University at Buffalo The State University of New York | REACHING OTHERS

CVPR 2013

PDF YouCook

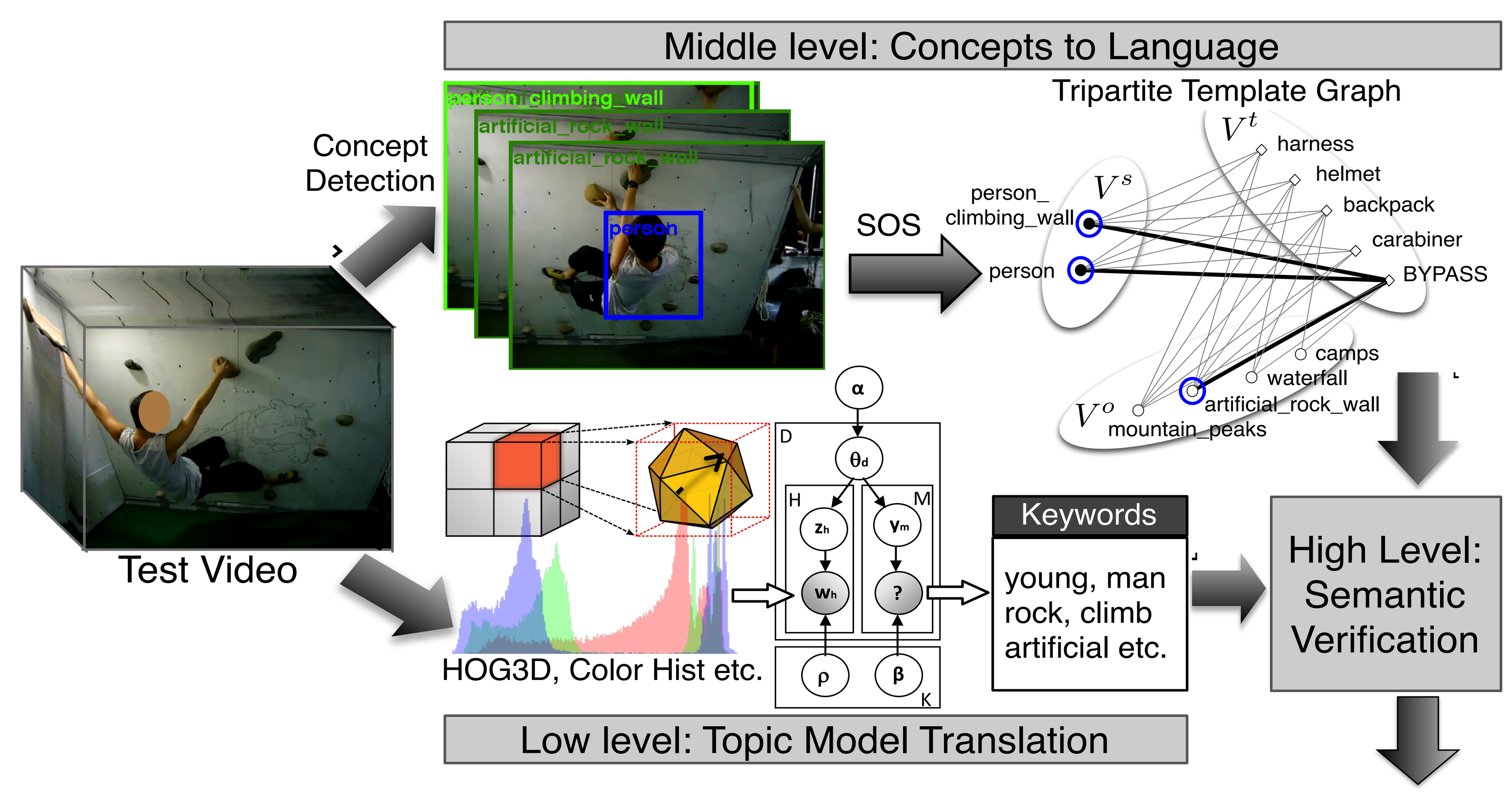
# A Thousand Frames in Just a Few Words:
## Lingual Description of Videos through Latent Topics and Sparse Object Stitching

Pradipto Das*, Chenliang Xu*, Richard F. Doell and Jason J. Corso
Department of Computer Science and Engineering - SUNY at Buffalo, Buffalo, NY
(* denotes equal contribution)

## Objective:

- Generate natural language descriptions of a video that incorporate fine-grained information extraction.
- Improve relevance of descriptions in a hybrid framework that leverages bottom-up keyword prediction semantically verified by top-down concept detection and tri-partite template graphs.

