# Grounding Action Descriptions in Videos

Michaela Regneri, Marcus Rohrbach,
Dominikus Wetzel, Stefan Thater, Bernt Schiele
and Manfred Pinkal (2013)

# 0. Content

# 1.1. Background (Previous Research)

A substantial amount of resources combining **natural language** and **visual information** has been produced over the past decade, e.g.:

- *ESP game* data (Ahn and Dabbish, 2004)
- Microsoft Video Description Corpus (Chen and Dolan, 2011)
- *Restaurant Game* data (Orkin and Roy, 2009)

Some **limitations** remain:

- Low quality and short duration of video sequences
- Brevity and short span of textual descriptions
- No grounding in real-world actions

# 1.1. Background (Previous Research)

A substantial amount of resources combining **natural language** and **visual information** has been produced over the past decade, e.g.:

- *ESP game* data (Ahn and Dabbish, 2004)
- Microsoft Video Description Corpus (Chen and Dolan, 2011)
- *Restaurant Game* data (Orkin and Roy, 2009)

Some **limitations** remain:

- Low quality and short duration of video sequences
- Brevity and short span of textual descriptions
- No grounding in real-world actions

# 1.2. Background (Current Objectives)

Address such shortcomings by providing:

1) A new **multimodal corpus** (TACOS)

➔ High-resolution video + sentence-level annotations

2) Visually grounded **action-description dataset** (ASim)

➔ Gold-standard for similarity model evaluation

➔ Subset of TACOS

# 2.1. TACOS (Overview)

What is **TACOS**?

# 2.1. TACOS (Overview)

What is **TACOS**?

Saarbrücken

Corpus of

**T**extually

**A**nnotated

**Co**oking

**S**cenes

# 2.1. TACOS (Overview)

What is **TACOS**? What does **TACOS** offer?

Saarbrücken

Corpus of

**T**extually

**A**nnotated

**Co**oking

**S**cenes

# 2.1. TACOS (Overview)

What is **TACOS**?

Saarbrücken

Corpus of

**T**extually

**A**nnotated

**Co**oking

**S**cenes

What does **TACOS** offer?

- Textual descriptions of **cooking activities**
- Alignment of **sentences & video segments**
- Alignment of **complex and low-level actions**
- **Paraphrase** collection
- 17.334 action descriptions realizing 11.796 different sentences

