Integrating language and vision to generate descriptions for videos

A person is slicing an onion in the kitchen.

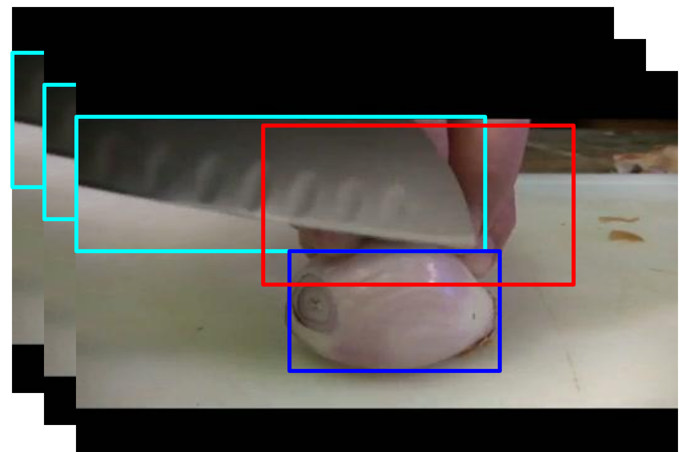
Subhashini Venugopalan

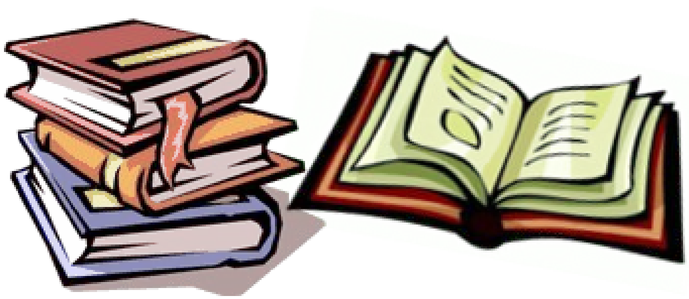
Oct 27, 2014

Problem Statement

Generate descriptions for events depicted in videos.

- Visually identify entities.
- Extract knowledge from text.
- Integrate language and vision.
- Generate description.





A person is slicing an onion in the kitchen.

Video

https://www.youtube.com/watch?v=hpkImroltgo

Pure Vision:

A person is slicing an egg in the kitchen.

Vision+Text (our system):

A person is slicing an onion in the kitchen.

Motivation

Grounding language in perception

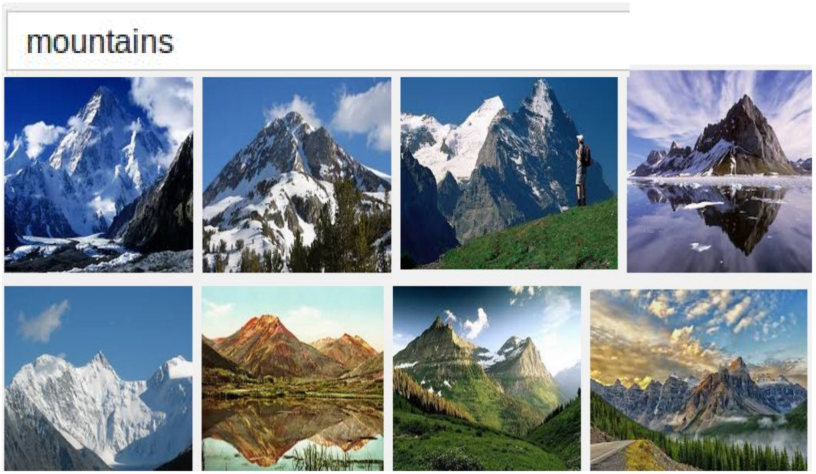
- understand the meaning of language
- relate words to actions in the world

Source: Busy Beaver teaching colors to kids.

Integrating language (NLP) and vision (CV) is important.

Applications

Image and video retrieval by content.



Video description service.

