. In addition, we demonstrate results against the following baselines

- global only: uses global video features only
- global+actions: uses only action classification features in addition to global video features
- global+roles: uses only role classification features in addition to global video features

From Tab. 3, we observe that our methods using either the action or role features outperform an SVM trained only with global video features. Our full model using both action and role scores achieves the maximum mean AP. Thus, our action and role models trained only with natural language descriptions matches state-of-the-art methods from [8], which uses ground truth spatio temporal action annotations for training. This supports the utility of the our action and role models learned with very weak supervision.

5. Conclusion

We have presented a method to learn atomic action and role models based on easily available natural language video descriptions. We proposed a language topic model based semantic relatedness measure to identify positive and negative training examples. These labels were used to train a CRF model with posterior regularization, which makes latent action and role assignments to human tracklets. Outliers introduced by the SR measure were handled through a self-paced scheme. The action and role models were used to achieve state-of-the-art event classification performance on the TRECVID-MED11 event kits. We demonstrated the efficacy of the topic model based SR measure in identifying training labels as well as the gain due to the posterior regularized method in a weakly supervised setting without temporal annotations. Further, such SR measures could also be used to tackle the problem of converting video content to natural language descriptions as proposed in [26].

Event	global only	[8] * SVM	[8] * joint CRF	global + action	global + roles	full model
Boarding trick	**0.8766**	0.7560	0.7570	0.8276	0.8625	0.8402
Feeding animal	0.4535	**0.5820**	0.5650	0.4490	0.3958	0.4595
Landing fish	0.6612	**0.7410**	0.7220	0.6612	0.6811	0.6593
Wedding	0.4729	0.6650	0.6750	0.5942	0.7555	**0.7871**
Woodworking project	0.2227	0.5760	**0.6530**	0.3697	0.2086	0.3568
Birthday party	0.9083	0.7090	0.7820	**0.9207**	0.9041	0.9008
Changing tire	0.5100	0.4650	0.4770	**0.5200**	0.4977	0.5012
Flash mob	**0.9301**	0.8590	0.9190	0.9273	0.9248	0.9240
Vehicle unstuck	0.6288	0.6610	**0.6910**	0.6212	0.5862	0.6173
Grooming animal	0.3881	0.4570	0.5100	0.3914	0.3927	**0.5415**
Making sandwich	0.5604	0.3560	0.4190	**0.5739**	0.5442	0.5704
Parade	**0.7462**	0.6570	0.7240	0.7283	0.6582	0.7335
Parkour	0.5426	0.5340	**0.6640**	0.6211	0.5681	0.6144
Repairing appliance	0.8025	**0.8080**	0.7820	0.7989	0.7692	0.7840
Sewing project	0.6579	0.5690	0.5750	0.6563	0.6286	**0.6688**
mean	0.6241	0.6263	0.6610	0.6441	0.6252	**0.6639**

Table 3. Event classification results. The highest score for a class is shown in bold font. * Unlike our method, [8] uses extensive ground truth spatio temporal annotations for training separate action classifiers to aid event classification.

Acknowledgements

We thank O. Russakovsky, K. Tang and B. Yao for helpful comments. This research is partially supported by Intel, the NFS grant IIS-1115493 and DARPA-Mind's Eye grant.

References

[1] Trecvid multimedia event detection track, 2011. 1

[2] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and charateriztion from noisy web data. In *ECCV*, 2010. 2

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 1

[4] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *ECCV*, 2008. 2

[5] A. Fathi, J. K. Hoggins, and J. M. Rehg. Social interactions: A first person perspective. In *CVPR*, 2012. 2