# 2.2. TACOS (Ingredients)

**TACOS**

# 2.2. TACOS (Ingredients)

**TACOS**

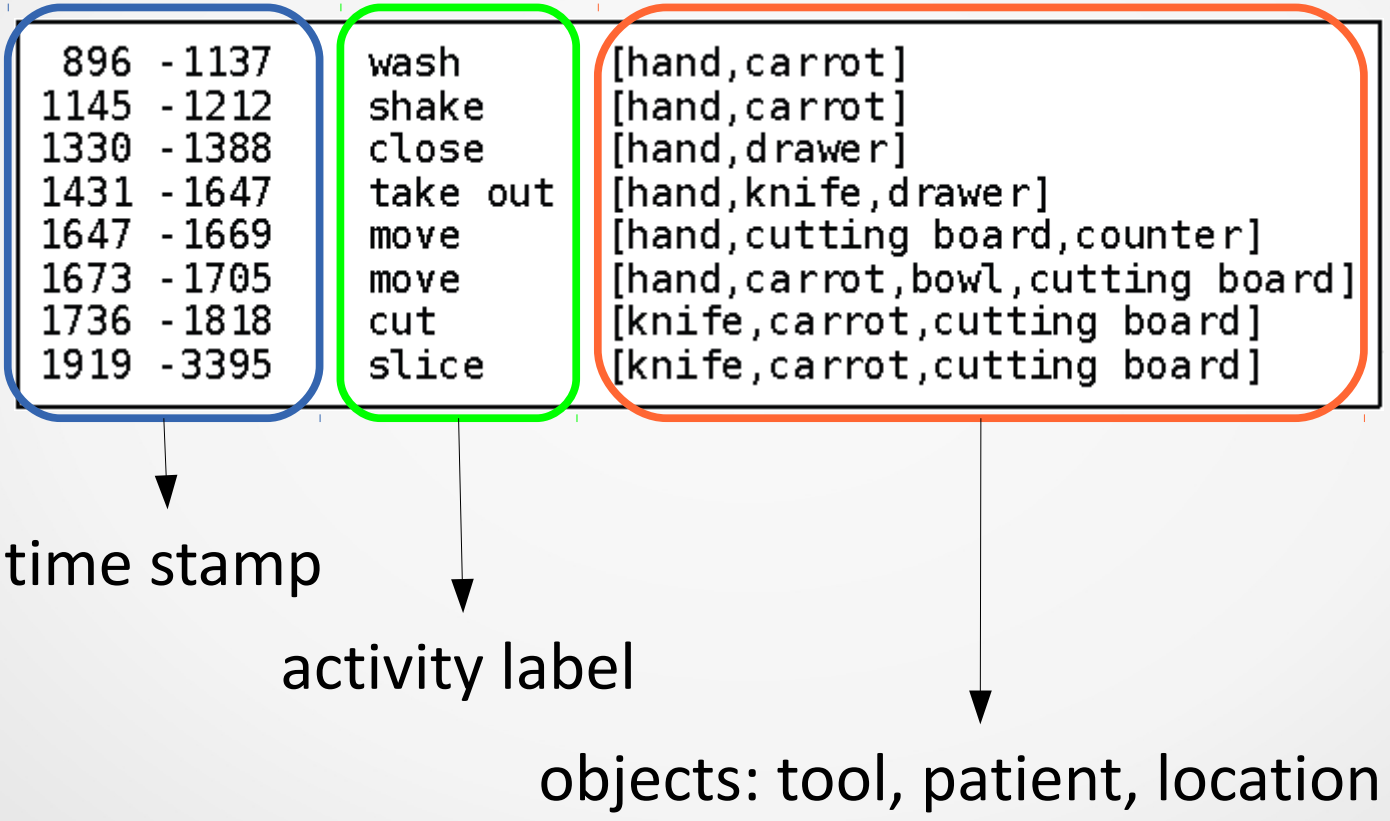
Video Corpus

**MPII Cooking Composite Activities**:

- 212 high-resolution recordings (1-23 min.)
- **41 basic cooking tasks** (each recorded 4-8 times)
- Expert annotations of **low-level activity tags**
  - 60 activity labels (e.g. STIR)

# 2.2. TACOS (Ingredients)

**Low-level annotations:**



# 2.2. TACOS (Ingredients)

**TACOS**

Video Corpus

**MPII Cooking Composite Activities**:

- 212 high-resolution recordings (1-23 min.)
- **41 basic cooking tasks** (each recorded 4-8 times)
- Expert annotations of **low-level activity tags**
  - 60 activity labels (e.g. STIR)