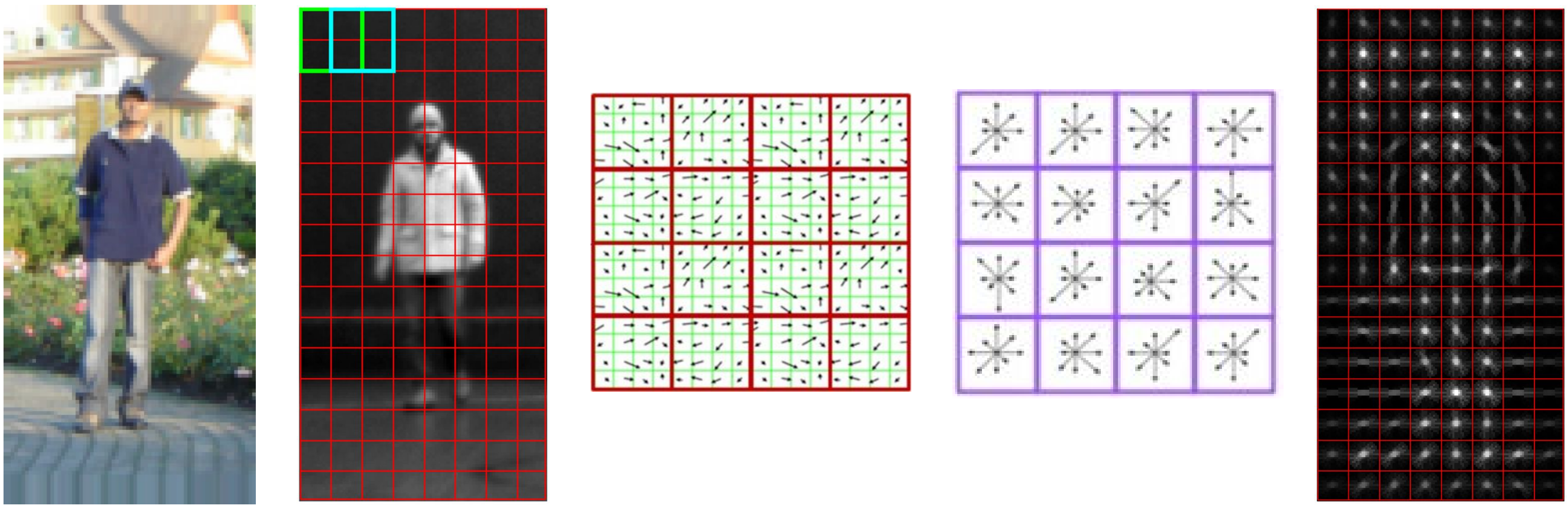
Human Robot Interaction

Video surveillance

Outline

- Related work
- Approach
- Experiments
- Demo

Background: Entity recognition

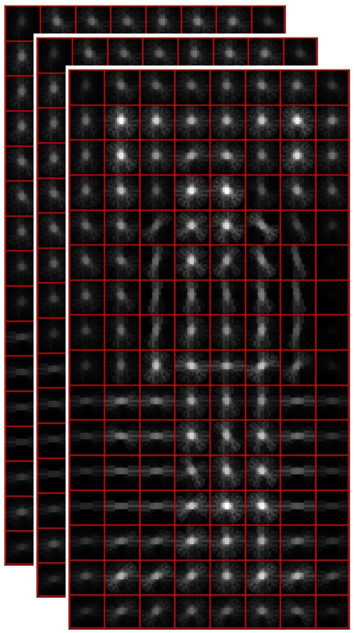


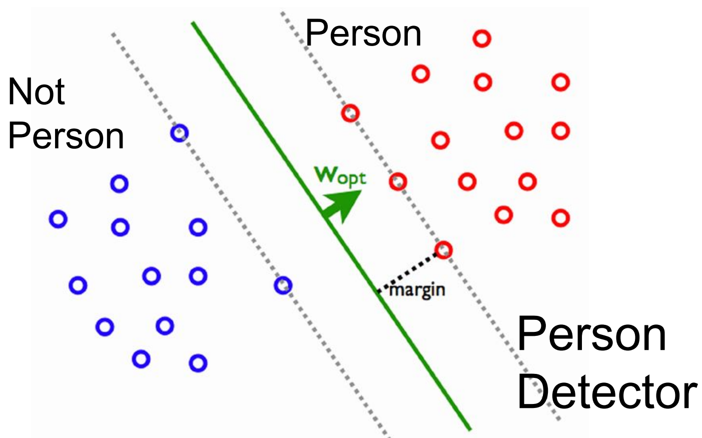
Visual feature (HoG)

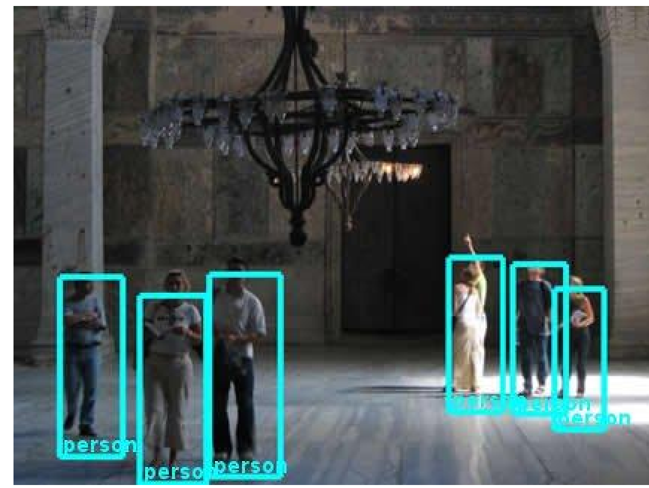
[Dalal & Triggs CVPR’05]

- Histogram of Oriented Gradients is one type of visual feature.
- Visual features are used to identify objects, scenes, and actions.

Background: Entity recognition







Support Vector Machine (**SVM**)

Features from many images are used to train a classifier.

Given visual features, this concept can be extended to classify multiple objects.

Related Work: Describing images

Farhadi et al. ECCV'10

(pet, sleep, ground)
(dog, sleep, ground)
(animal, sleep, ground)
(animal, stand, ground)
(goat, stand, ground)

Kulkarni et al. CVPR'11

There are one cow and one sky. The golden cow is by the blue sky.

Kuznetsova et al. ACL'12
ACL'13, TACL'14

I think this is a boy's bike lied in saltwater for quite a while.

Others: **Yang et al. EMNLP'11, Mitchell et al. EACL'12**

Need videos for semantics of wider range of actions.

Related Work: Describing Videos

Barbu et al. UAI’12, Yu and Siskind ACL’12

The narrow person snatched an object from something.

Others: **Khan & Gotoh EACL’12, Cao et al. CVPR’13**

+ interaction between objects
- limited vocabulary, grammar

Related Work: Describing Videos

SUBJECT
person

VERB
ride

OBJECT
motorbike

A person is riding a motorbike.

Krishnamoorthy et al. AAAI'13

Background: Language Model

A language model (LM) assigns a probability to a sequence of m words. $P(w_1, \ldots, w_m)$

E.g. A 5-gram language model is a PDF over five word combinations.

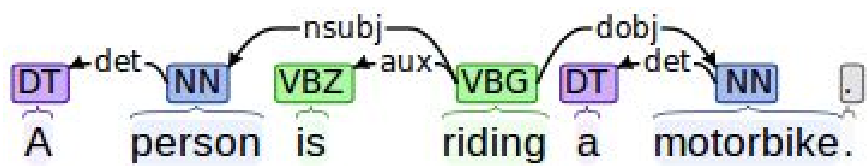
Autocomplete features in search websites use a statistical language model

Background: Language Model

Krishnamoorthy et al. use a Subject-Verb-Object (SVO) language model.

Consider the dependency parse of a sentence.

