Johns Hopkins CLSP Summer Workshop

Finding Objects and Actions in Videos with the Help of Accompanying Text

Final Presentation – 07/29/2010

J. Neumann, C. Fermueller, J. Kosecka, E. Tzoukermann, R. Chaudhry, F.Ferraro, H. He, Y. Li, I. Perera, B. Sapp, G. Singh, C.L. Teo, X. Yi, Y. Aloimonos, G. Hager, R. Vidal

The Team

- Senior Members
  - C. Fermueller (UMD), J. Kosecka (GMU), J. Neumann (Comcast), E. Tzoukermann (Comcast)
  - Affiliated members: Y. Aloimonos (UMD), G. Hager (JHU), R. Vidal (JHU)
- Graduate Students
  - R. Chaudhry (JHU), Y. Li (UMD), B. Sapp (UPenn), G. Singh (GMU), X. Yu (UMD), C. L. Teo (UMD)
- Undergraduates
  - F. Ferraro (URochester), I. Perera (UPenn), H. He (Hongkong Polytech Univ)

Human action analysis: Motivation

- Huge amount of video is available and growing (YouTube (24 hrs of new videos/min), cell phones, …)
- Human actions are major events in movies, TV news, personal video – we care about what someone is ***doing***, not just how they ***look***!





Pictures courtesy of Ivan Laptev, Inria

Action recognition useful for:

- Content-based browsing

  e.g. fast-forward to the next goal scoring scene
- Video indexing and search

  e.g. find "Bush shaking hands with Putin"
- Robotics

  e.g. help a robot to recognize an action when observing it

What are human actions?

Most current work:

- **Full body motion**

  • actions defined by large body parts in motion (e.g running, jumping, waving, …)
  • people interacting with each other (kissing, hugging, …) or leaving/entering cars, doors, using a telephone, …

Our focus:

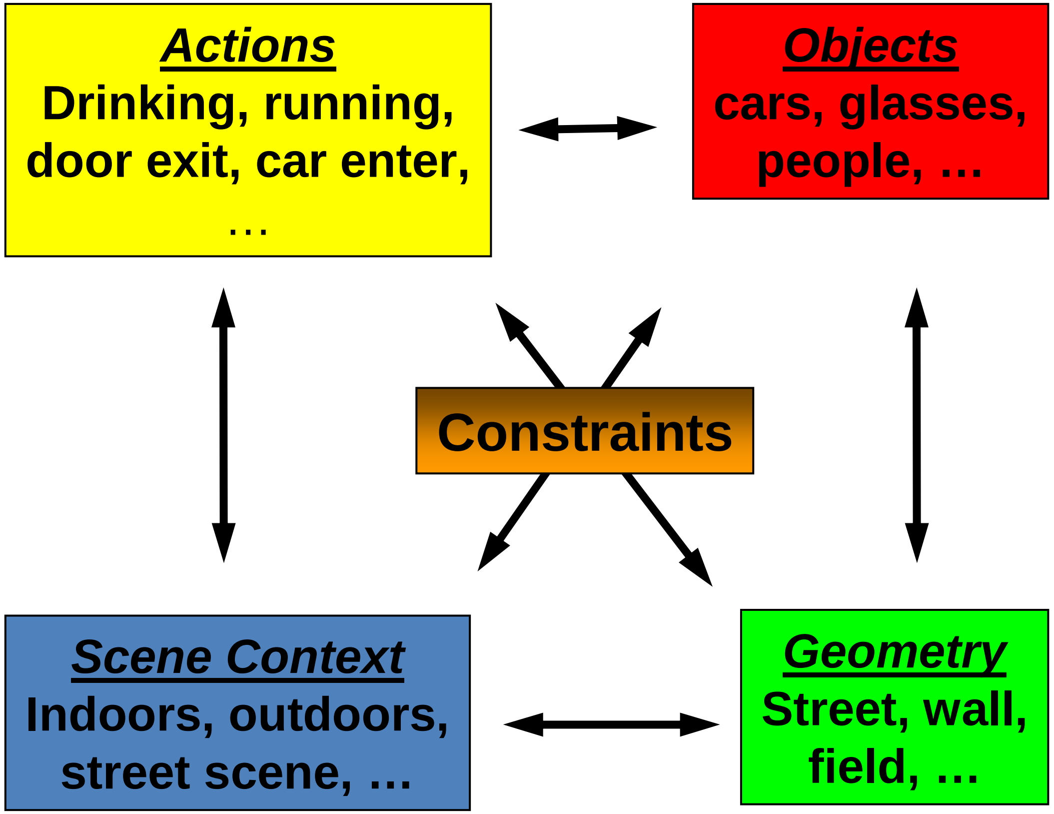
- **Interaction with environment for a specific purpose**

  same physical motion -- different actions depending on the context

Pictures courtesy of Ivan Laptev, Inria

Complexity of Visual Scene Understanding

Pictures courtesy of Ivan Laptev, Inria

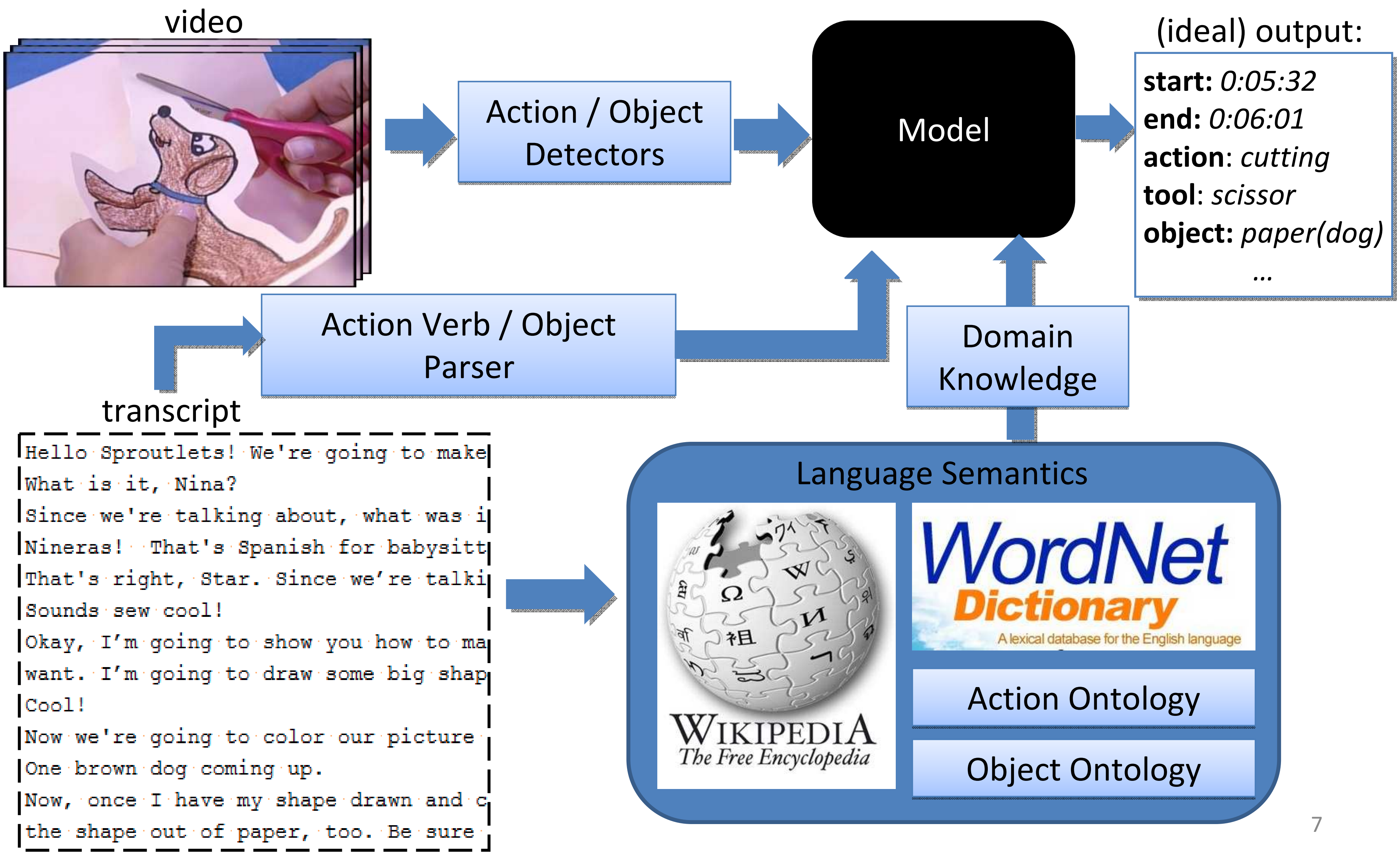


Need to utilize domain knowledge to leverage appropriate subset of constraints!

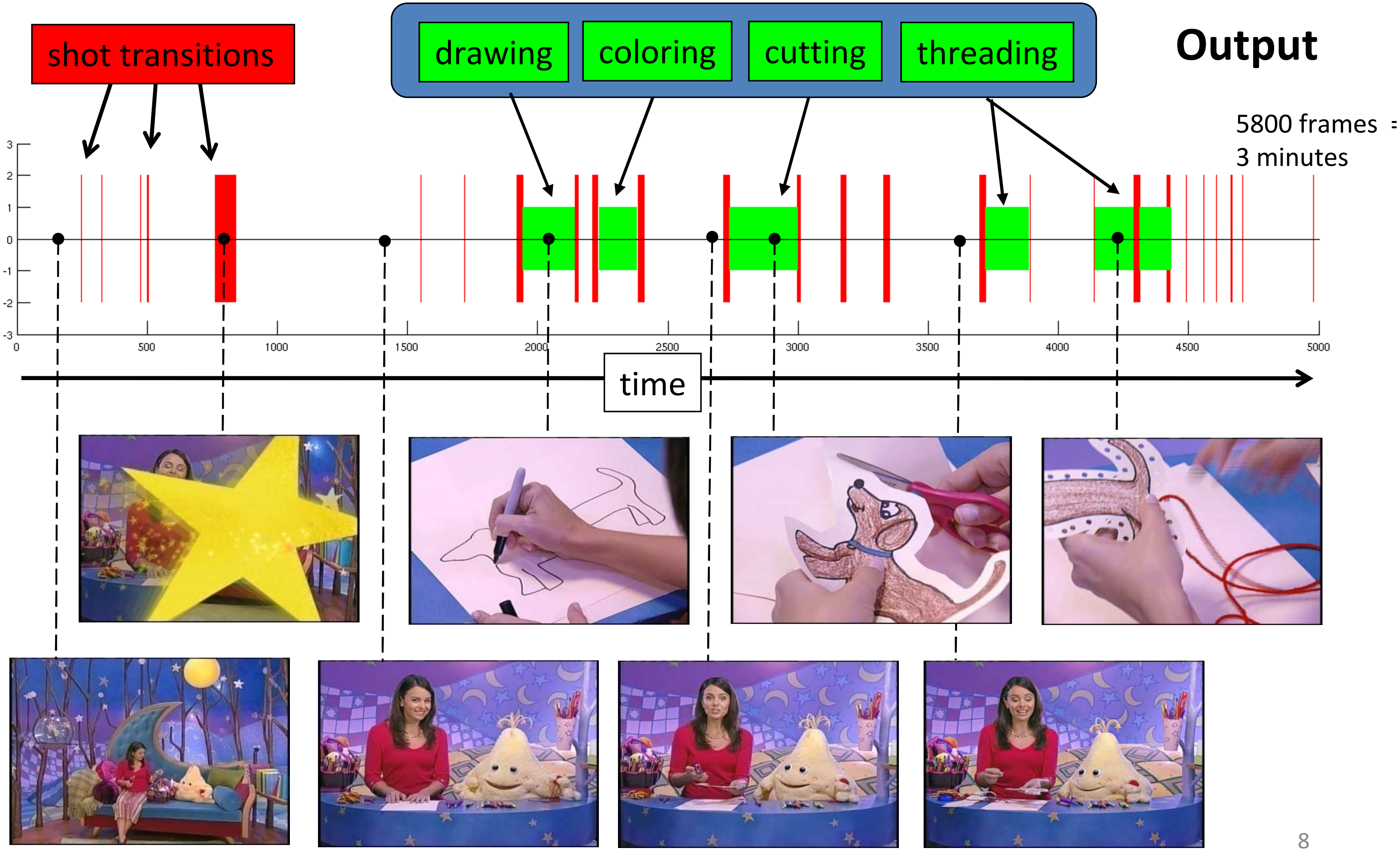
What role can NLP play in Action Recognition?

1. Provide semantic information
   - Parse the phrasal constituents to determine **action type** and human interaction through **objects**, **instruments**, and other **contextual information**
   - Describe **properties** of objects and their **spatial, temporal, and semantic relationships** (e.g. adjectives, adverbs, prepositions)
   - **Relate** entities to "outside world" (e.g. named entity recognition)
2. Provide temporal information
   - In **what order** are the actions happening?
   - **When** is the action being described? (if transcript is time aligned, e.g. closed captions, SR)

Our Approach

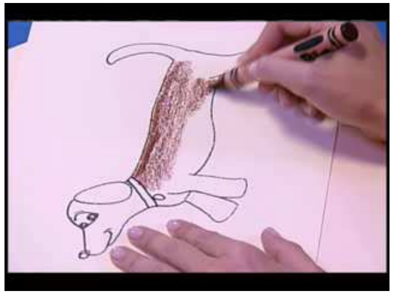


Example Video: *"Babysitter's Animal Sewing Cards"*, PBS Sprout TV

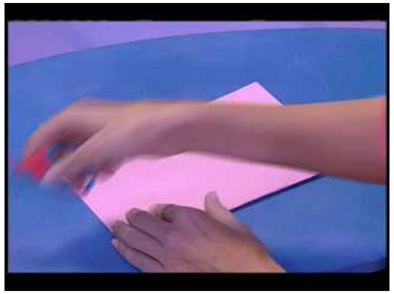


New data set: PBS Sprout Crafts

Bending

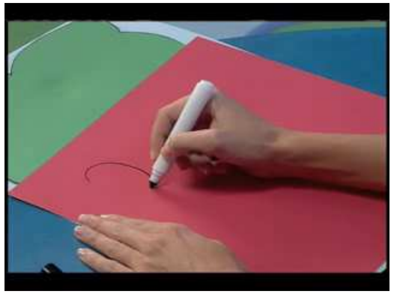


Coloring



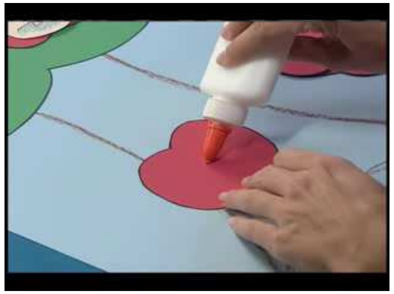
Cutting

Decorating



Drawing

Folding



Gluing

Painting

Placing

Taping

Threading

Properties of New Data Set

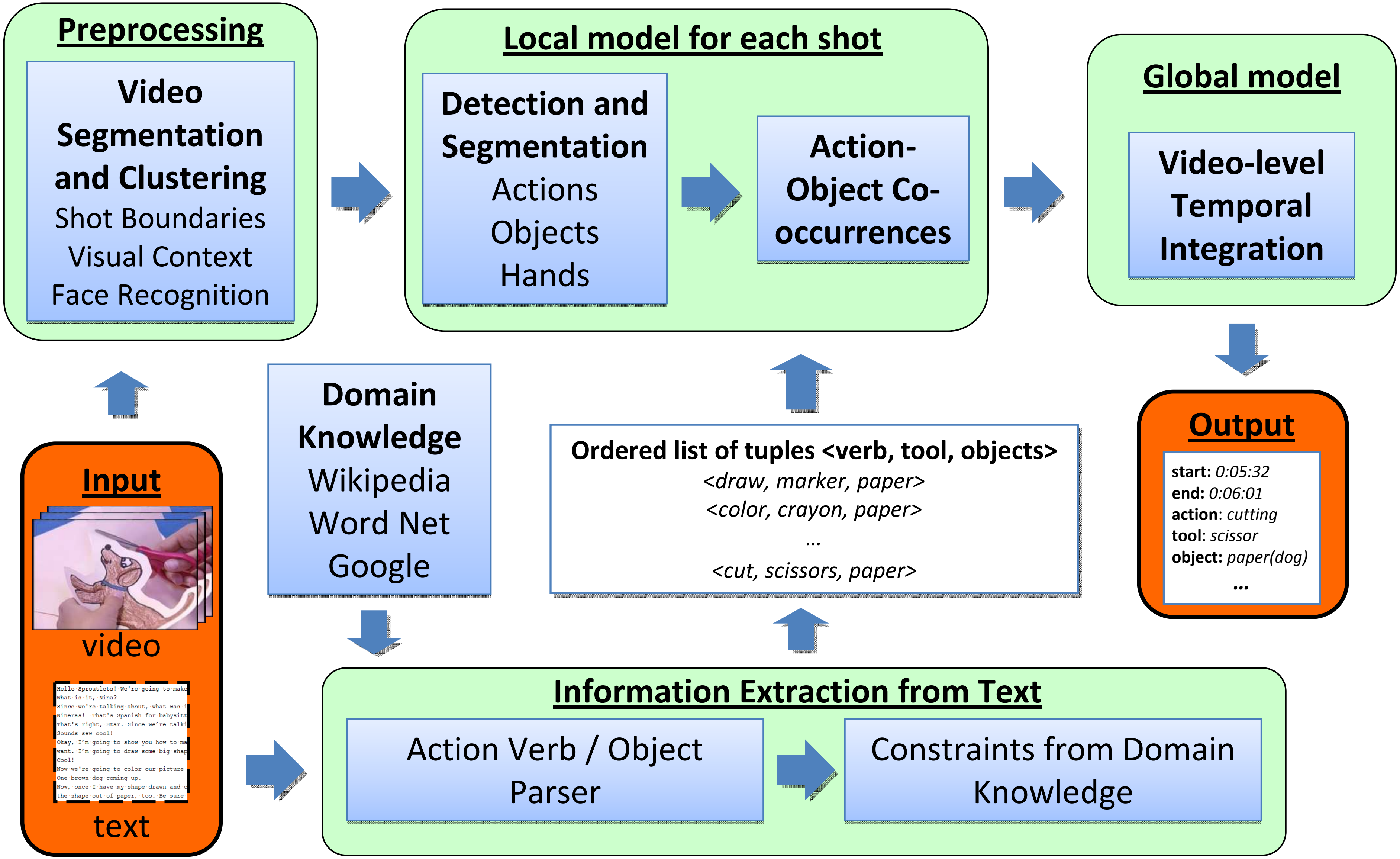
- Source: PBS Sprout
- 27 videos
  - 3 min each (130K frames)
  - 220 shots with actions (1s-25s each, 43K frames total)
  - 11 actions with more than 5 occurrences
  - Transcript (non-aligned) and list of instructions and materials available for each video
- Manual annotations
  - Actions and object presence
  - Shot transitions
  - Camera viewpoint
- Data and annotations will be publicly available to establish a new benchmark dataset

Name	Freq	Name	Freq
Bending	4	Painting	11
Coloring	12	Placing	32
Cracking	1	Pouring	2
Creasing	1	Pressing	1
Crumpling	1	Ripping	1
Cutting	38	Rolling	1
Decorating	5	Separating	1
Detailing	1	Shaping	1
Drawing	42	Spooning	1
Flattening	1	Sprinkling	1
Folding	10	Taping	6
Gluing	20	Threading	6
Hole Punching	5	Tying	1
Writing	1	Unfolding	1
Inserting	1	Wrapping	1

Accomplishments

- Created a **new baseline data set** for research into recognition of complex manipulation actions
  - Benchmark for future research
- Created an **end-to-end system** that annotates real-world broadcast videos with the presence of actions and objects
  - Will be publicly available, reducing barrier of entry for further research
  - Demonstrates how non-visual semantic and temporal information can be integrated to **improve action recognition**
  - Demonstrates how this information can be **automatically extracted from text and unstructured domain knowledge** (Wikipedia, Google)
- Numbers later in the presentation since not meaningful without further context

System Overview



Time Line

- 1:30 pm Overview (Jan Neumann)
- 1:40 pm Vision and NLP (Jana Kosecka)
- 1:55 pm Information Extraction from NLP (Evelyne Tzoukermann)
- 2:05 pm Extracting actions and verbs from text (Frank Ferraro)
- 2:15 pm Extracting domain knowledge from the web (Ian Perera)
- 2:25 pm Action recognition (Rizwan Chaudry)
- 2:45 pm Object recognition (Gautam Singh)
- 3:00 pm Break
- 3:15 pm Joint models for actions, objects and text (Ben Sapp)
- 3:35 pm Temporal modeling (Xiadong Yu)
- 3:45 pm Segmentation and object attributes (Cornelia Fermueller)
- 4:00 pm Closing Remarks (Jan Neumann)
- 4:05 pm Questions & Discussion

Topic Areas: Language, Vision, Language+Vision

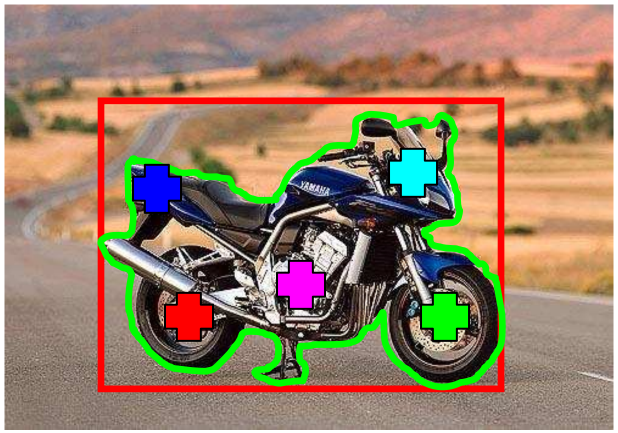
Sources and Types of Semantic Information in Image and Video

General problem:

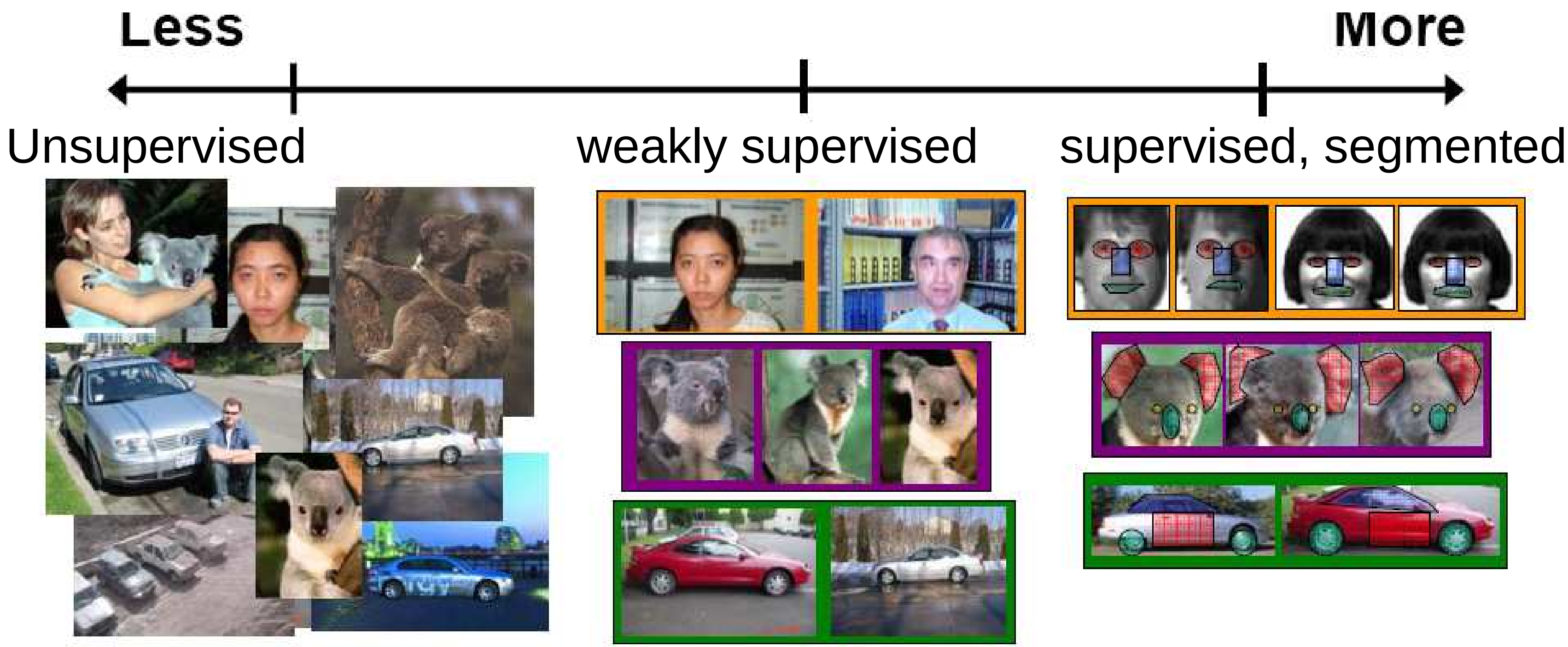
- Given an image/video find the most likely assignment of semantic labels (classes) to data
- Various levels of supervision

  tags, bounding boxes, pixel accurate segmentations

motorbike



Spectrum of Supervision



Choice depends of the task

What ? annotation: car, road

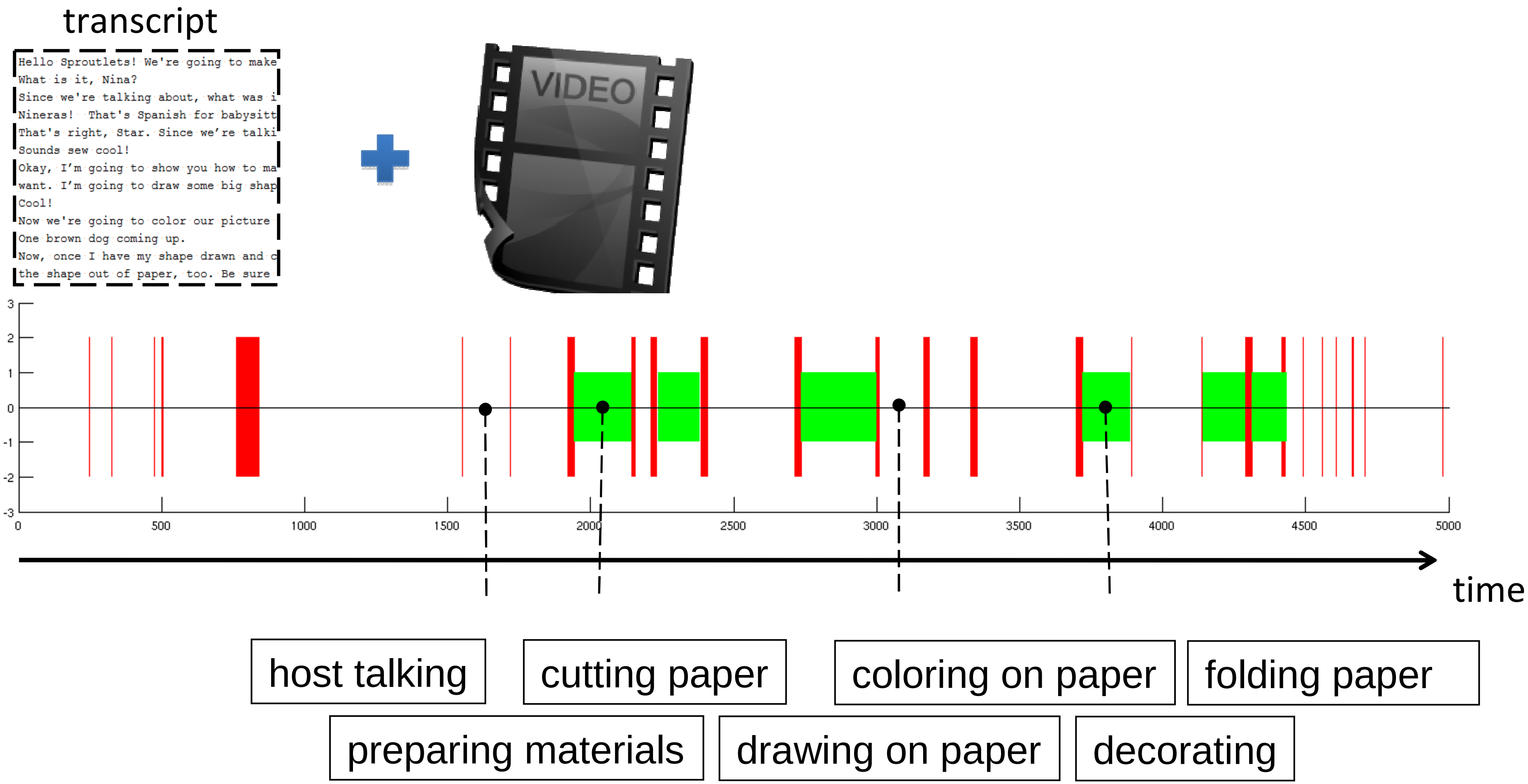
Where ? segmentation: car, road

- Associating semantic labels with images is costly
- Video annotation: image based + label propagation

Courtesy S. Lazebnik

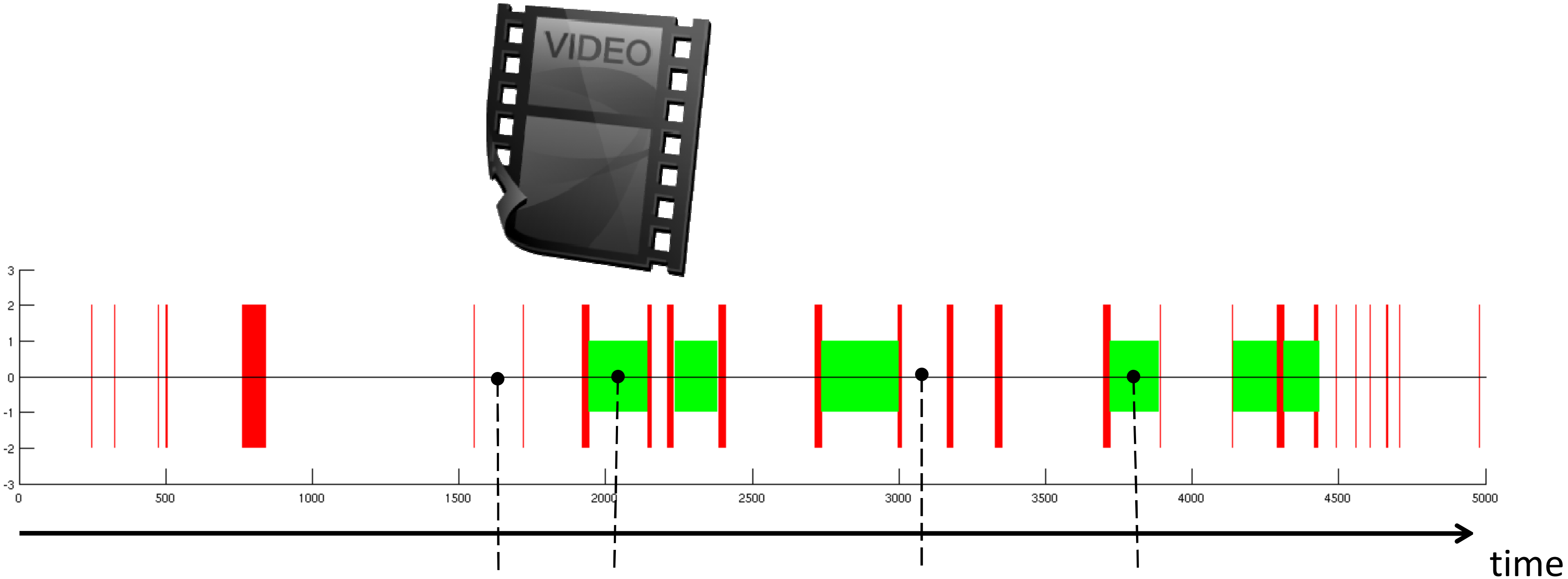
The task

- Automated annotations of videos
- Domain: Arts and Crafts PBS kids shows
- Video and transcript available



The task

- Automated annotations of videos
- Novel video



host talking | cutting paper | coloring on paper | folding paper

preparing materials | drawing on paper | decorating

Language and Image/Video Analysis

- Tags to weakly annotate data
- Given large database of images with tags
- Learn how to associate names with regions

Sky, sunset, beach

Sky, grass, bush, tiger

Sky, buildings, car grass

Solve the optimal assignment problem:
Match sought for concepts/names with visual attributes
Same concepts/tags have share similar patterns in visual representation space (large databases, relatively small number of concepts)

K. Barnard etal. Matching Words and pictures, JMLR, 2003
A.Gupta, L. Davis: Beyond nouns, exploiting prepositions and adjectives for learning vis. Classifiers, ECCV'08

Language and Image Analysis

- Image Captions and faces
- Less structured text, reliable face detectors

Courtesy of T. Berg et al. Names and Faces

- Given news captions
- Named entity recognition

- Exploits reliable face detection
- Formulate the problem as optimal assignment

- Deals with the ambiguities
  there are detected faces not mentioned in the captions
  there are names in the captions which are not detected
- 30,000 images, ~200 names

T. Berg et al. Names and Faces, CVPR'04

Language and Image Analysis

- Screenplays and videos

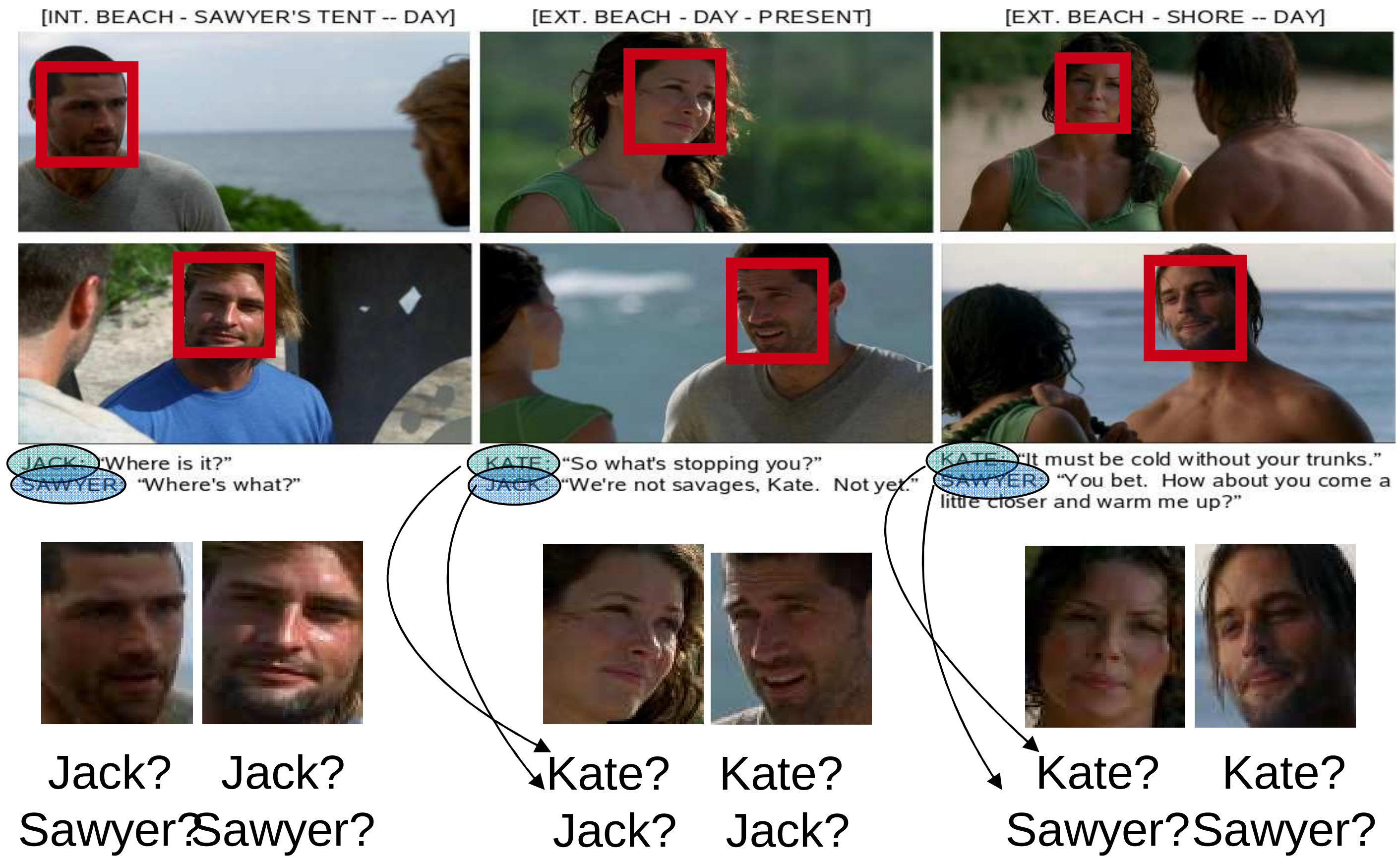


Image courtesy: Talking Pictures: temporal groping and dialog supervised person recognition. T. Cour, B. Sapp, A. Nagle and B. Taskar, CVPR 2010

Language and Image/Video Analysis

- Names and verbs are extracted from captions
- Faces and poses are extracted from images

(a) Four sets ... **Roger Federer** prepares to *hit a backhand* in a quarter-final match with **Andy Roddick** at the US Open.