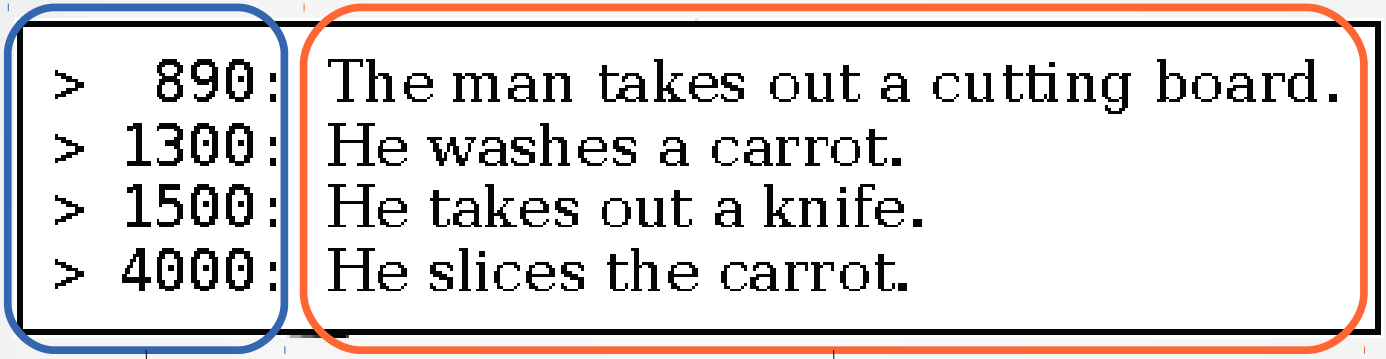
Text Data

Obtained via Amazon **Mechanical Turk**:

- For 26 tasks (>= 4 recordings)
  - Involve manipulation of cooking ingredients
- 5-15 English sentences per video
- Annotation onset = **ending timestamp** of described action

# 2.2. TACOS (Ingredients)

**MTurk data:**




annotation onset
/ end of action

natural language action description

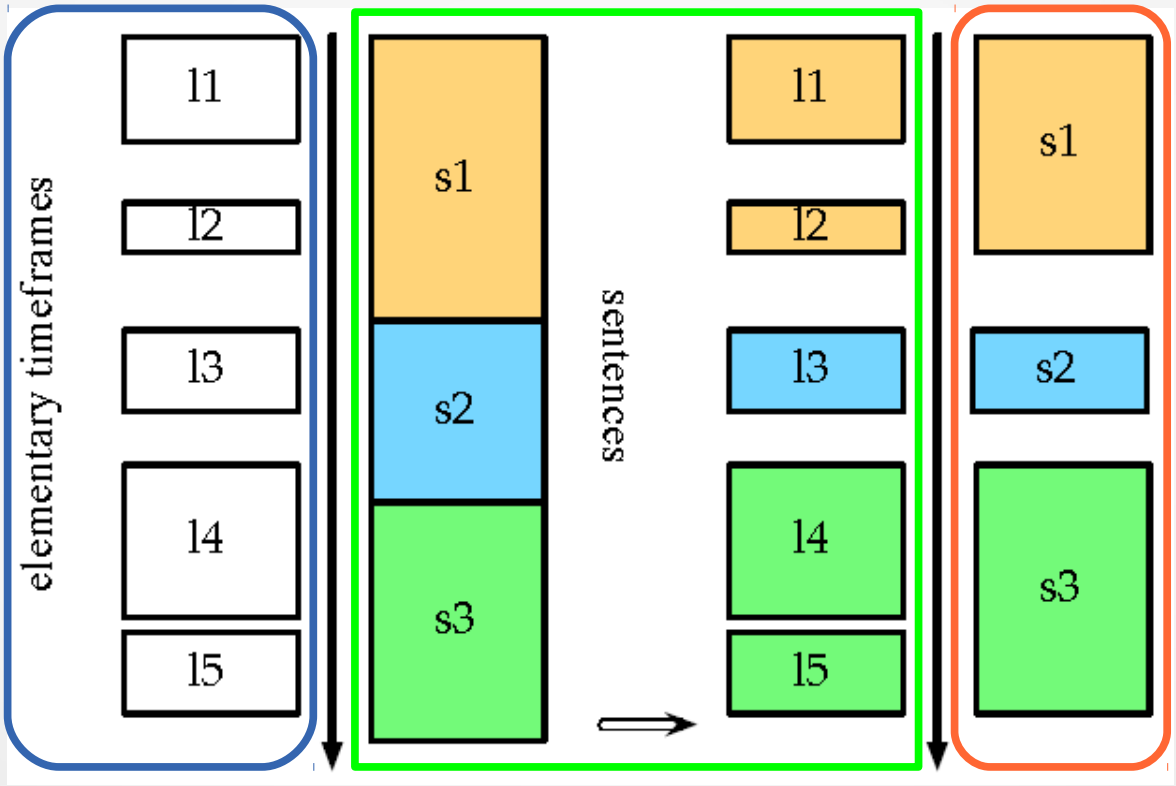
# 2.3. TACOS (Recipe)