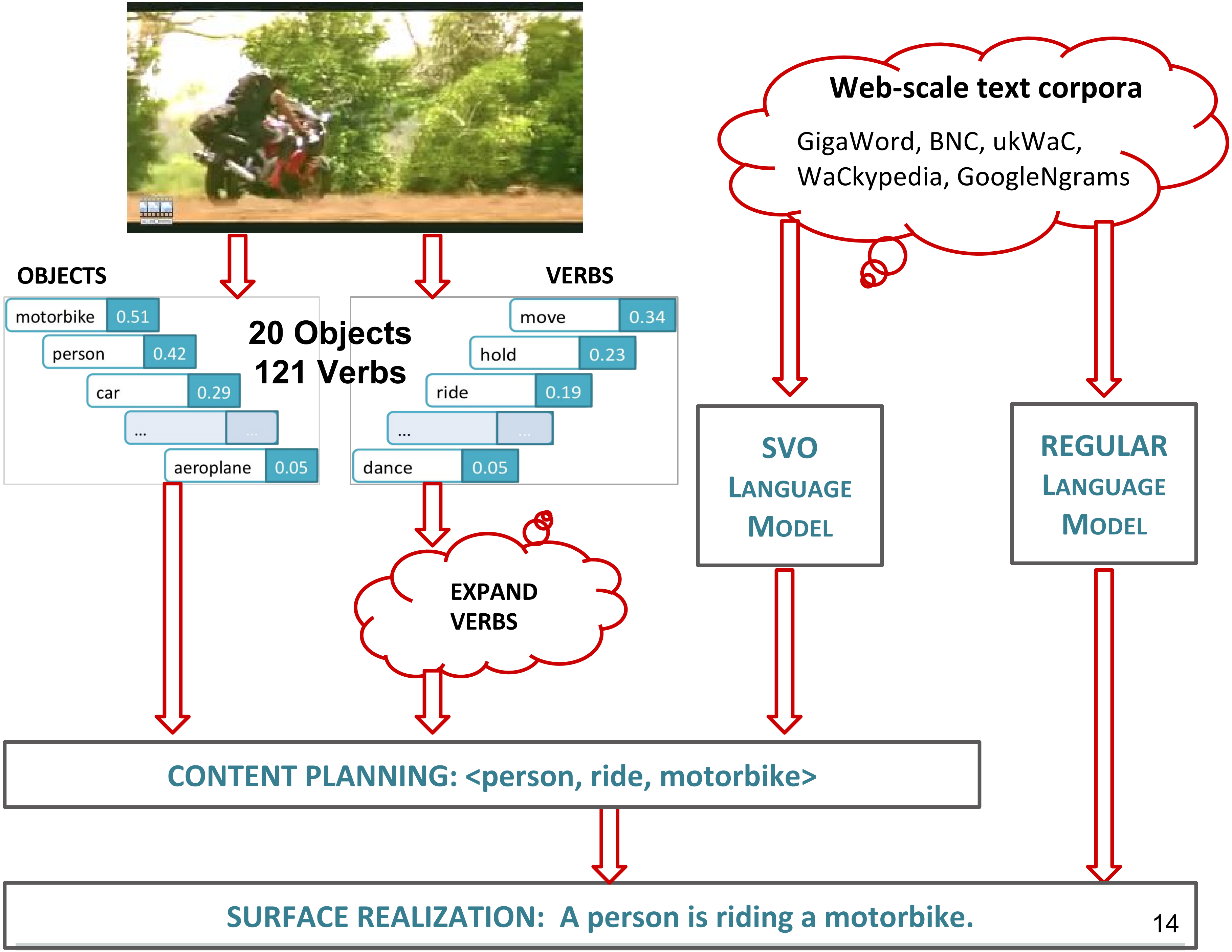
```
det(person-2, A-1)
nsubj(riding-4, person-2)
aux(riding-4, is-3)
root(ROOT-0, riding-4)
det(motorbike-6, a-5)
dobj(riding-4, motorbike-6)
```



Extract Subject, Verb, Object.

(person, ride, motorbike)

Learn SVO-LM.



OBJECTS
motorbike 0.51
person 0.42
car 0.29
... ...
aeroplane 0.05
20 Objects
121 Verbs
VERBS
move 0.34
hold 0.23
ride 0.19
... ...
dance 0.05
Web-scale text corpora
GigaWord, BNC, ukWaC,
WaCkypedia, GoogleNgrams
EXPAND VERBS
SVO LANGUAGE MODEL
REGULAR LANGUAGE MODEL
CONTENT PLANNING: <person, ride, motorbike>
SURFACE REALIZATION: A person is riding a motorbike.

This work

Generate descriptions for events depicted in videos.

- Identify more entities
  - 45 Subjects, 218 Actions, 241 Objects
- Add scenes
  - 12 scenes (Places)
- Use prior knowledge from text
- Integrate language and vision systematically
  - *content selection* using factor graph model
- Generate a description (*surface realization*)
  - simple template

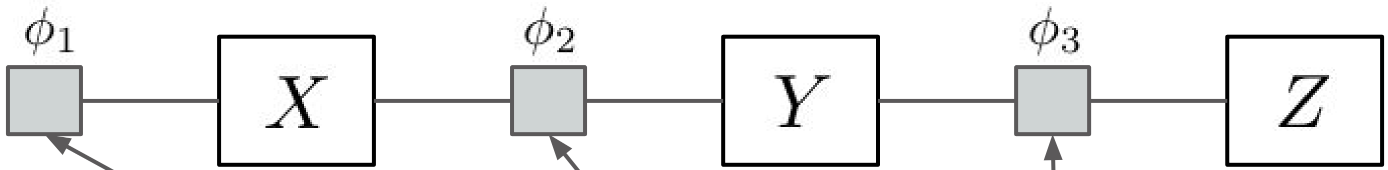
Generating Natural Language Descriptions for Videos

SUBJECT	**V**ERB	**O**BJECT	**P**LACE
person	slice	onion	kitchen

A person is slicing an onion in the kitchen.

Background: Factor Graphs

Relate observed measurements (factors) to quantities of interest (variables).



ϕ_i denote factors (interaction) between the variables they connect.

Factors need not be probabilities themselves, they *determine* probabilities.

$$P(x, y, z) = \frac{1}{Z} \phi_1(x) \phi_2(x, y) \phi_3(y, z)$$

$$Z = \sum_{x,y,z} \phi_1(x) \phi_2(x, y) \phi_3(y, z)$$

normalization constant

Factors ϕ_i are also called potentials.

Background: Factor Graphs

Inference in factor graph:

- Estimate the most likely assignment for the variables.