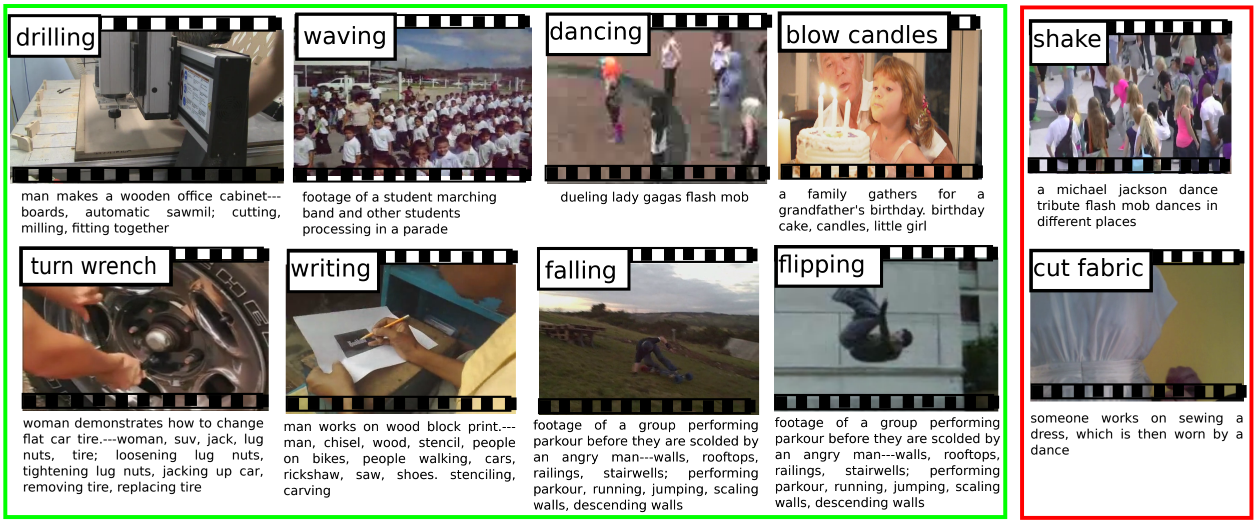

drilling
man makes a wooden office cabinet---boards, automatic sawmil; cutting, milling, fitting together
waving
footage of a student marching band and other students processing in a parade
dancing
dueling lady gagas flash mob
blow candles
a family gathers for a grandfather's birthday. birthday cake, candles, little girl
shake
a michael jackson dance tribute flash mob dances in different places
turn wrench
woman demonstrates how to change flat car tire.---woman, suv, jack, lug nuts, tire; loosening lug nuts, tightening lug nuts, jacking up car, removing tire, replacing tire
writing
man works on wood block print.---man, chisel, wood, stencil, people on bikes, people walking, cars, rickshaw, saw, shoes. stenciling, carving
falling
footage of a group performing parkour before they are scolded by an angry man---walls, rooftops, railings, stairwells; performing parkour, running, jumping, scaling walls, descending walls
flipping
footage of a group performing parkour before they are scolded by an angry man---walls, rooftops, railings, stairwells; performing parkour, running, jumping, scaling walls, descending walls
cut fabric
someone works on sewing a dress, which is then worn by a dance

Figure 7. Videos with highest SR measure added as positives by our full method for different actions are shown. The last column shows wrong videos which are added by the method. The video descriptions are shown below it. Note that they do not contain the action label.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010. 3

[7] K. Ganchev, J. Graca, J. Gillewater, and B. Taskar. Posterior regularization for structured latent variable models. *JMLR*, 11, 2010. 1, 3

[8] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012. 1, 2, 4, 5, 7

[9] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, 2011. 2

[10] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008. 3

[11] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010. 4

[12] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012. 1, 2, 3

[13] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2

[14] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 2

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2004. 3

[16] M. Marszaek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2

[17] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Trans. Multimedia*, 14:88–101, 2012. 1

[18] T. S. Motwani and R. J. Mooney. Improving video activity recognition using object recognition and text mining. In *ECAI*, 2012. 2

[19] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011. 2

[20] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993. 3

[21] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level vide representations. In *ECCV*, 2012. 2

[22] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *TACL*, 1:25–36, 2013. 1

[23] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *ECCV*, 2012. 1, 2

[24] M. Rohrbach, M. Stark, G. Szarvas, and B. Schiele. Combining language sources and robust semantic relatedness for attribute-based knowledge transfer. In *ECCV-PnA*, 2010. 2

[25] M. Rohrbach, M. Stark, G. Szarvas, and B. Schiele. What helps where – and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010. 2, 4, 5, 6

[26] M. Rohrbach, Q. Wei, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013. 7

[27] S. Satkin and M. Hebert. Modelling the temporal extent of action. In *ECCV*, 2010. 2

[28] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2

[29] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 2

[30] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: recognizing people and social relationships. In *ECCV*, 2010. 2

[31] W. Yang and G. Toderici. Discriminative tag learning on youtube videos with latent sub-tags. In *CVPR*, 2011. 2

[32] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *NIPS*, 2007. 3