**Core Idea:** Align data by matching time stamps

# 2.3. TACOS (Recipe)

**Core Idea:** Align data by matching time stamps

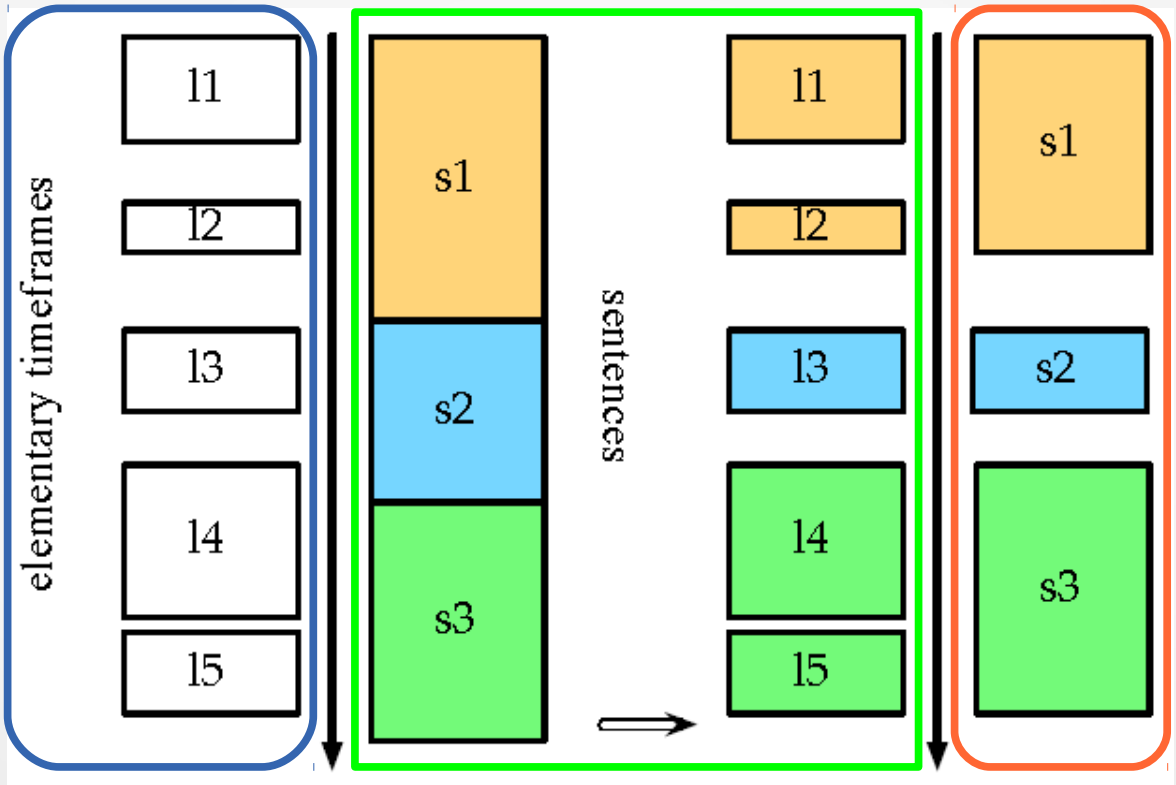
frames containing low-level actions (MPII CCA)



# 2.3. TACOS (Recipe)

**Core Idea:** Align data by matching time stamps

frames containing low-level actions (MPII CCA)

