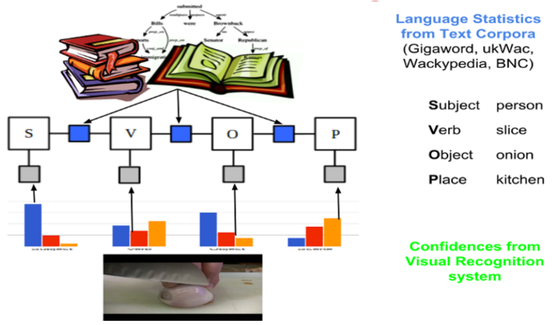
Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild

Subhashini Venugopalan[1], Jesse Thomason[1],
Raymond Mooney[1], Sergio Guadarrama[2], Kate Saenko[3]
[1] UT-Austin [2] UC-Berkeley [3] UMass-Lowell

Goals

Given a short YouTube video, output a natural language sentence that describes the event depicted in the video.

A person is slicing an onion in the kitchen.



We present a strategy to describe videos by using a factor graph to combine visual detection confidences with language statistics. We introduce a graphical model for integrating statistical linguistic knowledge mined from large text corpora with uncertain detections produced by computer vision in order to select content.

YouTube Dataset

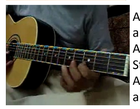
We demonstrate our approach on a large, realistic collection of YouTube videos.

(a) Hollywood (8 actions)

(b) TRECVID MED (6 actions)

A woman is cooking onions.
Someone is cooking in a pan.
someone preparing something
a person coking.
racipe for katsu curry

A girl is ballet dancing.
A girl is dancing on a stage.
A girl is performing as a ballerina.
A woman dances.



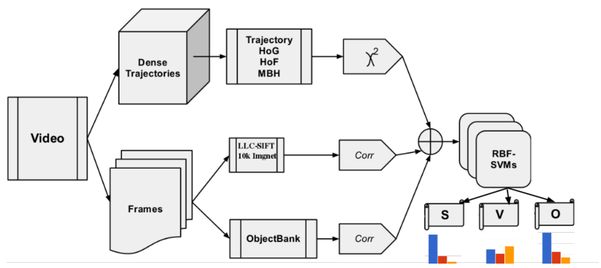
A man is sitting and playing a guitar
A man is playing guitar
Street artists play guitar.
A man is playing a guitar.
a lady is playing the guitar.

A train is rolling by.
A train passes by Mount Fuji.
A bullet train zooms through the countryside.
A train is coming down the tracks.

(c) YouTube (218 actions)

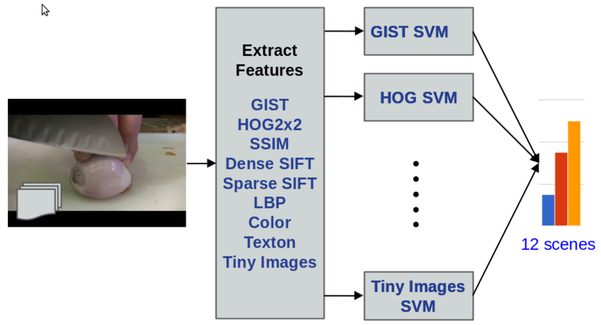
The YouTube dataset, collected by (Chen and Dolan, ACL 2011) consists of 1970 videos, where each video is accompanied by about 41 human descriptions (sentences), see (c) above. This new dataset (c) contains a wide variety of animate and inanimate objects, different scenarios, and many more actions than the other previously used activity datasets (a-b).

Vision Pipeline

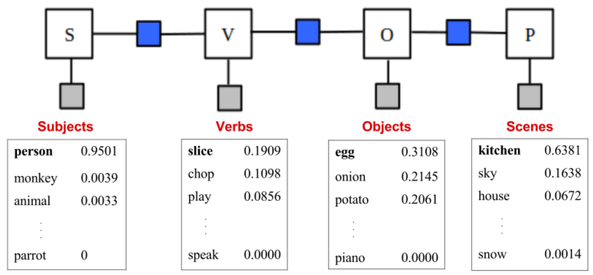


$$K(x_i, x_j) = \exp\Big(-\sum \frac{1}{A^c} D_c(x_i^c, x_j^c) \Big) \quad (1)$$

Scene confidences



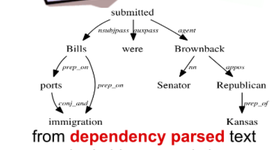
The outputs are confidences over subjects, verbs, objects, and scenes. These are input potentials for nodes in the factor graph.