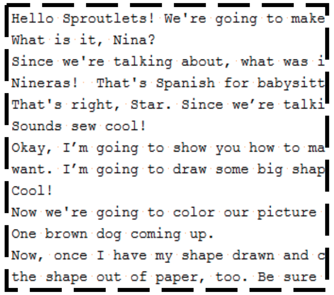
(b) US Democratic presidential candidate Senator **Barack Obama** *waves* to supporters together with his wife **Michelle Obama** *standing* beside him at his North Carolina and Indiana primary election night rally in Raleigh.

- Prior work exploits reliable human pose/face detectors, region detectors

Image courtesy: L. Jie, B. Caputo and V. Ferrari et. Al. Who is doing what ? Joint modeling of Names and Verbs for simultaneous face and pose annotation, NIPS 2009

The ingredients – Our domain

transcript



Hello Sproutlets! We're going to make
What is it, Nina?
Since we're talking about, what was i
Nineras! That's Spanish for babysitt
That's right, Star. Since we're talki
Sounds sew cool!
Okay, I'm going to show you how to ma
want. I'm going to draw some big shap
Cool!
Now we're going to color our picture
One brown dog coming up.
Now, once I have my shape drawn and c
the shape out of paper, too. Be sure

Annotated video

Action Verb, Object Parser

- Language input is less structured
- Correctly identify manipulation actions use additional domain resources

Representations of actions, objects

- Challenges of representations action, object, hand detectors

Global Model

- Learning and Classification approach

The ingredients – Our domain

transcript

```
Hello Sproutlets! We're going to make
What is it, Nina?
Since we're talking about, what was i
Nineras!  That's Spanish for babysitt
That's right, Star. Since we're talki
Sounds sew cool!
Okay, I'm going to show you how to ma
want. I'm going to draw some big shap
Cool!
Now we're going to color our picture
One brown dog coming up.
Now, once I have my shape drawn and c
the shape out of paper, too. Be sure
```

Annotated video

Action Verb,
Object Parser

- Language input is less structured
- Correctly identify manipulation actions use additional domain resources

Representations of
actions, objects

- Challenges of representations action, object, hand detectors

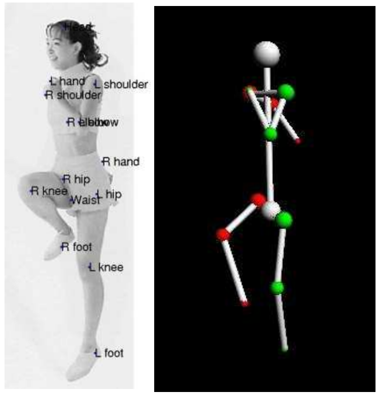
Global Model

- Learning and Classification approach

Recognition of Actions, Activities

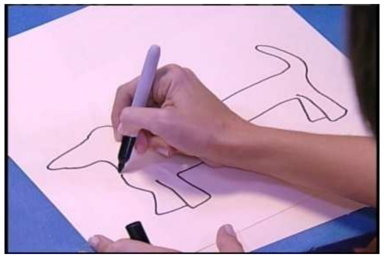
- Movement and posture change

  walk, run, jump, hop, skate, kneel, swim …



- Manipulation actions (object manipulation)

  eat, drink, draw, cut, stir, write, pick, carry, place, bike, play instrument



- Conversational Actions, Sign Language
- Activities involve some (partial) order of individual actions

Challenges of Action Recognition

- Large number of action categories (verbs)
- Large Intra-Category Variation

  viewpoint, illumination, scale, style, person performing the action

- Inter-Category variation (eating vs drinking)

  often the object or context disambiguates the action

- Similar to the object recognition, it is critical to study action recognition

  In context of the activities (Arts and Crafts, Cooking, Ice-skating)

  If applicable in interactions with objects

Object Recognition

- Large number of object categories ~10,000
- Object detectors typically trained in discriminative setting (select region, compute features, train classifiers)
- For large number of categories, the labeled data is sparse
- heavy tail distribution
- Challenges:

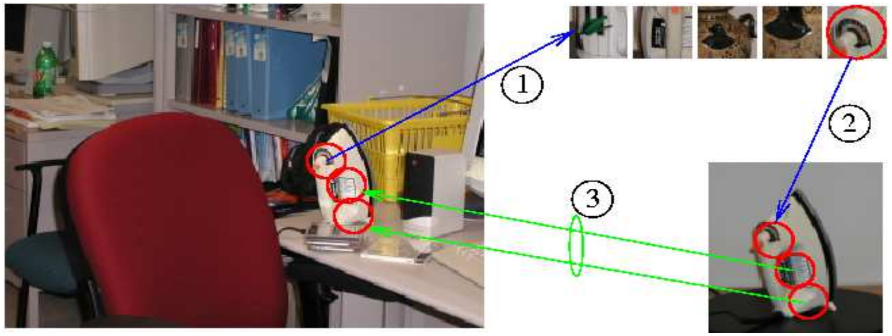
  Large viewpoint and scale changes, Intra-class variation (cups – object affordances), Inter-class variations (apples-pears), Deformable and transformable objects
- Visual only representations are highly ambiguous
- Great opportunity for language to ground the representations, provide context about objects and domain
- Video great opportunity of learning representations from video streams

Object Recognition

- Local features - combining *local* appearance, spatial constraints, invariants, and classification techniques
- Shape based representations, implicit shape models, contours
- Template Based representations, objects as templates

  sliding window approach for detection
- Part based models, object collections of parts and spatial relationships between them

Local features



Shape Based models

Part based models

Sliding window template based

Object Recognition

- Local features - combining *local* appearance, spatial constraints, invariants, and classification techniques
- Shape based representations, implicit shape models, contours
- Template Based representations, objects as templates
  sliding window approach for detection
- Part based models, object collections of parts and spatial relationships between them

- We use existing detectors combining part based models and template based models
- Parts, templates and their spatial relationships are learned automatically in supervised setting

The ingredients – Our domain

transcript

Hello Sproutlets! We're going to make
What is it, Nina?
Since we're talking about, what was i
Nineras! That's Spanish for babysitt
That's right, Star. Since we're talki
Sounds sew cool!
Okay, I'm going to show you how to ma
want. I'm going to draw some big shap
Cool!
Now we're going to color our picture
One brown dog coming up.
Now, once I have my shape drawn and c
the shape out of paper, too. Be sure

Annotated video

Action Verb, Object Parser

Representations of actions, objects

Global Model

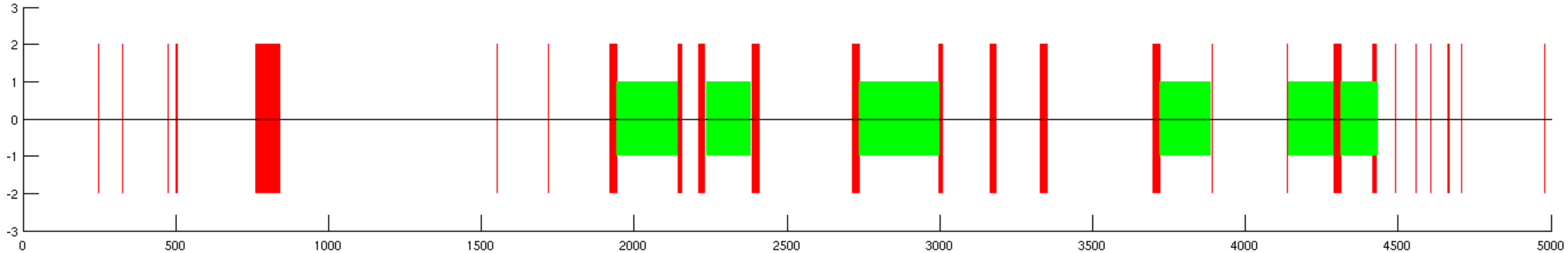
- Language input is less structured
- Correctly identify manipulation actions use additional domain resources

- Challenges of representations action, object, hand detectors

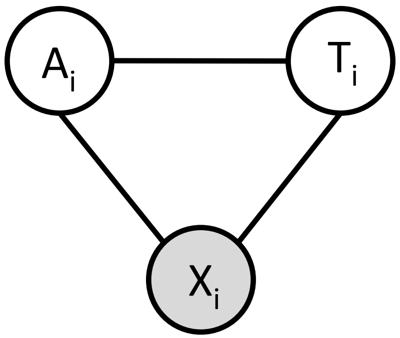
- Learning and Classification approach

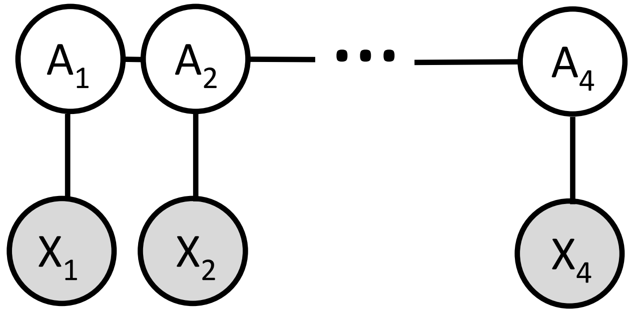
Global model

- Given segmentation of video into shots



$$X_i = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{array}{l} \} \text{ Action} \\ \} \text{ Tool} \\ \} \text{ Hand pose} \end{array}$$





- Discriminative training of action and joint action/object classifiers
- Undirected graphical models CRF to directly exploit structure of action/tool co-ocurrence learned from language, single shot classification
- Temporal model CRF model of the whole video clip and exploit partial order of verbs actions learned from transcript

Labeling Aspects

transcript

```
Hello Sproutlets! We're going to make
What is it, Nina?
Since we're talking about, what was i
Nineras!  That's Spanish for babysitt
That's right, Star. Since we're talki
Sounds sew cool!
Okay, I'm going to show you how to ma
want. I'm going to draw some big shap
Cool!
Now we're going to color our picture
One brown dog coming up.
Now, once I have my shape drawn and c
the shape out of paper, too. Be sure
```

Annotated video

Action Verb,
Object Parser

Representations of
actions, objects

Global Model

- Language input is less structured
- To correctly identify manipulation actions additional domain resources are used

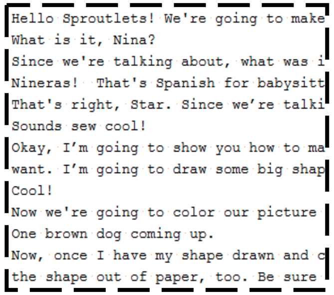
Fully supervised setting
Using hand annotated video

- Challenges of representations
  State of the art action, object, hand detectors

- Train discriminative classifiers for individual features
- Learn single clip structured model CRF explicit interaction between action and tool features
- Temporal models: exploit temporal order of actions determined from transcript

Labeling Aspects

transcript


Hello Sproutlets! We're going to make
What is it, Nina?
Since we're talking about, what was i
Nineras! That's Spanish for babysitt
That's right, Star. Since we're talki
Sounds sew cool!
Okay, I'm going to show you how to ma
want. I'm going to draw some big shap
Cool!
Now we're going to color our picture
One brown dog coming up.
Now, once I have my shape drawn and c
the shape out of paper, too. Be sure

Annotated video

Action Verb, Object Parser

Representations of actions, objects

Global Model

- Language input is less structured
- To correctly identify manipulation actions additional domain resources are used

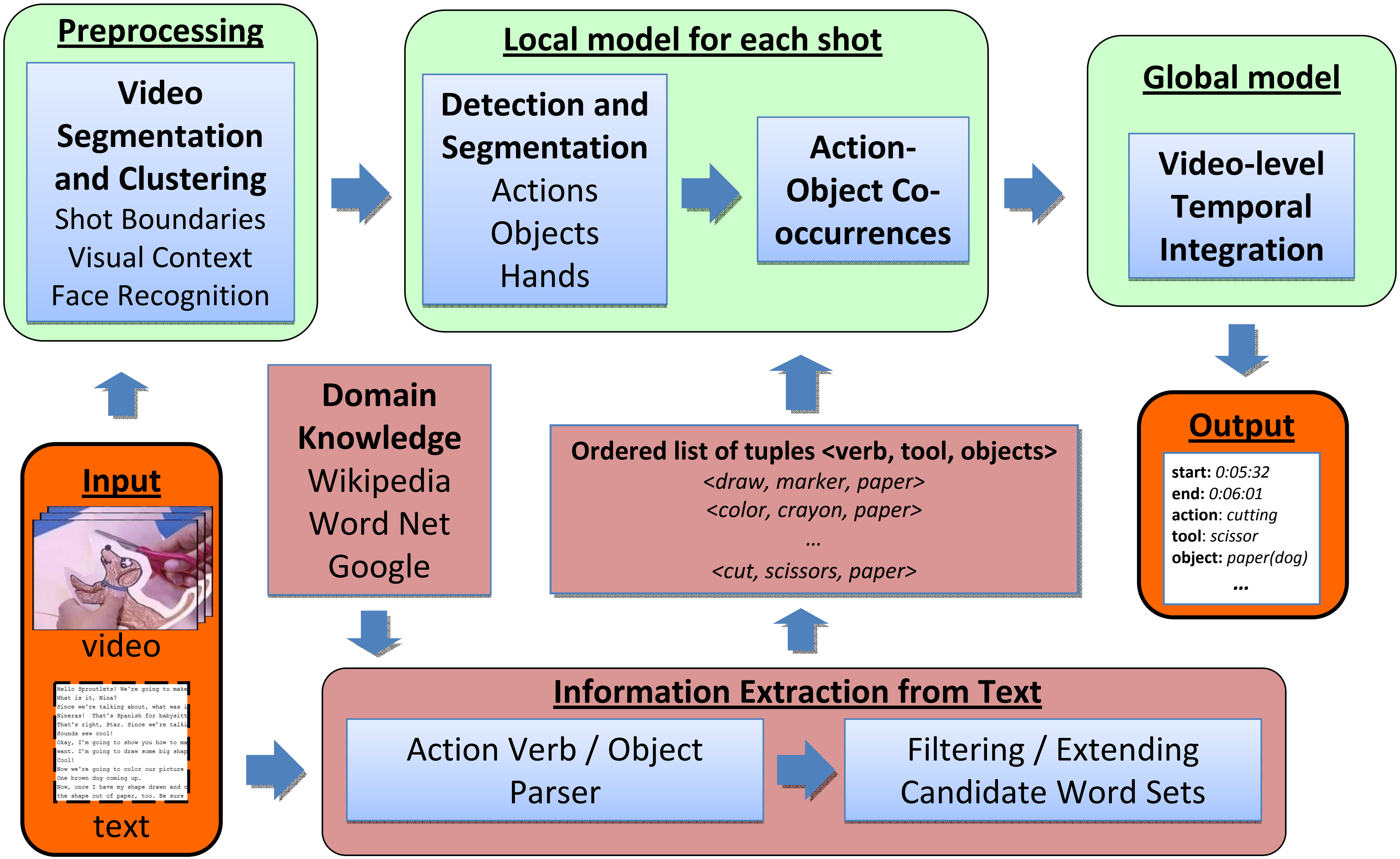
Multiple Instance Learning

automatic assignment of semantic concepts to features/measurements

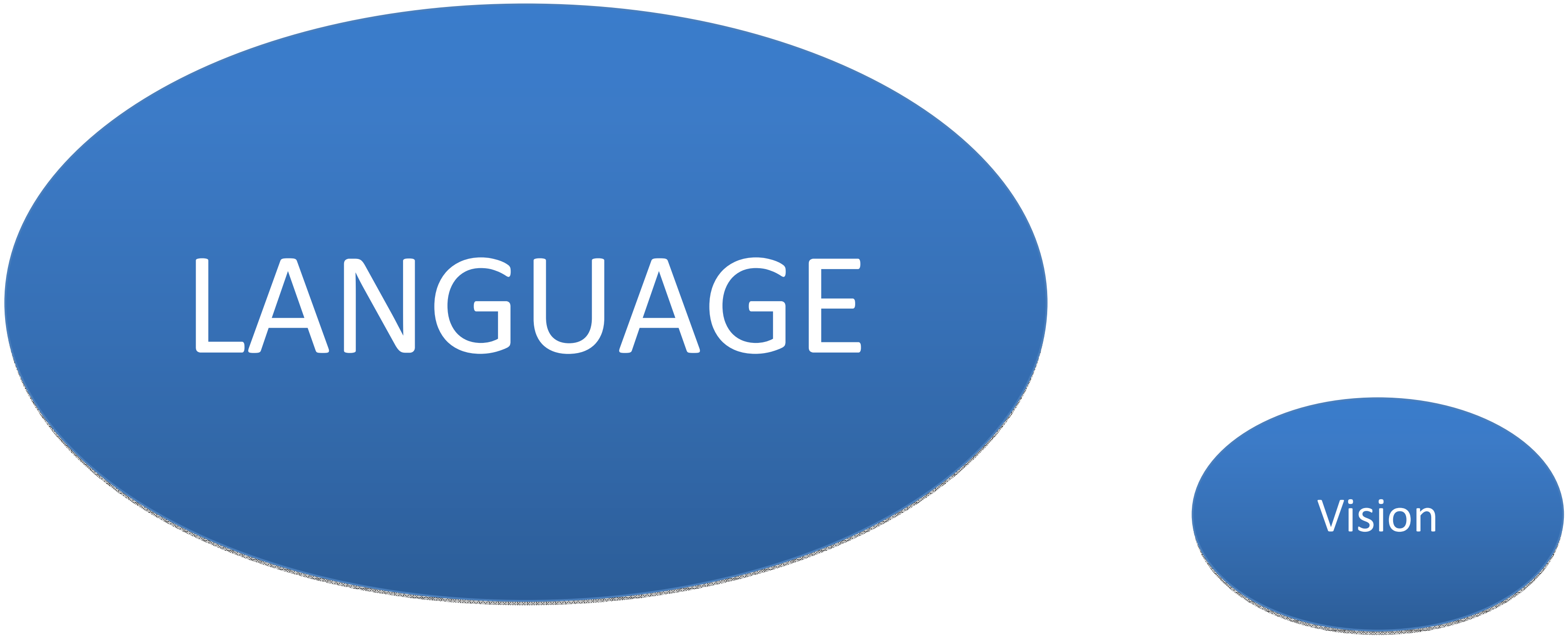
- Challenges of representations State of the art action, object, hand detectors

- Train discriminative classifiers for individual features
- Learn single clip structured model CRF explicit interaction between action and tool features
- Temporal models: exploit temporal order of actions determined from transcript

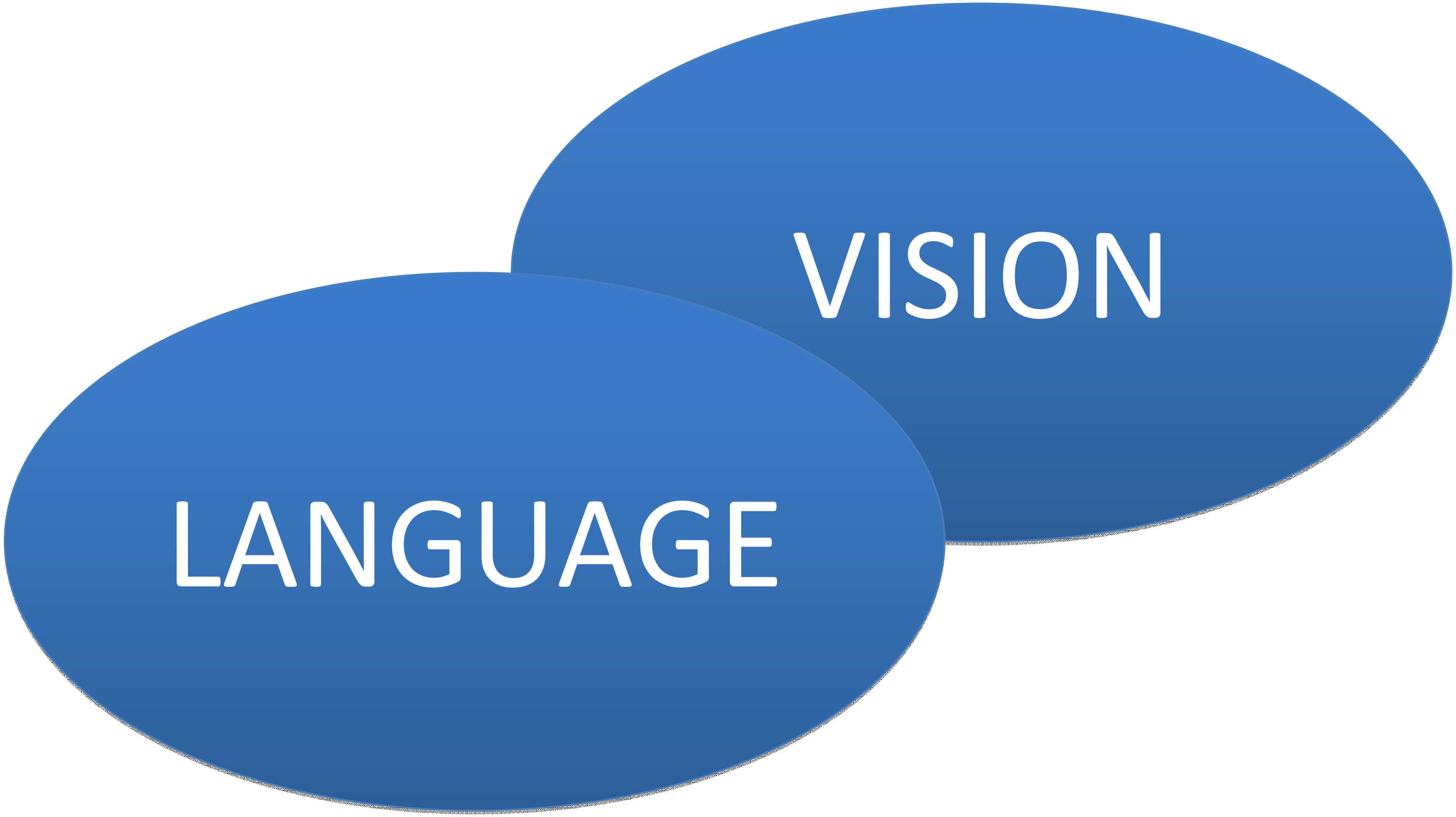
Natural Language for Action Recognition – *NLP is*



Food and Drink = Drink and Food?

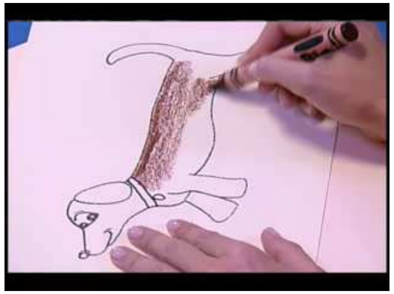

LANGUAGE
Vision

Language and Action Recognition in Video



Language is Key to Video Analysis

- **Verbs**: meaning of actions



- **Objects and Tools**: what is the interaction about?

- **Adverbs**: speed, manner…
- **Adjectives**: texture, color, size…

- **Prepositions**: spatial, temporal relations

What is the contribution of Language in this project? (1/2)

1. Annotations of videos

   – human annotator watches video and marks action verb and dependencies

   – for arts and crafts (11 action types)
   – cooking domain (53 actions on longer videos)

How to make an Eggshell Planter

Verb	**Direct Object**	**Instrument**	**Human Interaction**	**Location**	**Begin Time**	**End Time**	**Duration**
To crack	Egg	Spoon	Both Hands	Workspace	01:11.5	01:21.0	00:09.5
To crack	Egg	Spoon	Both Hands	Workspace	01:21.3	01:23.0	00:01.7
To spoon	Dirt	Spoon	Both Hands	Egg	01:45.4	01:51.6	00:06.2
To sprinkle	Grass Seed	Hands	Both Hands	Egg	01:56.4	02:02.5	00:06.1
To draw	Egg	Pen	Both Hands	Egg	02:09.4	02:10.7	00:01.3
To draw	Egg	Pen	Both Hands	Egg	02:13.7	02:20.6	00:06.9
To draw	Egg	Pen	Both Hands	Egg	02:21.9	02:25.7	00:03.8
To place	Bottle Cap	Hands	Both Hands	Bottlecap	02:30.0	02:32.2	00:02.2

What is the contribution of Language in this project? (2/2)

1. Automatic processing of text transcripts
   a) Perform syntactic analysis
   – Stanford probabilistic parser for dependency relations,
   – Adaptation of Stanford Named Entity Recognizer (CRF)

   b) Determine semantic relatedness of words
   - Verb – object
   - Object – instrument

   → matrices of co-occurrences to feed action recognition

Research Questions

- What is the best way to represent Actions with Language?
- What is the role of Language
  - in capturing entities,
  - in capturing actions over these entities
- How can vision and language be tightly integrated into the overall framework?

Related Work

- **“What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer”** Rohrbach, Stark, György Szarvas, I. Gurevych, B. Schiele (CVPR 2010)
  - knowledge transfer for object class recognition using Wikipedia, WordNet, Yahoo, Flickr
- **“Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions”** Atsuhiro Kojima , Takeshi Tamura and Kunio Fukunaga (2002)
  - generates textual descriptions from position and body orientation
  - recognizes position and orientation of human head, position of hands and interaction with objects from video images.

Historical Basis for Actions

- Case Frame Theory (Fillmore -1968)
- Hierarchy of actions
- Each action has a series of cases
  - the verb "give" requires an Agent (A) and Object (O), and a Beneficiary (B)
  - "Jones (A) gave money (O) to the school (B).

- Fillmore remains the authoritative reference for case analysis of meaning
- Framenet: lexical database describing objects, states, and events

Proposed semantic primitives about movements and states

- '6

Enhanced Information Extraction Approach

1. Standard Information Extraction:
   – extract structured information from unstructured machine-readable documents
   – Usually template driven (find who, what , where)
   – Narrow set of categories (named entities, locations)
2. *Enhanced* Information Extraction
   – Extends basic approach to incorporate syntax and semantics
   – Capture Verb – Object relations
   – More than just Entities: Verb, Object, Instrument, Prep, Adverb, Target Location, Human Interaction

Let’s make something new, (SONG)

Nina: Welcome back, Sproutlets! Since tonight we're talking about, what was it, star?

Star: Donede vivimos, that's where we live in Spanish.

Nina: Great remembering, Star! Let's make something that you can grow no matter where you live! It's an eggshell planter!

Star: A planter? I love to plant things! Let's get started, Nina!

Nina: Sproutlets I’ll show you how to make an eggshell planter and maybe tomorrow you can make one of your own! First, I’m going to take an egg, and use a spoon to carefully crack it open. Usually, you crack an egg right in the middle, but I’m going to crack this egg near the top, because I want save the larger piece at the bottom for our planter. You'll want a grownup sprout to help you with this, because it might be a little tricky. You just tap the egg all the way around the top of the shell, and once you've finished, you can just pull the top right off and then you'll want to rinse the egg shell in some water, just like this.

Star: So that it won't be all egg inside, right?

Nina: That's right, Star. And now I’m going to carefully fill the eggshell with some soil, you can just use a spoon. Next, I’m going to sprinkle some grass seed on the soil. Just like this.

Star: Nina, your planter looks like a face to me.

Nina: It does, doesn't it, Star? And that's the next step. I’m going to use some markers to draw a face on this egg!

Star: You have to be very careful with that eggshell, though.

Nina: Once the grass starts growing, our eggshell friend will have lots of pretty green hair, and I’m going to put a nice red smiley face, and now I’m going to put the planter down on a bottle cap, so we can display it nicely, and wait for the grass to grow. Tada!   This is your egg shell planter! I made this one a few weeks ago so you could see how it looks, isn't it cute?

Star: It is.

Nina: I'm so glad you like it, Star! Sproutlets, if you'd like to make this craft tomorrow you and a grown up can visit us online to find out how to make your very own egg shell planter!

Star: I can't wait to watch his green hair grow. I really like it.

First, I'm going to take an egg, and use a spoon to carefully crack it open. Usually, you crack an egg right in the middle, but I'm going to crack this egg near the top, because I want save the larger piece at the bottom for our planter.
You just tap the egg all the way around the top of the shell, and once you've finished, you can just pull the top right off and then you'll want to rinse the egg shell in some water, just like this.
And now I'm going to carefully fill the eggshell with some soil, you can just use a spoon. Next, I'm going to sprinkle some grass seed on the soil. And that's the next step. I'm going to use some markers to draw a face on this egg!
Once the grass starts growing, our eggshell friend will have lots of pretty green hair, and I'm going to put a nice red smiley face, and now I'm going to put the planter down on a bottle cap, so we can display it nicely, and wait for the grass to grow. Tada!

→40% of words in action sentences describe an action
→ Syntactic analysis to capture VERB-OBJECT-INSTR

- ✓ Natural Language grounds video processing in providing
  - – Semantics of Actions
  - – Temporal Information
  - – Measure of word co-occurrences
- ✓ Proof of Concept with an end-to-end system
- ✓ We want to learn actions on a larger set of videos
  - ➢ and build detectors corresponding to actions
  - – Once vision is equipped with enough data and good discriminative models, we can address the following challenges:

Language and Vision

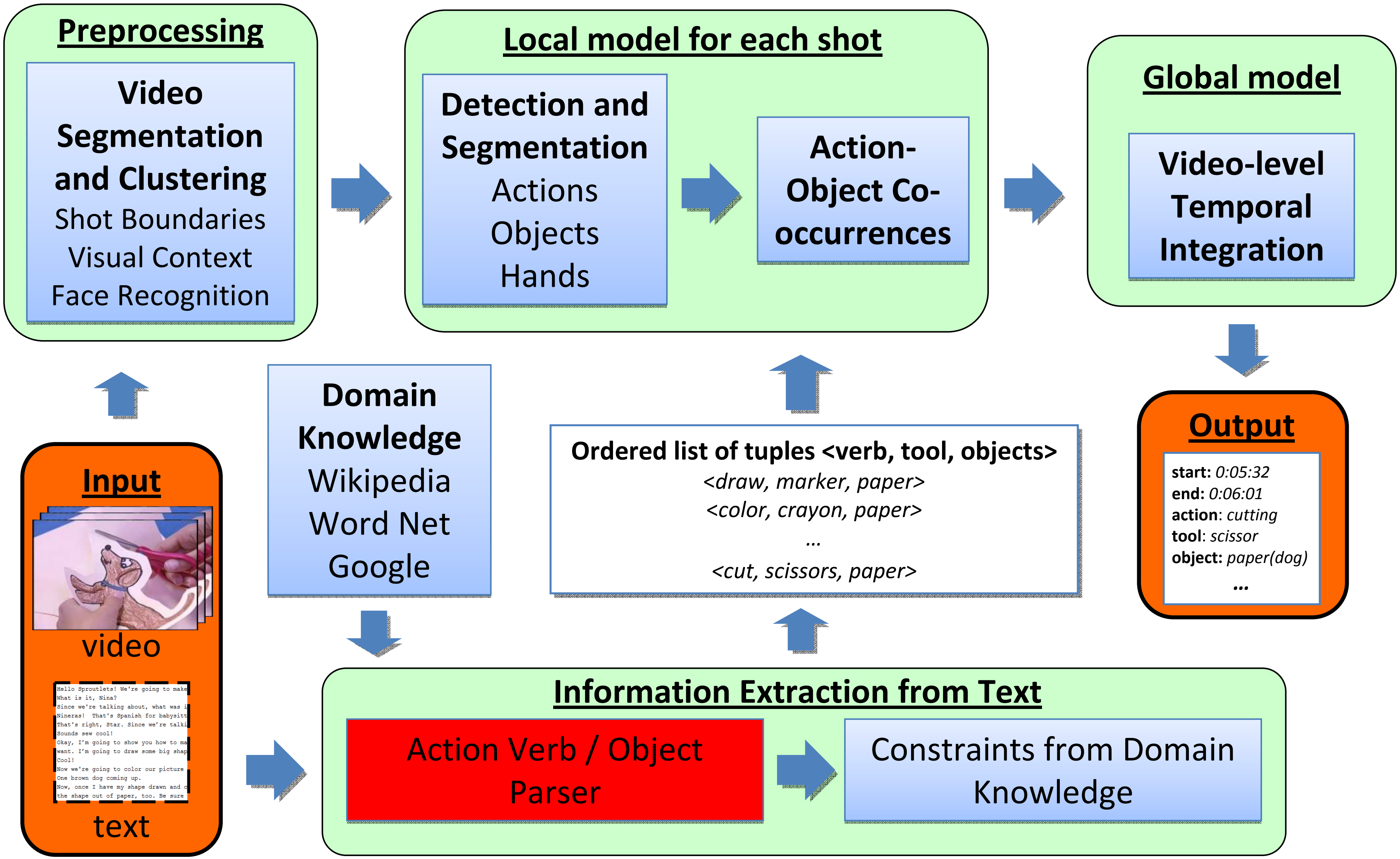
- CLSP workshops have focused on Speech and Language challenges
- Vision research is new in this community
- Combined data analysis promises deeper levels of processing

- Contribution: models where vision and language are intertwined

Vision, Language, and Challenges

	Types of Action: "draw" "cut" "glue"	Action Time beg_time end_time	Levels of complexity
	known	**known**	- Learn Action types with time information
	known	**unknown**	- Learn Action types without time - localization
	unknown	**unknown**	-Identify action -Localize objects

Information Extraction from Text



How Language Helps

- Transcript contains a lot of useful information
  - Provide seed information for targeting certain actions, objects and tools
- Even without time-aligned video, we can get relative, sequential information
  - This information is given to the global temporal model

Some Previous Work

- High-performing systems tend:
  - To have a lot of training data
    - DIRT (Lin, Pantel, SIGKDD01): 1GB of AP data
  - To use a “semantically dense” dataset (e.g. USP)
    - USP (Poon, Domingos, NAACL10): λ-reduction semantics with Markov Logic Network
    - Academic prose, PubMed abstracts, etc.
- We have neither with Sprouts transcripts

Sprouts Data

- Source: PBS Sprouts Craft TV
- Size: 27 shows with transcripts
- Gold standard: manual annotations *based on the video* (not necessarily the text)
- Problems
  - Very low semantic density; most clauses are irrelevant to project
  - No one-to-one correspondence between text and gold standard annotations

Sample Action Annotations

Nina: Now we're going to color our picture in.

Star: One brown dog coming up.

Nina: Now, once I have my shape drawn and colored in, I'm going to cut him out with safety scissors. Always have a grown up sprout with you when you're cutting. But you can tear the shape out of paper, too. Be sure to leave lots of room around the edges so you have room to sew later on.

Transcript

Number:	2
Action Verb:	Coloring
Objects:	Paper, Crayon
Description:	Hands color in drawing
Camera Angle:	Full, Tight
Start Time:	01:15.0
End Time:	01:20.0
Duration:	00:05.0
Number:	3
Action Verb:	Cutting
Objects:	Paper, Scissors
Description:	Hands cut out drawing
Camera Angle:	Full, Tight
Start Time:	01:32.0
End Time:	01:40.0
Duration:	00:08.0

Manual gold-standard annotations

Parser

- Stanford probabilistic parser (Klein and Manning, ACL 2003)
  - POS tags ...   Color/VB our/PRP picture/NN
  - Dependencies ...   dobj(color-8, picture-10)
  - Parse tree ...   VP[color-18] (color-18/VB
    NP[picture-22] (…

Approach 1: Bag-of-Words

- For every sentence in the transcript:
  - Match certain key phrases
  - Use a list of domain-specific action words
  - Use POS tags to certify verbs
  - Use dependencies to find direct objects (and sometimes tools)

	Against visual annotations	Against text transcript
Recall	85%	85%
Precision	88%	89%

Limitations of Approach 1

- Parser fails to tag imperatives correctly

  “Once you've done that, $\underbrace{\text{tape}}_{\text{Noun}}$ or $\underbrace{\text{glue}}_{\text{Noun}}$ the two ends together.”

- Inherent difficulty

  e.g. “We're going to do this now” to describe cutting paper

- How do we get our seed action words?

Crafts from the Web

- Hundreds of craft instructions mined from four websites
  - Initially had 121 crafts, recently received another 299 for total of 420 crafts
  - Ages 3-13
- Imperative and narrative form
  - Semantic density ranges from very low to high
  - Rich vocabulary

Adapting a Named Entity Recognizer

- Stanford CRF NER (Finkel et al., ACL 2005)
- Given input words and a set of labels L, give each word the most appropriate label from L

L={verb, object, tool, mod, prep, adv, other} in action domain (not necessarily grammatical)

"Have your kids cut the shapes with scissors and then paste them."

Have your kids	cut	the	shapes	with	scissors	and then	paste	them.
	verb		object	prep	tool		verb	object

Diagnostic Results from CRF

- 70/30 training/test split on 121 crafts; tested on Web Crafts
- Correct 90% of the time in identifying semantic relevance

Class	Accuracy	Recall	Precision	F1
Other	89.80	92.64	91.79	92.21
Relevant (average of 6)	97.91	75.13	85.03	78.31

- 70/30 training/test split on 121 crafts; tested on Sprouts transcripts
- Correct 95% of the time in identifying semantic relevance

Class	Accuracy	Recall	Precision	F1
Other	94.81	96.70	97.52	97.11
Relevant (average of 6)	99.00	67.81	73.68	69.80

Approach 2: Add CRF to Bag-of-Words

- In addition to bag-of-words approach, use CRF output:
  - To compensate for parser errors
  - To verify whether a word is an action verb
  - To find new action verbs that are not in bag-of-words
- Use CRF data on verb frequencies, CRF output and parser output to calculate certainty (ϵ [0,1]) of a given verb actually being a correct action
  - Nearly all false detections have very low certainty; many correct detections have high (> 0.5) certainty

	Against visual annotations	Against text transcript
Recall	92%	99%
Precision	65%	69%

Results

Nina: Now we're going to color our picture in.
Star: One brown dog coming up.
Nina: Now, once I have my shape drawn and colored in, I'm going to cut him out with safety scissors. Always have a grown up sprout with you when you're cutting. But you can tear the shape out of paper, too. Be sure to leave lots of room around the edges so you have room to sew latear on.