## System Overview:




Output from our system: 1. A person in on artificial rock wall 2. A person climbing a wall is on artificial rock wall 3. Person climbs rock wall indoors 4. Young man tries to climb artificial rock wall 5. A man demonstrates how to climb a rock wall

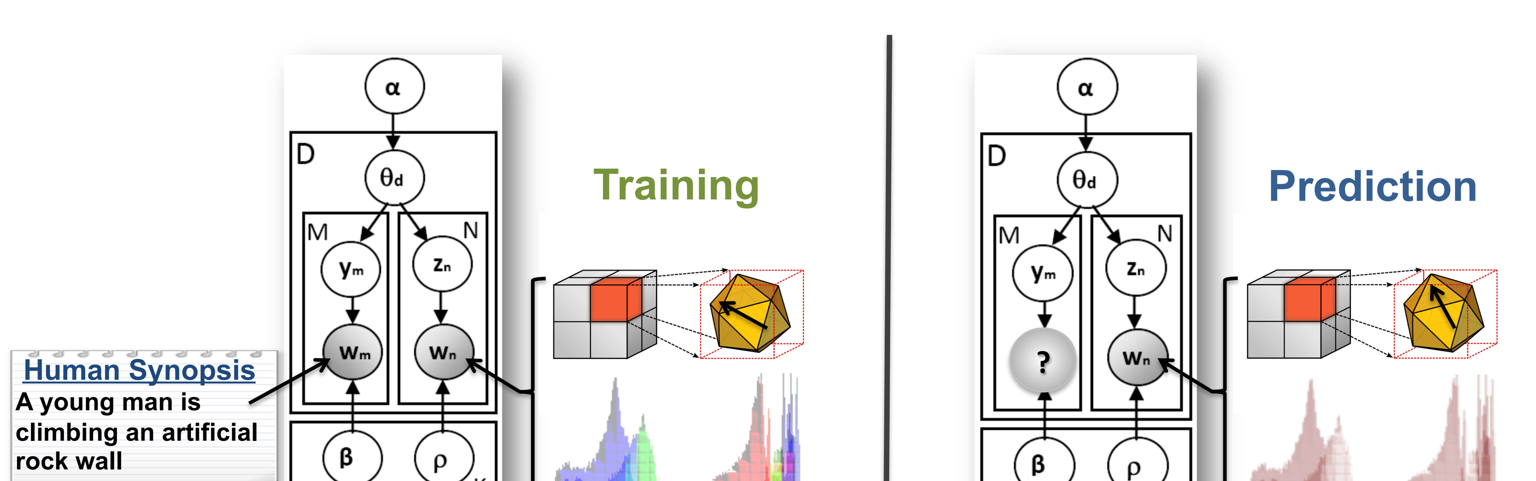
## Low Level: Topic Model Translation

- **Translation model** – capture semantic correlations from low level feature codebooks to bag-of-words
- Two families of multimodal (MM) topic models (LDA):
  - **Correspondence MMLDA** enforces a stronger constraint: Topic sparsity on low level features enforces stronger correspondence to text (Computationally expensive)
  - **MMLDA** operates more diffusely and focuses on *semantic summarization of multiple views* based on observation frequencies