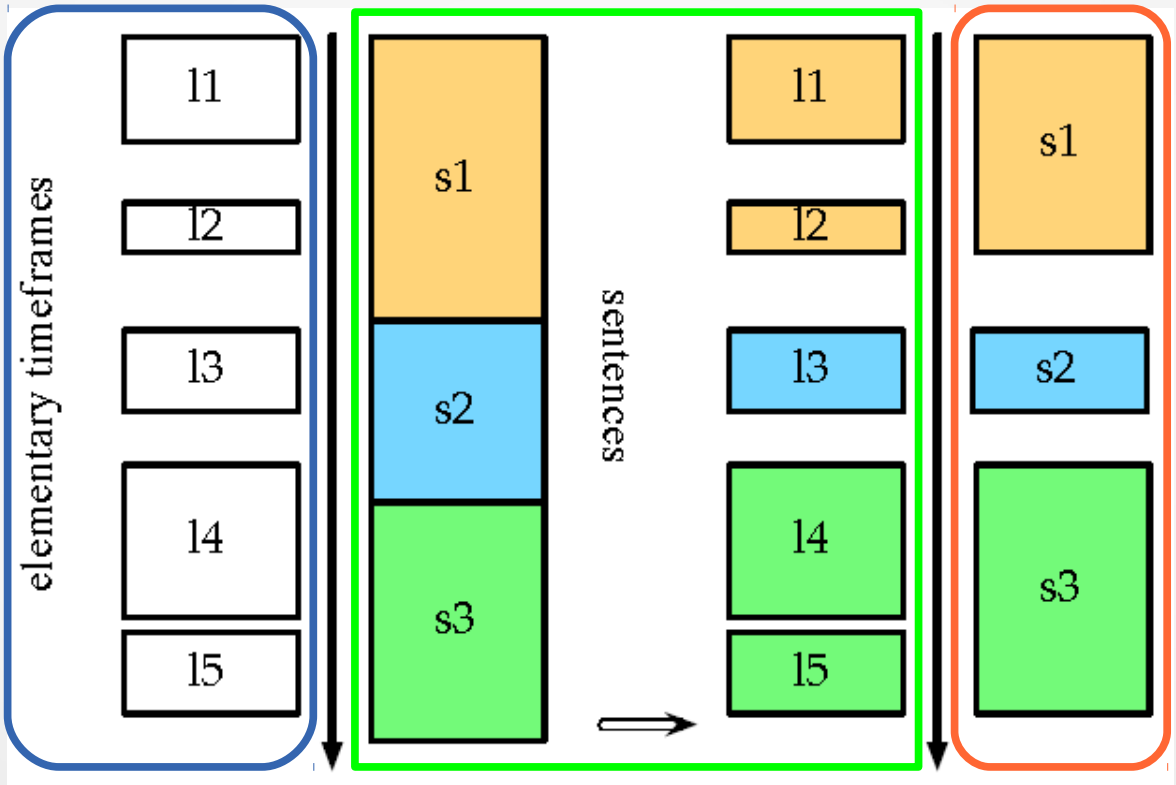



elementary frames mapped onto sentence frames (iff **>=1/2** of former covered by latter)

# 2.3. TACOS (Recipe)

**Core Idea:** Align data by matching time stamps

frames containing low-level actions (MPII CCA)




final sentence frames (less noise)

elementary frames mapped onto sentence frames (iff **>=1/2** of former covered by latter)

# 2.3. TACOS (Recipe)

## Description to low-level activity ratios:

| Sample frame | Start | End | Action | Participants | NL Sequence 1 | NL Sequence 2 | NL Sequence 3 |
|---|---|---|---|---|---|---|---|
|  | 743 | 911 | wash | hand, carrot | He washed carrot | The person rinses the carrot. | He rinses the carrot from the faucet. |
|  | 982 | 1090 | cut | knife, carrot, cutting board | He cut off ends of carrots 1:1 | The person cuts off the ends of the carrot. | He cuts off the two edges. |
|  | 1164 | 1257 | open | hand, drawer | | 1:3 | 1:2 |
|  | 1679 | 1718 | close | hand, drawer | | | He searches for something in the drawer, failed attempt, he throws away the edges in trash. |
|  | 1746 | 1799 | trash | hand, carrot | | The person searches for the trash can, then throws the ends of the carrot away. | |

# 2. TACOS

- Varying **degrees of granularity** in the description of activities indicated by the number of **subsumed low-level tasks** (1 to 10+)

- **High linguistic variance** of collected sentence-level descriptions: 58 low-level activities realized by 435 verb lemmas

- Minor alignment errors, but high accuracy in general

# 3.1. ASim (Overview)

What is **ASim**?