Number:	7
Verb:	Color
Object:	Picture
Certainty:	0.696

Number:	8
Verb:	Shape
Certainty:	0.237

Number:	9
Verb:	Cut
Object:	Him
Tool:	Scissors
Certainty:	0.966

Transcript

Our output

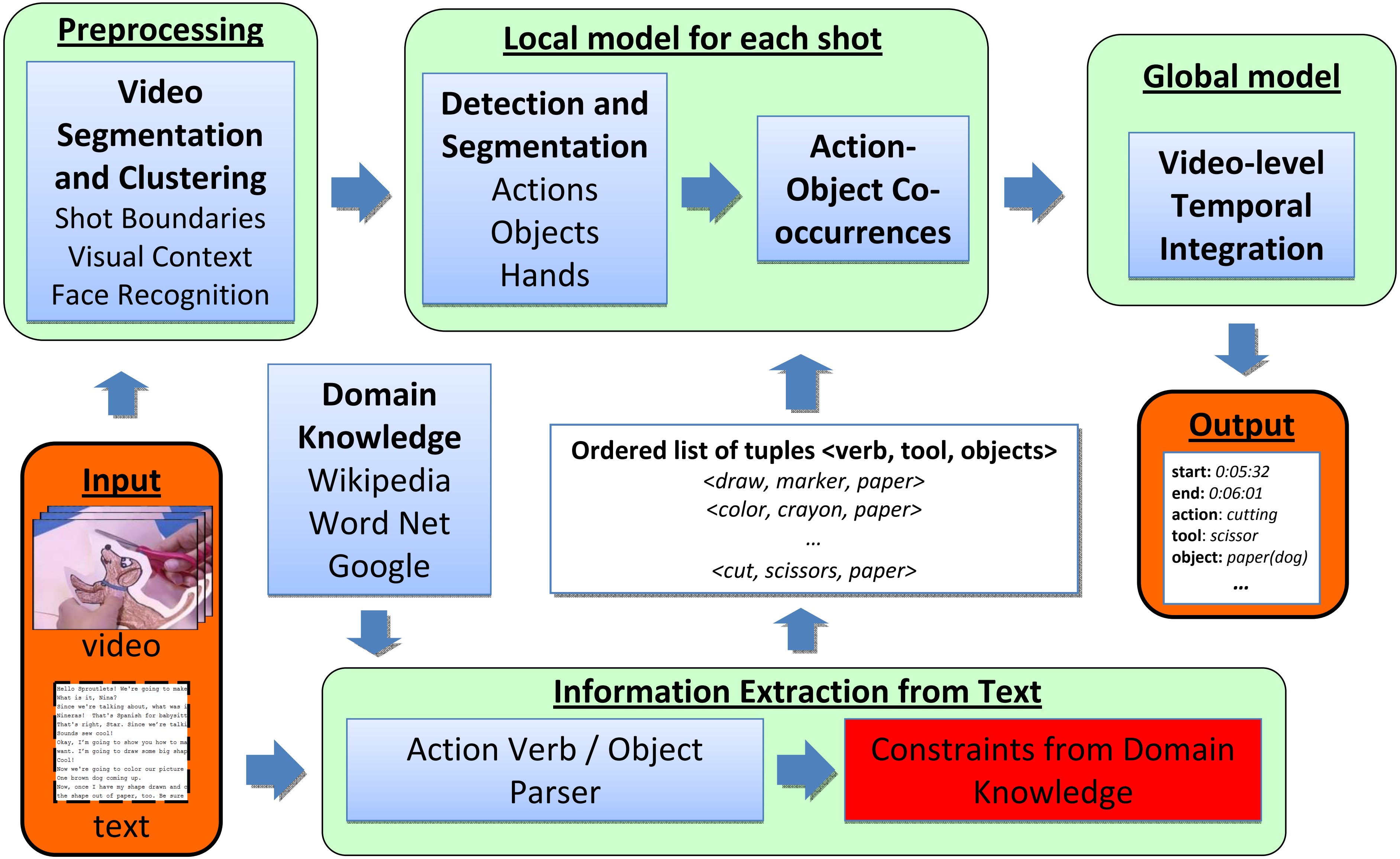
Summary:
Comparing Approaches 1 and 2

- Approach 1 (bag-of-words):
  - Against visual annotation
    - 85% recall, 88% precision
  - Against transcript
    - 85% recall, 89% precision
- Approach 2 (bag-of-words + adapted CRF NER):
  - Against visual annotation
    - 92% recall, 65% precision
  - Against transcript
    - 99% recall, 69% precision
- CRF helps extract relevant actions

Summary: Additional Benefits of the CRF

- Addresses bag-of-words generation problem
  - Up to verb stemming, CRF data has all of the relevant action verbs in the bag-of-words approach
- Scalable: can crawl web to obtain more domain-specific action words
- Provides data for more analysis
  - Frequencies, heuristics, action n-grams

Using Domain Knowledge to Aid in Tool-Action Recognition



Co-occurrence Problem

Problem: Model co-occurrences of actions and tools in video to predict action-tool pairs

- **But**: small training set
  - We can't foresee all possible matches
  - We would also like to avoid relying on labeled training data
- How can we find general knowledge to give us these co-occurrences without seeing them in training first?

Domain Knowledge

Solution: Use domain-specific knowledge to predict action-tool co-occurrences

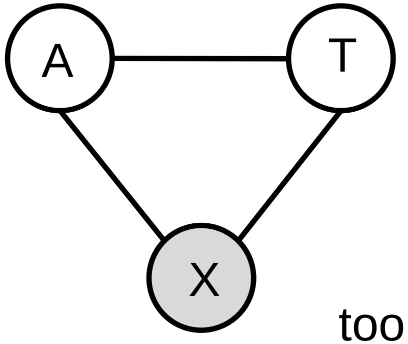
- Find action-tool relationships that are "common sense" to people
  - You cut with scissors
  - You paint with a brush
- Assumption: These action-tool pairs are likely to show up in the video at the same time

Domain Knowledge Implementation

- Create co-occurrence matrices to indicate that certain objects or tools are likely to appear with certain actions
- Three sources:
  - **Wikipedia**
    - With some help from Wordnet
  - **ConceptNet**[1]
  - WWW (**Google Similarity Distance**[2])
    - Could also use Pointwise Mutual Information, similar results

[1]Havasi, C., Speer, R. & Alonso, J. (2007) "ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge." Proceedings of Recent Advances in Natural Languges Processing 2007.
[2]Rudi L. Cilibrasi, Paul M.B. Vitanyi, "The Google Similarity Distance," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, pp. 370-383, Mar. 2007, doi:10.1109/TKDE.2007.48

Modeling Action-Tool Interaction



action A ϵ { Cut, Draw, Color, Glue, Paint }
tool T ϵ { Marker, Scissors, …}
data X = image *and* text features

tool / score

	scissors	pencil	crayon	brus[illegible]	[illegible]
	10	20	50	10	1[illegible]

action / score

action	score					
cut	10	80	0	0	20	0
draw	20	0	80	10	5	5
color	0	0	20	50	30	0
[illegible]	[illegible]5	0	10	1[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	0	0	[illegible]	[illegible]	[illegible]

tool score **from vision**

action score **from vision**

action-tool co-occurences **from text**

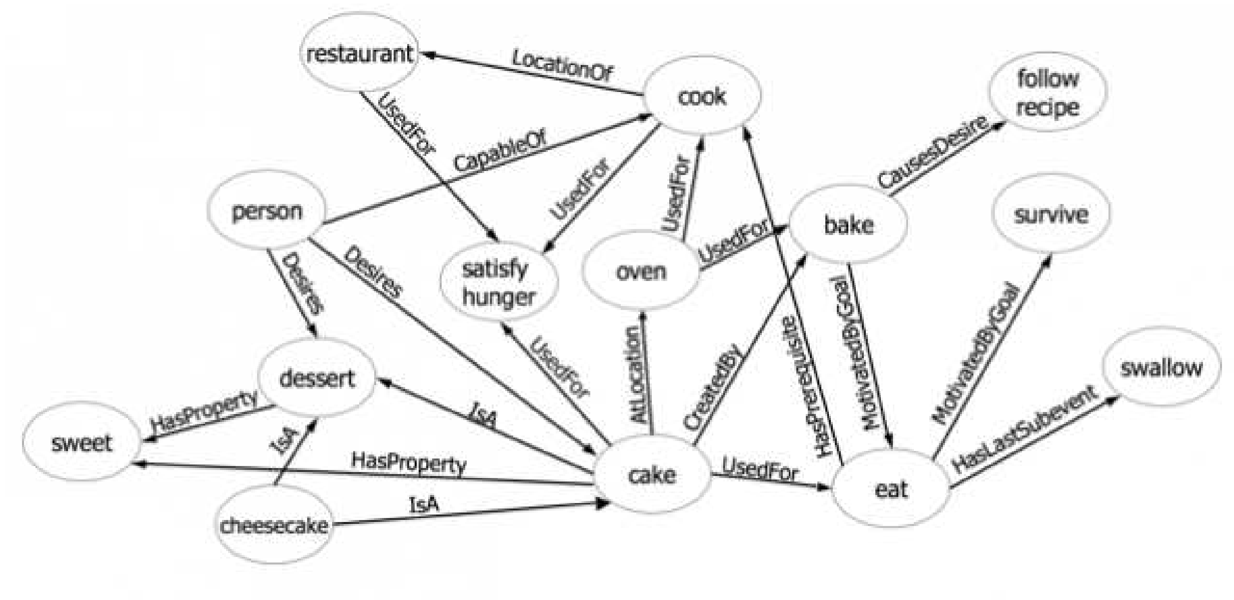
$$a^*, t^* = \mathbf{argmax}_{a,t}\ \text{score}(a) + \text{score}(t) + \text{score}(a,t)$$

Binary Matrices - 1

- An (object x action) matrix with a '1' if the object and action are related or '0' if not
- **Wikipedia**
  - Find the Wikipedia page associated with the desired action
  - Retrieve nouns that fit into Wordnet's 'tool' or 'implement' category
  - High recall, moderate precision (high with tool list)

Binary Matrices - 2

- **ConceptNet**
  - Open user-edited common sense semantic network
  - Query for “usedFor” relationship
  - Very low recall, high precision
  - Not used in final project, low coverage



Wikipedia Matrix

	coloring	cutting	drawing	gluing	painting	placing
brush	0	0	1	0	1	0
writing implement	1	0	1	0	0	0
glue	0	0	0	1	0	0
scissors	0	1	0	0	0	0

- "Writing implement" = logical OR of the results of pen, pencil, crayon, and marker

Semantic Distance Matrix

- Normalized Google Distance measures the semantic distance between two terms using Information Content defined by search results from Google (or Yahoo in our case)

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{(\log N - \min\{\log f(x), \log f(y)\})}$$

$$f(x) = \#\ of\ results\ returned\ for\ search\ term\ x$$

$$f(x,y) = \#\ of\ results\ returned\ for\ search\ terms\ x\ AND\ y$$

$$N = \#\ of\ pages\ indexed\ by\ search\ engine$$

- Undefined if any f(x) is 0 (and we ignore low numbers)
- Example Results for “brush”: (lower number = more related)

  color: 1.92, **cut**:2.72, **draw**:2.74, **glue**:1.61, **paint**:1.11

Modifications to NGD - 1

- **Adding domain to search query**
  - *paint brush "arts and crafts"*
  - Small push towards domain-specific relations
  - Restricts possible word senses
    - Partially addresses shortcoming of NGD being sense-unaware
- **Adding –ing to verbs**
  - Disambiguates between verb and noun forms
  - Removed glue-scissor confusion

Modifications - 2

- **Word proximity (pattern matching)**
  - Related words often appear near each other in a document
  - Use * to allow for any one word in a phrase
  - <u>Example</u>: *“painting brush” OR “painting * brush” OR “brush painting” OR “brush * painting”*
    - Matches “painting brush” and “brush for painting”
    - Can have up to 5 *’s in a row

Normalized Google Distance Matrix

	coloring	cutting	drawing	gluing	painting	placing
brush	2.51	2.11	2.4	INF	1.85	INF
writing implement	2.12	3.51	1.72	INF	2.08	INF
glue	2.51	2.51	2.51	1.2	2.44	INF
scissors	2.47	1.76	2.36	INF	2.68	INF

- "Writing implement" = average distance of pen, pencil, marker and crayon
- Co-occurrence was defined as within two words of each other
- INF values were smoothed to 2x max for input to model

Training Co-occurrences

	coloring	cutting	drawing	gluing	painting	placing
brush	0	0	0	1	8	0
writing implement	12	0	42	0	0	0
glue	0	0	0	20	0	0
scissors	0	38	0	0	0	0

- "Writing implement" = logical OR of the results of pen, pencil, crayon, and marker

Wikipedia Matrix

	coloring	cutting	drawing	gluing	painting	placing
brush	0	0	1	0	1	0
writing implement	1	0	1	0	0	0
glue	0	0	0	1	0	0
scissors	0	1	0	0	0	0

- “Writing implement” = logical OR of the results of pen, pencil, crayon, and marker

Modifications - 3

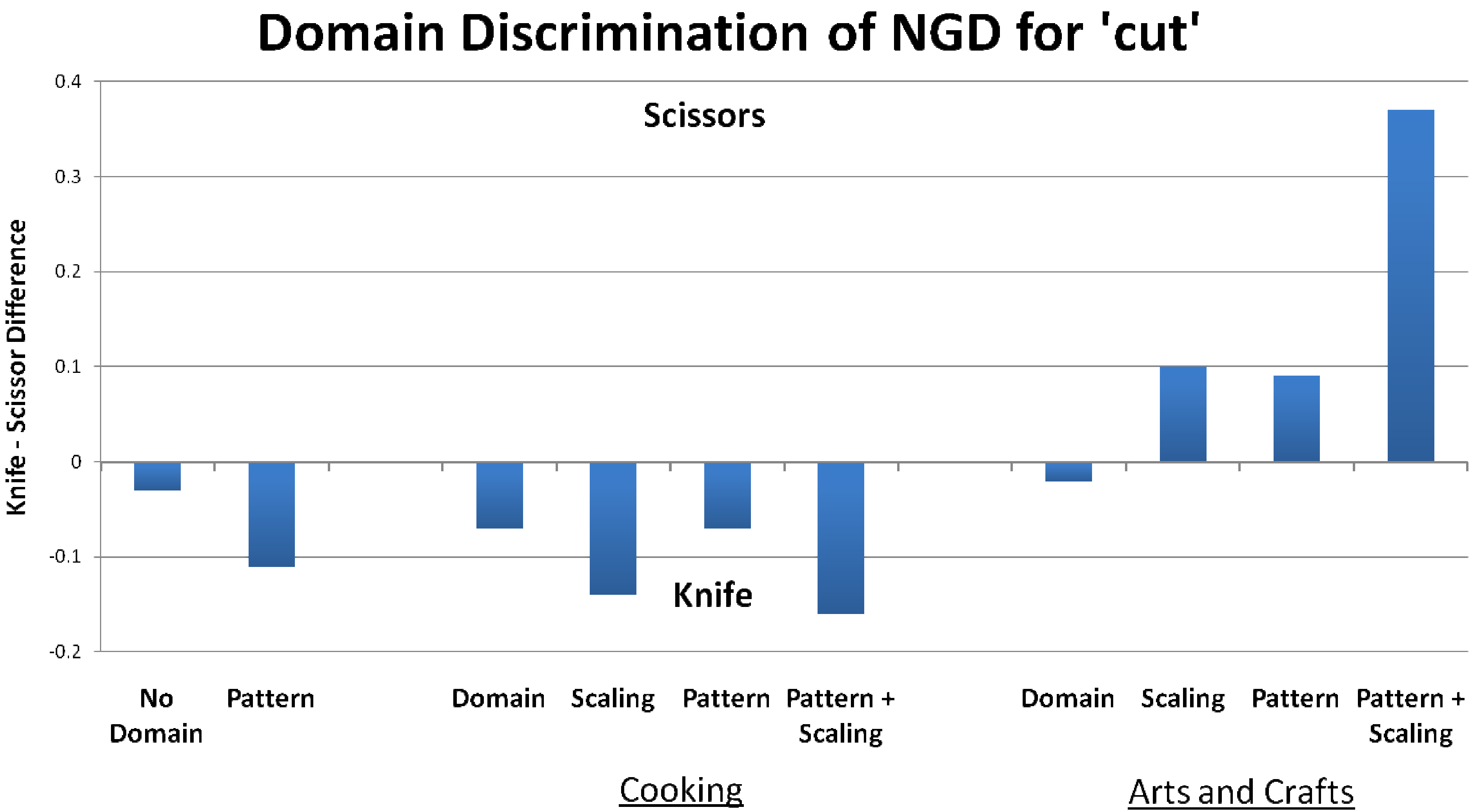
- **Domain scaling**
  - If tools could be from different domains, such as from Wikipedia tool search

$$SNGD(x, y, domain) = NGD(x, y) \times NGD(y, domain)$$

  - *x* is an action, *y* is a tool, *domain* is a domain such as "arts and crafts" or "cooking"
  - Further bias towards tools common to a particular domain
  - Empirically based

Other Uses of Domain Knowledge

- **Objects unknown** – look up objects listed in each action's Wikipedia page
- **Actions unknown** – look up actions listed in each object's Wikipedia page
- **Refine results** – use modified Google Distance to only find objects or actions relevant to the domain



Domain: … AND “arts and crafts” / ”cooking”
Scaling: SNGD(x,y,domain) = NGD(x,y) * NGD(y,domain)
Pattern: “scissors * cut” OR “cut * scissors” …

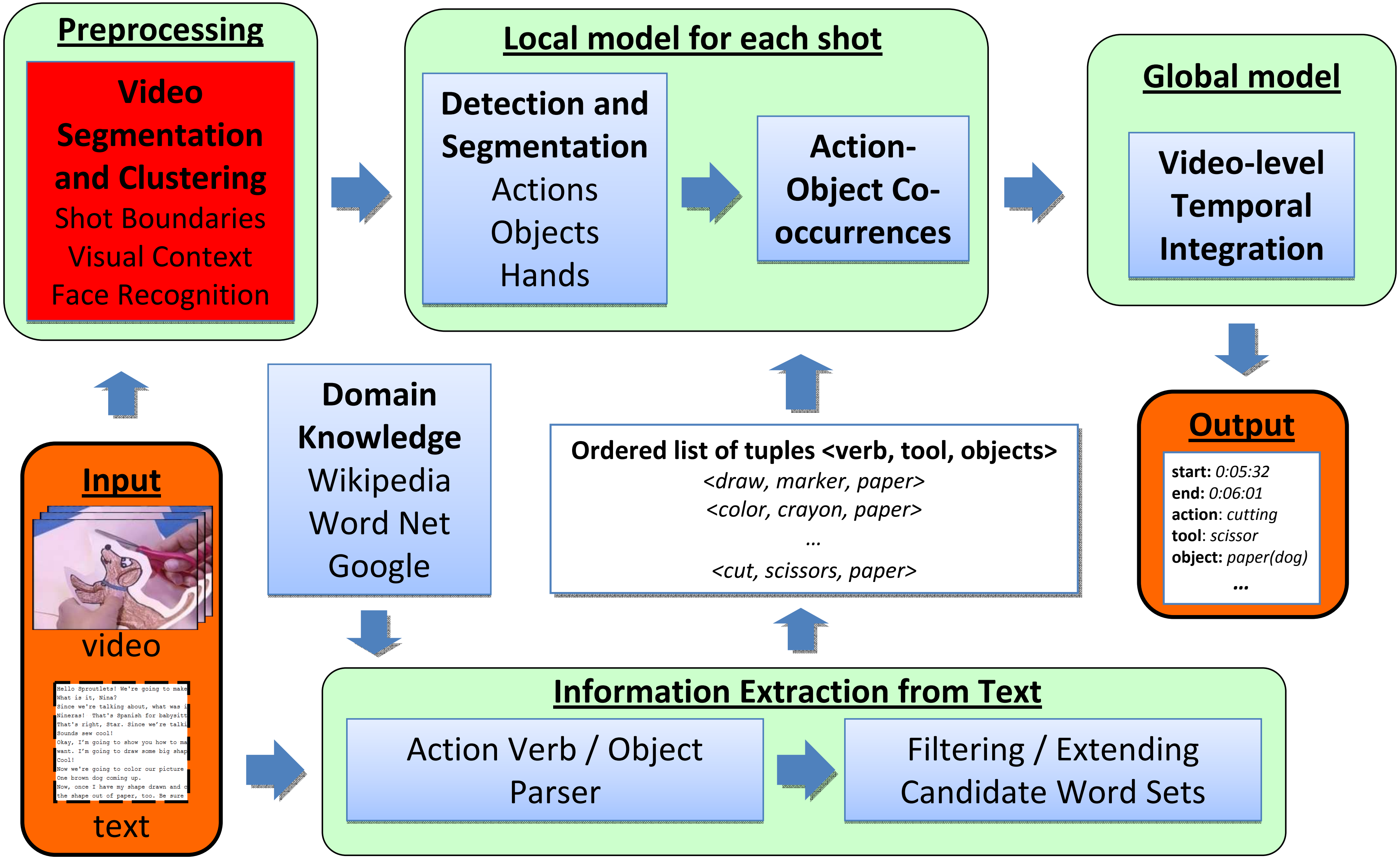
Future work

- Extract physical characteristics from web and Wikipedia to aid in unsupervised object detection

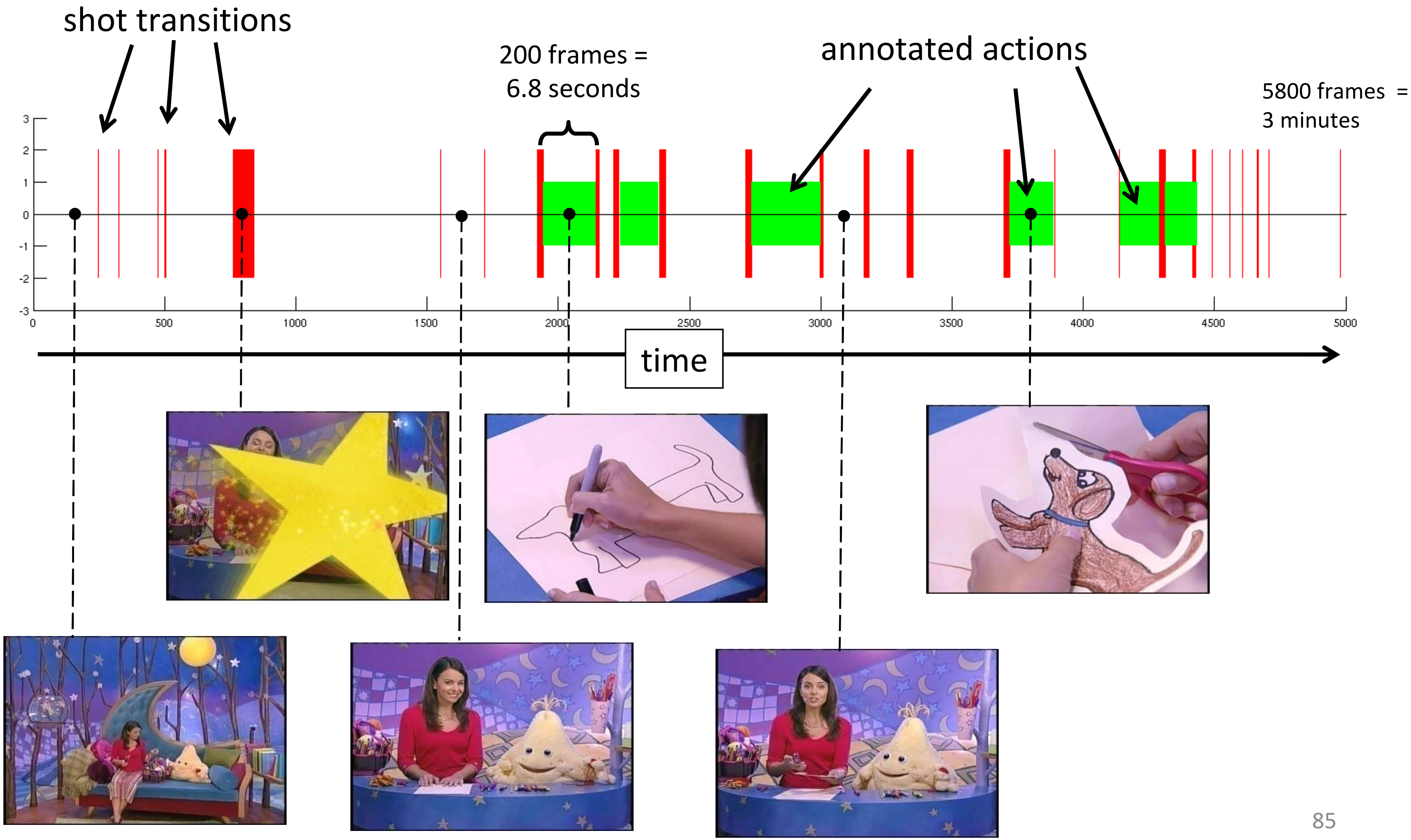
	crayon	marker	brush	scissors	glue
color	other	other	silver	silver	white
bristles	no	no	yes	no	no
elongated	yes	yes	yes	no	no
convex	yes	yes	yes	no	yes

‘bristles’, ‘elongated’,

Preprocessing



Episode timeline: *“Babysitter’s Animal Sewing Cards”,* PBS Sprout TV



Motivation

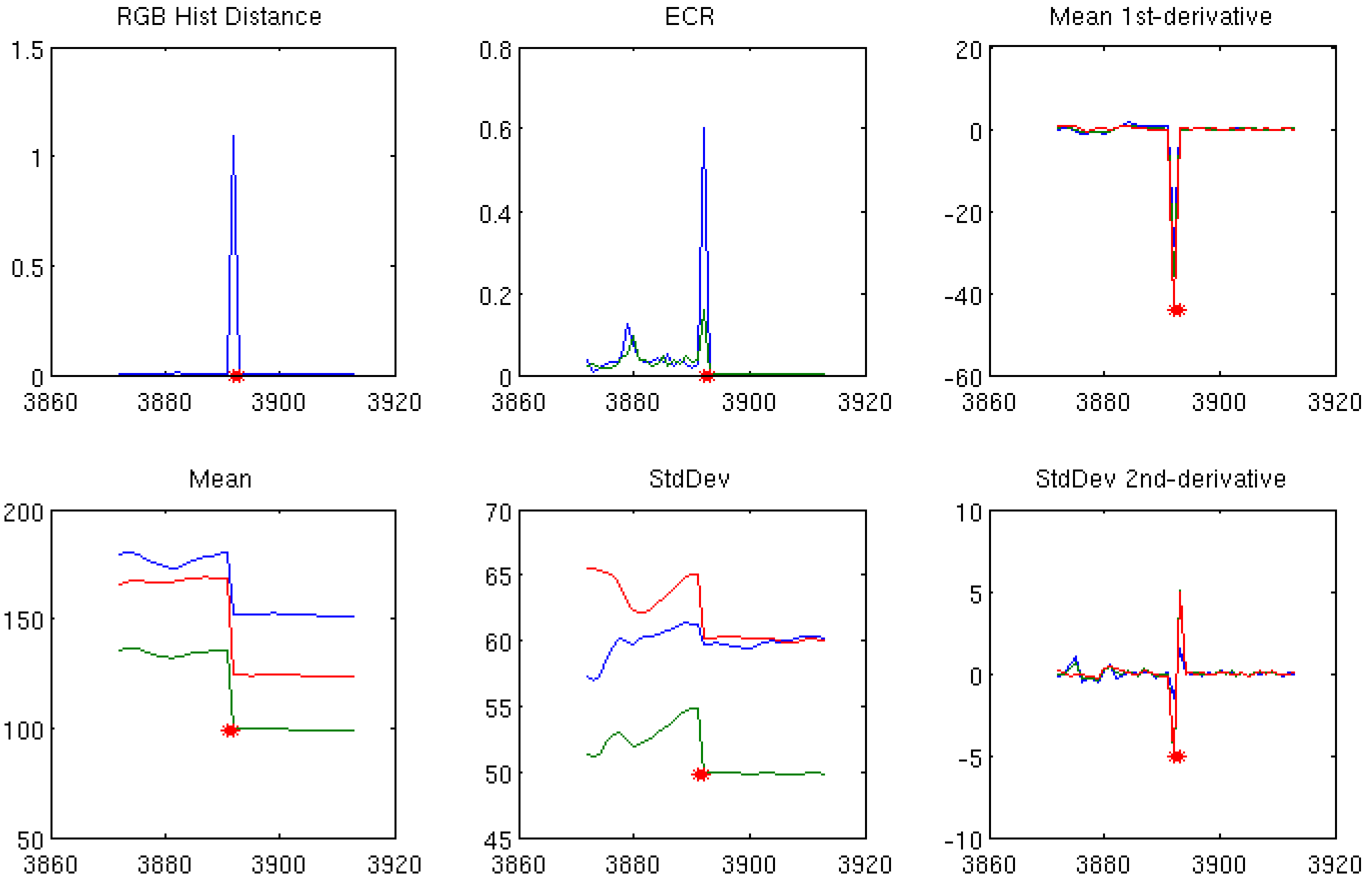
- A broadcast video consists of a sequence of "shots" that are separated by transitions
- Type of transition indicates semantic changes (or not) – Grammar of the Film Language (Arijon, 91)
  - Cut: semantic change
  - Dissolve: change in time or place, but action continues
- Segment and cluster the video into semantic subdivision ("shots") based on shot boundary detection and clustering based on visual similarity

Previous Work

- Shot boundary estimation
  - **Reliable Transition Detection In Videos: A Survey and Practitioner's Guide** (R. Lienhart, 2001)
- Shot clustering
  - **Identification Of Film Takes For Cinematic Analysis (**B.Truong, S. Venkatesh & C. Dorai, 2005)
  - **Movie/Script: Alignment and Parsing of Video and Text Transcription** (Cour et. al., *2008)*
  - **Taxonomy of Directing Semantics for Film Shot Classification** (Wang & Cheong, 2009)

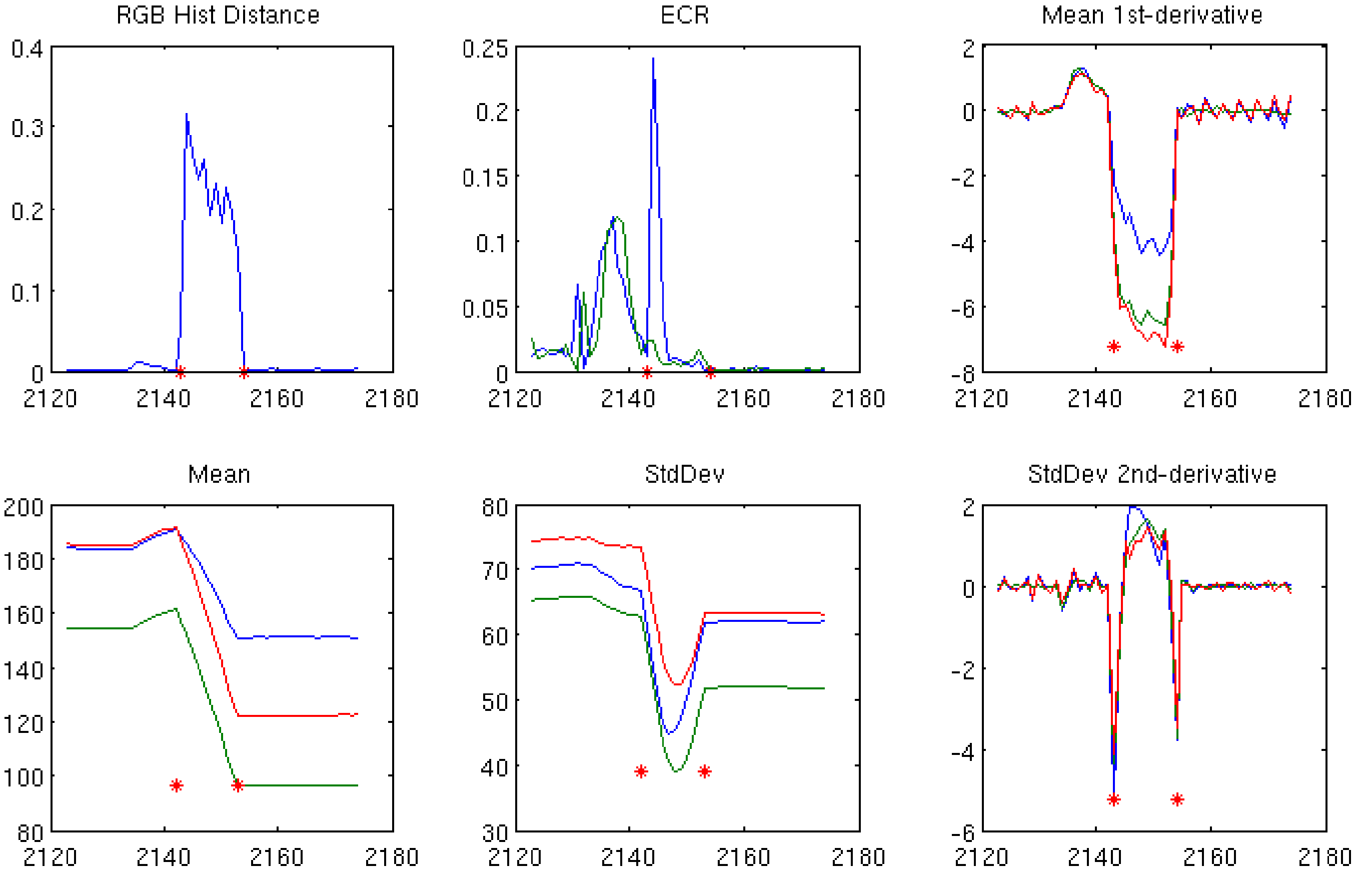
Hard cut shot boundary

- Threshold RGB Color Histogram Frame Differences



Dissolve shot boundary

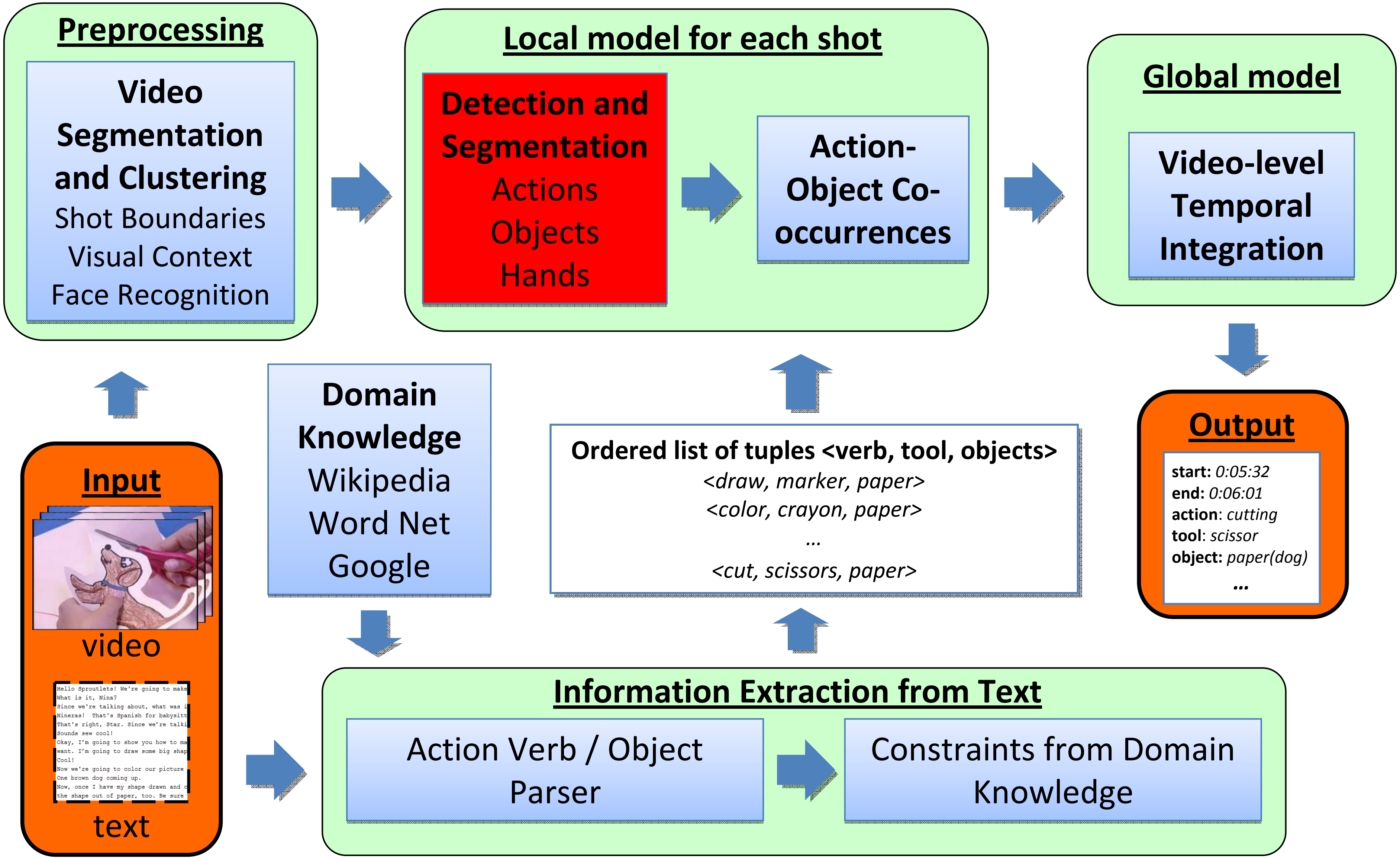
- Discont in 1st Deriv. of Mean and 2nd Deriv of StdDev.