**Why we need to build separate models for separate events that depend only on the event specific vocabularies?**

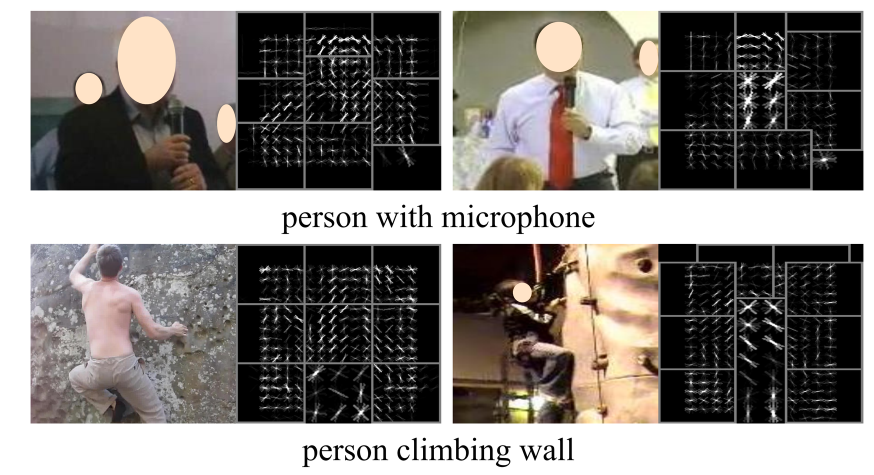
| Model | Actual Synopsis | Predicted Words (top 10) |
|---|---|---|
| Correspondence MMLDA | *man feeds fish bread* | fish jump bread fishing skateboard pole machine car dog cat |
| MMLDA | *man feeds fish bread* | bread shampoo sit condiment place fill plate jump pole fishing |

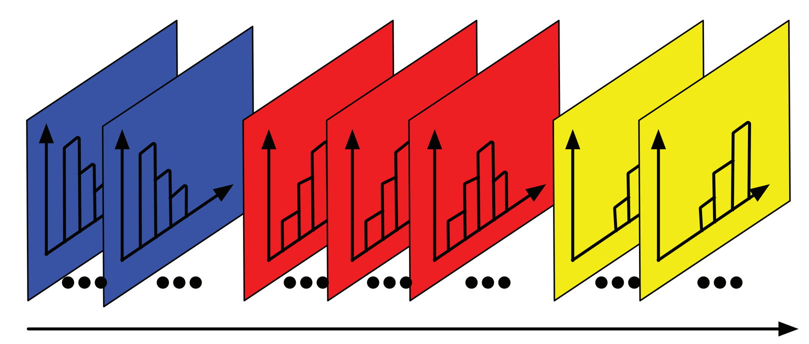
- Multiple low level descriptors are shared across many events (common actions)
  - Leads to misleading semantic correspondence to text
  - **Prediction *quality* degenerates rapidly!**




## Middle Level: Concepts to Language

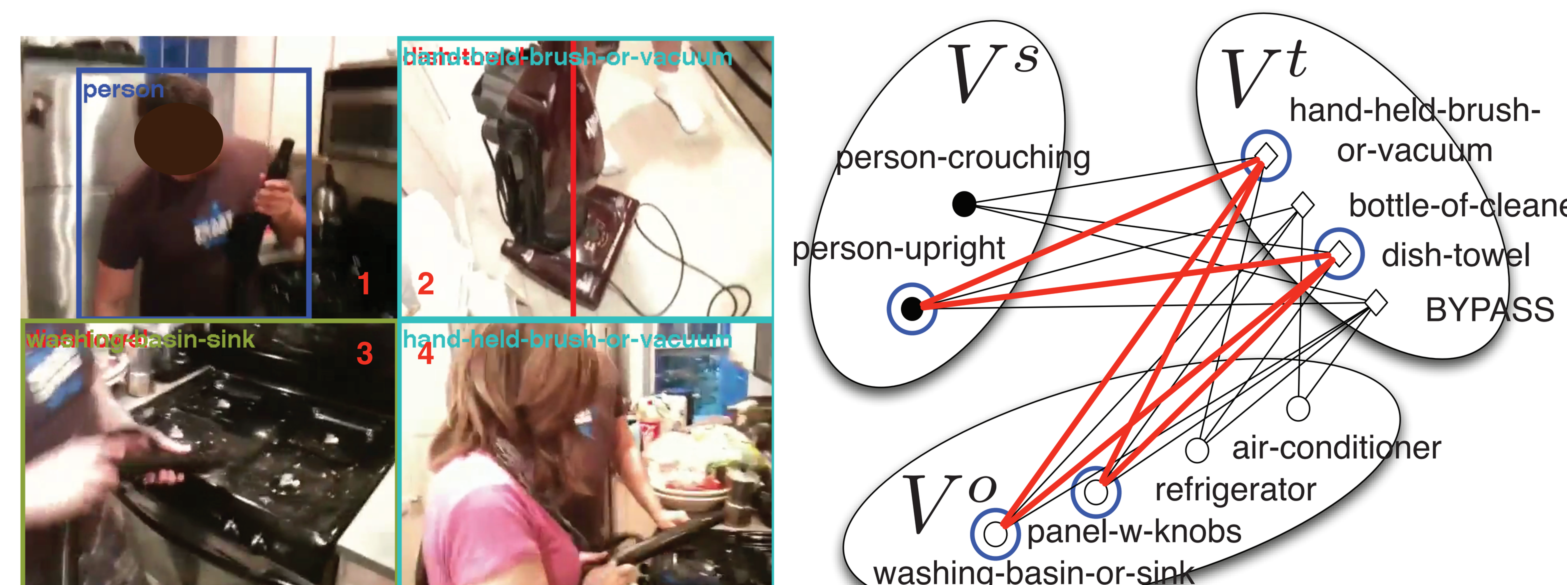
- Concept Detectors
  - -- rich semantics from object, action and scene level
  - -- reduce visual complexity
  - -- similar to "visual phrases"
  - -- deformable parts models are the base detector