**A**ction

**Sim**ilarity

Dataset

# 3.1. ASim (Overview)

What is **ASim**?

**A**ction

**Sim**ilarity

Dataset

What does **ASim** offer?

- Subset of TACOS:
  - Restricted to **ingredient manipulation**
  - Limited to **reasonably frequent** activities
- **Sentence pairs (+ video segments)** with human annotator assigned **similarity scores**
  - Assessment of **activity similarity**
- A **gold standard** for evaluation of visually grounded models of action similarity

# 3.2. ASim (Structure)

- Three subsets within ASim:
  - *Different activity, same object*:
    - Models **semantic relations between actions**
  - *Same activity, same object*:
    - Head verbs may not always agree
    - Differing degrees of **underspecification** in action descriptions
  - *Same activity & verb, different object*:
    - Contribution of the **object's meaning** to the complete action

- Similarity judgments obtained from native English speakers

# 3.2. ASim (Structure)

| Part of Gold Standard | Sim | $\sigma$ | $\rho$ |
|---|---|---|---|
| DIFF. ACTIVITY, SAME OBJECT | 2.20 | 1.07 | 0.73 |
| SAME ACTIVITY, SAME OBJECT | 4.19 | 1.04 | 0.73 |
| ALL WITH SAME OBJECT | 3.20 | 1.44 | 0.84 |
| SAME VERB, DIFF. OBJECT | 3.34 | 0.69 | 0.43 |
| COMPLETE DATASET | 3.27 | 1.15 | 0.73 |

Spearman's ρ: Evaluates how samples are rated **relative to each other**

# 3.2. ASim (Structure)

| Part of Gold Standard | Sim | $\sigma$ | $\rho$ |
|---|---|---|---|
| DIFF. ACTIVITY, SAME OBJECT | 2.20 | 1.07 | 0.73 |
| SAME ACTIVITY, SAME OBJECT | 4.19 | 1.04 | 0.73 |
| ALL WITH SAME OBJECT | 3.20 | 1.44 | 0.84 |
| SAME VERB, DIFF. OBJECT | 3.34 | 0.69 | 0.43 |
| COMPLETE DATASET | 3.27 | 1.15 | 0.73 |

Spearman's ρ: Evaluates how samples are rated **relative to each other**

Similarity assessment for **different variants** of the **same activity** is hard even for human annotators

# 4.1. Experimental Evaluation (Setup)

**Task: Action similarity evaluation**

**Text-based models:**

**Jaccard Coefficient**: Ratio between the size of the intersection and the union of distinct words in two sentences

**Vector model (Thater et al. 2011)**: Cosine similarity of two sentence vectors (each a sum of contextualized vectors of all content words in the sentence)

**Video Based models:**

**Raw Visual Features (Wang et al. 2011):** Intersection of histograms encoding video features (greadients, flow etc.)

**Visual Classifiers (Rohrbach et al. 2012)**: Cosine similarity of two classifier output vectors (components = likelihood of an activity or object to have appeared in a video segment); supervised

# 4.1. Experimental Evaluation (Setup)

**Task: Action similarity evaluation**

**Text-based models:**

**Jaccard Coefficient**: Ratio between the size of the intersection and the union of distinct words in two sentences

**Vector model (Thater et al. 2011)**: Cosine similarity of two sentence vectors (each a sum of contextualized vectors of all content words in the sentence)

**Video Based models:**

**Raw Visual Features (Wang et al. 2011):** Intersection of histograms encoding video features (gradients, flow etc.)

**Visual Classifiers (Rohrbach et al. 2012)**: Cosine similarity of two classifier output vectors (components = likelihood of an activity or object to have appeared in a video segment); supervised