Our Approach for Visual Context Detection

- Features used
  - Face Recognition (Pittsburgh Pattern Recognition or OpenCV)
  - Gist features (Oliva & Torralba, 2001)
  - Color SIFT (van de Sande, Gevers and Snoek,2010)
- Cluster shots into zoomed-in (="*action*") and zoomed-out (="*conversation*") shots
- 97% accuracy to distinguish zoomed-in/zoomed-out shots

Action Recognition



Previous Work

- Very active research topic:
  - CVPR 2010: ~10% ECCV 2010: > 15%*
- Common approaches:
  - Skeletal models
  - Appearance and motion statistics
  - Local vs Global models
  - Frame-level vs shot-level
- Common challenges:
  - Scene and Self occlusions, …
  - Environmental affects: Lighting, clothing, carry-on accessories, …
  - Video size, sampling rate, camera motion, …
  - Other challenges: Multiple actions/humans, human/object interactions, semantic interpretations, …

*Estimated from officially published CVPR and ECCV statistics on Action and Event recognition

Previous Work

- Global approaches
  - Optical Flow histograms [Efros 03, Chaudhry 09]
  - Flow and/or Shape [Tran 08, Gorelick 07, Yilmaz 05]
  - System theoretic with skeletons [Bissacco 01 06, Ali 07]
- Local approaches
  - Spatio-temporal features [Dollar 05, Laptev 08, Willems 08]
  - Bag of features
  - Limb motion models [Ikizler, 08]

Our Approach

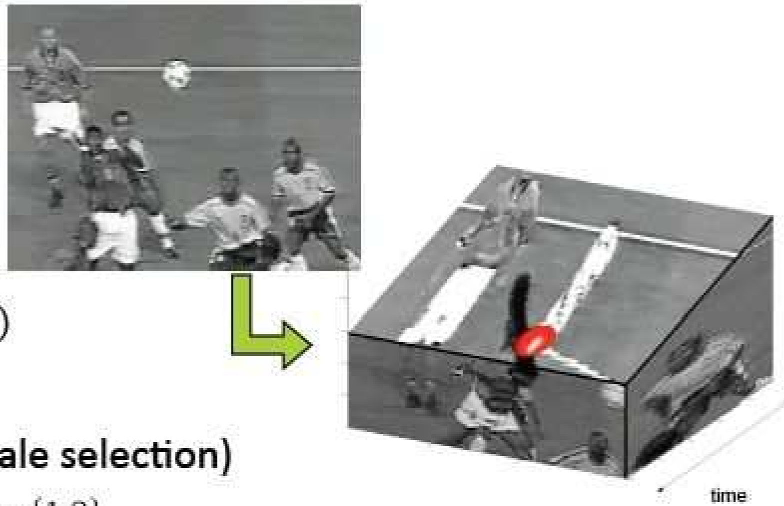
- Supervised action learning
  - Global Histograms of Oriented Optical Flow (HOOF) [Chaudhry 09]
  - Spatial Temporal Interest Points [Laptev 08]
    - Histograms of Gradients (HOG)
    - Histograms of Flow (HOF)
  - Local Histograms of Oriented Optical Flow
- Unsupervised Multiple Instance Learning
  - Automatic action label learning

Feature extraction

- Space-time corner detector
  [Laptev, IJCV 2005]

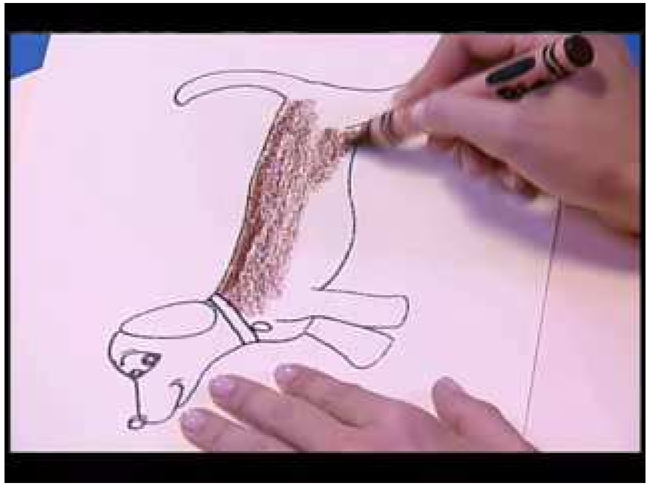
$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$



- Dense scale sampling (no explicit scale selection)

$$(\sigma^2, \tau^2) = \mathcal{S} \times \mathcal{T},\ \mathcal{S} = 2^{\{2,\ldots,6\}}, \mathcal{T} = 2^{\{1,2\}}$$

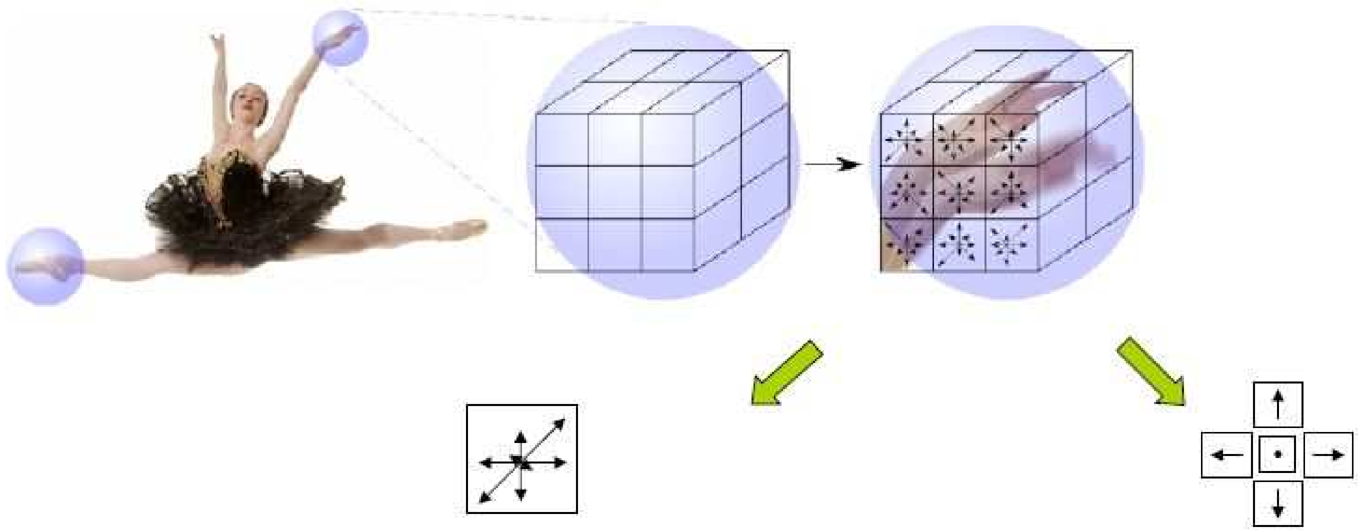


Coloring

Bending

Feature descriptor

Space-Time Features: Descriptor



Histogram of oriented gradient (HOG)

Histogram of oriented flow (HOF)

[Laptev 07]

Experiments

- Use all manually annotated sequences
  - Sequence-level features
    - HOG+HOF – 72+90 = 162 dimensions
    - 100/4000 clusters (codewords)
    - Compute Term-Frequency for each codeword
    - Chi-squared distance
  - Setup:
    - Zoomed-in view = 186 seq
    - 50 % Training, 50% Test
    - 1-NN classification
    - SVM results later

Results

- Confusion matrices – HOG+HOF

HOG/HOF features - 4000 clusters

	Bending	Coloring	Cutting	Decorating	Drawing	Folding	Gluing	Painting	Placing	Taping	Threading	
Bending									1.0			2
Coloring		.67			.17				.17			6
Cutting	.11		.84						.05			19
Decorating					.50				.50			2
Drawing			.14		.38				.43		.05	21
Folding			.20			.20			.40	.20		5
Gluing			.10				.50	.10	.20		.10	10
Painting			.20		.20				.60			5
Placing	.13		.31		.06	.13			.31		.06	16
Taping			.67						.33			3
Threading			.33						.67			3

Recognition: 42.39%

HOG/HOF features - 100 clusters

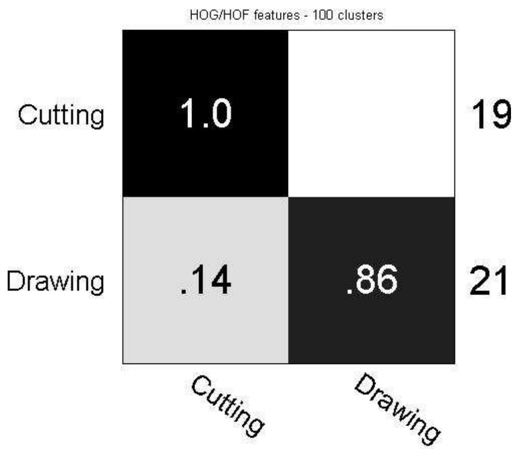
	Bending	Coloring	Cutting	Decorating	Drawing	Folding	Gluing	Painting	Placing	Taping	Threading	
Bending			.50						.50			2
Coloring		.67			.33							6
Cutting	.16		.63						.11	.05	.05	19
Decorating							.50		.50			2
Drawing			.05		.62		.05		.24	.05		21
Folding			.40						.20	.40		5
Gluing		.10			.30	.10	.30		.20			10
Painting			.20		.40			.20	.20			5
Placing	.06		.25		.06	.06			.44	.06	.06	16
Taping			.67						.33			3
Threading			.33						.67			3

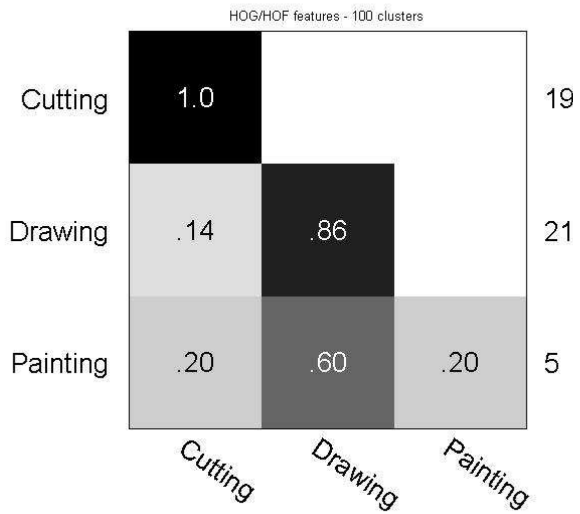
Recognition: 43.48%

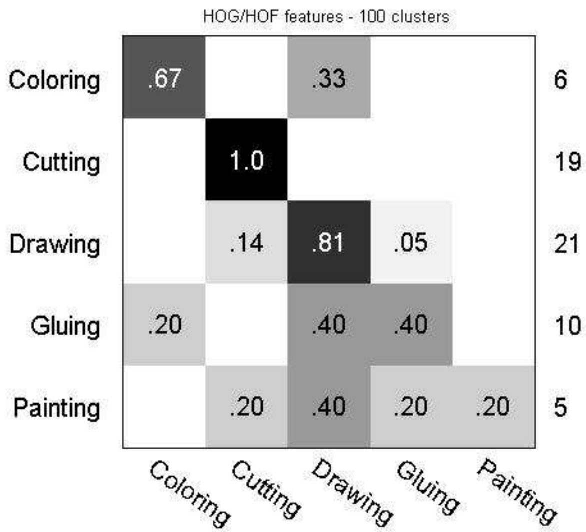
[Assigning label of most frequent class in training set: 22.83%]

Results

- Most frequent classes:







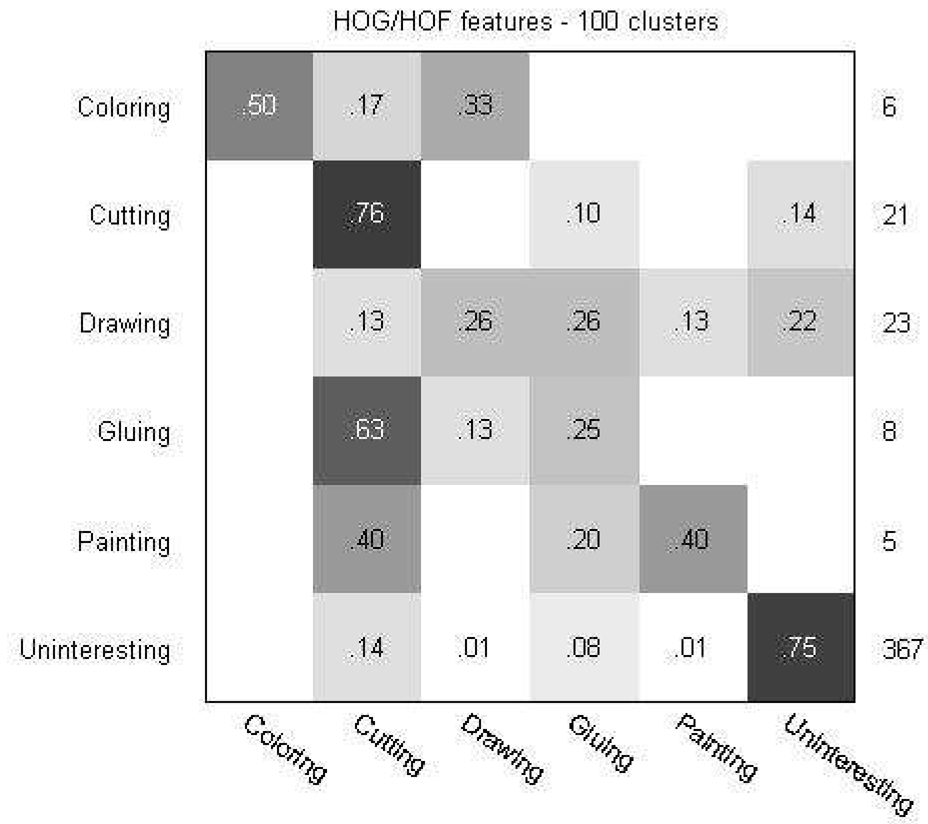
Recognition: 92.5%   Recognition: 84.44%   Recognition: 73.77%

Experiments

- Fully automatic shot segmentation
- 13/27 episodes used for training action features
  - Ground-truth annotations transferred to shots
    - Naturally overlap
  - 5 class + 1 ‘Uninteresting’ or ‘Other’ class
    - Zoomed out sequences with actors talking
    - Rare actions
  - Testing on remaining 14/27 episodes

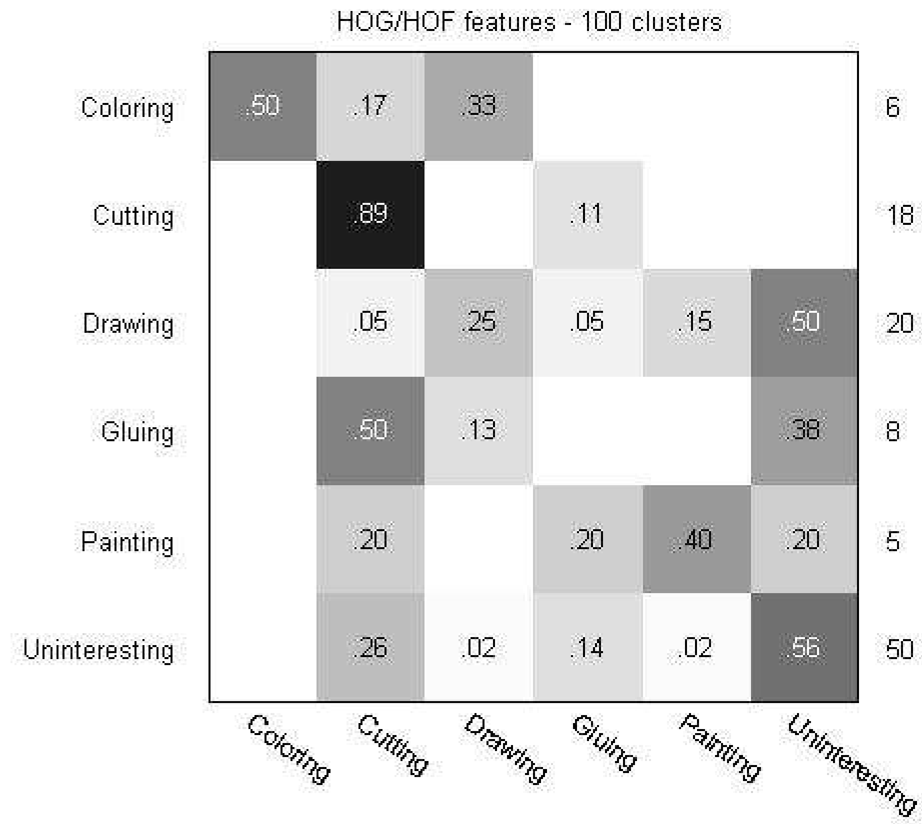
Results

- Train on all shots
  - Zoom outs included
- Include all uninteresting shots
- Find codewords by equally sampling from all classes
- Train classifier on same order of sequences for all classes
- Recognition rate = 71%
  - (most freq class level = 85%)
- Average class-level recognition rate = 48.67%
  - (random choice = 17%)



Results

- Train on all shots
  - Zoom outs **excluded**
- Find codewords by equally sampling from all classes
- Train classifier on same order of sequences for all classes
- Detect and discard zoom out shots
- Recognition rate = 50%
  - (most freq class level = 47%)
- Average class-level recognition rate = 43.33%
  - (random choice = 17%)



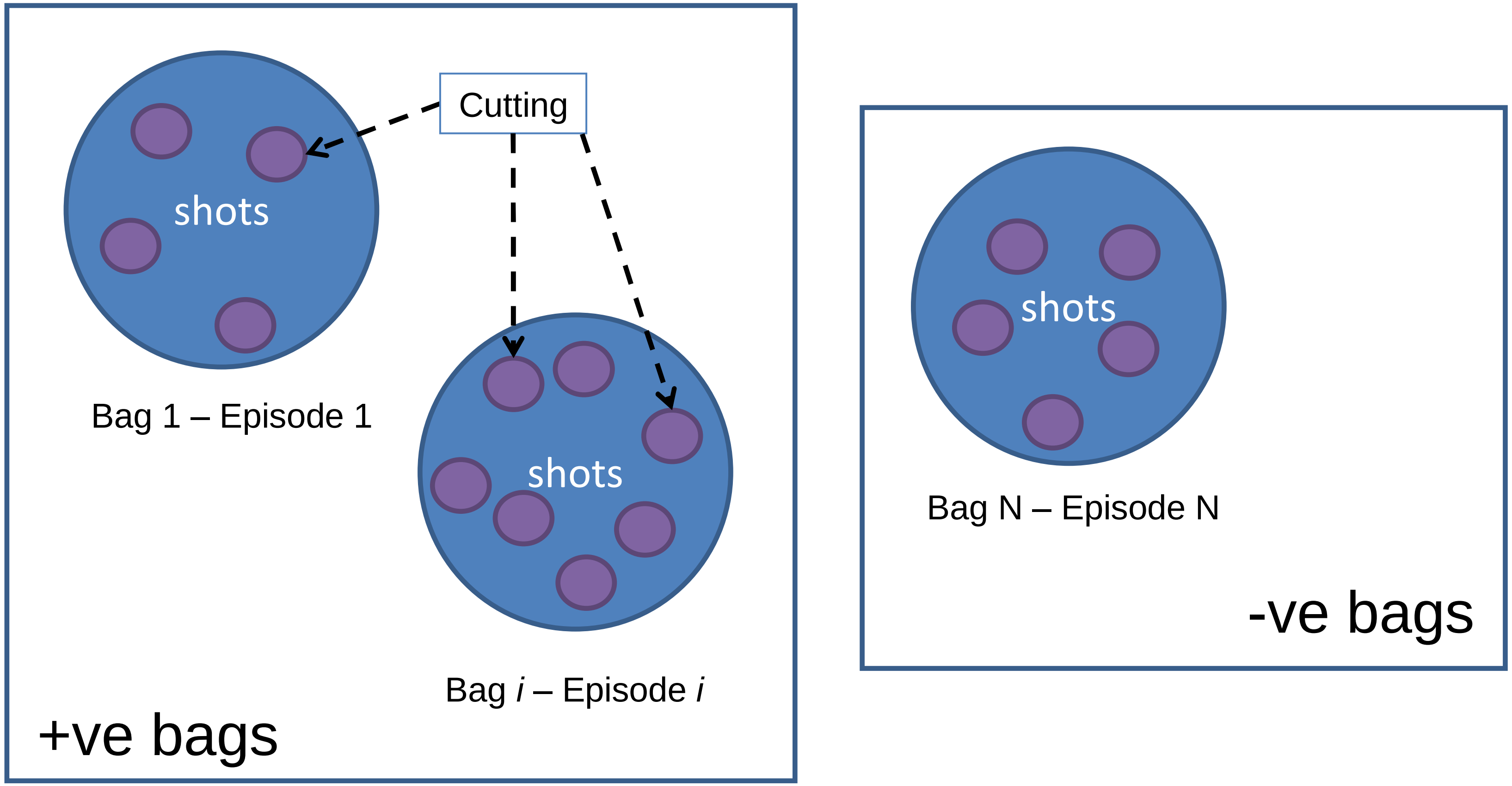
Our Approach

- Supervised action learning
  - Global Histograms of Oriented Optical Flow (HOOF) [Chaudhry 09]
  - Spatial Temporal Interest Points [Laptev 08]
    - Histograms of Gradients (HOG)
    - Histograms of Flow (HOF)
  - Local Histograms of Oriented Optical Flow
- Unsupervised Multiple Instance Learning
  - Automatic action label learning

Multiple Instance Learning

- Instance level labeling is costly
- Movie level labeling can be guessed from text
- Can we get instance level labeling from movie-level labels?
- Create bags of instances such that
  - Positive bags: at least one positive instance
  - Negative bags: No positive instance
  - Automatically learn the best feature weighting and label all instances

Multiple Instance Learning



Learn label of all instances given bag labels

Experiments

- Diverse Density [Maron 98]
  - Find regions in feature space that have
    - high density of positive examples
    - low density of negative examples
  - Positive should lie *close* to these regions
  - $\mathrm{argmax}_t \, \mathrm{Prob}(t \mid \{P_1, \ldots, P_n\},\{N_1, \ldots, N_m\})$
  - Gradient ascent to optimize t
  - MIL Library Toolkit [http://www.cs.cmu.edu/~juny/MILL]

Experiments

- Setup
  - Fully annotated dataset
  - 13/27 training, 14/27 test
  - 10 starting points
  - Average bag-level and instance-level accuracies
  - 1 vs all action classification
- Observations
  - Binary classification inconclusive
  - Data size too small

Accuracy (%)	Bag level	Instance level
Coloring	71	94
Cutting	50	20
Drawing	57	22
Gluing	50	10
Painting	86	95

Experiments

- Setup
  - Fully annotated dataset
  - Full dataset
  - 10 starting points
  - Average bag-level and instance-level accuracies
  - 1 vs all action classification
- Observations
  - Not comparable with previous results
  - Promising for automatic labeling

Action	# +ve bags Total = 27	Accuracy (%)
Coloring	6	94
Cutting	17	80
Drawing	18	77
Gluing	13	89
Painting	4	94

Summary

- State-of-the-art action recognition approaches do not scale well
  - Number of classes
  - Different number of sequences per class
  - Unknown action models
  - Across different contexts and domains
- Need for integrating context and domain knowledge
  - Hand and object (tool)
  - Text, temporal order

Later steps and Future Work

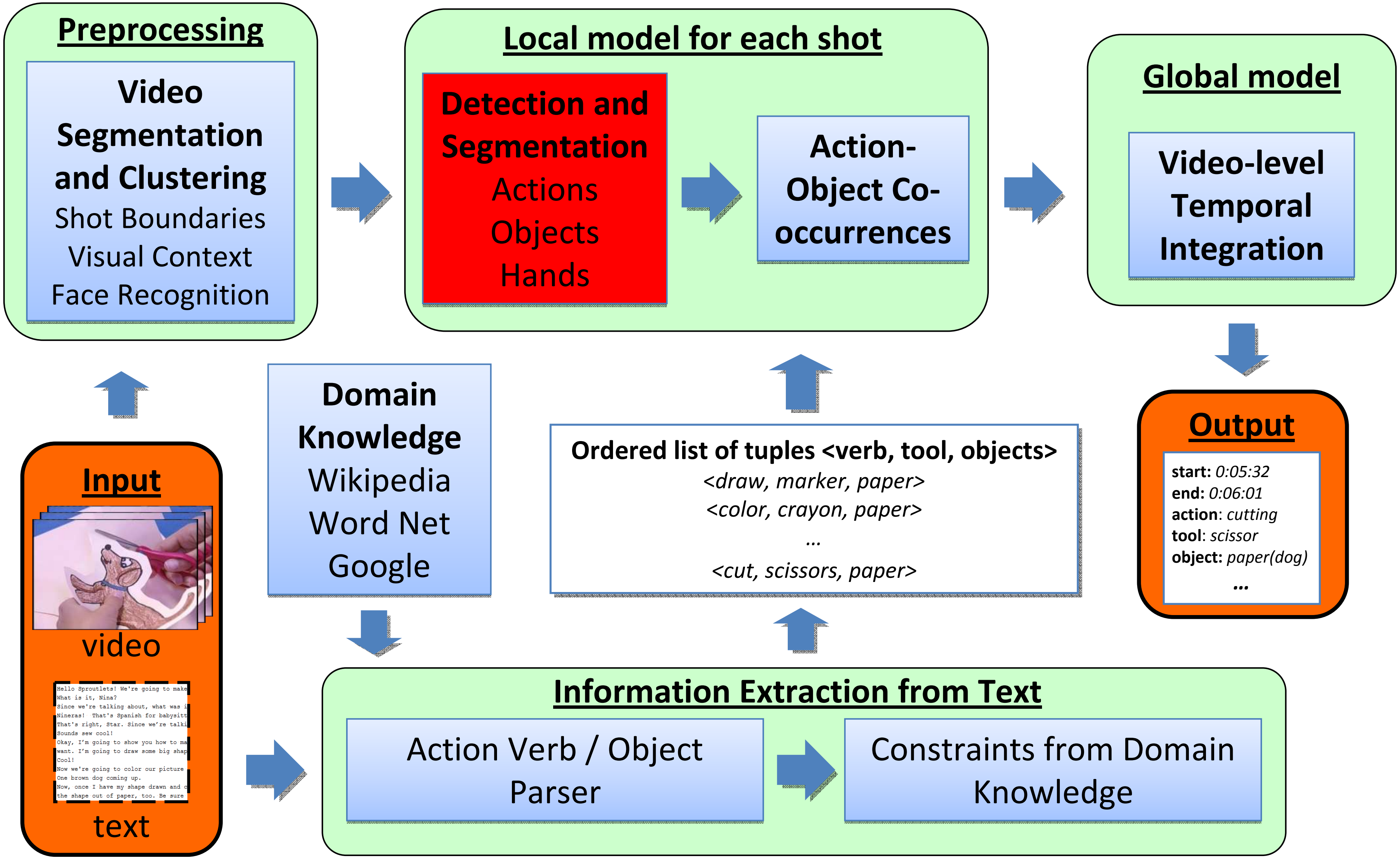
- Next steps
  - STIP HOG+HOF provides good action representation
  - Combined with Textual and Object and Hand features
- Future work
  - The best feature for action representation?
  - Other Combinations of flow and texture feature distributions over time, motion trajectories
  - Train using labels extracted using MIL and action-names from textual analysis

Time Line

- 1:30 pm Overview (Jan Neumann)
- 1:40 pm Vision and NLP (Jana Kosecka)
- 1:55 pm Information Extraction from NLP (Evelyne Tzoukermann)
- 2:05 pm Extracting actions and verbs from text (Frank Ferraro)
- 2:15 pm Extracting domain knowledge from the web (Ian Perera)
- 2:25 pm Action recognition (Rizwan Chaudry)
- **3:20 pm Break**
- **3:30 pm Object recognition (Gautam Singh)**
- **3:45 pm Joint models for actions, objects and text (Ben Sapp)**
- **4:05 pm Temporal modeling (Xiadong Yu)**
- **4:15 pm Segmentation and object attributes (Cornelia Fermueller)**
- **4:30 pm Closing Remarks (Jan Neumann)**
- **4:35 pm Questions & Discussion**

Topic Areas: Language, Vision, Language+Vision

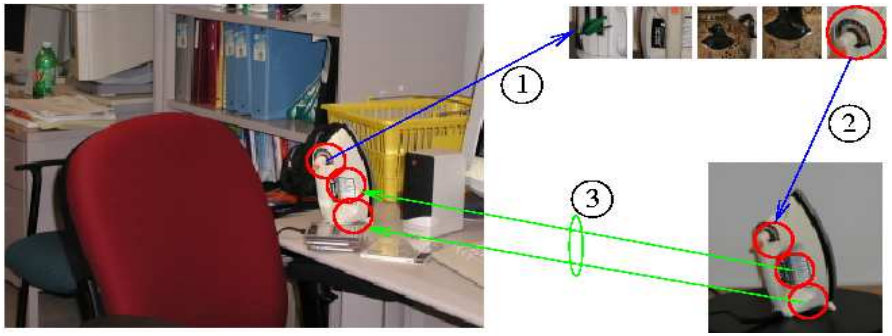
Object Detection



Object Detection

- Presence of certain objects in video provide an indication of the possible action being performed in them
- Possible challenges:
  - Viewpoint Variation
  - Illumination
  - Occlusion
  - Scale
  - Intra-class Variation
- Common Models:
  - Shape-based
  - Part-based
  - Sliding window template based
  - Local features based

Local features



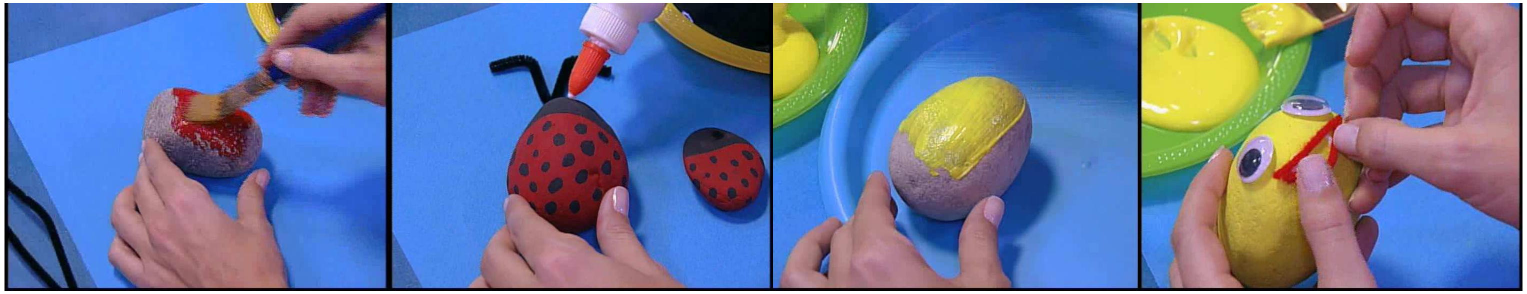
Shape Based models

Part based models

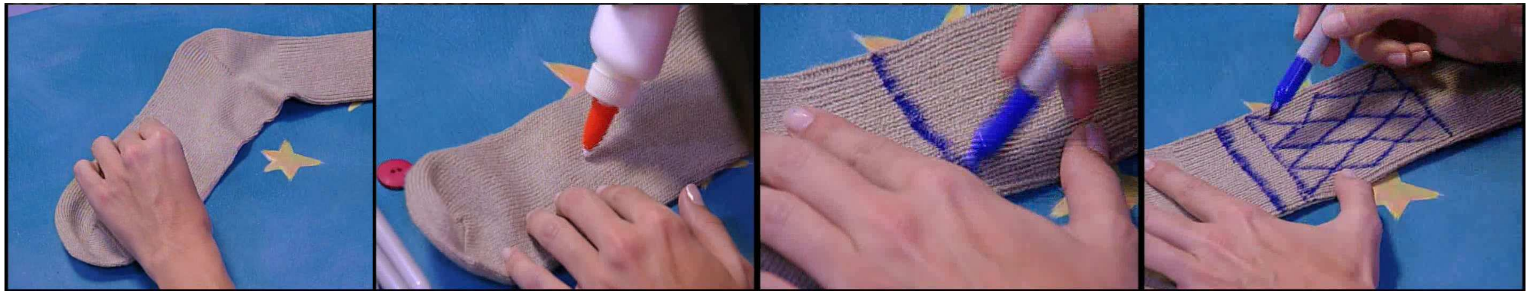
Sliding window template based

Sample Objects

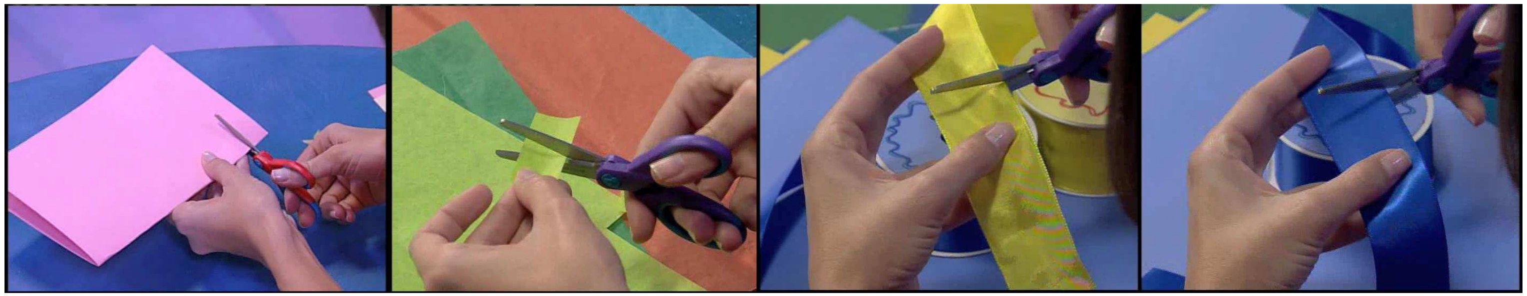
Rock



Sock



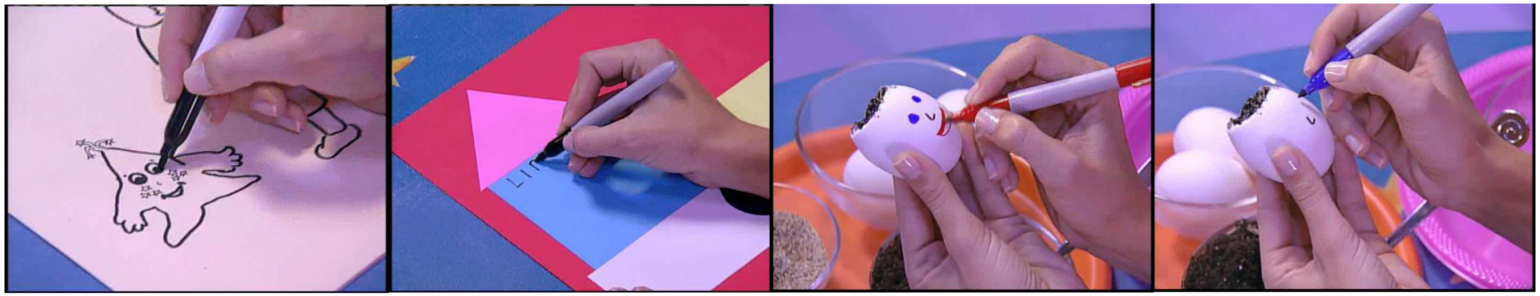
Paper/
Ribbon



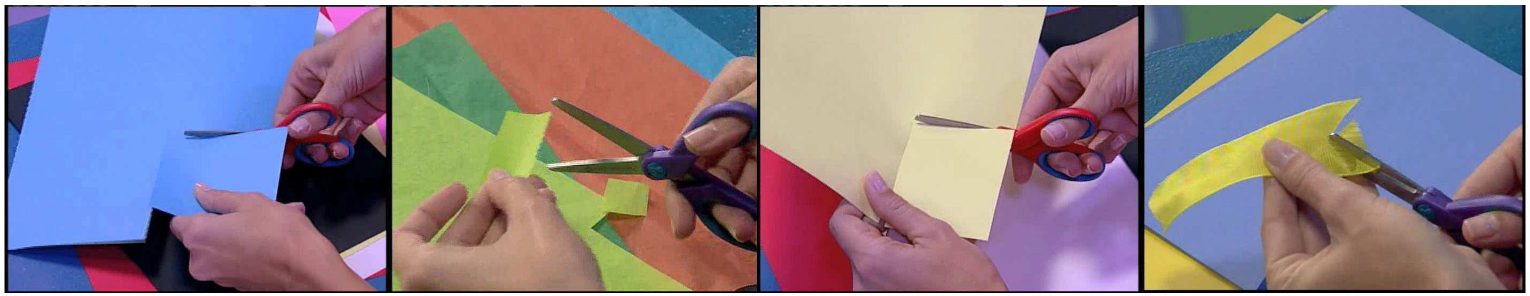
Sample Objects

Brush

Pen



Scissor



Objects

- Divided into two categories:

  Tools-

    - can be used to perform particular actions
    - consistent visual appearance

  Others-

    - may undergo transformation during an action
    - visual appearance may change over the course of the action

Tools List

Name	Name
Bottlecap	Papercutouts
Brush	Paperfigure
Button	Paperplate
Clay	Papershapes
Coffeefilter	Pen
ContainerofGlitter	Pencil
Crayon	Pietin
Cutout	Pipecleaner
Doily	Plasticeye
Egg	Ribbon
Figurine	Rock
Fuzzyredpompom	Scissors
Glitterpen	Sock
GlueBottle	Sponge
Jar	Tape
Marker	Thread
Paint	Tube
Paper	

Name
Brush
Crayon/ Marker/ Pen/Pencil
GlueBottle
Scissor

Discriminatively Trained Part Based Models

[Felzenszwalb et al 2010]

- Combines part based and template based models
- Parts and their spatial relationships learned automatically
- System represents object using a mixture of multi-scale part-based models
  - Each model component has a root filter and a set of part filters
  - Filters analogous to templates
  - Coarse root filter covers entire object
  - Higher resolution part filters cover smaller sections
  - Mixture of models useful for viewpoint invariance
  - Achieves state-of-the-art results on the PASCAL Visual Object Challenge
- Useful for tool detection problem
  - Part filters allow for tolerance to occlusion
  - Able to model deformation

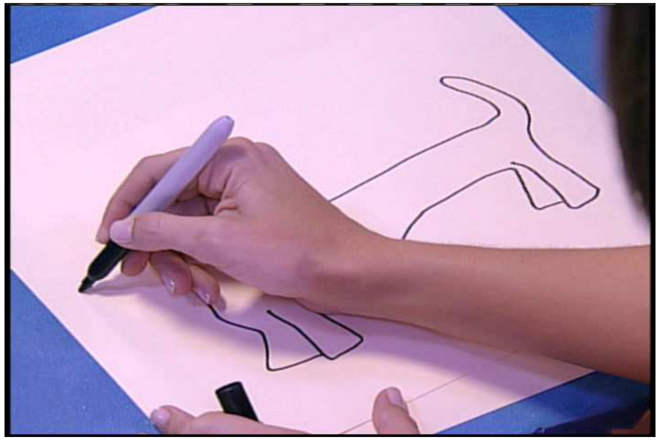
Discriminatively Trained Part Based Models

[Felzenszwalb et al 2010]

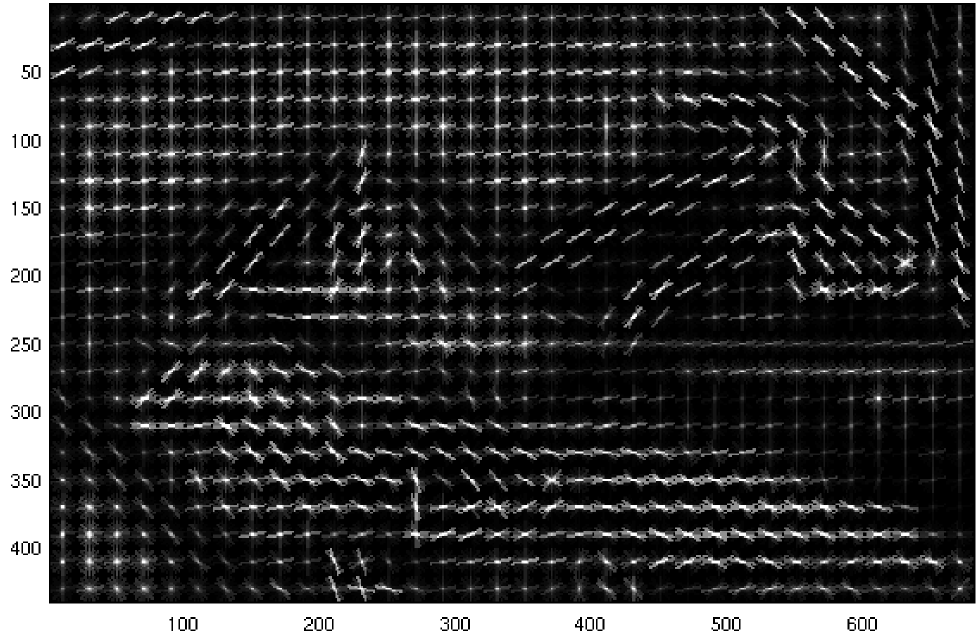
- Uses Histogram of Oriented Gradients (HOG) features as visual descriptors for an image
- Automatically learns parameters for individual model components
  - user specify number of components and parts before training
- Object hypothesis score computed as sum of response to individual filters minus deformation costs (for the parts)

Histogram of Oriented Gradients (HOG) representation

Image



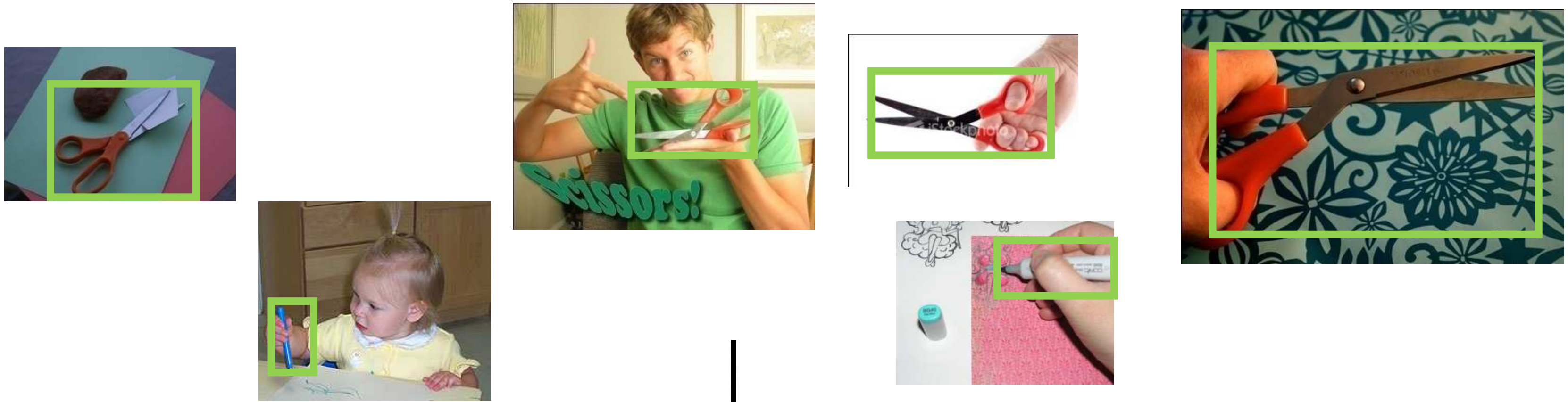
HOG



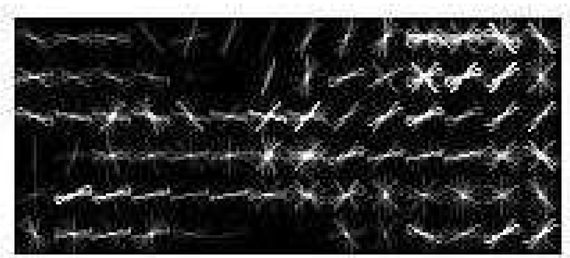
- Compact representation of image as 9 quantized edge orientations
- Invariant to extreme changes in lighting and color
- Invariant to slight changes in translation and rotation

Training

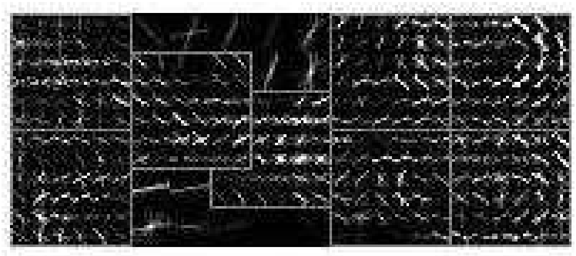
- Images obtained from the web
- Manually annotate with bounding boxes
- Training data includes positive and negative examples



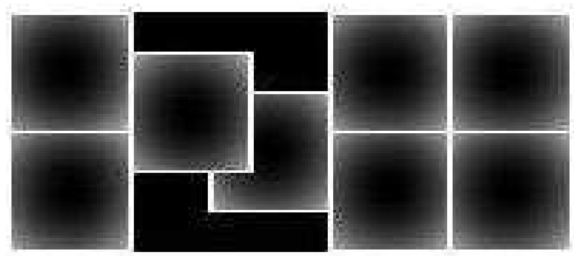
Learned parts-based model
(One component visualized)



Root Filter



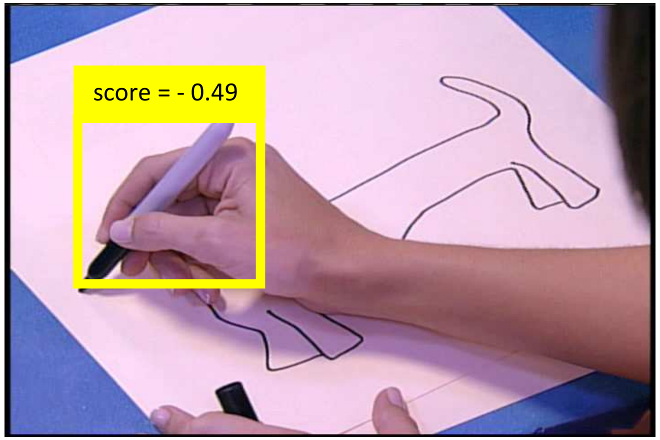
Part Filters



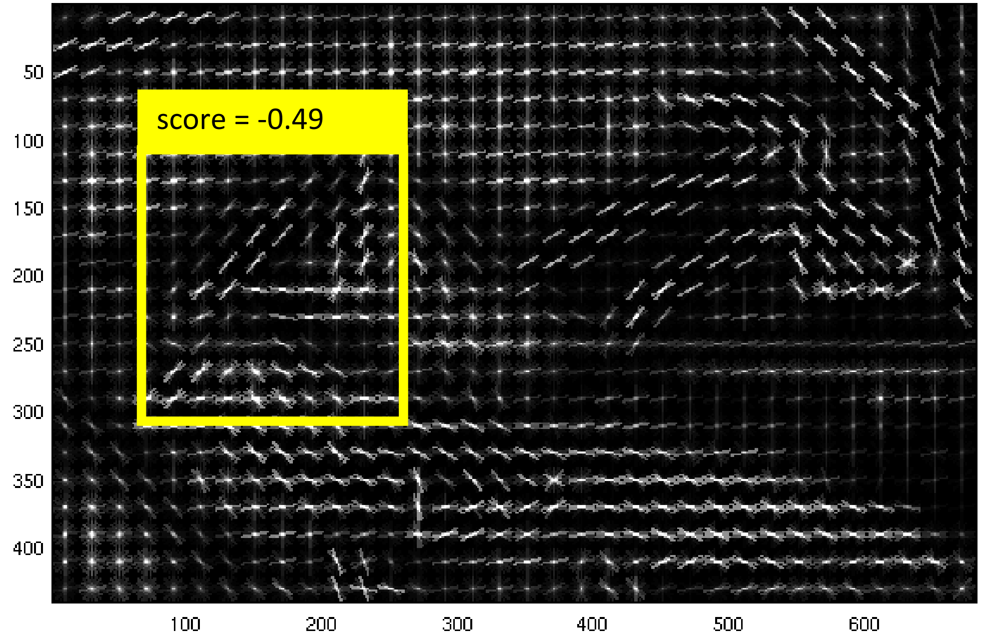
Spatial Model

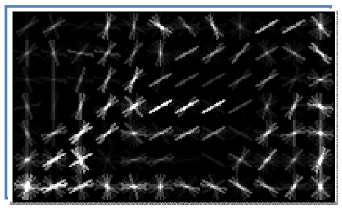
Matching

Image



HOG Feature Map



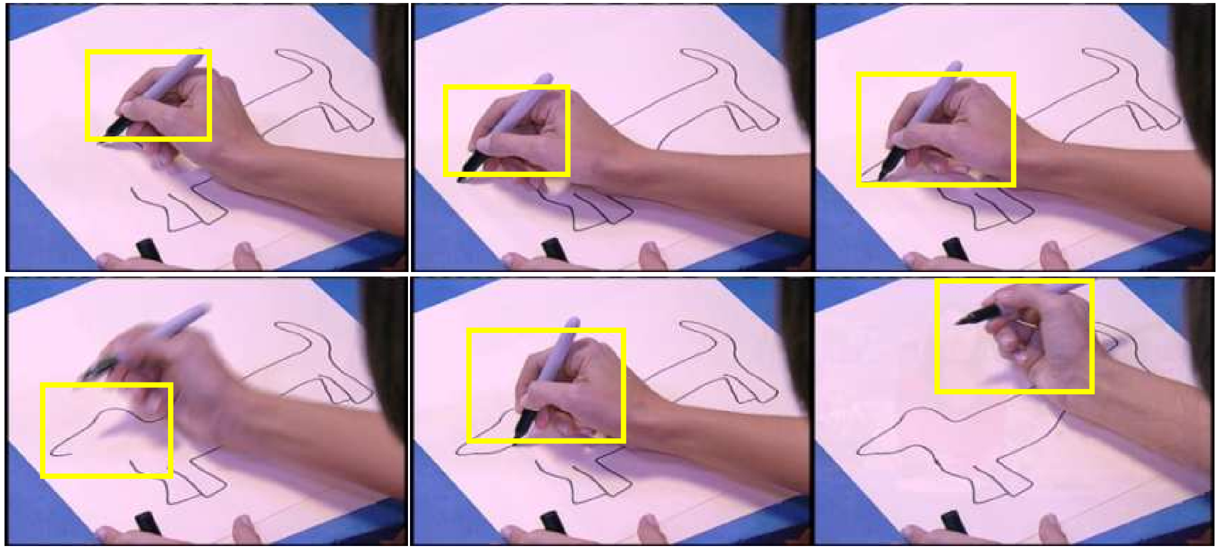


Root Filter

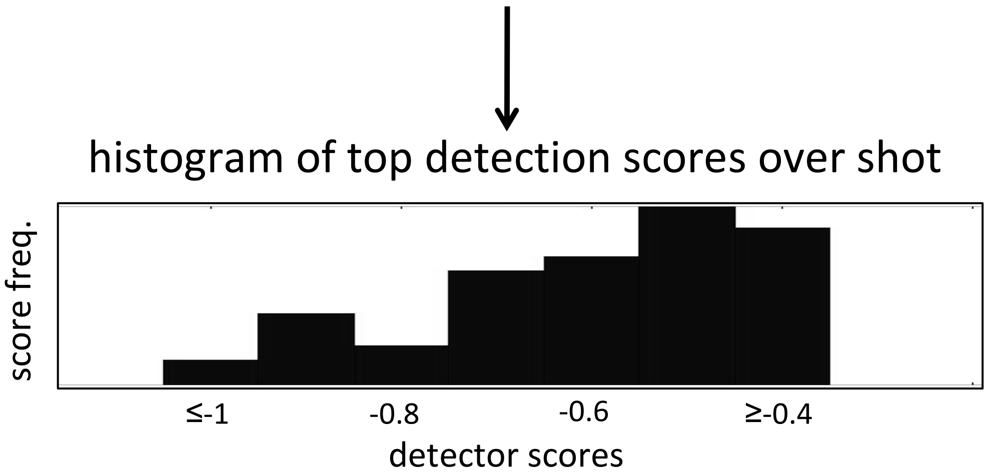
Discriminatively Trained Part Based Models

[Felzenszwalb et al 2010]

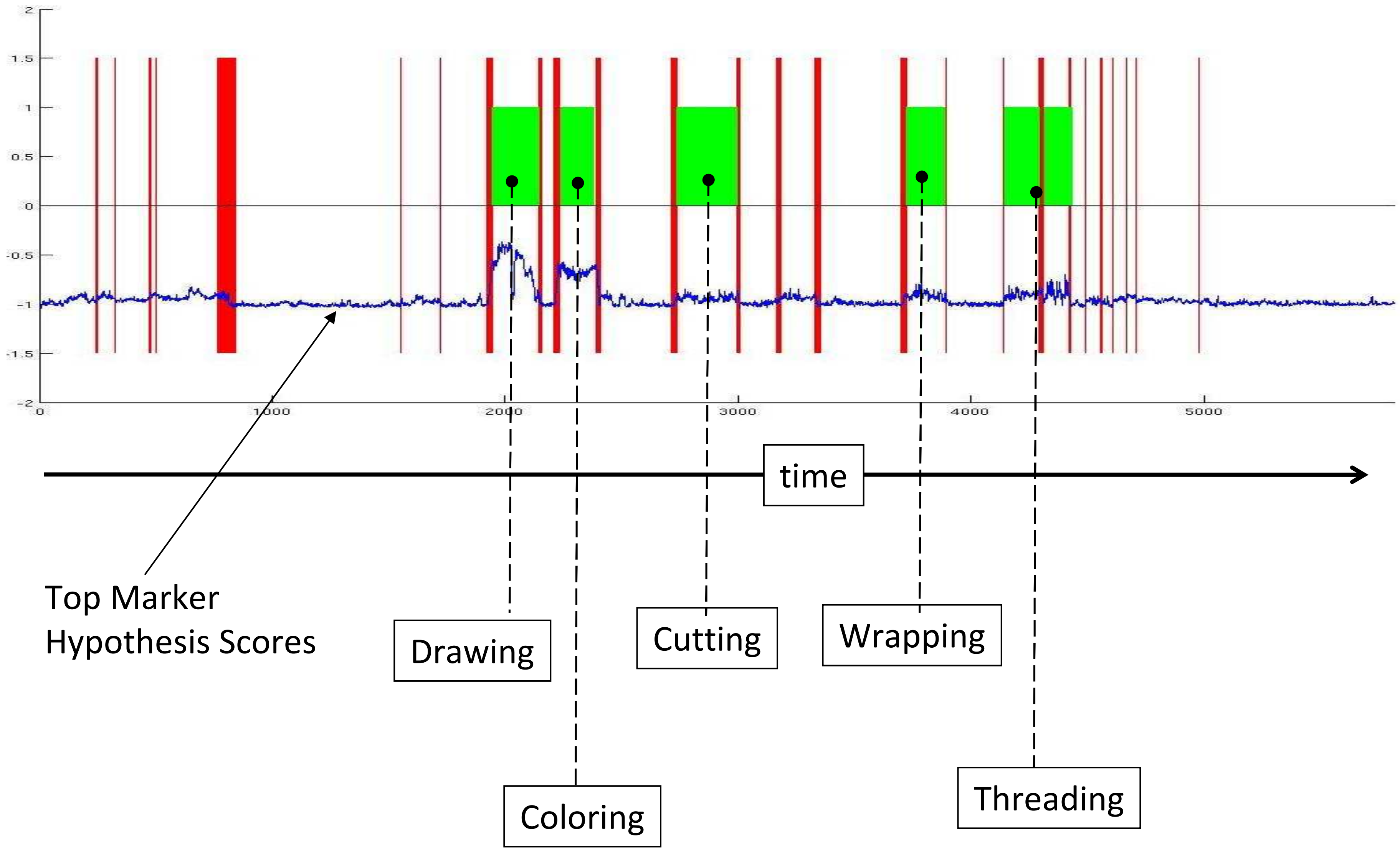
200 frame action shot



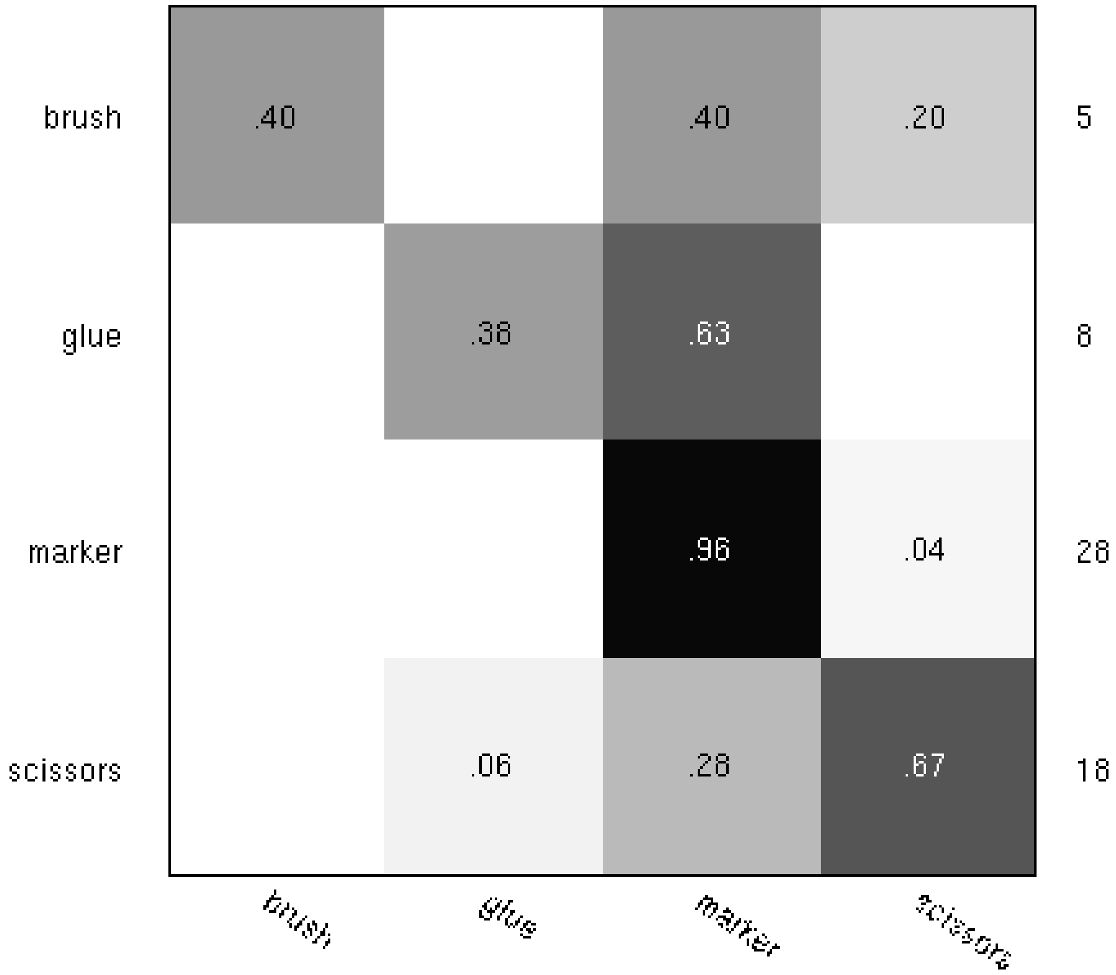
histogram of top detection scores over shot



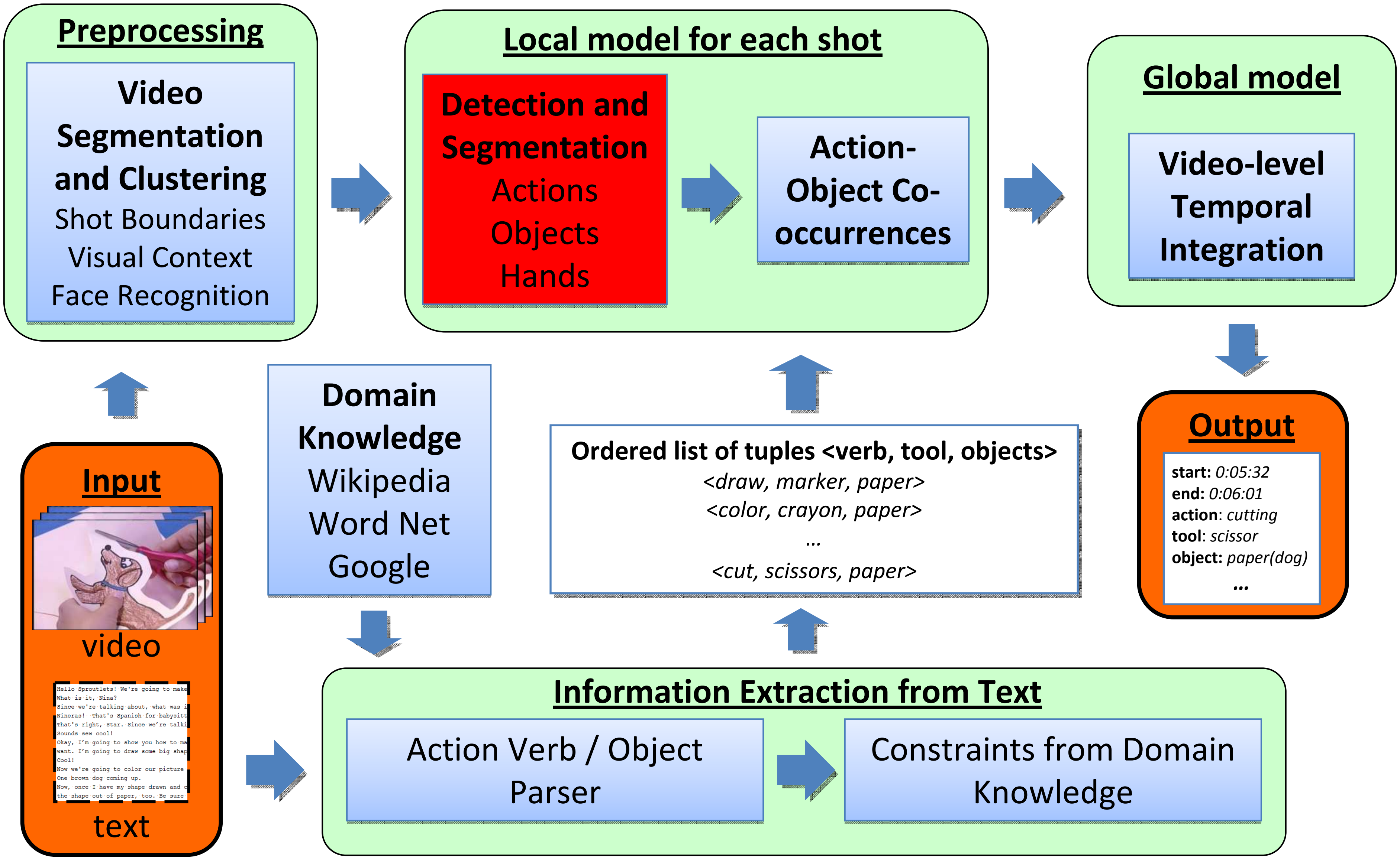
Episode timeline: *"Babysitter's Animal Sewing Cards"*, PBS Sprout TV



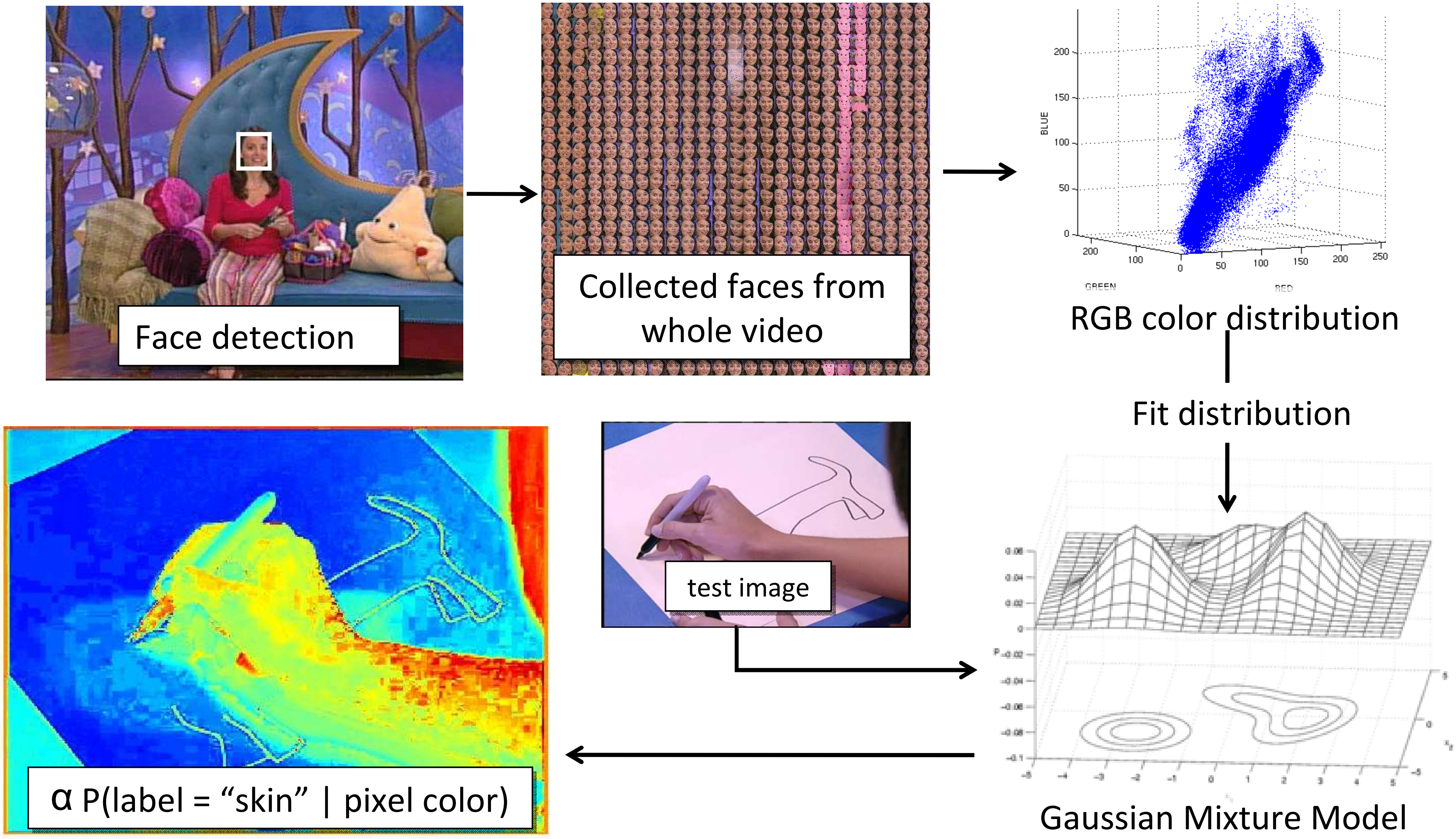
Tool Classification Confusion Matrix



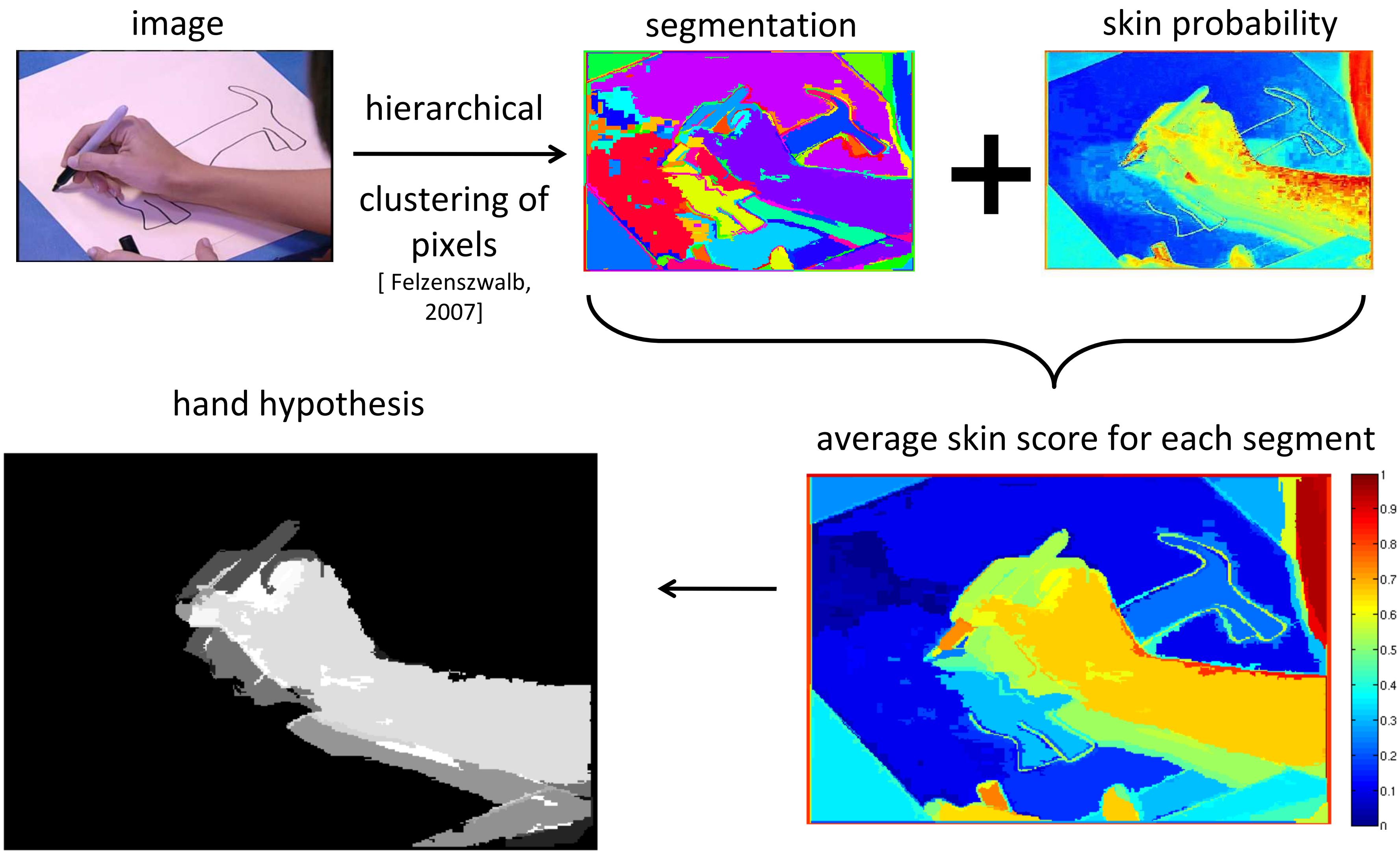
Hand Pose Detection



Skin color model pipeline



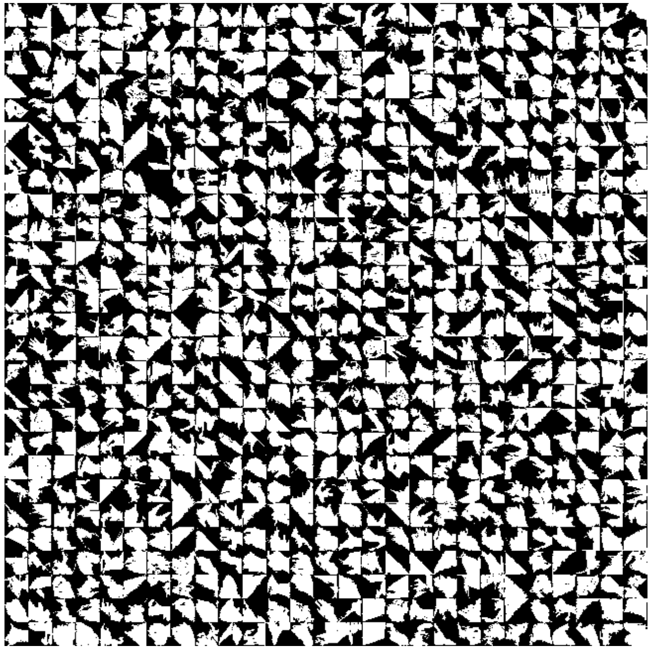
Skin model to hand detection



Hand pose words

hand hypotheses collected from all clips

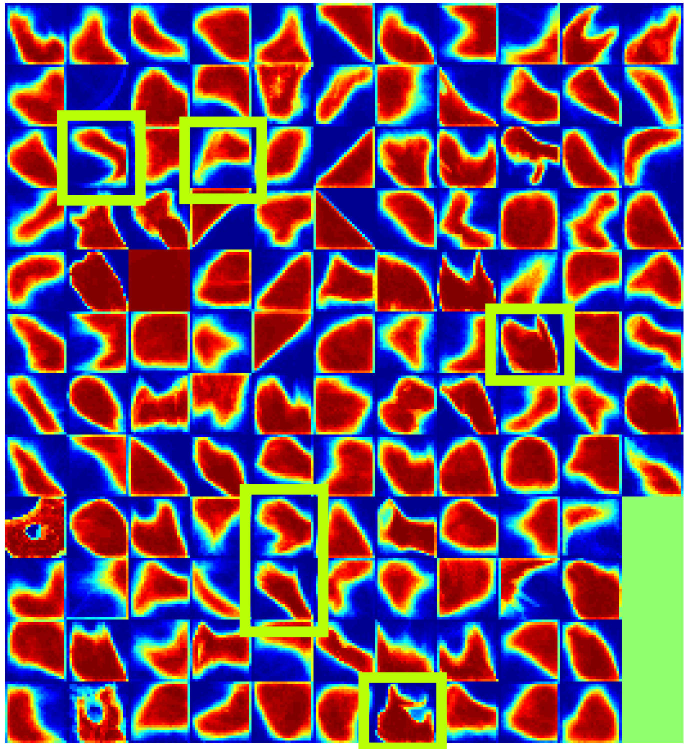
resized to 24x24 = 576 dimensional samples