- Sparse Object Stitching
  - -- segment video into a set of concept shots
  - -- Record distribution of detected concepts per shot
  - -- avoids need to do expensive dense detection and tracking

- Tripartite Template Graph for sentence generation.




Template Language Output:
**([a person])** *is using* **([dish towel]** *and* **[hand held brush or vacuum])** *to clean* **([panel with knobs]** *and* **[washing basin or sink])**

## High Level: Semantic Verification

- Rank nearest neighbor sentences from the training synopses by a ranking function
$$r_s = bh(w_1 x_{s_1} + w_2 x_{s_2})$$
  - $b$ (boolean): at least two semantically verified semantically verified
  - $h$ (boolean): at least one human subject
  - $x_{s_1}$ (real): ratio of *the total # of matches* to *the # of words in the sentence*
  - $x_{s_2}$ (real): sum of the weights of the predicted words from the topic model

---

- Run MMLDA on a vocabulary of training synopses and training concept annotations
- **Semantic Verification**: computing # of topic rank inversions for two ranked lists

$L_{synopsis}$
$$= \left\langle \left\{ k: \sum_{j=1}^{V} \sum_{m=1}^{P} p(w_m|\beta_k)\delta(w_m^j) \right\}^{\uparrow} \right\rangle$$
$L_{concept}$
$$= \left\langle \left\{ k: \sum_{j=1}^{corrV} \sum_{n=1}^{C} p(w_n|\rho_k)\delta(w_n^j) \right\}^{\uparrow} \right\rangle$$

- $P$: number of word predictions; $C$ : number of positive concept detections

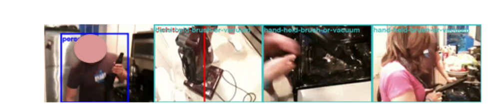
- ***A person is speaking to a large group of standing people and a small group of standing people with board in the back and a camera man*** **[Sentence from SOS]**
- **Representative speaks to crowd at town hall meeting**
- **A man speaks at a town hall meeting about health care**
- **Man opens town hall meeting care**
- **People get angry at town hall meeting on health care**
- **Politician gives speech at town hall meeting**

**Concept Objects**
Upright camera-man, group of sitting person, group of standing person, man with microphone

**hall/OBJ town/NOUN meeting/VERB**
**man/SUBJ-HUMAN speaks/VERB**
**microphone/OBJ talking/VERB**
**representative/SUBJ-HUMAN health/NOUN**
**care/NOUN politician/SUBJ-HUMAN**
**chairs/NOUN flags/OBJ people/OBJ crowd/OBJ**

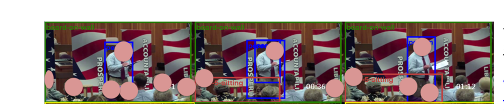
## TRECVID MER12 Dataset:

- Videos are in 5 Events
- Annotation: Human descriptions/synopses & concept bounding boxes
- Train:
  - -- Concept Detector: 200 Videos/Event
  - -- Topic Model: 120 Videos/Event (only positive instances)
- Test:
  - -- 6 Videos/Event (MER12 Test Set for Recounting)



Cleaning an appliance

**Keywords**: refrigerator/OBJ cleans/VERB man/SUBJ-HUMAN clean/VERB blender/OBJ cleaning/VERB woman/SUBJ-HUMAN person/SUBJ-HUMAN stove/OBJ microwave/OBJ sponge/NOUN food/OBJ home/OBJ hose/OBJ oven/OBJ
**Sentences from Our System** 1. *A person is using dish towel and hand held brush or vacuum to clean panel with knobs and washing basin or sink* 2: Man cleaning a refrigerator. 3: Man cleans his blender. 4: Woman cleans old food out of refrigerator. 5: Man cleans top of microwave with sponge.
**Human Synopsis**: Two standing persons clean a stove top with a vacuum clean with a hose.