# 4.2. Experimental Evaluation (Results)

| | Model | Same Object | Same Verb | Overall |
|---|---|---|---|---|
| Text | Jaccard | 0.28 | 0.25 | 0.25 |
| | Textual Vectors | 0.30 | 0.25 | 0.27 |
| | Text Combined | 0.39 | **0.35** | 0.36 |
| Video | Visual Raw Vectors | 0.53 | -0.08 | 0.35 |
| | Visual Classifier | 0.60 | 0.03 | 0.44 |
| | Video Combined | 0.61 | -0.04 | 0.44 |
| Mix | All Unsupervised | 0.58 | 0.32 | 0.48 |
| | All Combined | **0.67** | 0.28 | **0.55** |
| Upper Bound | | 0.84 | 0.43 | 0.73 |

Measure: Spearman's ρ

Combined: Average over normalized scores

# 4.2. Experimental Evaluation (Results)

| | Model | Same Object | Same Verb | Overall |
|---|---|---|---|---|
| Text | Jaccard | 0.28 | 0.25 | 0.25 |
| | Textual Vectors | 0.30 | 0.25 | 0.27 |
| | Text Combined | 0.39 | **0.35** | 0.36 |
| Video | Visual Raw Vectors | 0.53 | -0.08 | 0.35 |
| | Visual Classifier | 0.60 | 0.03 | 0.44 |
| | Video Combined | 0.61 | -0.04 | 0.44 |
| Mix | All Unsupervised | 0.58 | 0.32 | 0.48 |
| | All Combined | **0.67** | 0.28 | **0.55** |
| Upper Bound | | 0.84 | 0.43 | 0.73 |

Measure: Spearman's ρ

Combined: Average over normalized scores

Visual model **captures genuine action similarity**,
but **disregards object similarity**

# 4.2. Experimental Evaluation (Results)

| | Model (Same Object) | *same action* | *diff. action* |
|---|---|---|---|
| TEXT | JACCARD | 0.44 | 0.14 |
| | TEXT VECTORS | 0.42 | 0.05 |
| | TEXT COMBINED | **0.52** | 0.14 |
| VIDEO | VIS. RAW VECTORS | 0.21 | 0.23 |
| | VIS. CLASSIFIER | 0.21 | **0.45** |
| | VIDEO COMBINED | 0.26 | 0.38 |
| MIX | ALL UNSUPERVISED | 0.49 | 0.24 |
| | ALL COMBINED | 0.48 | 0.41 |
| UPPER BOUND | | 0.73 | 0.73 |

Measure: Spearman's ρ

Combined: Average over normalized scores

Visual model **captures genuine action similarity**, but **disregards object similarity**

# 4.2. Experimental Evaluation (Results)

- **Visual models** are better suited for actions with **different activity types**
- **Textual models** perform better for closely related activities involving **different objects**
- Supervision helpful, but not necessary to achieve good results following integration of visual information (i.e. significant correlation with ASim gold standard)

# 4.2. Conclusions

- **TACOS**: Offers coherent sentence-level textual descriptions aligned with high-quality video of cooking domain activities
  - **ASim:** Subset containing sentence-pairs annotated with similarity scores
- Both resources were used successfully in an **action similarity assessment task** for the **evaluation of visually grounded models**
  - All models performed well, despite a lack of optimization and simplistic combination methods
- Creation of further corpora of similar quality (and subsequent application of the presented models) presupposes availability of **videos of high quality** and a **means to align visual information with action descriptions**

# Thanks!

# Questions

- What other methods could be used to align video data with textual descriptions than the one employed in the creation of TACOS? How else could textual annotations be obtained?

- Are there some activity domains which would not yield themselves as easily to the corpus creation method described in the paper as the cooking domain? Why?

- What other methods of combining the predictions of individual models are feasible? What advantages do they offer over the presented method (i.e. averaging over individual scores)?

# References


- M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, M. Pinkal (2013) Grounding Action Descriptions in Videos. Transactions of ACL.
- Marcus Rohrbach, Michaela Regneri, Micha Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In Proceedings of ECCV 2012.
- Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In Proceedings of IJCNLP 2011.
- Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action Recognition by Dense Trajectories. In Proceedings of CVPR 2011.