$$\underset{x,y,z}{\operatorname{argmax}}\, P(x, y, z)$$

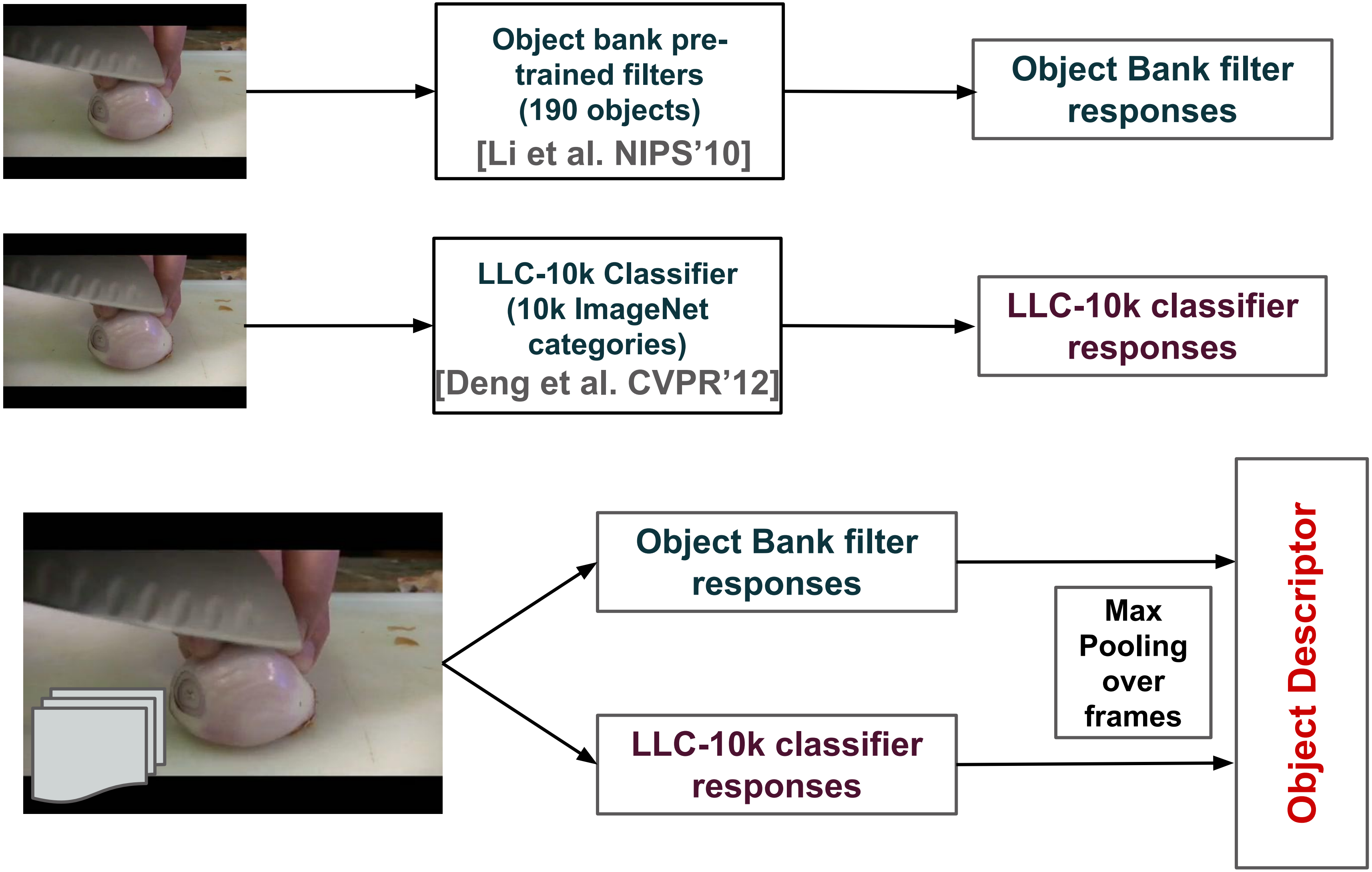
- Exhaustive search: $\mathcal{O}(S^N)$ [S:#states, N:#variables]

Belief Propagation:

- $\mathcal{O}(S^2)$
- exact inference on trees.

Visual Confidences

Object Descriptors



Visual Confidences

Action Descriptors

Dense Trajectories

[Wang et al. CVPR’11]

Descriptors

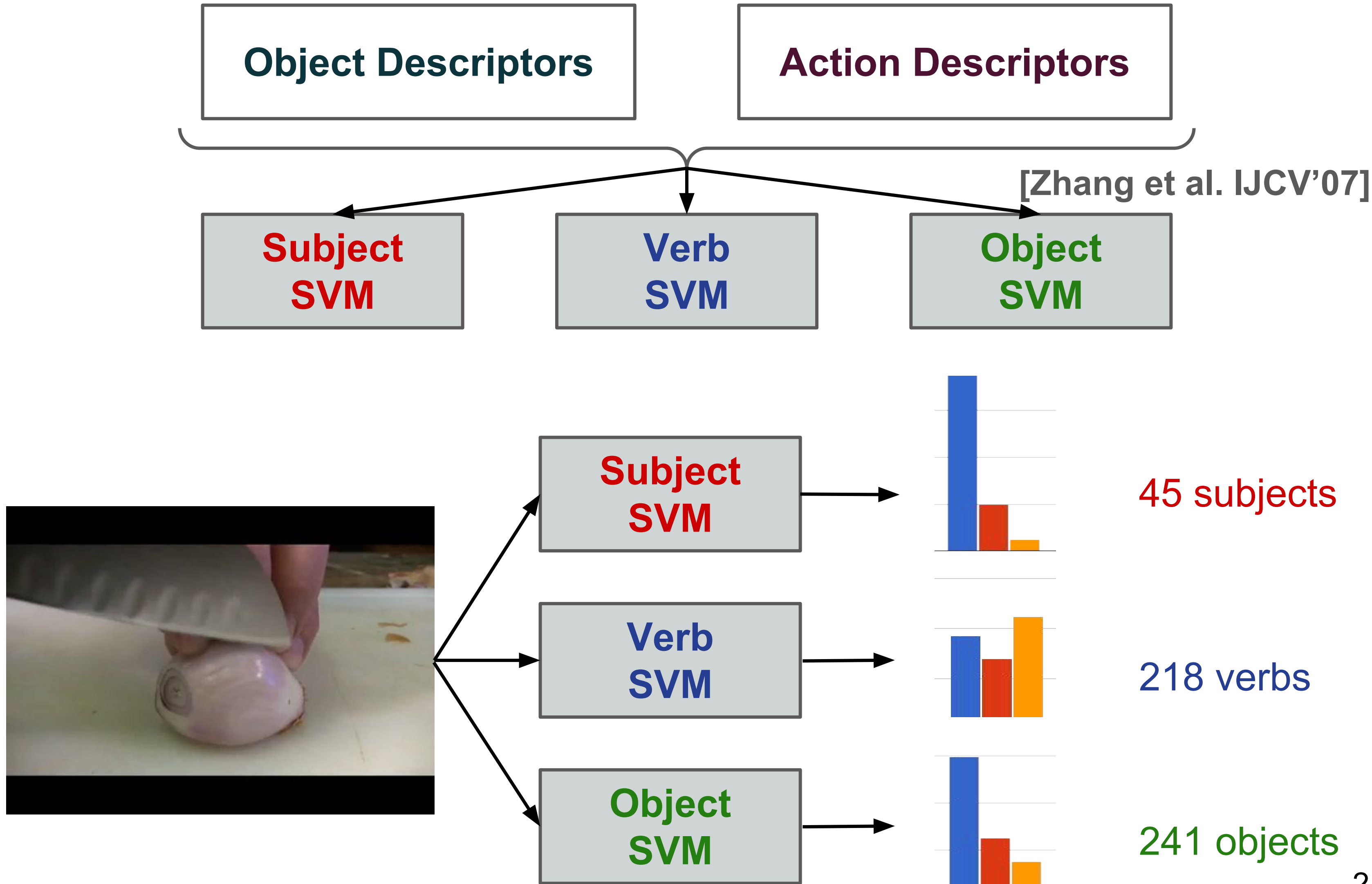
Trajectory
Histogram of Gradients (HoG)
Histogram of Flow (HoF)
Motion Boundary Histograms (MBH)