Language Statistics

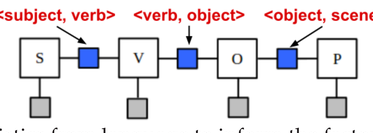
External Corpora
ukWac, Wackypedia, Gigaword, BNC

In-domain text
Textual descriptions accompanying the training videos



from **dependency parsed** text
obtain bigram statistics

<subject, verb> **<verb, object>** **<object, scene>**



Extract statistics from language to inform the factor graph.

Inference

We perform MAP inference on the factor graph to generate the most likely subject, verb, object, scene quadruple.

Surface Realization

Generate sentences based on a simple template and rank them using an n-gram language model.

(A, The) (is) (prep)(a, an, the) | **subject** | **verb** | **object** | **scene**

(prep)(a, an, the)

verb tense is present or present continuous

n-gram LM ranking

A person is slicing the onion in the kitchen.
A person slices the onion in the kitchen.
A person is slicing the onion.
A person slices the onion.
A person is in the kitchen.
.
.
.

⇒ **A person is slicing the onion in the kitchen.**

Binary 0-1 Accuracy

Most	S%	V%	O%	[P]%	SVO%	SVOP%
n-gram	76.57	11.04	11.19	18.30	2.39	1.86
HVC	76.57	+22.24	11.94	17.24	+4.33	+2.92
FGM	76.42	+21.34	12.39	19.89	+**5.67**	+3.71
Any						
n-gram	86.87	19.25	21.94	21.75	5.67	2.65
HVC	86.57	+38.66	22.09	21.22	+10.15	+4.24
FGM	86.27	+37.16	+**24.63**	**24.67**	+10.45	+6.10

n-gram: Previous best system, HVC: model using just the Highest Vision Confidence, FGM: our factor graph model.
Most: Most frequent S,V,O,P Any: Any S,V,O,P in the human descriptions.

Comparison of WUP Similarity

Most	S%	V%	O%	[P]%	SVO%	SVOP%
n-gram	89.00	41.56	44.01	**57.62**	17.53	10.83
HVC	89.09	+48.85	43.99	56.00	+20.82	+12.95
FGM	89.01	+47.05	+**45.29**	+**59.64**	+21.54	+**14.50**
Any						
n-gram	96.60	55.08	65.52	**61.98**	35.70	22.84
HVC	96.54	+65.61	65.32	60.67	+42.53	+27.75
FGM	96.32	+63.49	+**67.52**	+**64.68**	+42.43	+29.34

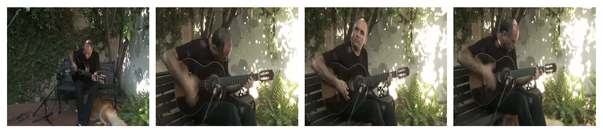
Conclusions

We presented a system that takes a short video clip "in-the-wild" and outputs a brief sentence that sums up the event in the video, such as the actor, the action, its object and location.
Our approach achieves modest improvements over a pure vision system and significantly improves over previous methods in jointly predicting the complete SVO and SVOP tuples.

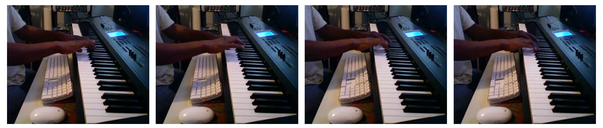
Qualitative Results

GT: *A person is slicing an onion.*
HVC: A person is slicing the egg in the kitchen.
FGM: **A person is slicing the onion in the kitchen.**

GT: *Men are racing on a track.*
HVC: A person is riding in a race to the ground.
FGM: **A person is running a race on the road.**



GT: *A man is playing a guitar.*
HVC: A person is playing in the water.
FGM: **A person is playing the guitar on the stage.**



GT: *A person is playing on piano.*
HVC: A person is playing on the keyboard in the kitchen.
FGM: **A person is playing a piano in the house.**

GT: *A boy is playing a guitar.*
HVC: A person is pouring the chili in the kitchen.
FGM: **A person is playing the guitar in the house.**

Examples where pure vision outperforms FGM.

GT: *A man is lifting a car.*
HVC: **A person is lifting a car on the road.**
FGM: A person is driving a car on the road.