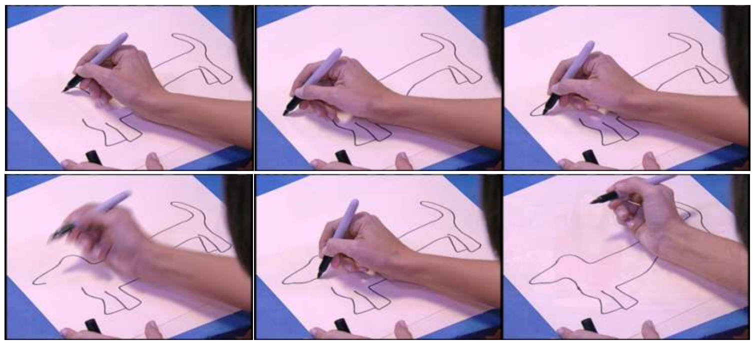
k-means clustering

on 576-dimensional patch vectors

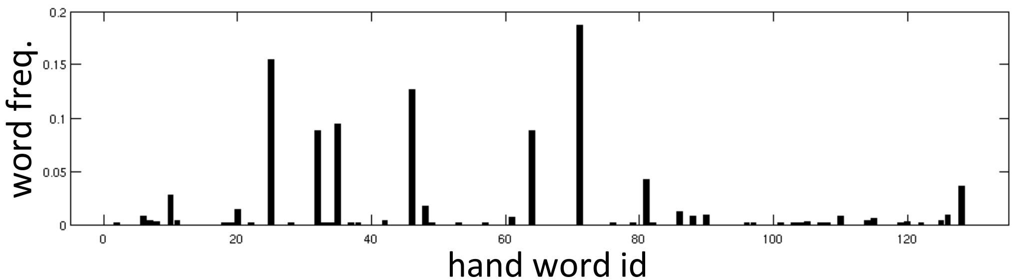
128 hand pose “words”



action shot

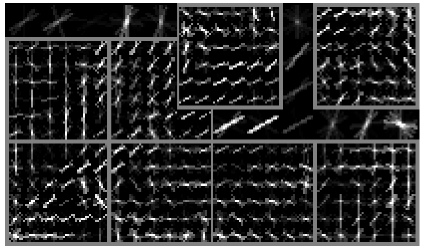


histogram of hand words

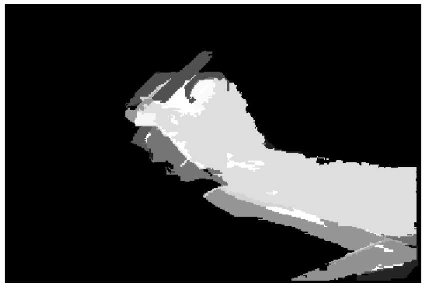


Visual Features Recap

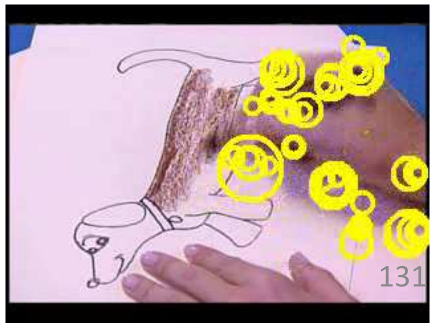
- **Tool detection features**
  - Histogram of object detector scores
  - 4 tool detectors (*writing tool, scissors, glue bottle, paint brush)*
  - 10 bins



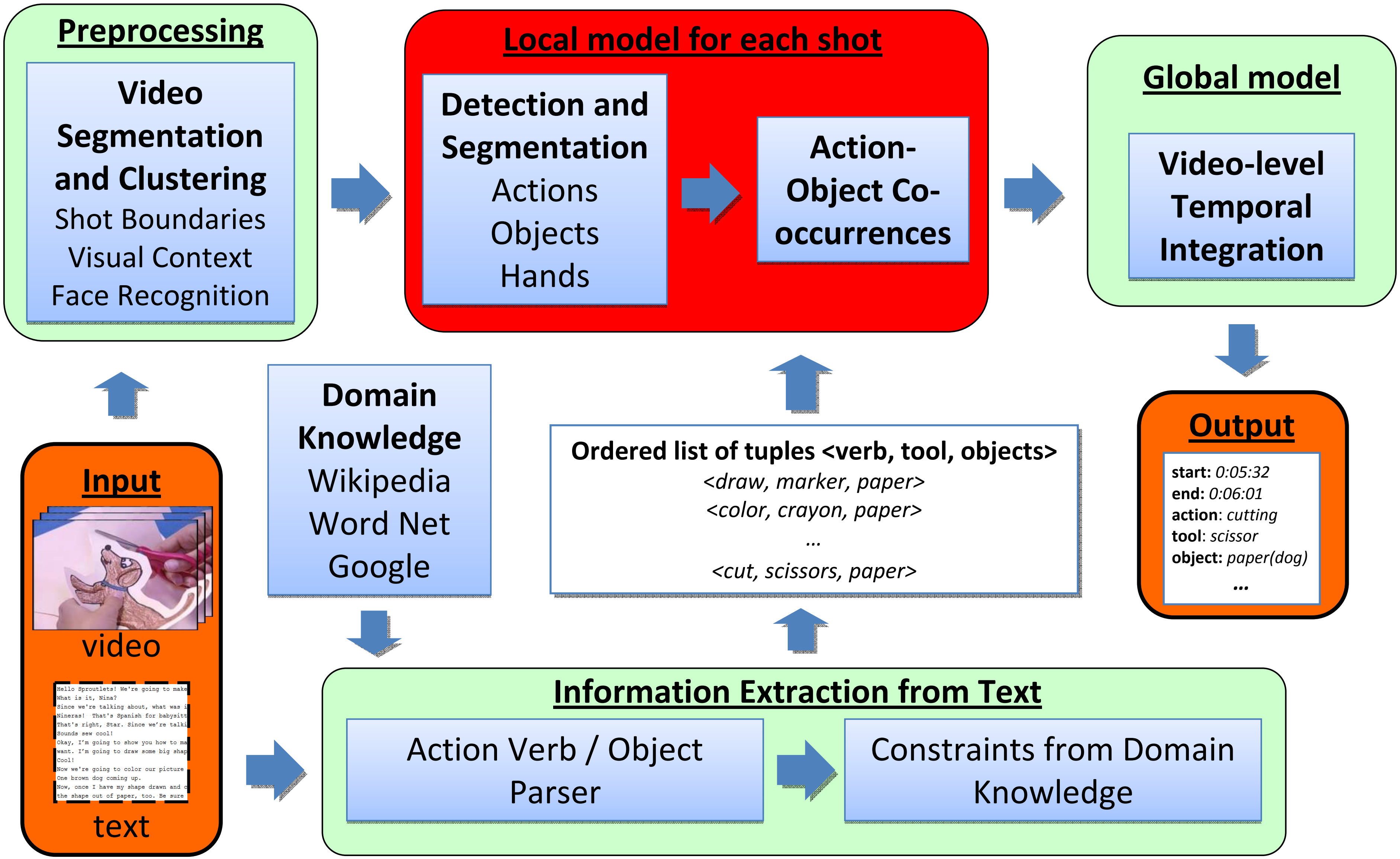
- **Hand pose features**
  - Histogram of 128 hand pose words



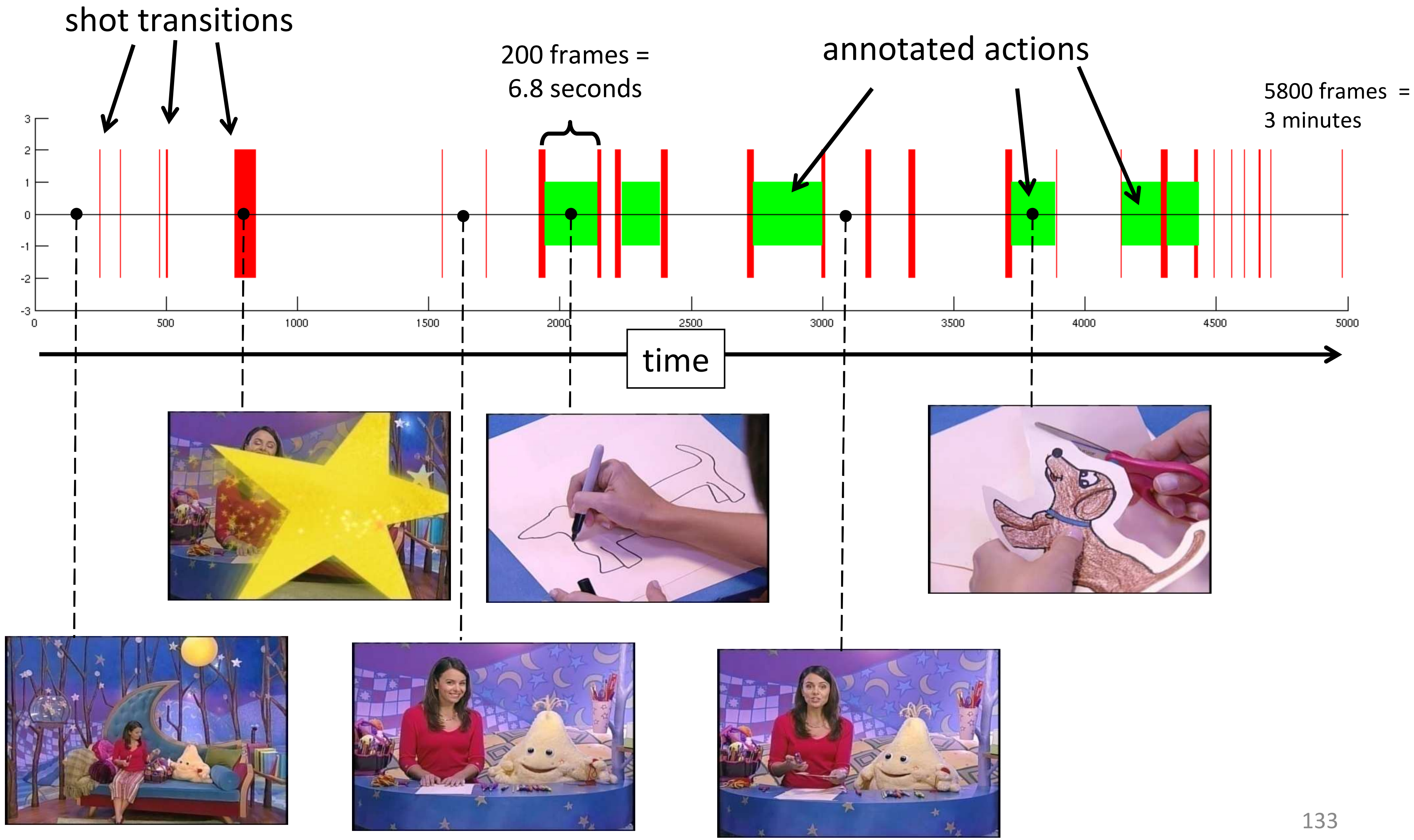
- **Global motion features**
  - Histogram of 100 STIP words



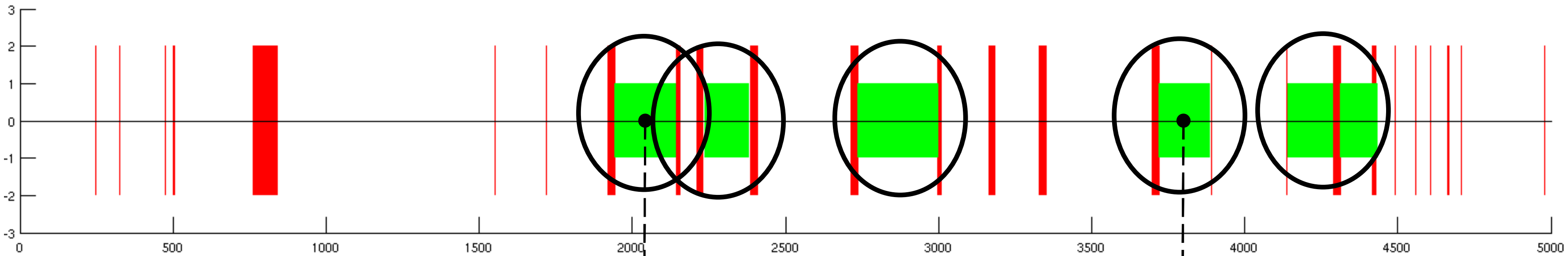
Joint models for Actions, Objects and Text

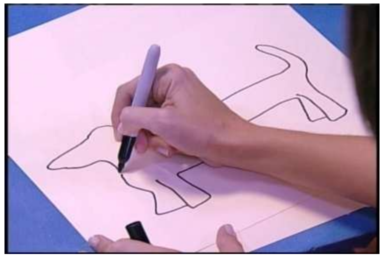


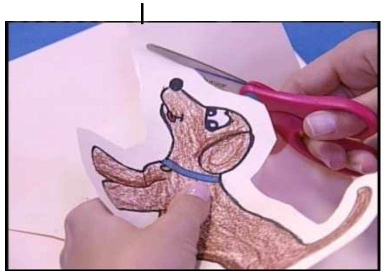
Episode timeline: *"Babysitter's Animal Sewing Cards",* PBS Sprout TV



This talk: *Multi-class action and tool classification*





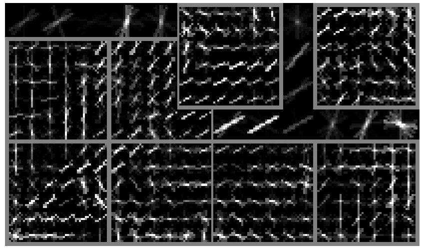


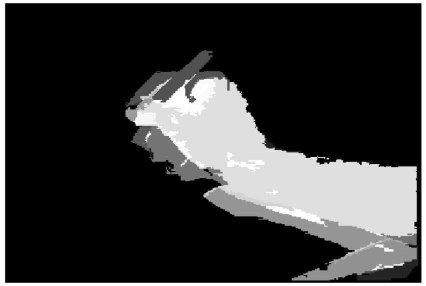
cut?
draw?
glue?
paint?
color?
other?

scissors?
marker?
glue bottle?
paint brush?
none?

Visual Features Recap

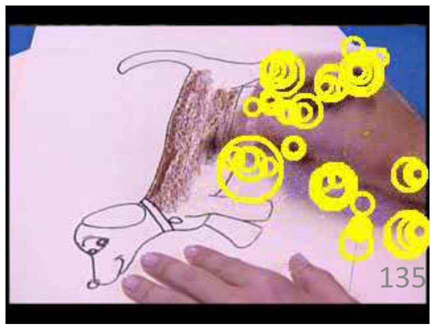
- **Tool detection features**
  - Histogram of object detector scores
  - 4 tool detectors (*writing tool, scissors, glue bottle, paint brush)*
  - 10 bins



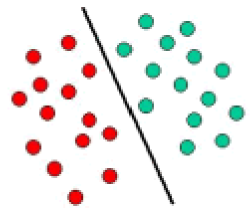


- **Hand pose features**
  - Histogram of 128 hand pose words

- **Global motion features**
  - Histogram of 100 STIP words



Multi-class action classification



- L_2-regularized, multi-class, logistic regression
  - liblinear matlab library (http://www.csie.ntu.edu.tw/~cjlin/liblinear/
  - found to work better than SVM (linear or kernelized)
  - 10-fold cross validation to select C (regularization tradeoff)

- Used 13/27 episodes for training, 14/27 for testing
  - Chosen to have an even distribution of actions across test/train split

- Accuracies reported are weighted by the frequency of each class
  - 10/20 class 1 and 100/100 class 2 --> report 75% accurate, not 91.67%

Multi-class action classification

class (# in class) > **normalized accuracy**	**cut** (18) **draw** (20)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5) **other** (50)
Hand Pose			
Tool Detectors			
STIP			
All combined			
Guess most frequent class			

Multi-class action classification

class (# in class) > **normalized accuracy**	**cut** (18) **draw** (20)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5) **other** (50)
Hand Pose	63.3		
Tool Detectors	91.7		
STIP	97.5		
All combined	97.5		
Guess most frequent class	*50.0*		

Multi-class action classification

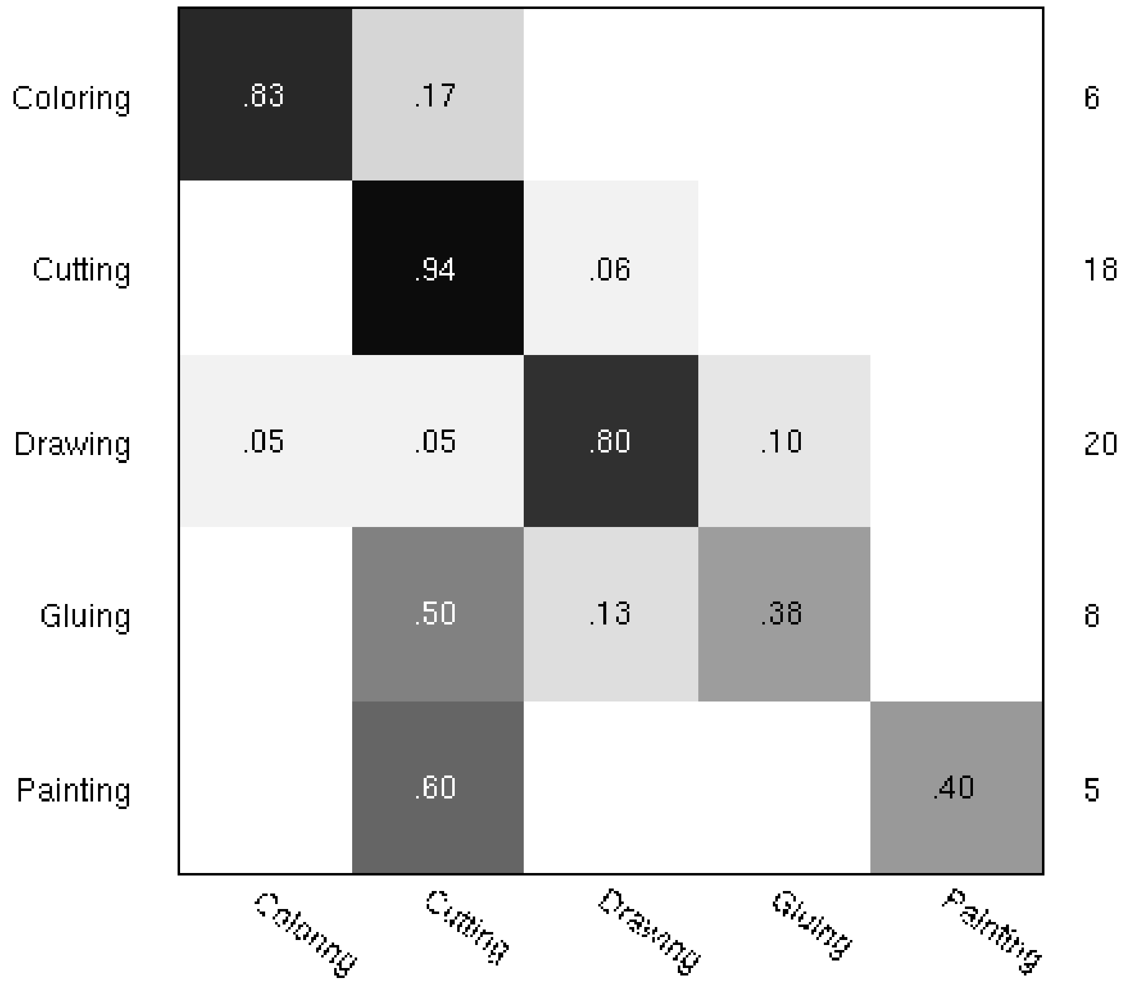
class (# in class) > **normalized accuracy**	**cut** (18) **draw** (20)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5) **other** (50)
Hand Pose	63.3	27.8	
Tool Detectors	91.7	42.9	
STIP	97.5	61.1	
All combined	97.5	67.1	
Guess most frequent class	*50.0*	*20.0*	

Multi-class action classification

class (# in class) > **normalized accuracy**	**cut** (18) **draw** (20)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5) **other** (50)
Hand Pose	63.3	27.8	20.5
Tool Detectors	91.7	42.9	37.1
STIP	97.5	61.1	42.1
All combined	97.5	67.1	47.0
Guess most frequent class	*50.0*	*20.0*	*16.7*

Multi-class action classification
5-class confusion matrix

normalized accuracy	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5) **other** (50)
All combined	67.1	47.0



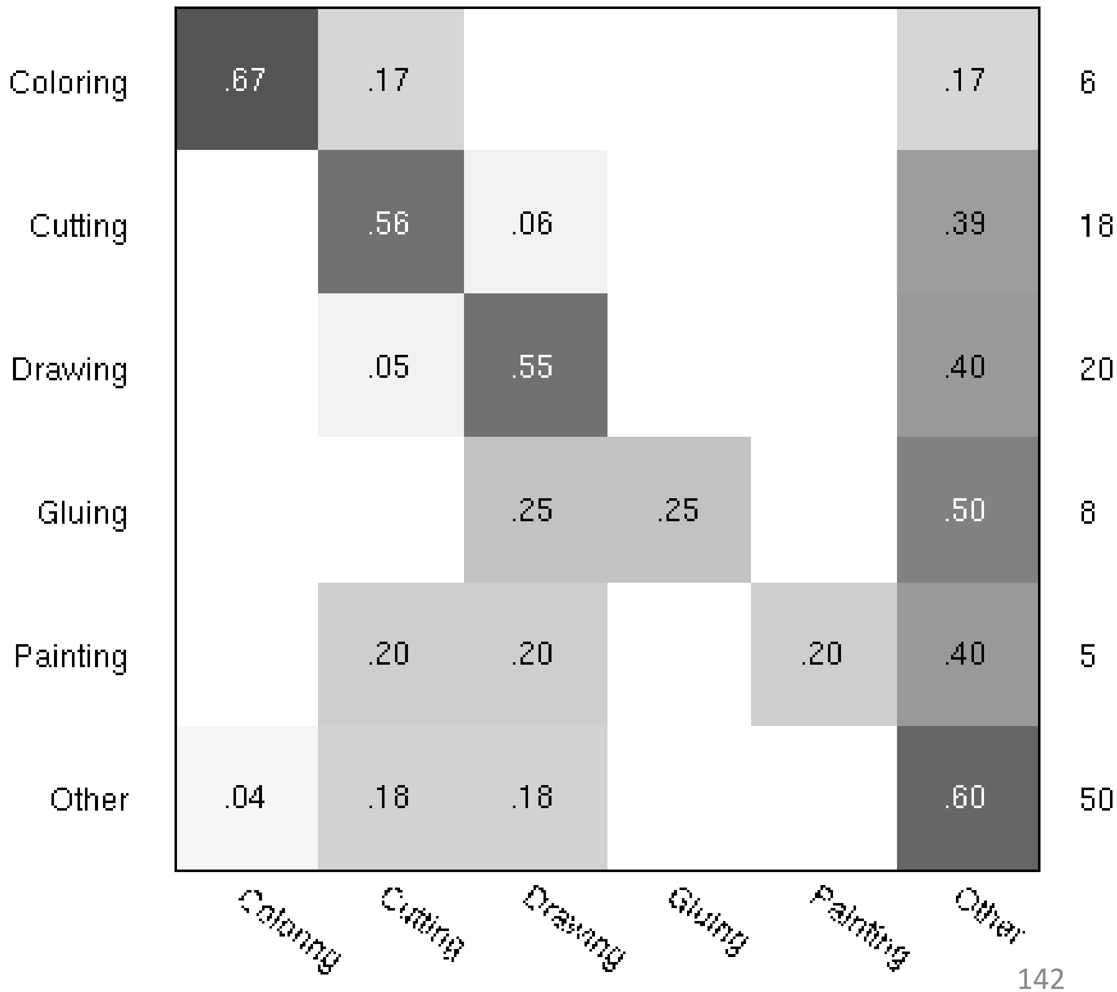
Multi-class action classification

5-class + “other” confusion matrix

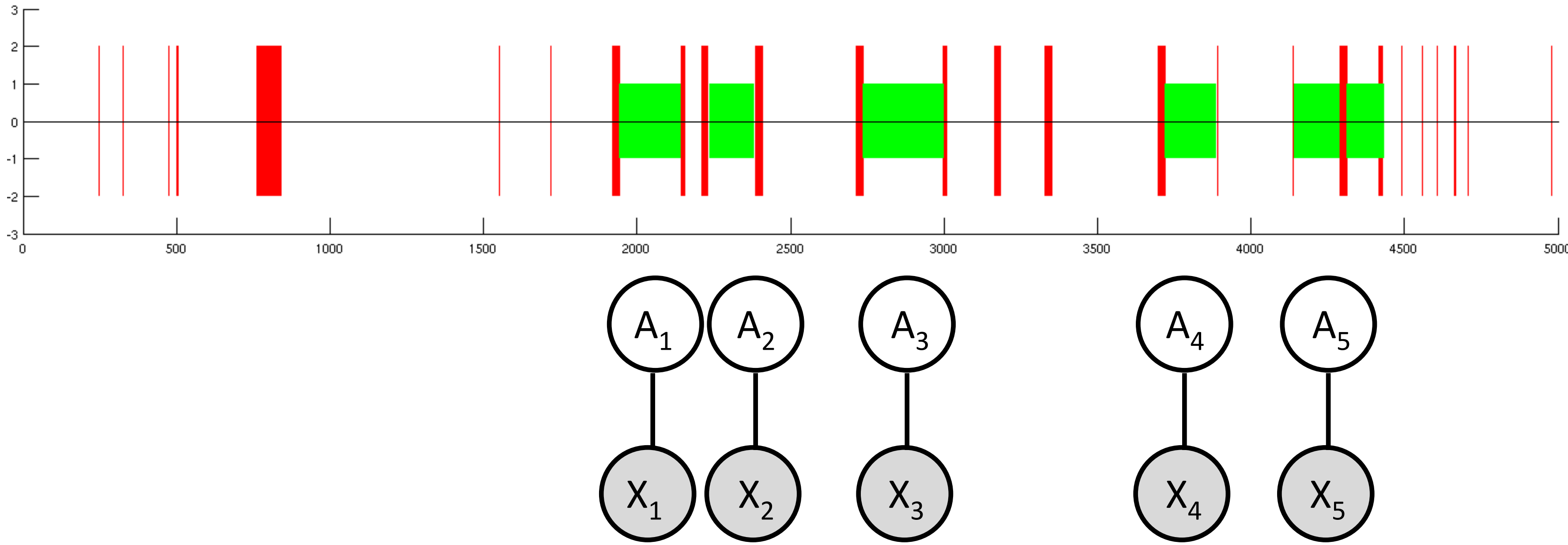
normalized accuracy	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5)	**color** (6) **cut** (18) **draw** (20) **glue** (8) **paint** (5) **other** (50)
All combined	67.1	47.0

- Heavy tail of misc. actions

- More training examples could help model more classes
- Using transcript text can narrow down the number of classes considered

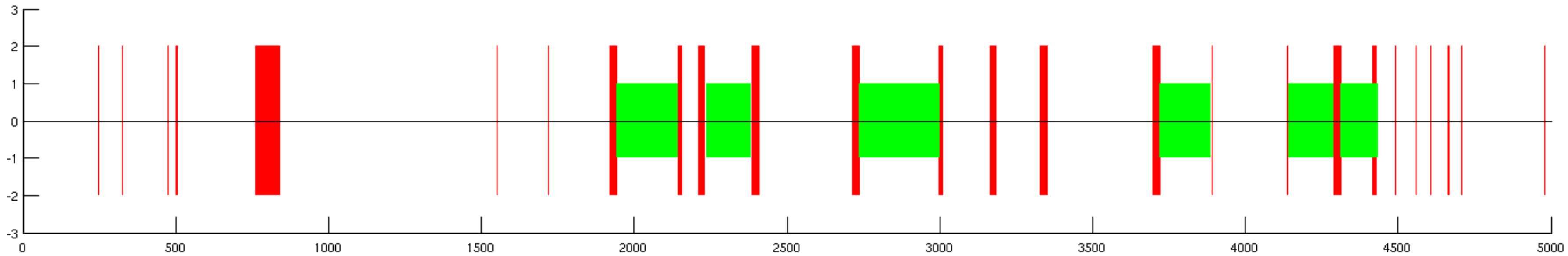


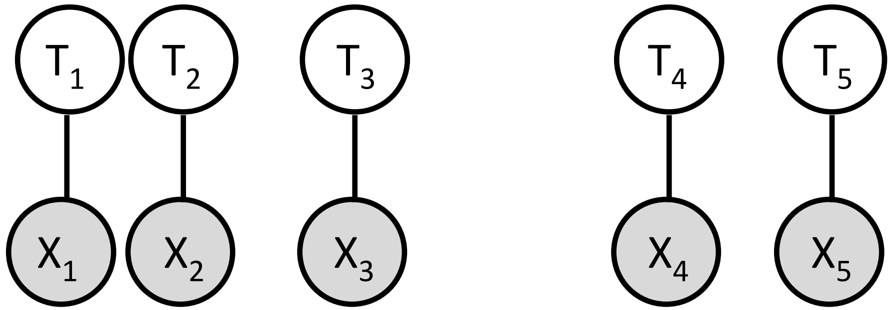
So far: Independent, multi-class action classification



action A_i ∈ { Cut, Draw, Color, Glue, Paint, Other }
data X_i = image features

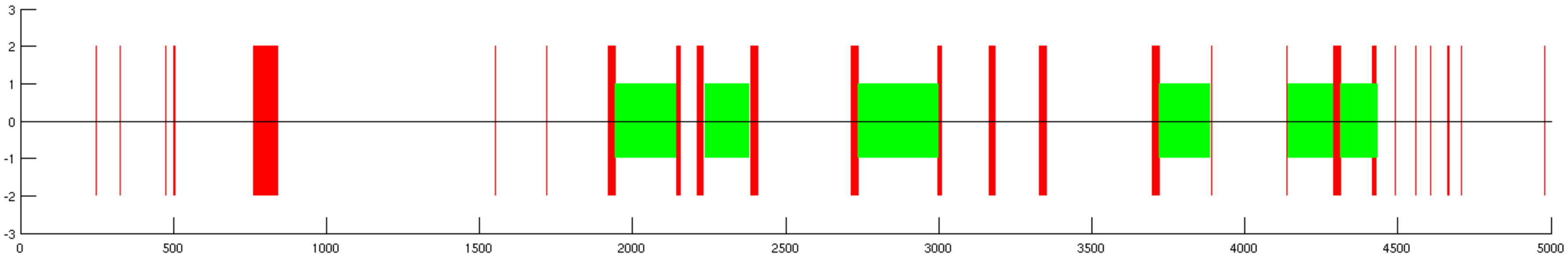
So far: Independent, multi-class tool classification

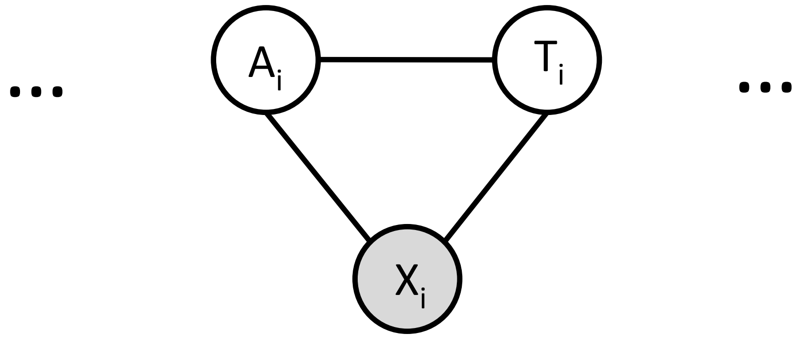




tool T_i $\in$ { Paint Brush, Glue Bottle, Writing Tool, Scissors, None}
data X_i = image features

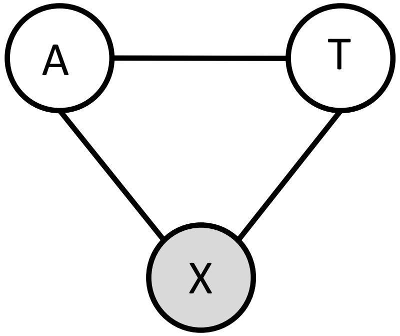
Modeling action-tool interaction





action A ∊ { Cut, Draw, Color, Glue, Paint, Other }
tool T ∊ { Paint Brush, Glue Bottle, Writing Tool, Scissors, None}
data X_i = image *and* text features

Modeling action-tool interaction: toy example



action A ϵ { Cut, Draw, Color, Glue, Paint, Other }
tool T ϵ { Paint Brush, Glue Bottle, Writing Tool, Scissors, None}

action / score	tool / score	scissors	**pencil**	*crayon*	br	
		10	20	50	10	
cut	10	80	0	0	20	0
draw	20	0	80	10	5	5
color	0	0	20	50	30	0
paint	45	0	10	10		
	25	0	0	0		

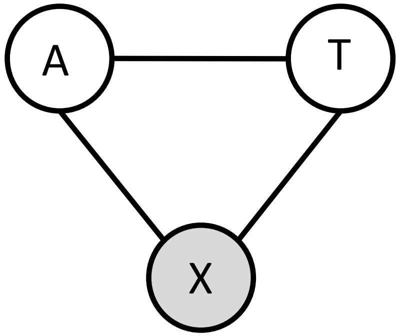
tool score **from vision**

action-tool co-occurences **from groundtruth or web**

action score **from vision**

$$a^*,t^* = \mathbf{argmax}_{a,t}\ \text{score}(a) + \text{score}(t) + \text{score}(a,t)$$

Modeling action-tool interaction



Conditional Random Field:

action A ϵ { Cut, Draw, Color, Glue, Paint, Other }
tool T ϵ { Paint Brush, Glue Bottle, Writing Tool, Scissors, None}
data X = image *and* text features

visual action features

visual tool features

action-tool
co-occurrence

$$p(A=a, T=t|x) \propto \exp\Big(w_A \cdot f_A(a,x) + w_T \cdot f_T(t,x) + w_{A,T} \cdot f_{A,T}(a,t) \Big)$$

learned, discriminative weights
(gradient descent on conditional likelihood with L_2 regularization)

MAP decision: $a^\star, t^\star = \arg \max_{a \in A, t \in T} p(A=a, T=t|x)$

Sources of action-tool co-occurrence

Dataset groundtruth

	color	cut	draw	glue	paint	place
brush	0	0	0	1	8	0
writing tool	12	0	42	0	0	0
glue	0	0	0	20	0	0
scissors	0	38	0	0	0	0

Domain knowledge from the web

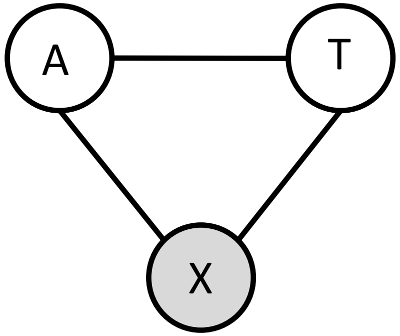
WIKIPEDIA
The Free Encyclopedia

	color	cut	draw	glue	paint	place
brush	0	0	1	0	1	0
writing tool	1	0	1	0	0	0
glue	0	0	0	1	0	0
scissors	0	1	0	0	0	0

Normalized Google Distance: Google

	color	cut	draw	glue	paint	place
brush	2.51	2.11	2.4	INF	1.85	INF
writing tool	2.12	3.51	1.72	INF	2.08	INF
glue	2.51	2.51	2.51	1.2	2.44	INF
scissors	2.47	1.76	2.36	INF	2.68	INF

Modeling action-tool interaction: Results

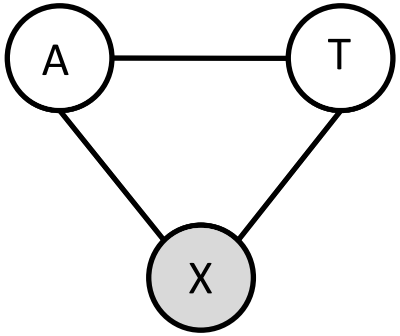


action A ϵ { Cut, Draw, Color, Glue, Paint, Other }
tool T ϵ { Paint Brush, Glue Bottle, Writing Tool, Scissors, None}

- Estimated action-tool domain knowledge obtained from Wikipedia and Normalized Google Distance (NGD)

	no joint modeling	groundtruth action-tool co-occurrence	domain knowledge co-occurrence from the web
action & tool both correct			
action			
tool			

Modeling action-tool interaction: Results



action A ϵ { Cut, Draw, Color, Glue, Paint, Other }
tool T ϵ { Paint Brush, Glue Bottle, Writing Tool, Scissors, None}

