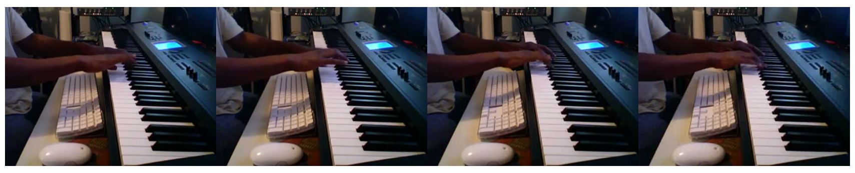
Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild

Jesse Thomason[1], Subhashini Venugopalan[1],
Sergio Guadarrama[2], Kate Saenko[3], Raymond Mooney[1]

(1) University of Texas at Austin,
(2) University of California Berkeley, (3) University of Massachusetts Lowell

Using Language Statistics to Clean Up Visual Detections

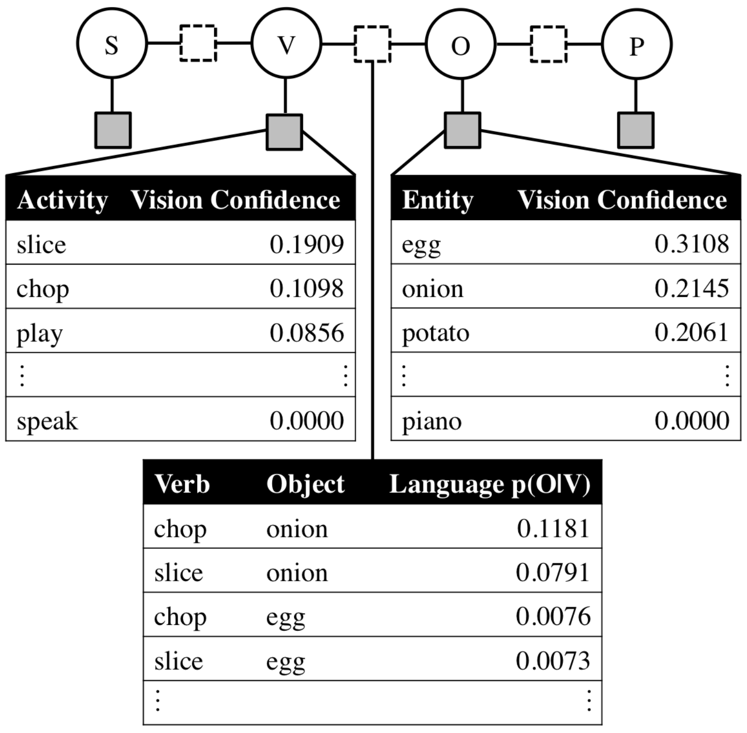
Consider the frames of the video below of a person playing a piano.



The state-of-the-art vision detection systems we use correctly identify a person in a kitchen engaged in a 'playing' activity. However, they also identify the computer keyboard in these frames as more salient than the piano. Using statistics mined from parsed corpora, our proposed system describes the video with "A person is playing the piano in the house," because language tells us that playing a piano is more felicitous than playing a computer keyboard.

Factor Graph Model for Integrating Evidence

We use the probabilistic factor-graph model shown below to combine visual and linguistic evidence to predict the best subject (S), verb (V), object (O), and place (P) for a sentence description of a video. Thinking generatively, we determine the set of descriptive words which are most likely to have produced the video information we observe.



Activity	Vision Confidence
slice	0.1909
chop	0.1098
play	0.0856
⋮	⋮
speak	0.0000

Entity	Vision Confidence
egg	0.3108
onion	0.2145
potato	0.2061
⋮	⋮
piano	0.0000

Verb	Object	Language p(O|V)
chop	onion	0.1181
slice	onion	0.0791
chop	egg	0.0076
slice	egg	0.0073
⋮		⋮

Sample frames from a video to be described (left), and the factor graph model used for content selection (right). Visual confidence values are observed (gray potentials) and inform sentence components. Language potentials (dashed) connect latent words between sentence components.