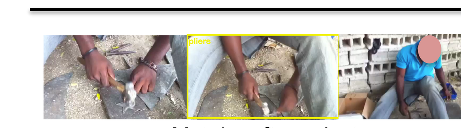


Town hall meeting

**Keywords**: meeting/VERB town/NOUN hall/OBJ microphone/OBJ talking/VERB people/OBJ podium/OBJ speech/OBJ woman/SUBJ-HUMAN man/SUBJ-HUMAN chairs/NOUN clapping/VERB speaks/VERB questions/VERB giving/VERB
**Sentences from Our System** 1: *A person is speaking to a small group of sitting people and a small group of standing people with board in the back.* 2: A person is speaking to a small group of standing people with board in the back 3: Man opens town hall meeting. 4: Woman speaks at town meeting. 5: Man gives speech on health care reform at a town hall meeting.
**Human Synopsis**: A man talks to a mob of sitting persons who clap at the end of his short speech.

Renovating home

**Keywords**: people/SUBJ-HUMAN, home/OBJ, group/OBJ, renovating/VERB, working/VERB, montage/OBJ, stop/VERB, motion/OBJ, appears/VERB, building/VERB, floor/OBJ, tiles/OBJ, floorboards/OTHER, man/SUBJ-HUMAN, laying/VERB
**Sentences from Our System**: 1. *A person is using power drill to renovate a house*. 2. *A crouching person is using power drill to renovate a house*. 3. A person is using trowel to renovate a house. 4: man lays out underlay for installing flooring. 5: A man lays a plywood floor in time lapsed video.
**Human Synopsis**: Time lapse video of people making a concrete porch with sanders, brooms, vacuums and other tools.



Metal crafts project

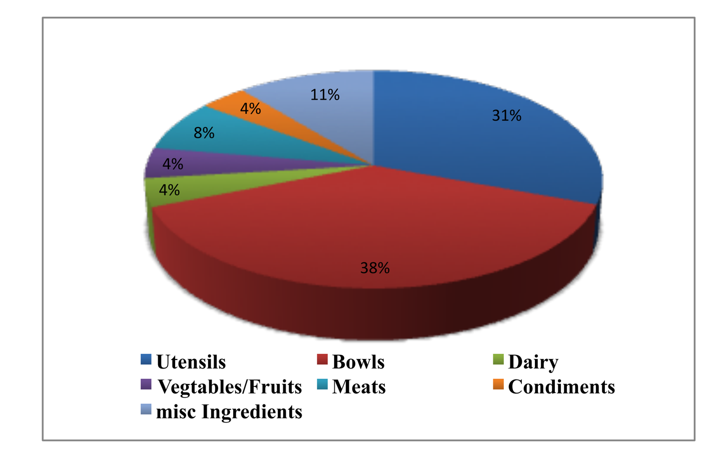
**Keywords**: metal/OBJ man/SUBJ-HUMAN bending/VERB hammer/VERB piece/OBJ tools/OBJ rods/OBJ hammering/VERB craft/VERB iron/OBJ workshop/OBJ holding/VERB works/VERB steel/OBJ bicycle/OBJ
**Sentences from Our System** 1. *A person is working with pliers.* 2 Man hammering metal. 3. Man bending metal in workshop. 4. Man works various pieces of metal. 5. A man works on a metal craft at a workshop.
**Human Synopsis**: A man is shaping a star with a hammer.