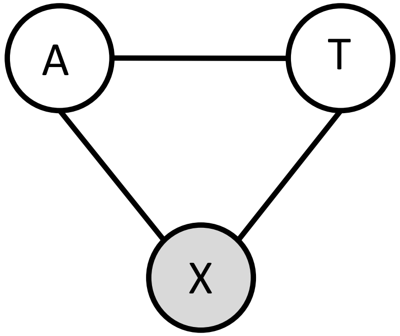
• Estimated action-tool domain knowledge obtained from Wikipedia and Normalized Google Distance (NGD)

Google

WIKIPEDIA
The Free Encyclopedia

	no joint modeling	groundtruth action-tool co-occurrence	domain knowledge co-occurrence from the web
action & tool both correct	28.0		
action	50.9		
tool	44.9		

Modeling action-tool interaction: Results

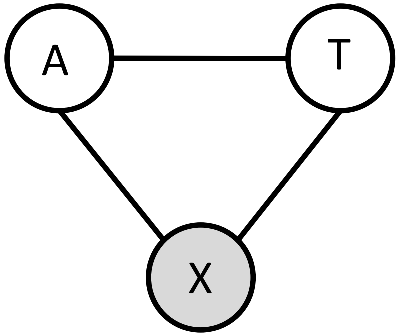


action A ϵ { Cut, Draw, Color, Glue, Paint, Other }
tool T ϵ { Paint Brush, Glue Bottle, Writing Tool, Scissors, None}

- Estimated action-tool domain knowledge obtained from Wikipedia and Normalized Google Distance (NGD)

	no joint modeling	groundtruth action-tool co-occurrence	domain knowledge co-occurrence from the web
action & tool both correct	28.0	40.7	
action	50.9	50.8	
tool	44.9	46.7	

Modeling action-tool interaction: Results



action A ϵ { Cut, Draw, Color, Glue, Paint, Other }
tool T ϵ { Paint Brush, Glue Bottle, Writing Tool, Scissors, None}

- Estimated action-tool domain knowledge obtained from Wikipedia and Normalized Google Distance (NGD)

Google

WIKIPEDIA
The Free Encyclopedia

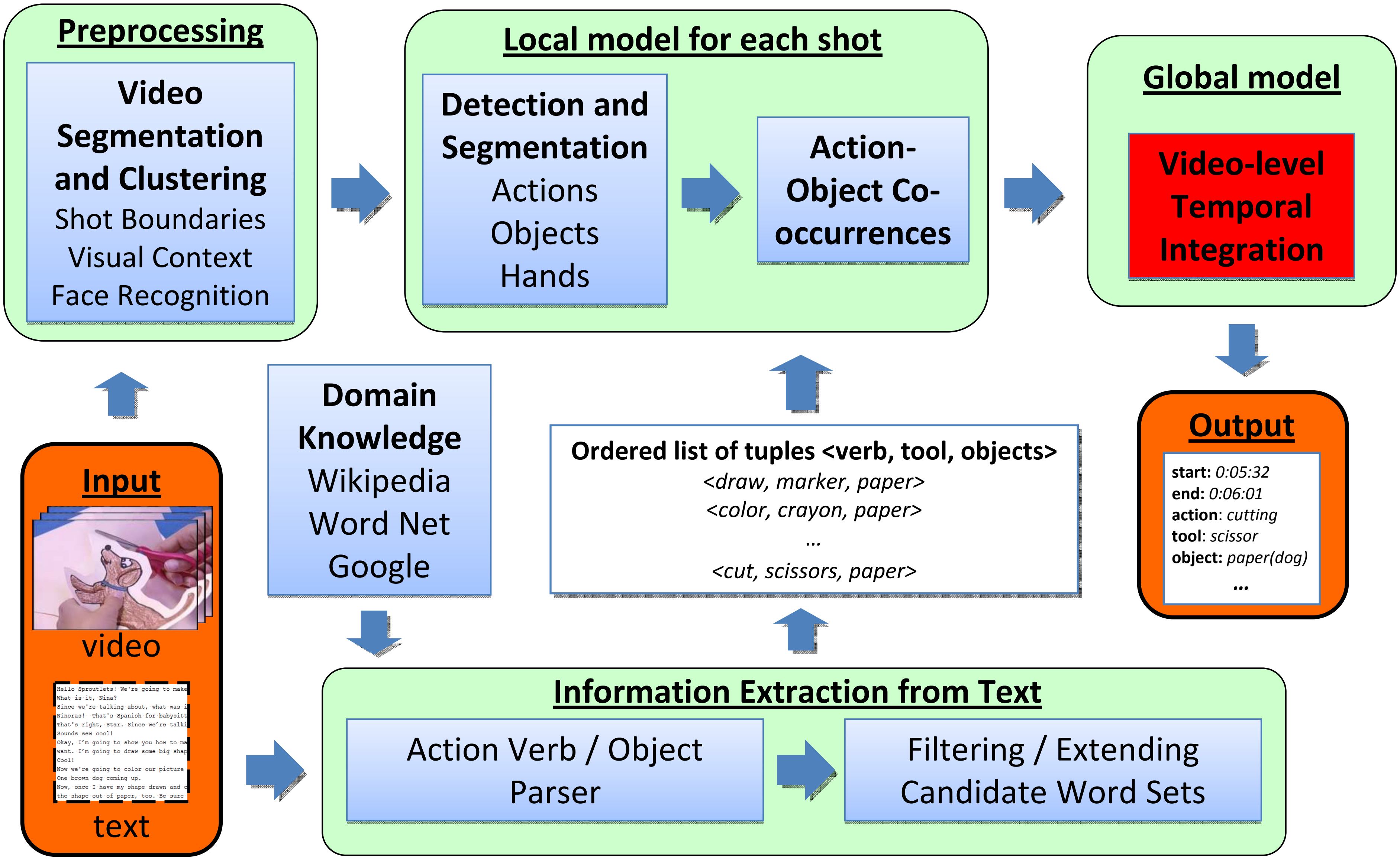
	no joint modeling	groundtruth action-tool co-occurrence	domain knowledge co-occurrence from the web
action & tool both correct	28.0	40.7	37.8
action	50.9	50.8	50.8
tool	44.9	46.7	48.3

Summary

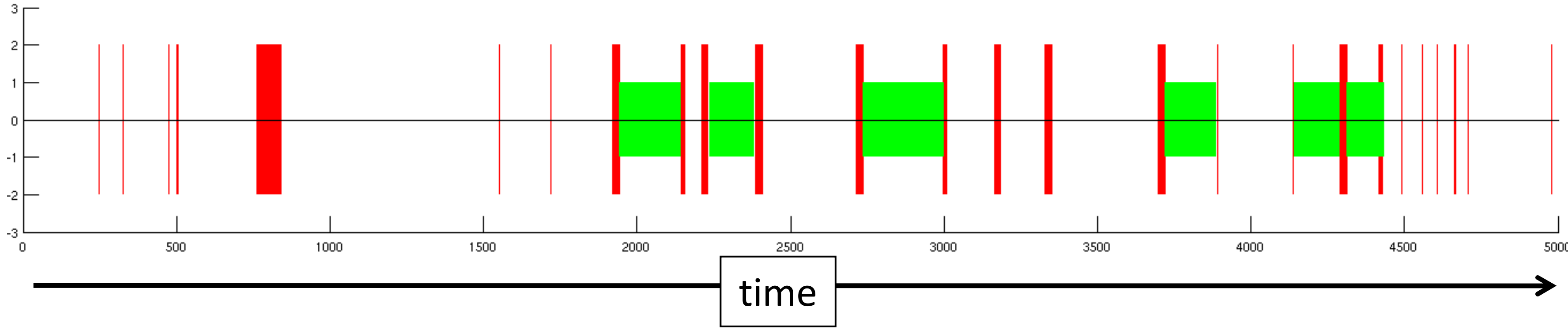
Joint models for actions, objects and text

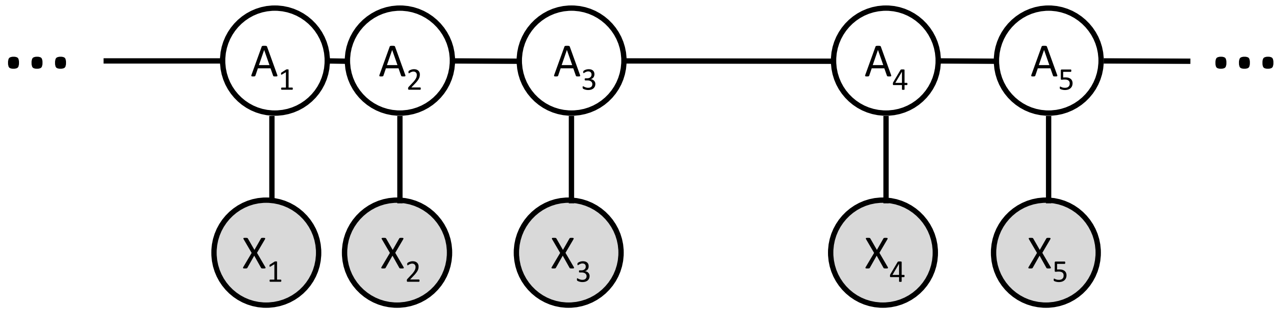
- We can improve upon standard action-recognition techniques (STIP) by modeling tool presence and hand pose
- Explicitly modeling the interactions between tools and actions improves performance
- Can leverage domain knowledge from the web as a substitute for labeled data

Temporal Constraints

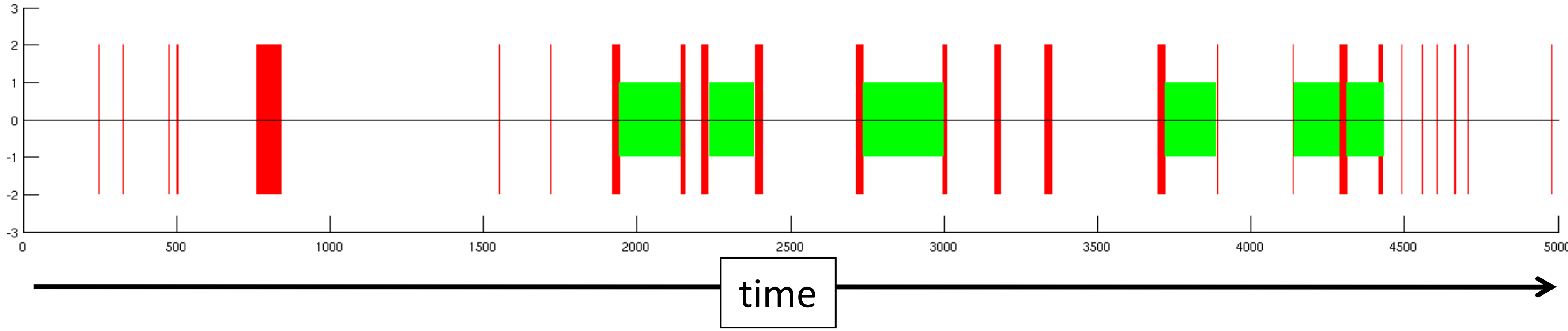


Incorporate temporal action ordering from text+vision

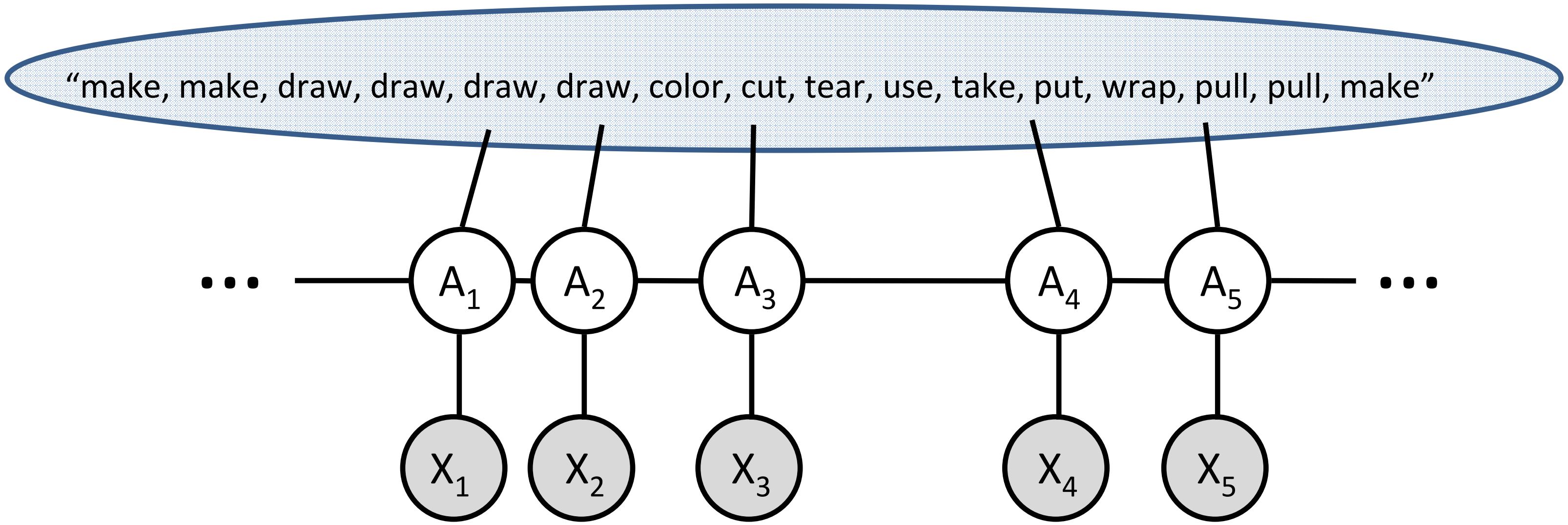




Incorporate temporal action ordering from text+vision



actions (in order) extracted from the transcript:



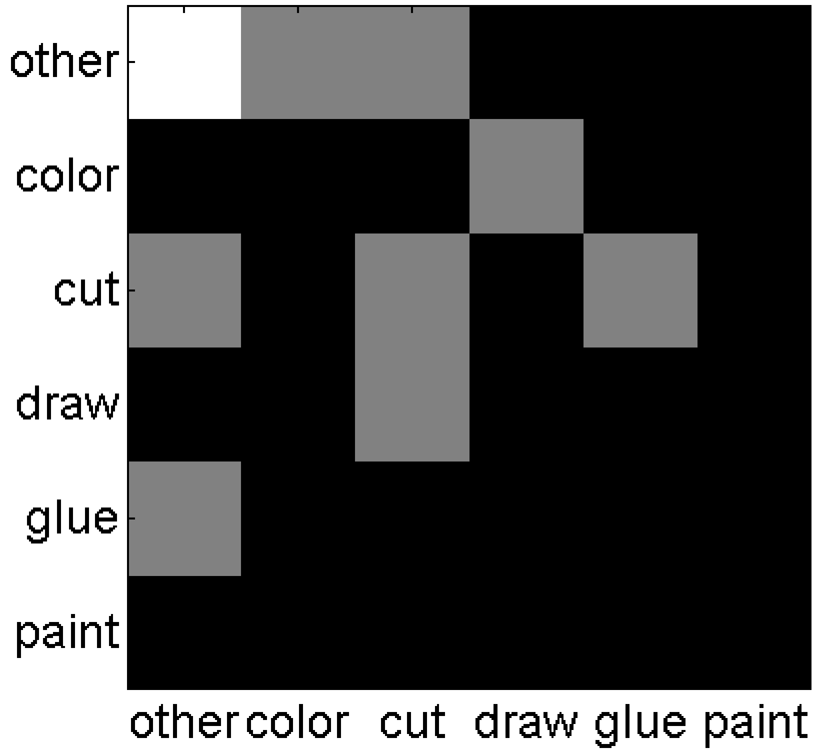
Verbs in Transcripts vs. Action Annotations

- Idea: the bigram of verb in the text may imply the partial order of actions in videos
  – If there is a verb bigram (v, w) in the text, the chance to find a corresponding video shot pair in the video sequence should be higher
- Verb bigram example:
  – Transcripts:
    make make draw draw draw draw color cut tear use take put wrap pull pull make
  – Action annotations:
    color cut draw thread thread wrap

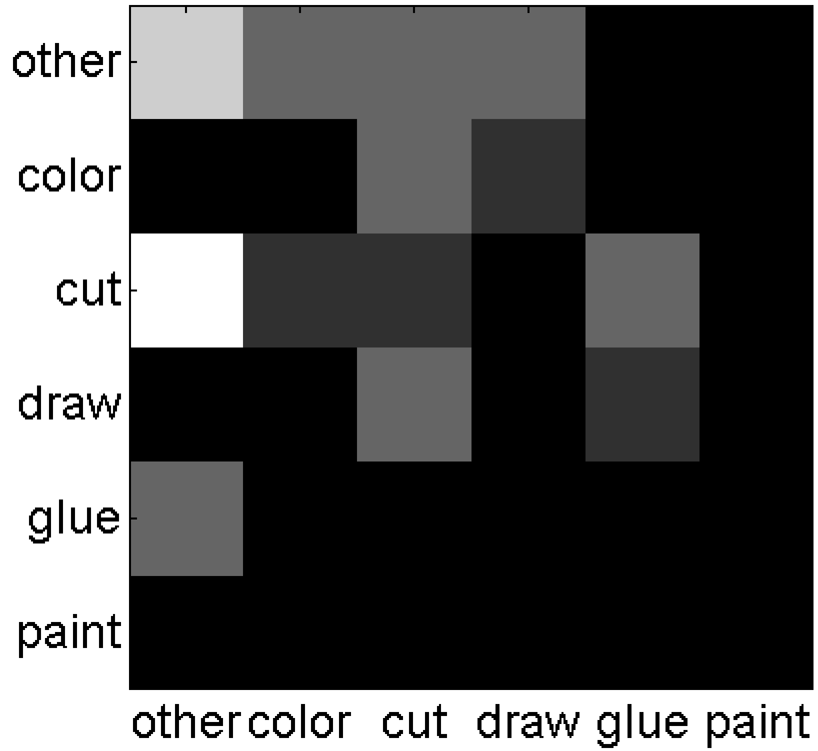
Verbs in Transcripts vs. Action Annotations

- Idea: the bigram of verb in the text may imply the partial order of actions in videos
  - Since the text and video are not strictly aligned, we further relax the bigram to incorporate verb pairs across up to two positions
- Relaxed verb bigram example:
  - Transcripts :
    use show cut tear cut make flatten take write
  - Action annotations:
    cut cut cut cut draw draw place place place

Sample Distributions of Verb Bigrams in Transcript

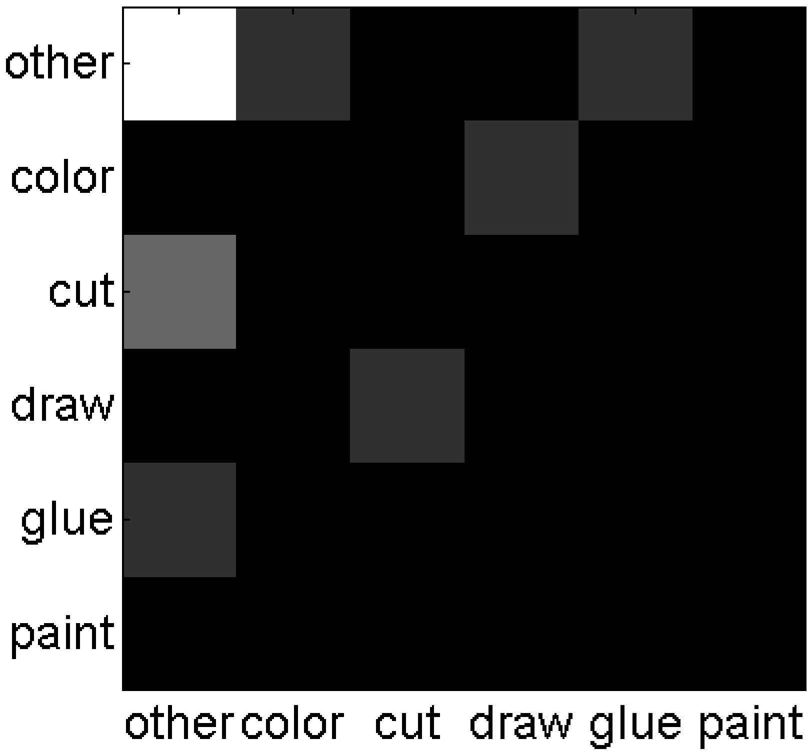


Bigram

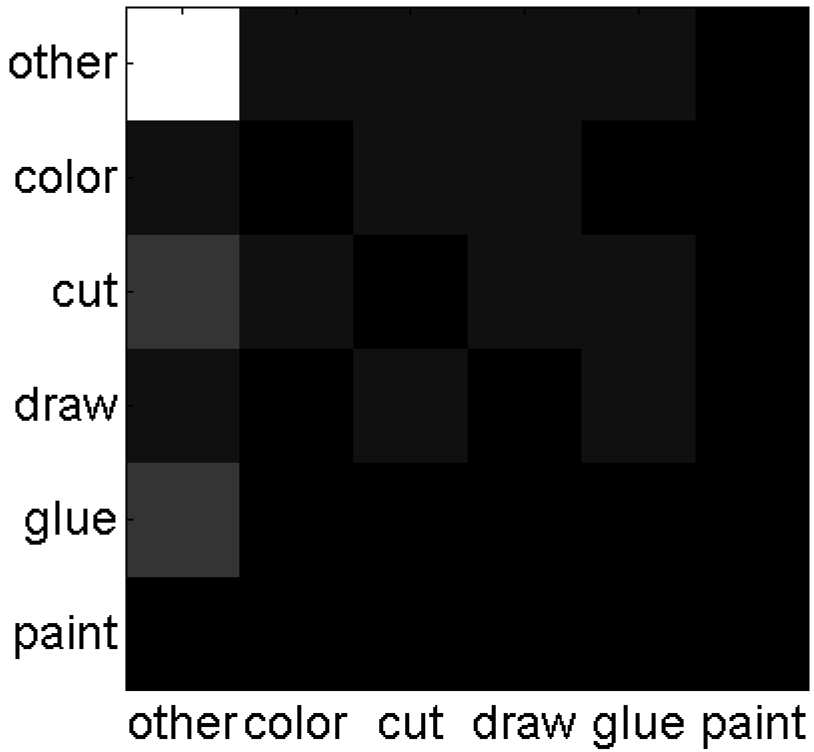


Relaxed bigram

Sample Distributions of Verb Bigrams in Online Instruction

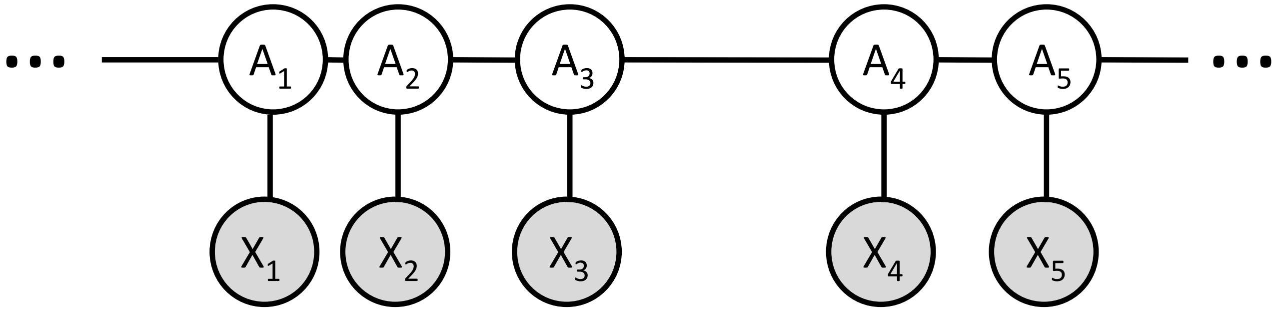


Bigram



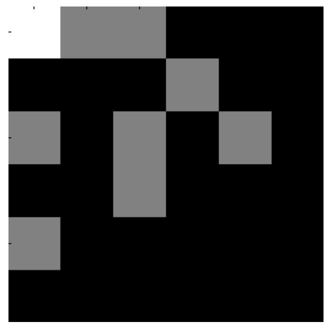
Relaxed bigram

Chain CRF Model



Node = single shot

Node Potential = score of action classification in single shot

Edge potential = exp(λ )

Results

Single Shot Action Recognition using STIP (SVM)	0.42
previous + Tool + Hand Feature	0.47
Single Shot Joint CRF Model (STIP+Tool+co-occurrence of verb and tool)	0.51
Sequence Model CRF with temporal constraints - extracted from transcripts (bigram)	0.52
Previous with relaxed bigram	0.52
Sequence Model CRF with temporal constraints extracted from online instructions (bigram)	0.53
Previous relaxed bigram	0.53

Summary

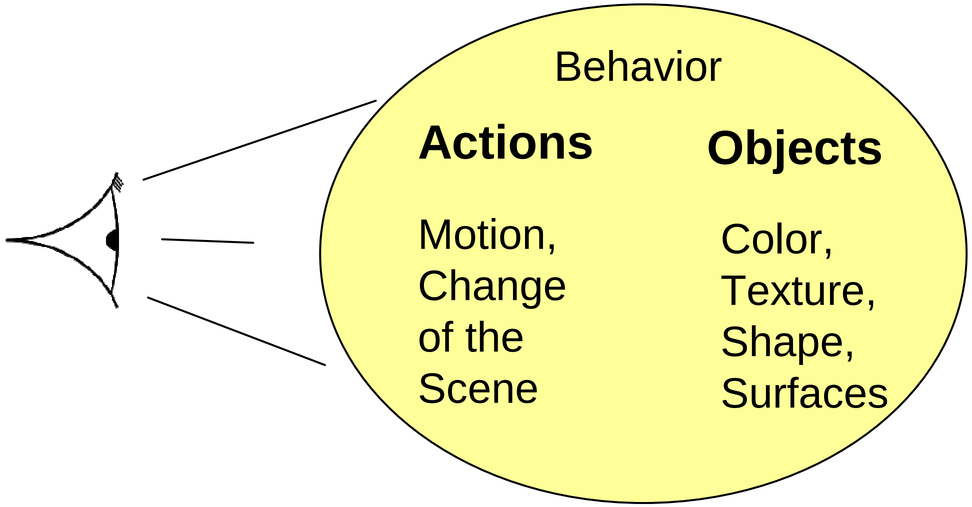
- The order of verbs in transcripts or instructions can be used as temporal constraints to the actions in videos
- The co-occurrence of verbs and tools can be used as semantic constraints to the actions in videos
- Both types of knowledge can be obtained either from transcripts or online using nature language processing techniques

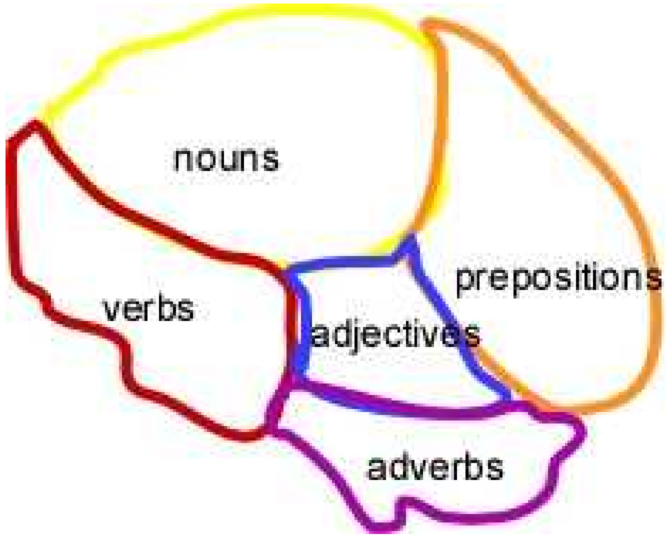
Attribute based object recognition

Ching Lik Teo, Yi Li, Cornelia Fermuller

Visual Space

Language Space





Parts of speech

Attributes of actions and objects

- Objects: **adjectives** (color, texture, shape)

  part descriptions (scissor blades, handle)
- Actions: **adverbs**

  decomposition into sub-actions

  (grasp the scissors, cut, put down the scissors)

  movements of body parts
- Objects and actions: **prepositions**

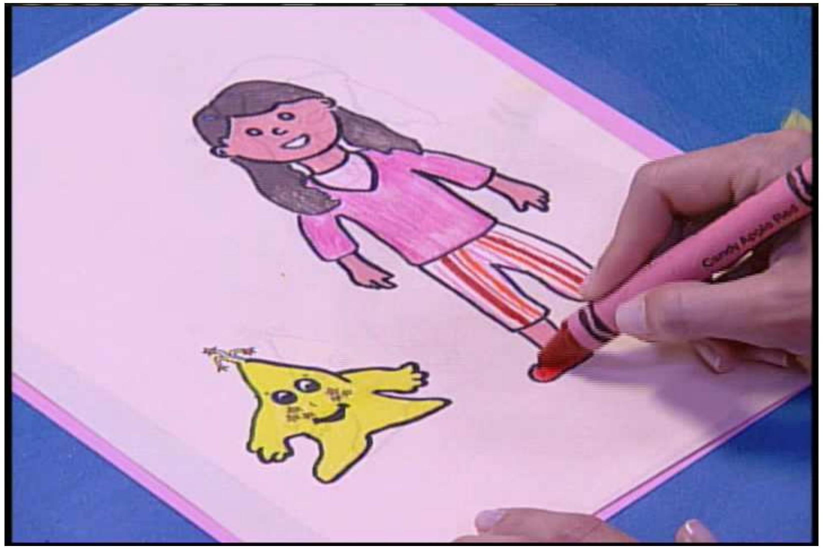
  temporal relationship (before, after)

  spatial relationship (on top, left, right, in between)

Segmentation for Manipulation

Attention based approach

Hand → manipulates a tool → touches an object

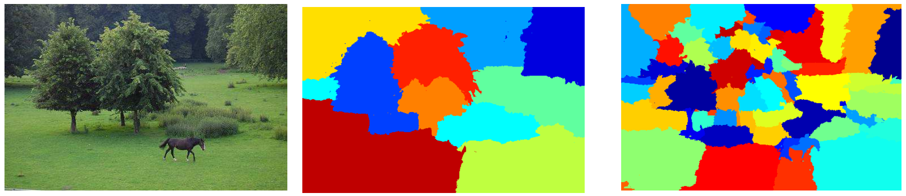


Hand draws with **crayon** on **paper**

Prerequisite: Segmentation

- Textbook definition:

  Division of the image into regions that have some *homogeneous* property?



How many regions?

Literature

Multi-label: Normalized Cut (Shi, Malik 2000), Mean Shift Clustering (Comaniciu Meer 2002), Graph cuts

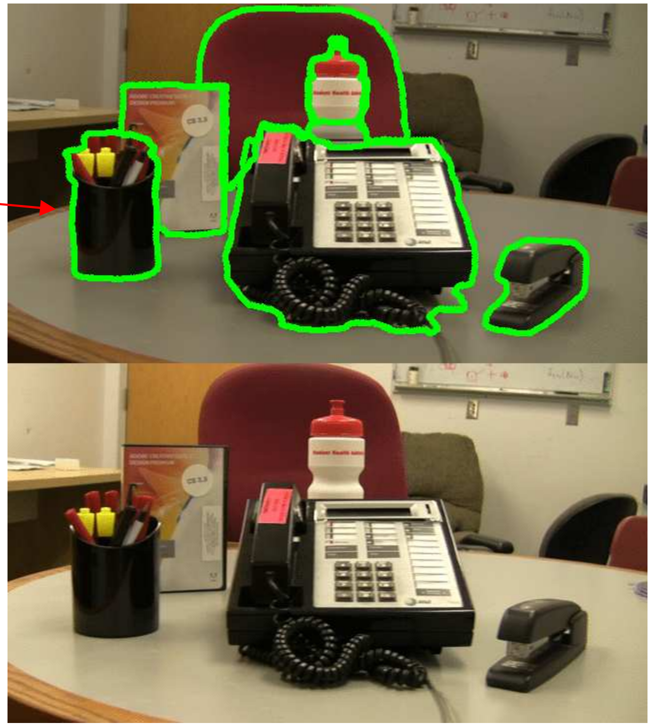
Two-label: Variational Minimization (Mumford Shah), Active contours (Kass, Witkin, Terzopoulos, 1988), Level Set methods (Tsai, Osher 2003),

Motion segmentation : 2D motion homogeneity, 3D rigid motion (Vidal, Tron, Hartley 2008)

Our definition of segmentation

Object - background segmentation: division into two regions, with the object region **bounded by a closed contour**, that contains some depth boundaries.