Action Descriptors

https://www.youtube.com/watch?v=mN5s3EHBz5I

Visual Confidences

Confidence Scores



Visual Confidences

Scene confidences

Extract Features

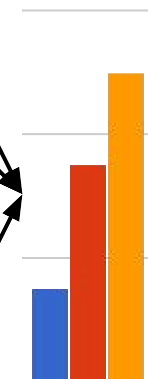
- GIST
- HOG2x2
- SSIM
- Dense SIFT
- Sparse SIFT
- LBP
- Color
- Texton
- Tiny Images

GIST SVM

HOG SVM

• • • • •

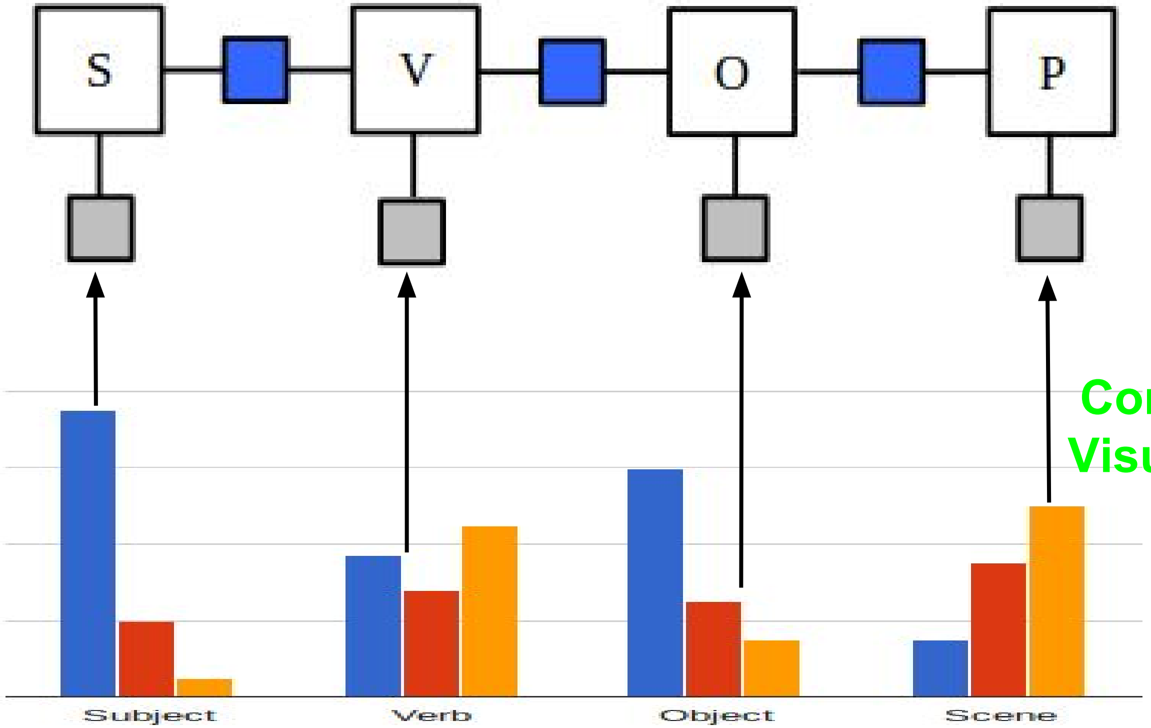
Tiny Images SVM



12 scenes

[Xiao et al. CVPR'10]

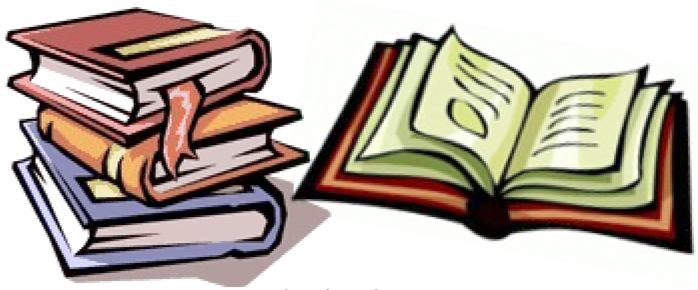
Observed Potentials



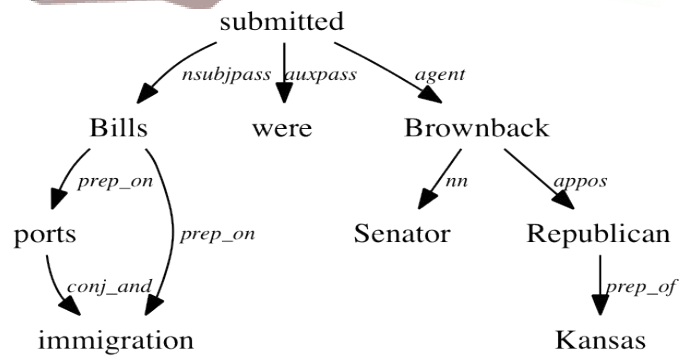
Confidences from Visual Recognition system