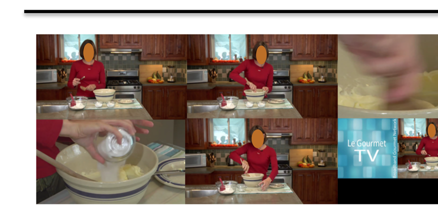
### Automatic Evaluation through ROUGE-1 Metric

| Events | Precision BASELINE | Precision OURS (Low, Middle and High) | | | Recall BASELINE | Recall OURS (Low, Middle and High) | | |
|---|---|---|---|---|---|---|---|---|
| Cleaning appliance | 20.03 | 17.52 | 11.69 | 10.68 | 19.16 | 32.6 | 35.76 | **48.15** |
| Renovating home | 6.66 | 15.29 | **12.55** | 9.99 | 7.31 | 43.41 | 30.67 | **49.52** |
| Rock climbing | 24.45 | 16.21 | 24.52 | 12.61 | 44.09 | 59.22 | 46.23 | **65.84** |
| Town hall meeting | 17.35 | 14.41 | **27.56** | 13.36 | 13.8 | 28.66 | 45.55 | **56.44** |
| Metal crafts project | 16.73 | 18.12 | **31.68** | 15.63 | 19.01 | 41.87 | 25.87 | **54.84** |

## YouCook Dataset:

Cooking videos downloaded from YouTube
Splits: Train 49 / Test 39
Annotations:
-- Human Descriptions from MTurk
-- Object and Action Bounding Boxes
ROUGE Benchmark Scoring



Cooking video 1

**Keywords**: bowl/OBJ pan/OBJ video/OBJ adds/VERB lady/OBJ pieces/OBJ ingredients/OBJ oil/OBJ glass/OBJ liquid/OBJ butter/SUBJ-HUMAN woman/SUBJ-HUMAN add/VERB stove/OBJ salt/OBJ
**Sentences from Our System**: 1. A *person is cooking butter with bowl and stovetop*. 2. In a pan add little butter. 3. She adds some oil and a piece of butter in the pan. 4. A woman holds up Bisquick flour and then adds several ingredients to a bowl. 5. A woman adds ingredients to a blender.
**Human Synopsis1**: A lady wearing red colored dress, blending (think butter) in a big sized bowl. Besides there is 2 small bowls containing white color powders. It may be maida flour and sugar. After she is mixing the both powders in that big bowl and blending together. **Human Synopsis2**: In this video, a woman first adds the ingredients from a plate to a large porcelain bowl. She then adds various other ingredients from various different bowls. She then mixes all the ingredients with a wooden spoon.

Cooking video 2

**Keywords**: bowl/OBJ pan/OBJ video/OBJ adds/VERB ingredients/OBJ lady/OBJ woman/SUBJ-HUMAN add/VERB pieces/OBJ stove/OBJ oil/OBJ put/VERB added/VERB mixes/VERB glass/OBJ
**Sentences from Our System**: 1. A *person is cooking with pan and bowl. 2. A person is cooking with pan*. 2. A woman adds ingredients to a blender. 2. In this video, a woman adds a few ingredients in a glass bowl and mixes them well. 3. In this video, a woman first adds the ingredients from a plate to a large porcelain bowl 4. The woman is mixing some ingredients in a bowl. 5. the woman in the video has a large glass bowl.
**Human Synopsis1**: The woman is giving directions on how to cook bacon omelette. She shows the ingredients for cooking and was frying the bacon, scrambling the egg, melting the butter and garnishing it with onions and placed some cheese on top. The woman then placed the scrambled egg and bacon to cook and then placed it on a dish. **Human Synopsis2**: In this video the woman takes bacon, eggs, cheese , onion in different containers. On a pan she cooks the bacon on low flame. Side by side she beats the eggs in a bowl. she removes the cooked bacon on a plate. In the pan she fries onions and then adds the beaten eggs. She sprinkles grated cheese on the pan and cooks well. She then adds the fried bacon on the eggs in the pan and cook well. She transfers the cooked egg with bacon to as serving plate.

| Precision Bi/Uni-gram BASELINE | | Precision Bi/Uni-gram OURS | | Recall Bi/Uni-gram BASELINE | | Recall Bi/Uni-gram OURS | |
|---|---|---|---|---|---|---|---|
| 0.0006 | 15.47 | **5.04** | **24.82** | 0.0006 | 19.02 | **6.81** | **34.2** |

**Acknowledgements.** This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0090), the DARPA Mind's Eye program (W911NF-10-2-0062), and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding 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, DARPA, ARO, NSF or the U.S. Government. We thank Simon Fraser Univ. & Kitware Inc. for support in feature extraction and the anonymous reviewers for their comments. We also thank the in-house annotators Philip Rosebrough,Yao Li, and Cody Boppert, without whom this project would not have been complete.