Depth boundary



Three ideas

1. Hand segmentation based on color, edges and motion

2. Fixation based object segmentation

   based on contours

3. Attention mechanism : object filters

Hand Segmentation CRF model

- Energy Terms:

$$E(\mathbf{x}) = \sum_{i \in v} \psi_i^{gmm}(x_i) + \sum_{(i,j) \in \varepsilon} \left(\psi_{ij}^{edge}(x_i, x_j) + \psi_{ij}^{col}(x_i, x_j) + \psi_{ij}^{flow}(x_i, x_j) \right)$$

edges

color edges

flow

Unary potential (color, learned)

Pairwise Potentials

Learning the GMM

Training data in L*a*b space

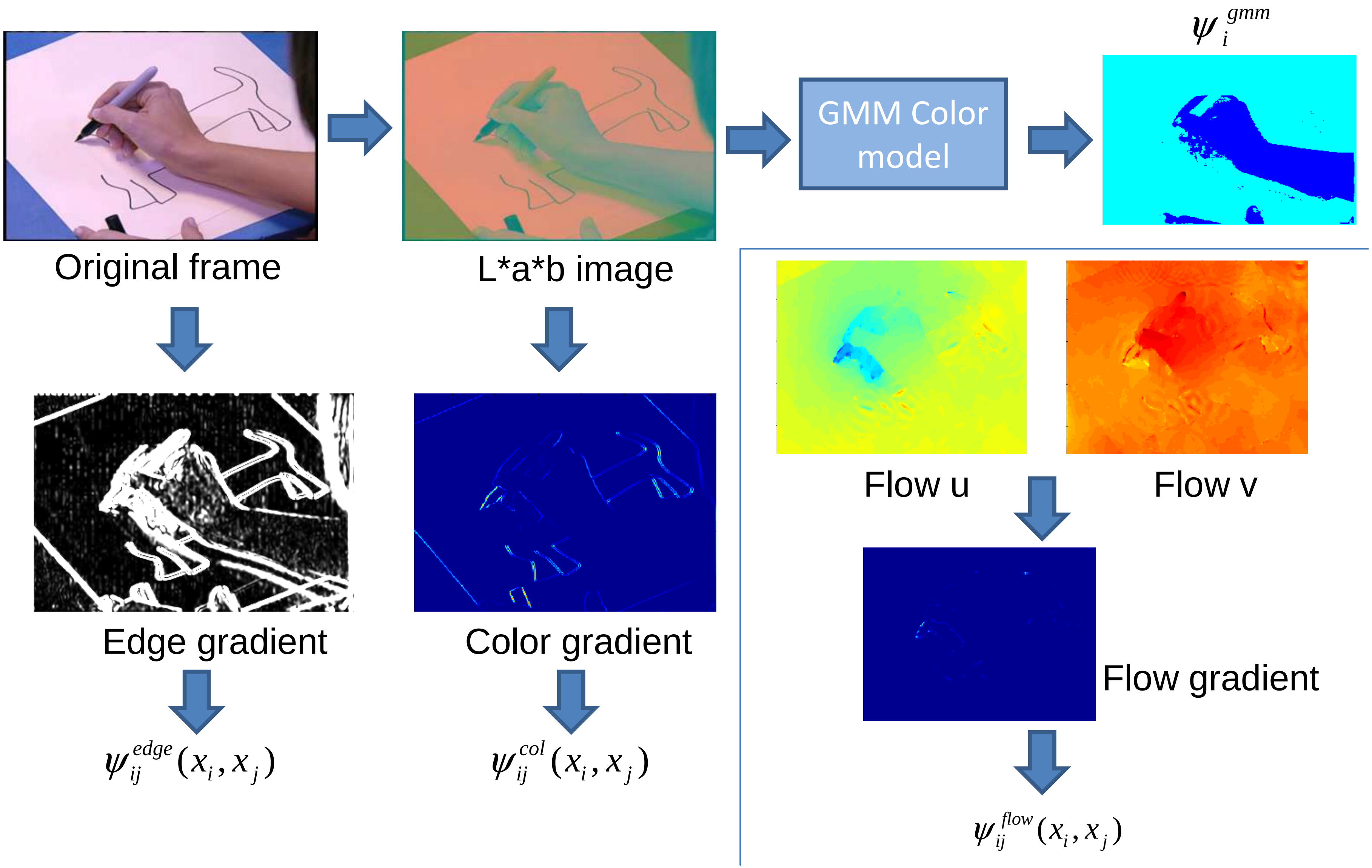
2 classes: Foreground/Background

Cluster pixel colors into *k* clusters

$$P(x \mid k) = N(x \mid \mu_k, \Sigma_k)$$

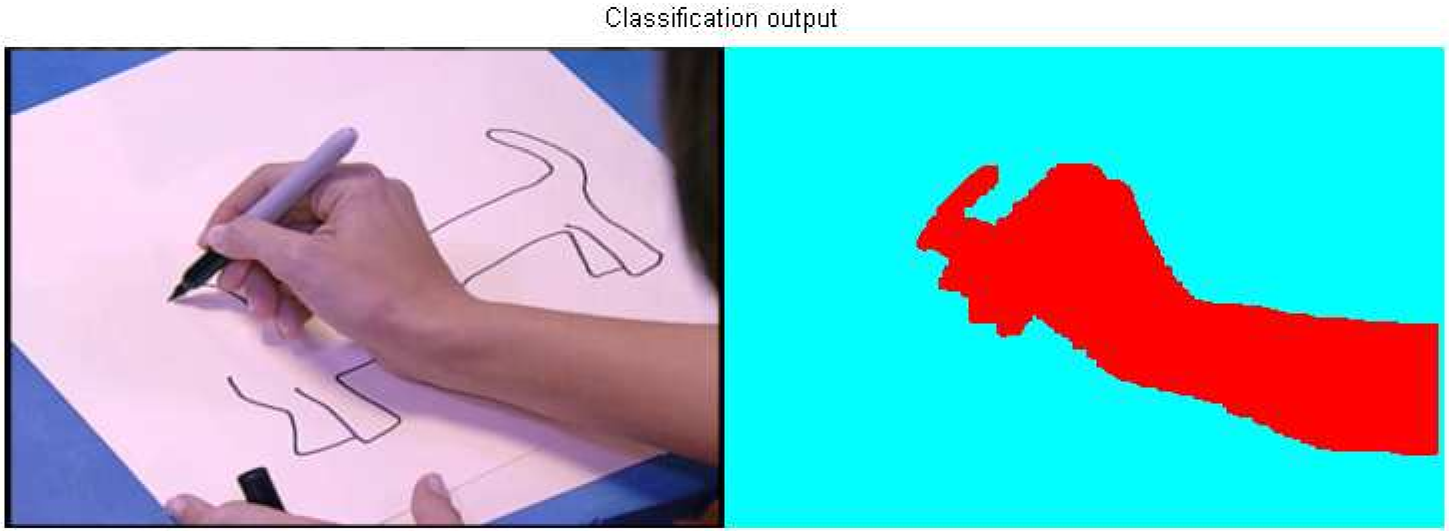
GMM Color model

Computing the potentials

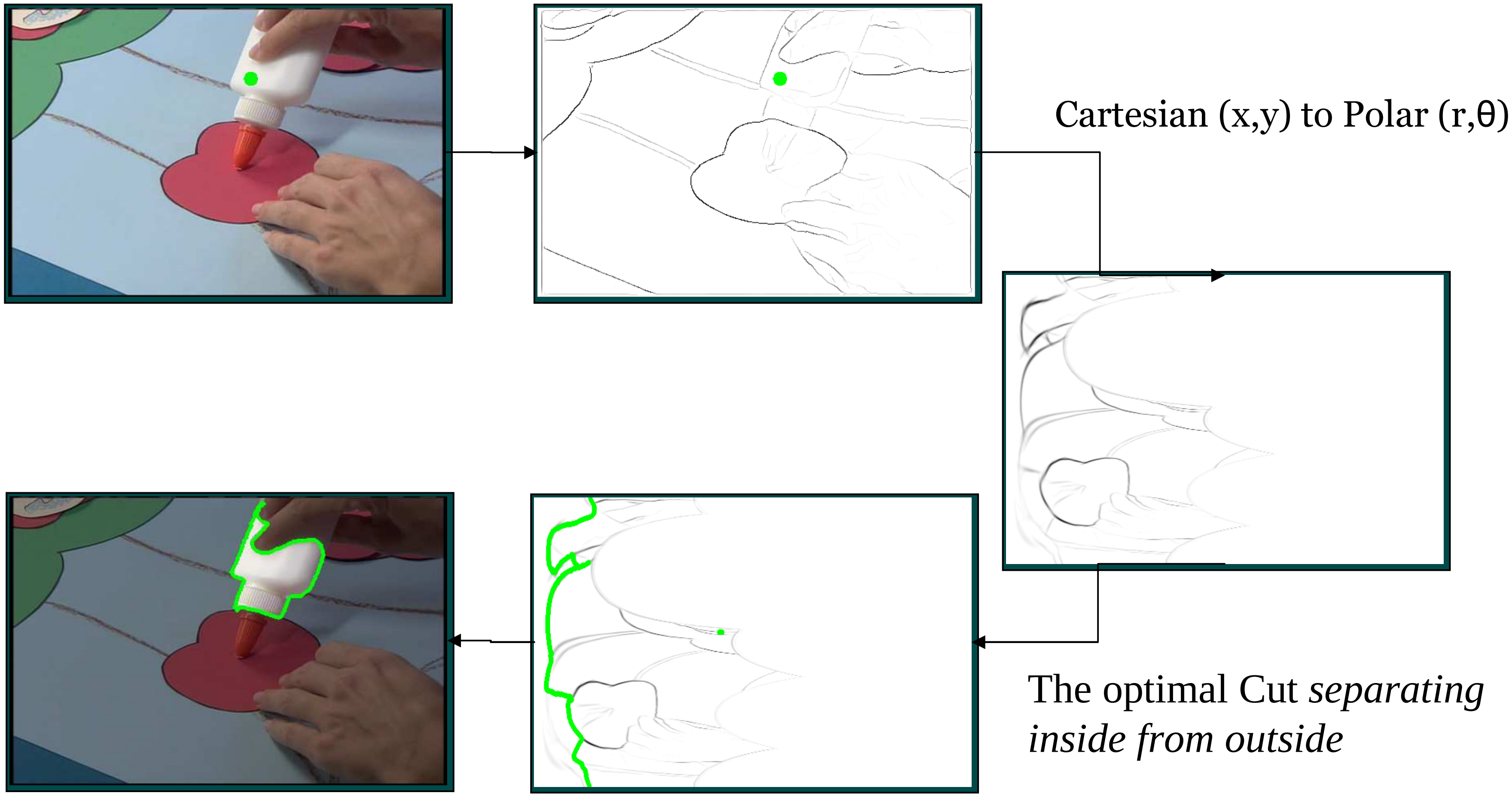


Inference

- MAP estimate using Graph-Cuts: $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbf{C}} E(\mathbf{x})$

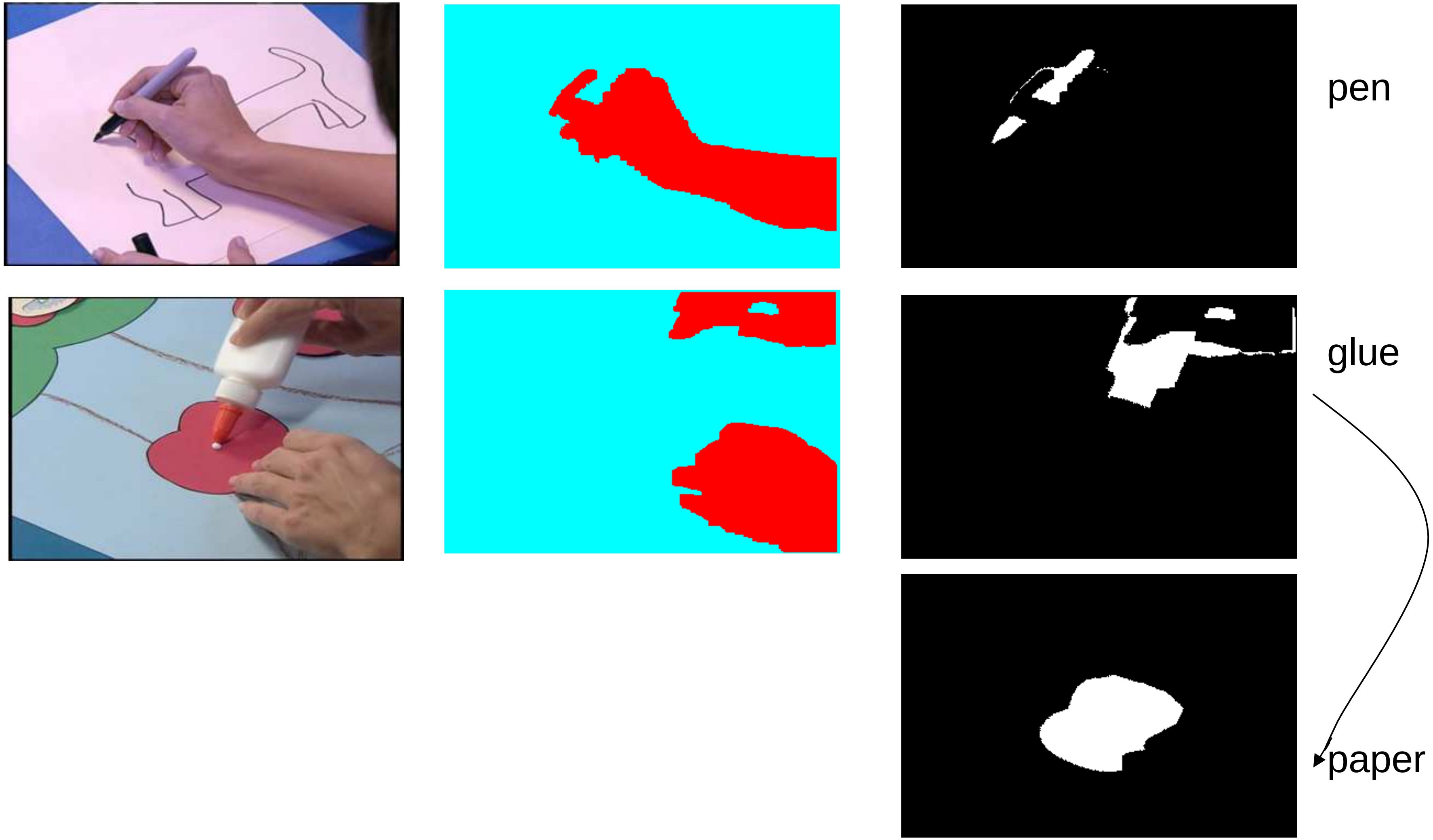


Fixation-based Algorithm



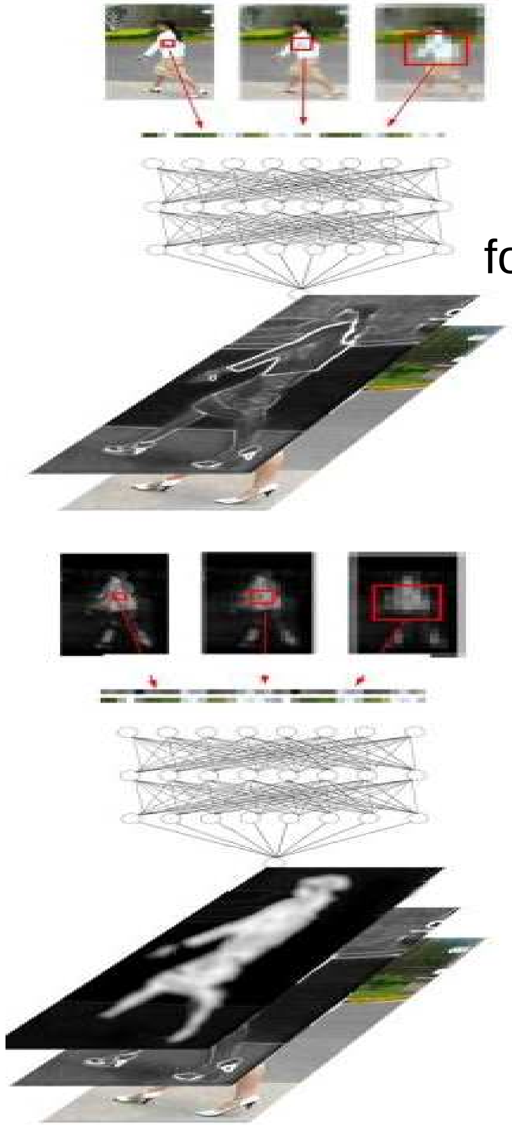
(Mishra et al, ICCV'09)

Examples



Object filters

(related to deep learning)



Algorithm:

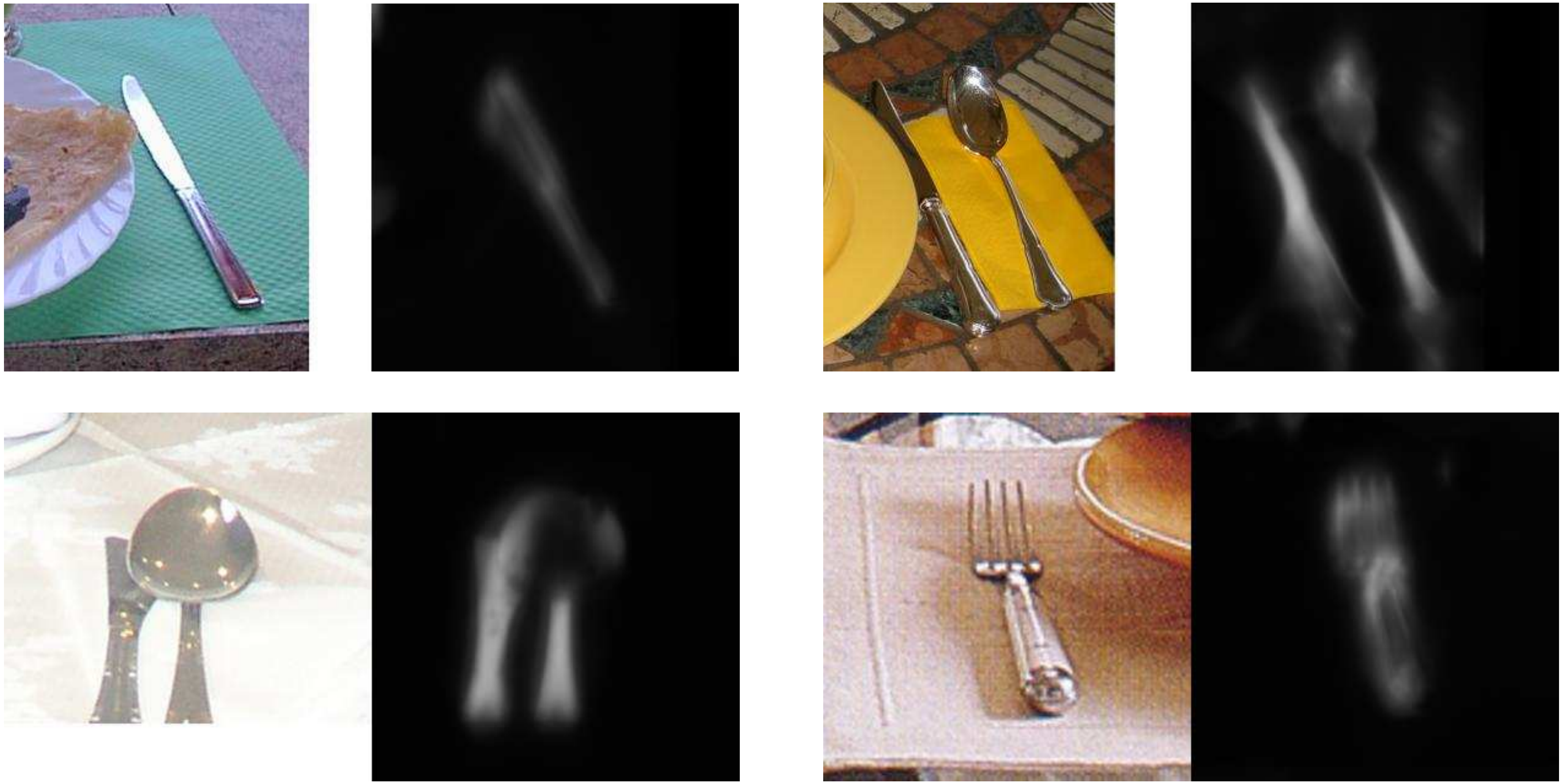
for i = 1.. 5

1. Compress using PCA.
2. Collect multi-scale patches.
3. Train a multilayer perceptron classifier.
4. Run the classifier on the images.
5. Return to step one an train a new classifier, but this time collect samples that include data from the original and the results of step 4.

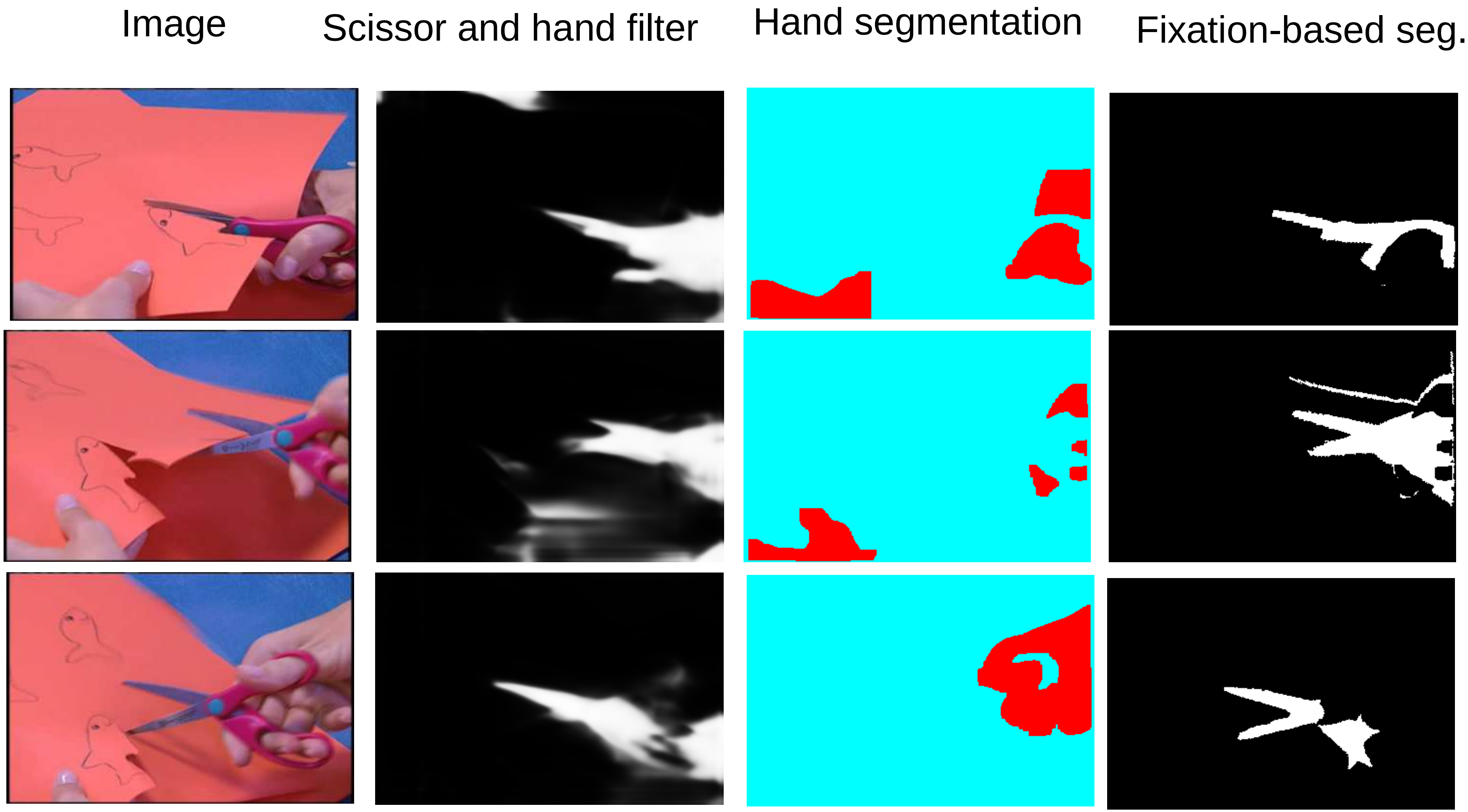
end for

(Human filter: Summerstay, Aloimonos 2010)

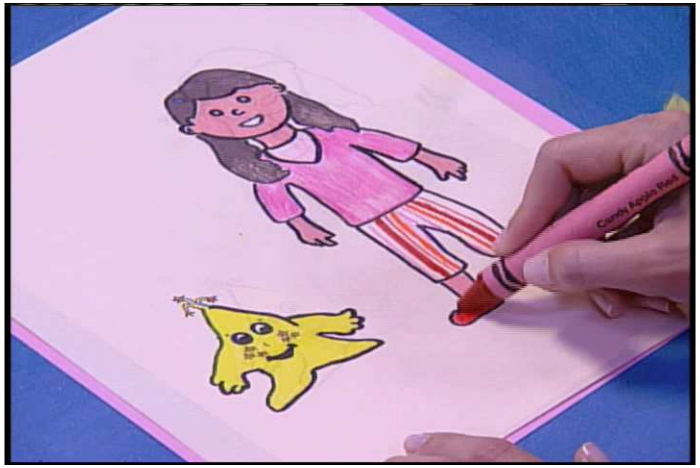
Silverware filters

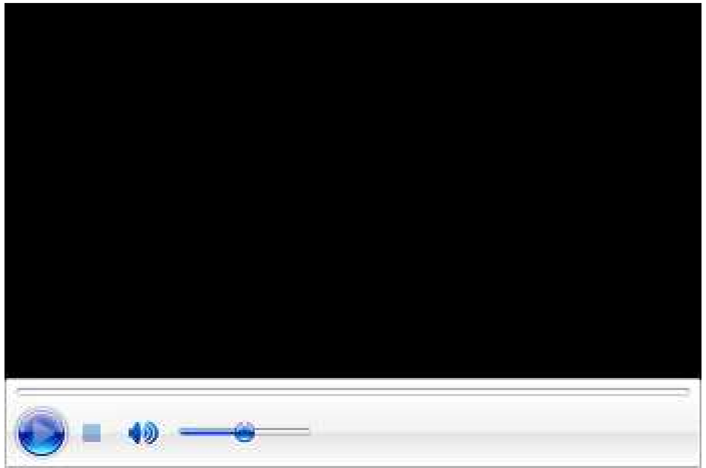


Scissors filter



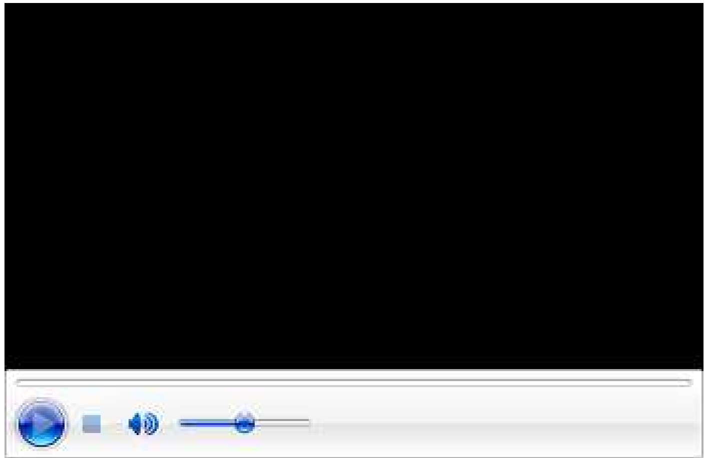
Segmentation results





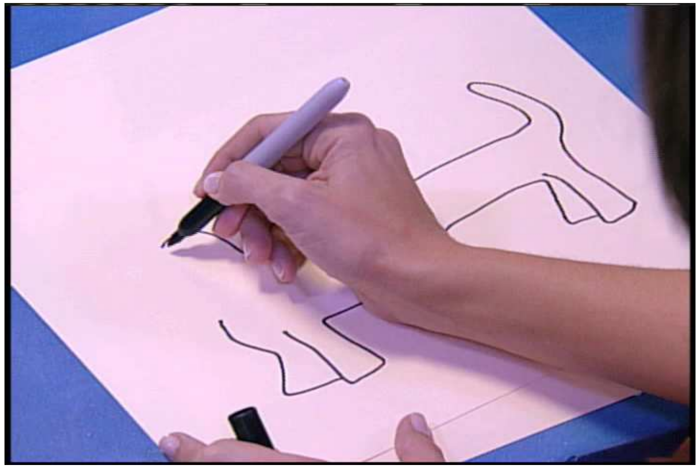
Crayon filter

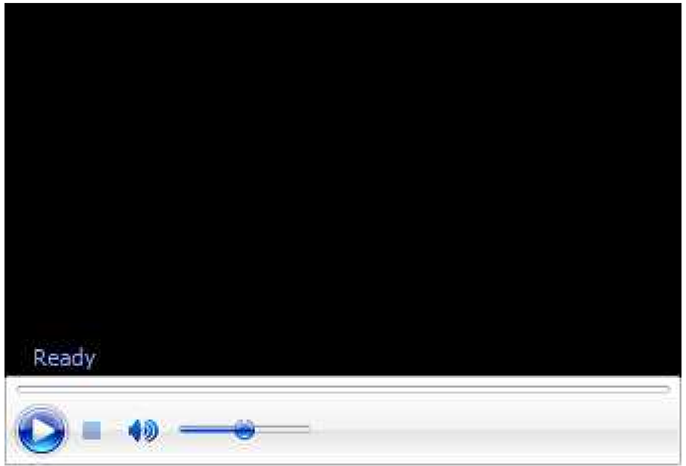
Hand segmentation



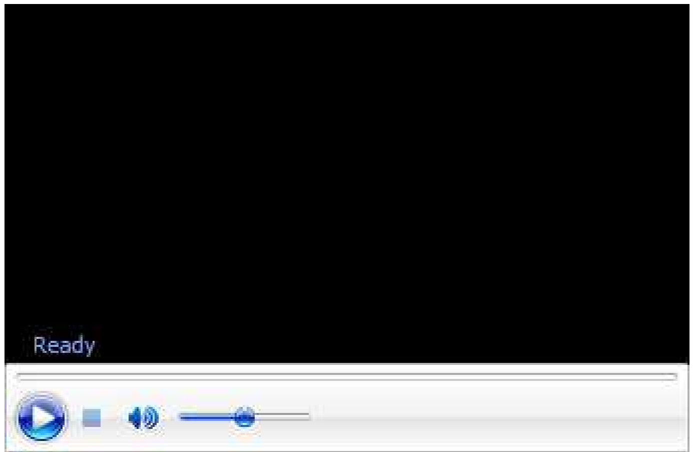
Object segmentation

Segmentation results

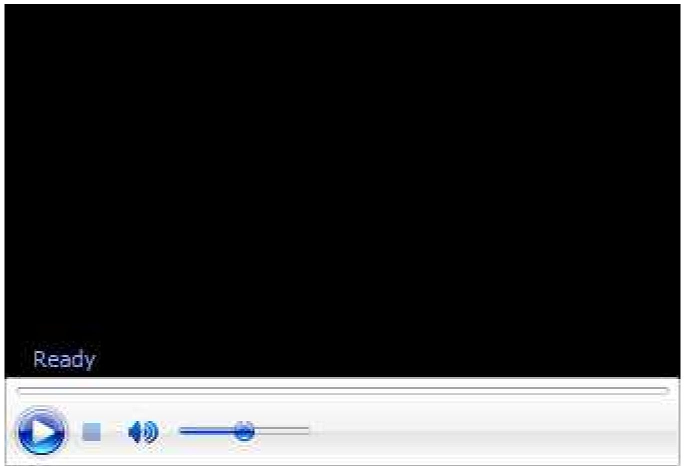




Marker filter

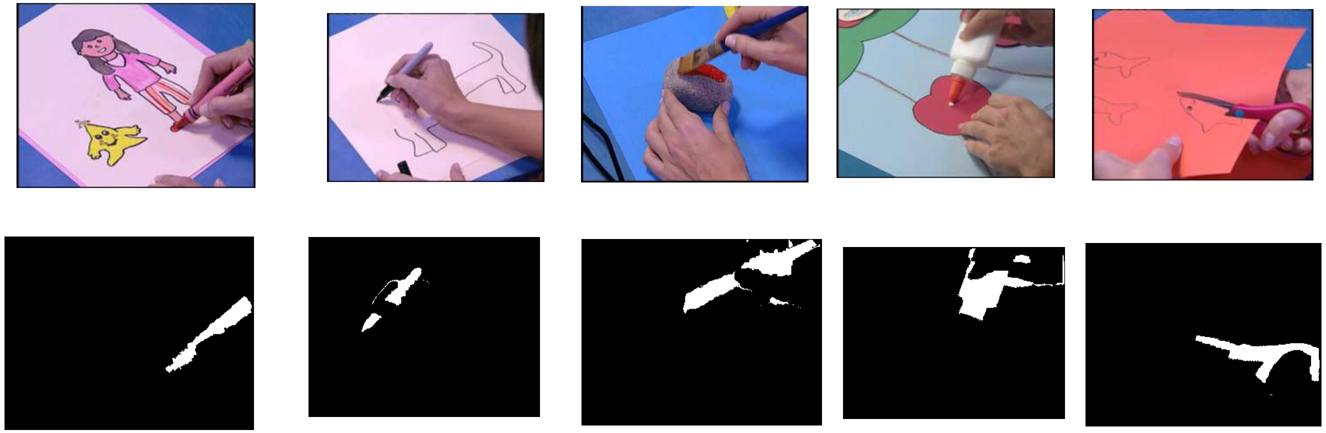


Hand segmentation



Object segmentation

Computed attribute description

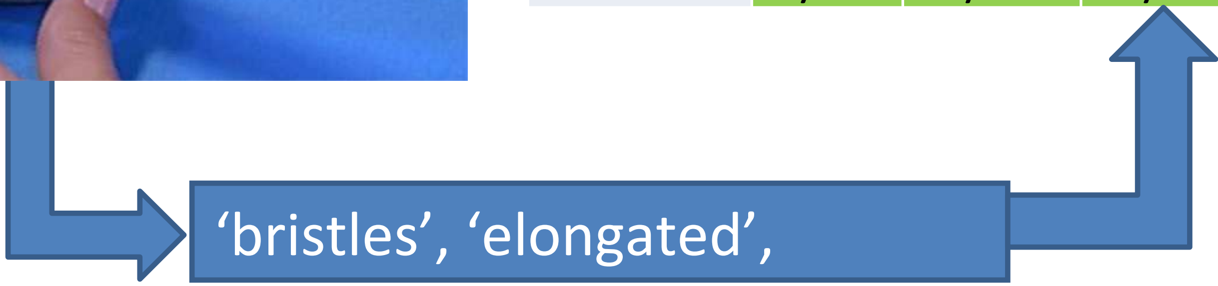


Color white, silver,other	other	other	silver	white	silver
Texture bristles (1D) yes no	no	no	yes	no	no
Shape elongated: yes no	yes	yes	yes	no	no
Shape convex: yes no	yes	yes	yes	yes	no

Ongoing NLP work

- Extract physical characteristics from web and Wikipedia to aid in unsupervised object recognition

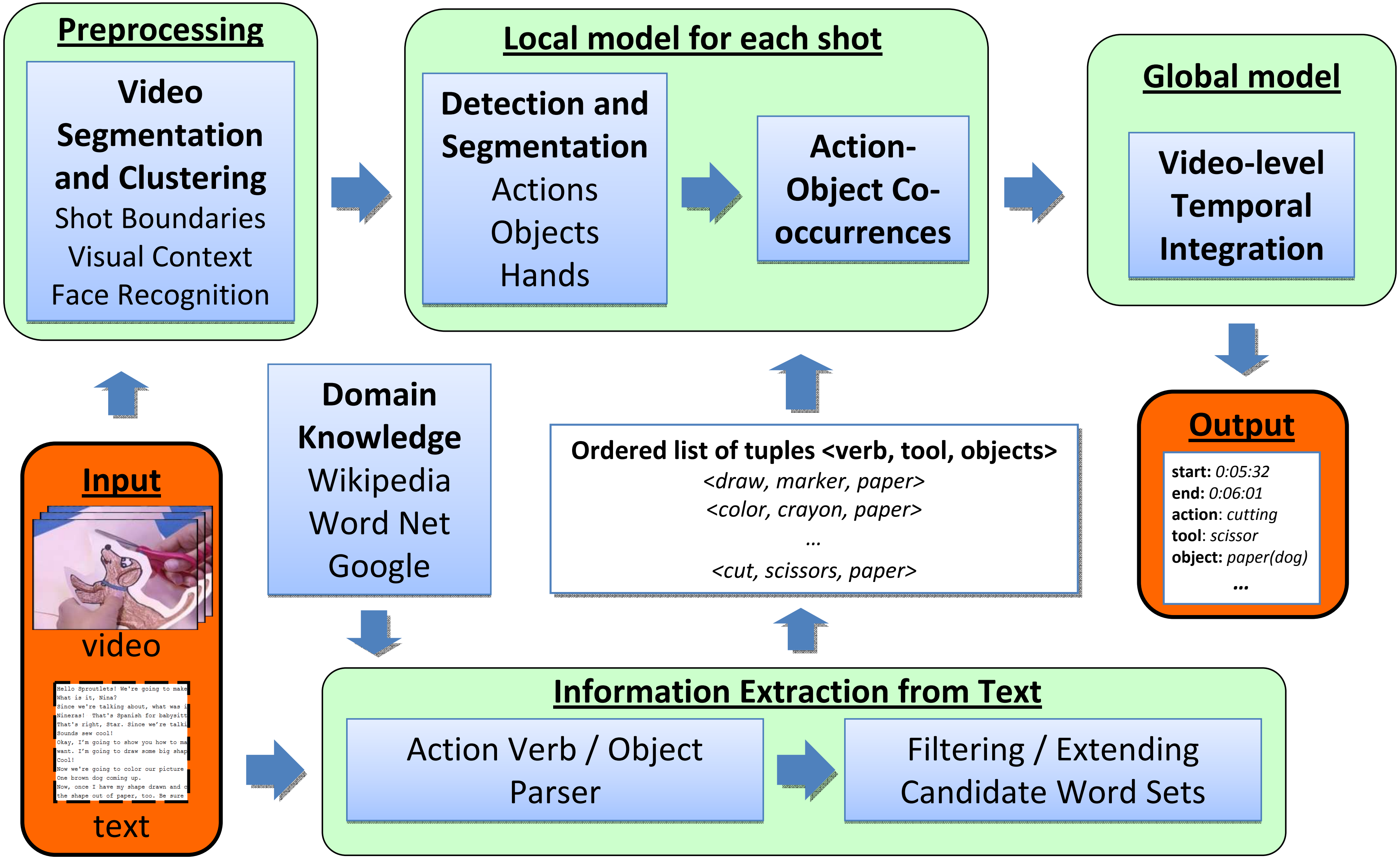
	crayon	marker	brush	scissors	glue
color	other	other	silver	silver	white
bristles	no	no	yes	no	no
elongated	yes	yes	yes	no	no
convex	yes	yes	yes	no	yes



Summary

- Unsupervised object recognition based on computing visual attributes derived from language
- Visual segmentation: attention based approach
- Proof of concept on a small set of videos

Recap


Preprocessing
Video Segmentation and Clustering
Shot Boundaries
Visual Context
Face Recognition
Local model for each shot
Detection and Segmentation
Actions
Objects
Hands
Action-Object Co-occurrences
Global model
Video-level Temporal Integration
Output
start: 0:05:32
end: 0:06:01
action: cutting
tool: scissor
object: paper(dog)
...
Input
video
text
Domain Knowledge
Wikipedia
Word Net
Google
Ordered list of tuples <verb, tool, objects>
<draw, marker, paper>
<color, crayon, paper>
...
<cut, scissors, paper>
Information Extraction from Text
Action Verb / Object Parser
Filtering / Extending Candidate Word Sets

Next steps

- ***Improve temporal modeling of videos***
  - sequence labeling with more complex temporal models for the text
  - use tracking to improve object detection
- ***Use more complex object-action models***
  - occlusion reasoning from the segmentation in training object classifiers,
  - model how an actions can transform an object's shape and appearance (cooking, cutting, painting, bending, …)
- ***Explore new object and action representations to deal with***
  - Large numbers of action and object categories (e.g. attribute-based representations?)
  - Large intra category variations (e.g. decorating, placing)
  - Transparent objects (glass),
  - Deformable objects
- ***Extend unsupervised learning approaches***
  - include temporal order of words in text into multiple instance learning
  - get suggestions for labels directly from text
- ***Apply approach to more complex videos and larger data sets***
  - cooking, home improvement, surveillance, …

Accomplishments

- Created a **new baseline data set** for research into recognition of complex manipulation actions
  - Benchmark for future research
- Created an **end-to-end system** that annotates real-world broadcast videos with the presence of actions and objects
  - Will be publicly available, reducing barrier of entry for further research
  - Demonstrates how non-visual semantic and temporal information can be integrated to **improve action recognition**
  - Demonstrates how this information can be **automatically extracted from text and unstructured domain knowledge** (Wikipedia, Google)

Accomplishments

- Created a **<u>new baseline data set</u>** for research into recognition of complex manipulation actions
  - Benchmark for future research
- Created an **<u>end-to-end system</u>** that annotates real-world broadcast videos with the presence of actions and objects
  - Will be publicly available, reducing barrier of entry for further research
  - Demonstrates how non-visual semantic and temporal information can be integrated to **<u>improve action recognition</u>**
  - Demonstrates how this information can be **<u>automatically extracted from text and unstructured domain knowledge</u>** (Wikipedia, Google)
- Results (Mean Recognition Rate across Classes)
  - 0.42 : Single Shot Action Recognition using STIP (SVM)
  - 0.47 : SSAR + Tool + Hand Feature
  - 0.51 : Single Shot Joint CRF Model (STIP+Tool+co-occurrence of verb and tool from text)
  - 0.52 : Sequence Model CRF with temporal text constraints

Outcomes for the research community

- Novel insights into
  - Leveraging NLP to improve visual scene understanding
  - Action recognition for human actions defined by interactions with the environment

- Software pipeline to annotate video with semantic information extracted from a text

- A publicly available data set of richly annotated videos with realistic action-object interactions
  - PBS Sprout: 27 craft shows with 8 to 11 individual actions each