Language Statistics

External Corpora
ukWac, Wackypedia, Gigaword, BNC

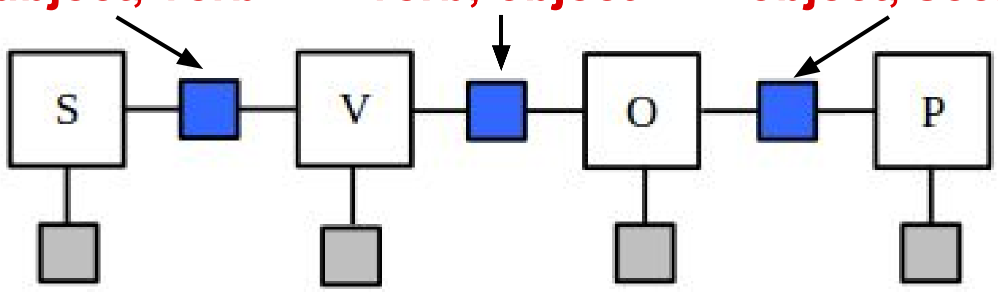


In-domain text
Textual descriptions accompanying the training videos

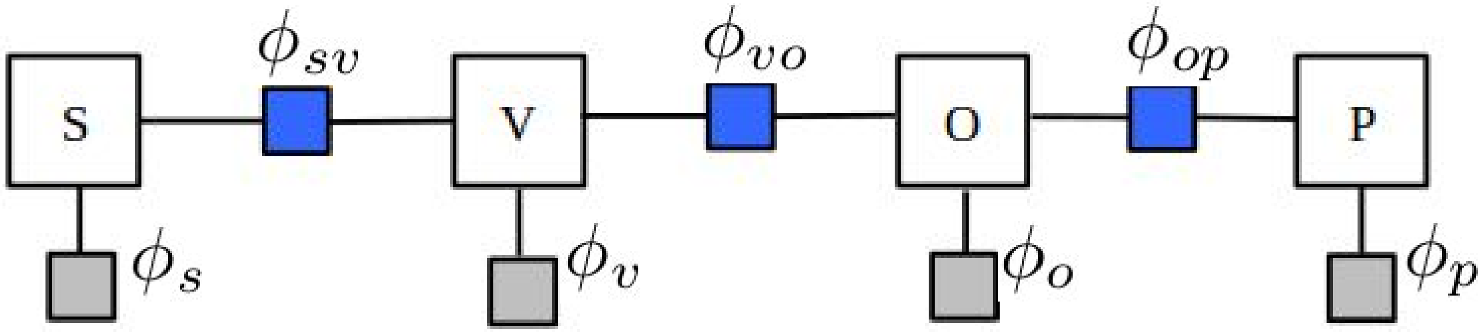


from **dependency parsed** text
obtain bigram statistics

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



Factor Graph



$$\phi_j(x) = C_j(x)$$ Vision confidence

$$j, k \in \{S, V, O, P\}$$

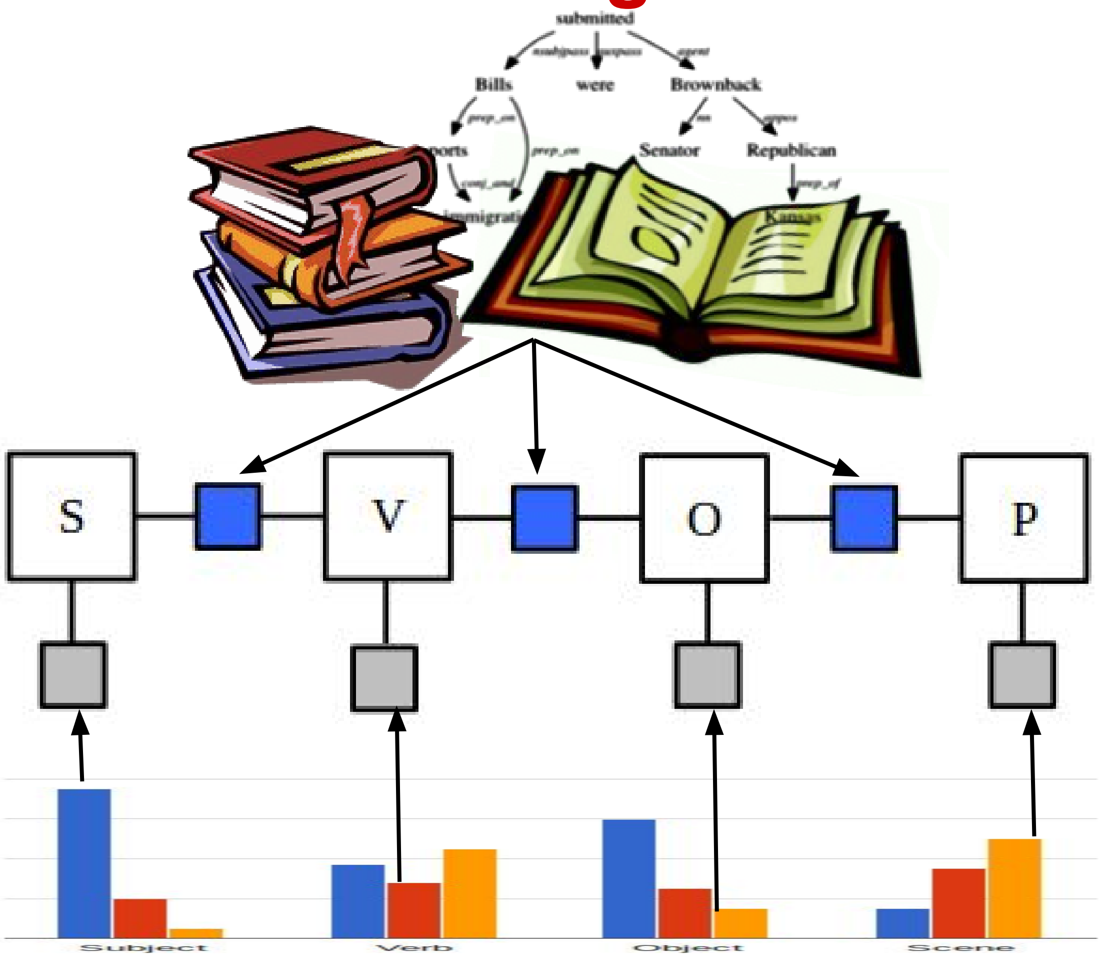
$$\phi_{j,k}(x, y) := P(i = y | j = x)$$ Language confidence

$$:= \alpha P_o(j = y | k = x) + (1 - \alpha) P_i(j = y | k = x)$$

P_o Out-of-domain P_i In-domain α - weight

Eg: $\phi_{V,O}(\text{ride}, \text{motorbike}) := p(\text{O=motorbike}|\text{V=ride})$ = 0.288

Content Planning: Inference on Factor Graph



Language Statistics from Text Corpora
(Gigaword, ukWac, Wackypedia, BNC)

Most likely

Subject

Verb

Object

Place

Confidences from Visual Recognition system

Surface Realization

(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.**

Experiments: Dataset

YouTube Videos [Chen & Dolan, ACL'11]

Link: http://www.cs.utexas.edu/users/ml/clamp/videoDescription/

- 1970 video snippets
  - 10-30s each
  - typically single activity
  - no dialogues
  - 1300 training, 670 test
- Annotations
  - Descriptions in multiple languages
  - ~40 English descriptions per video
  - descriptions and videos collected on AMT

Sample video and descriptions

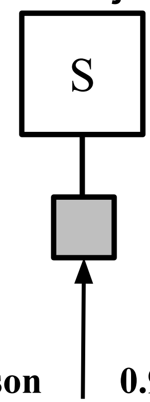
A man appears to be plowing a rice field with a plow being pulled by two oxen.
A man is plowing a mud field.
Domesticated livestock are helping a man plow.
A man leads a team of oxen down a muddy path.
A man is plowing with some oxen.
A man is tilling his land with an ox pulled plow.
Bulls are pulling an object.
Two oxen are plowing a field.
The farmer is tilling the soil.
A man in ploughing the field.

A man is walking on a rope.
A man is walking across a rope.
A man is balancing on a rope.
A man is balancing on a rope at the beach.
A man walks on a tightrope at the beach.
A man is balancing on a volleyball net.
"A man is walking on a rope held by poles
A man balanced on a wire.
The man is balancing on the wire.
A man is walking on a rope.
A man is standing in the sea shore.

Visual Confidences

Subjects

Extract entity descriptors and get visual confidence scores on 45 subjects.