Our Contributions

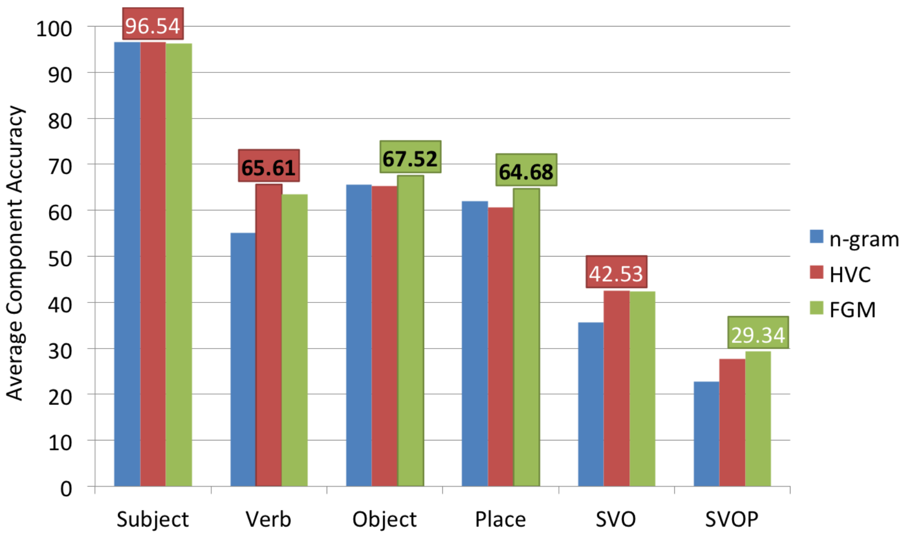
We present a new method to perform content selection by integrating visual and linguistic information to select the best subject-verb-object-place description of a video. This inclusion of scene information has not been addressed by previous video description works.

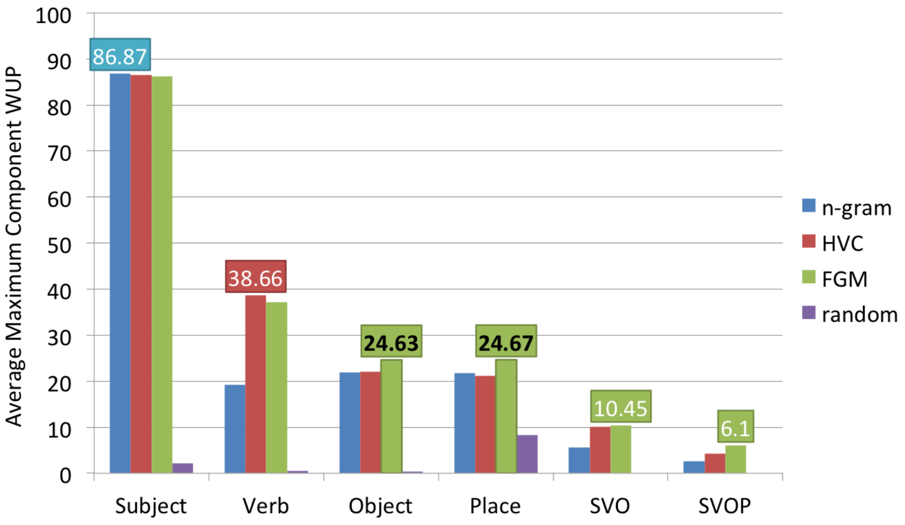
Videos	
Training	Testing
1297	670

Components			
Subjects	Verbs	Objects	Places
45	218	241	12

We explore the scalability of our factor graph model (**FGM**) by evaluating it on a large dataset (outlined in the table above) of naturally occurring videos from YouTube. We demonstrate that our model improves a highest vision confidence (**HVC**) baseline of state-of-the-art entity and activity recognition at the video description task.

Results





Bold averages are statistically significantly ($p < 0.05$) highest.

Examples

FGM improves over HVC

"A person is slicing the onion in the kitchen"

Gold: person, slice, onion, *(none)*
HVC: person, slice, egg, kitchen
FGM: person, slice, onion, kitchen

"A person is running a race on the road"

Gold: person, run, race, *(none)*
HVC: person, ride, race, ground
FGM: person, run, race, road

"A person is playing the guitar on the stage"

Gold: person, play, guitar, tree
HVC: person, play, water, kitchen
FGM: person, play, guitar, stage

"A person is playing a guitar in the house"

Gold: person, play, guitar, *(none)*
HVC: person, pour, chili, kitchen
FGM: person, play, guitar, house

HVC better alone

"A person is lifting a car on the road"

Gold: person, lift, car, ground
HVC: person, lift, car, road
FGM: person, drive, car, road