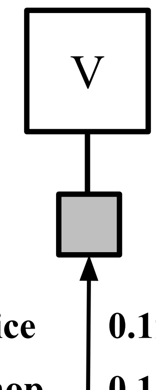


person	**0.95**
monkey	**0.0039**
animal	**0.0033**

Visual Confidences

Verb

Use spatio-temporal features to obtain visual confidence over 218 activities.

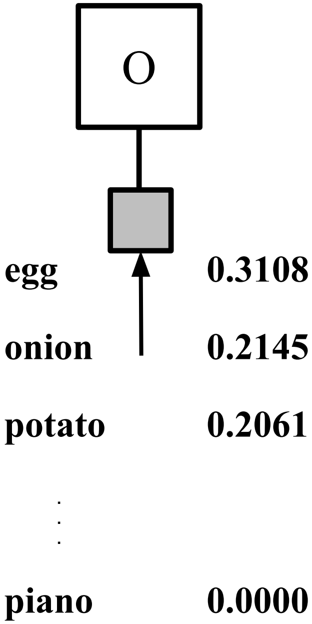


slice	**0.1909**
chop	**0.1098**
play	**0.0856**
⋮	
speak	**0.0000**

Visual Confidences

Objects

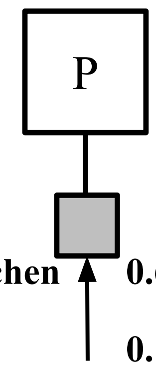
Extract entity descriptors and get visual confidence scores on 241 objects.



Visual Confidences

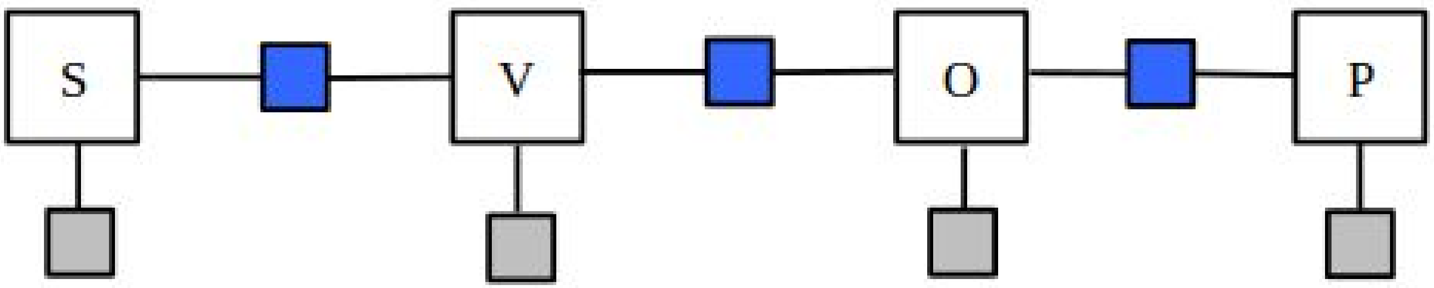
Scenes

Extract features (GIST, SIFT, HOG,..) and train classifiers for 12 scenes categories.



kitchen	**0.6381**
sky	**0.1638**
house	**0.0672**
⋮	
snow	**0.0014**

Visual Confidences



Subjects

person	0.9501
monkey	0.0039
animal	0.0033
⋮	
parrot	0

Verbs

slice	0.1909
chop	0.1098
play	0.0856
⋮	
speak	0.0000

Objects

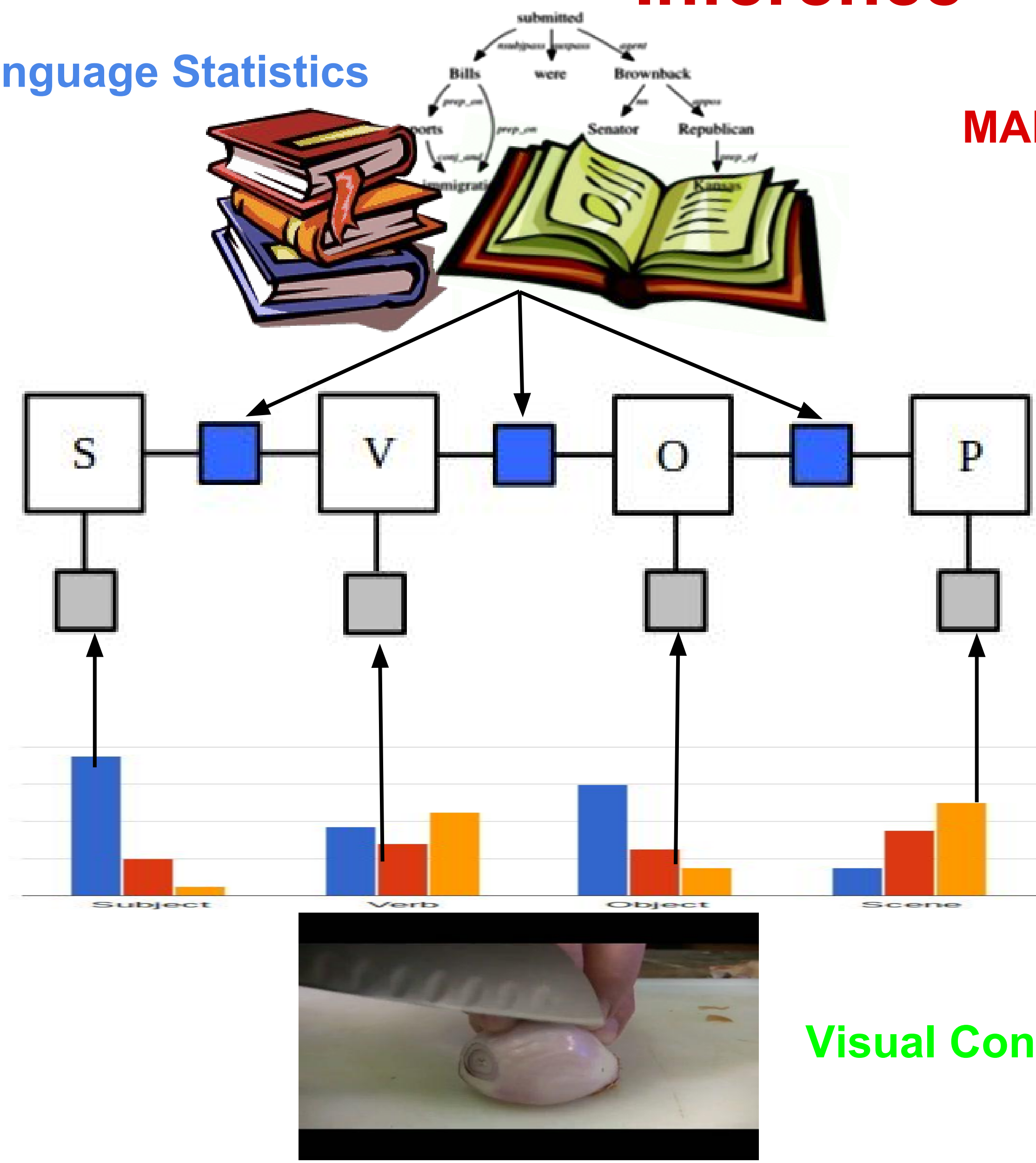
egg	0.3108
onion	0.2145
potato	0.2061
⋮	
piano	0.0000

Scenes

kitchen	0.6381
sky	0.1638
house	0.0672
⋮	
snow	0.0014

Inference

Language Statistics



Visual Confidences

MAP Inference on Factor Graph
estimates the most likely SVOP quadruple.

Subject	person
Verb	slice
Object	onion
Place	kitchen

Evaluation

Compare predicted subject, verb, object, scene with ground truth.

- ground truth (S,V,O,P) extracted by parsing.
- most frequent ground truth tuple
- any valid tuple

Binary accuracy: $s_{01}(v,l) = \mathbb{I}[v{=}{=}l]$

  - 1 if predicted equals ground truth, 0 otherwise

WUP similarity:

  - Partial credit

E.g : $s_{WUP}(\text{motorbike, dog}){=}0.10$ $s_{WUP}(\text{slice, chop}){=}0.80$

Results: Binary Accuracy

- n-gram: Similar to Krishnamoorthy et al.
- HVC: Highest Vision Confidence
- FGM: Factor Graph Model

Most	S%	V%	O%	[P]%	SVO%	SVO[P]%
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

bold: significantly better than HVC.
\+ :significantly better than n-gram.

Modest improvement over objects and scenes and overall tuple accuracy.

Results: WUP accuracy

Most	S%	V%	O%	[P]%	SVO%	SVO[P]%
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

bold: significantly better than HVC.
+ :significantly better than n-gram.
★ :significantly better than FGM

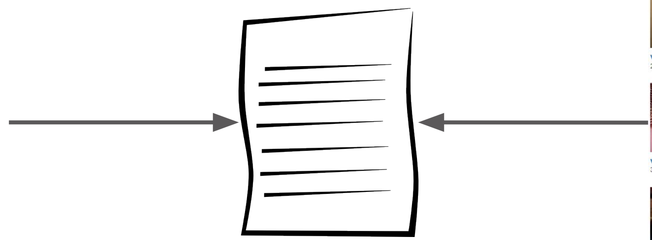
Modest improvements.

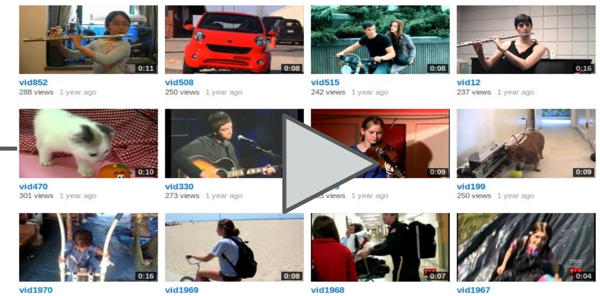
Language shows improvements when subject and object are detected reasonably well.

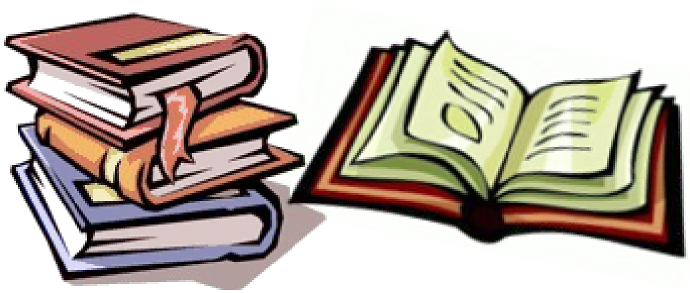
Demo Video

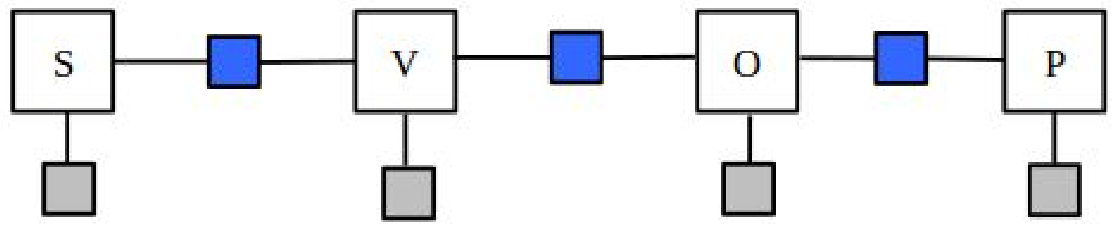
https://www.youtube.com/embed/pShM8CVAYxI

Summary/Conclusion

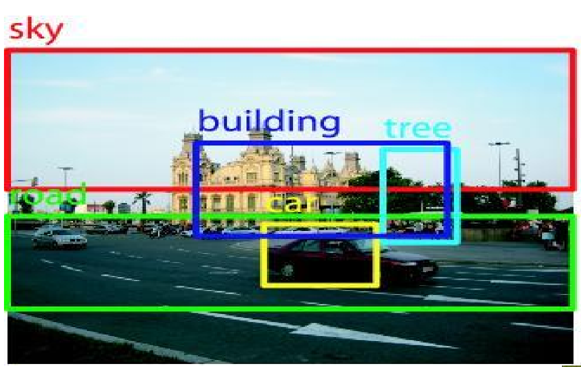


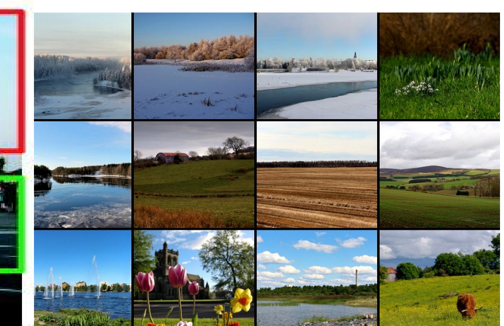






A person is slicing an onion in the kitchen.





Thank You

Project Page with Code: http://www.cs.utexas.edu/~vsub/fgm.html

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

Jesse Thomason*, Subhashini Venugopalan*, Sergio Guadarrama, Kate Saenko, Raymond Mooney
***equal contribution**

International Conference on Computational Linguistics, Dublin, Ireland, August 2014.
(COLING 2014)