

Categories of actions for story evaluation

Introduction

We want to collect *labels* for video segments in our benchmark that reflect different aspects of **video** and **story** generation.

For example, which entities are in the video? What kind of transitions are there between segments?

Instructions

We have grouped annotation labels into multiple **categories** (eg “camera movements”, “entities”; see below).

For each category, there are multiple **labels** which describe some properties that may be in the video (e.g, “the camera follows a person”).

Each label has two checkboxes that need to be checked independently according to:

- **Text:** The label is mentioned in the “segment caption”
- **Video:** The label is shown in the video


For a given category, select all labels that apply (or none when appropriate).

To better relate categories and labels to the segment video and caption, you can **click on a category name** (e.g. “Camera movements”) to collapse it, and click again on it to reveal it again. This will hopefully allow you to work on the next category without having to scroll.

NB: When we refer to **animate entities**, we mean: people, animals and other living organisms that are capable of performing actions.

Example user interface (truncated):

Segment caption: A brown-white dog is standing on the right side of a canal and trying to catch a white ball from the canal with its paw and moving its paw in the water again and again.




0:00 / 0:15

Camera movements		
static shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
pan	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tilt	Text <input type="checkbox"/>	Video <input type="checkbox"/>
zoom	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tracking shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
over-the-shoulder-shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
aerial shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
point-of-view-shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>

Entities		
people	Text <input type="checkbox"/>	Video <input type="checkbox"/>
animals	Text <input type="checkbox"/>	Video <input type="checkbox"/>
food/drinks	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tools	Text <input type="checkbox"/>	Video <input type="checkbox"/>
containers	Text <input type="checkbox"/>	Video <input type="checkbox"/>

Segment caption: The brown-white dog goes to the other side and picks the ball with its mouth.




0:00 / 0:08

Camera movements		
static shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
pan	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tilt	Text <input type="checkbox"/>	Video <input type="checkbox"/>
zoom	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tracking shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
over-the-shoulder-shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
aerial shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
point-of-view-shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>

Entities		
people	Text <input type="checkbox"/>	Video <input type="checkbox"/>
animals	Text <input type="checkbox"/>	Video <input type="checkbox"/>
food/drinks	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tools	Text <input type="checkbox"/>	Video <input type="checkbox"/>
containers	Text <input type="checkbox"/>	Video <input type="checkbox"/>

Segment caption: The brown-white dog jumps to the right and drops the ball in the canal.




0:00 / 0:01

Camera movements		
static shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
pan	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tilt	Text <input type="checkbox"/>	Video <input type="checkbox"/>
zoom	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tracking shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
over-the-shoulder-shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
aerial shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
point-of-view-shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>

Entities		
people	Text <input type="checkbox"/>	Video <input type="checkbox"/>
animals	Text <input type="checkbox"/>	Video <input type="checkbox"/>
food/drinks	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tools	Text <input type="checkbox"/>	Video <input type="checkbox"/>
containers	Text <input type="checkbox"/>	Video <input type="checkbox"/>

Segment caption: The brown-white dog is looking towards the ball.



0:00 / 0:03

Camera movements		
static shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
pan	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tilt	Text <input type="checkbox"/>	Video <input type="checkbox"/>
zoom	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tracking shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
over-the-shoulder-shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
aerial shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>
point-of-view-shot	Text <input type="checkbox"/>	Video <input type="checkbox"/>

Entities		
people	Text <input type="checkbox"/>	Video <input type="checkbox"/>
animals	Text <input type="checkbox"/>	Video <input type="checkbox"/>
food/drinks	Text <input type="checkbox"/>	Video <input type="checkbox"/>
tools	Text <input type="checkbox"/>	Video <input type="checkbox"/>
containers	Text <input type="checkbox"/>	Video <input type="checkbox"/>

First video segment on the left, second video segment to its right, and so on.

We recommend printing the following two pages and have them readily available during the annotation process.

Camera Movements

Which of these camera movements are shown in the video and/or described in the caption?

- Static Shot: A static shot has no camera movement at all (for most of the video, eg 80%).
- Pan: The camera pan directs a camera horizontally left or right.
- Tilt: direct a camera upward or downward.
- Zoom: The camera moves closer or further away.
- Tracking shot: The camera continuously follows one character.
- Aerial shot: The camera is located up above, overhead, capturing the action going on below (eg from a drone, from a crane/jib)
- Point-of-view shot: video shows what a character is looking at in the first person. In other words, the camera acts as the eyes of a character and the audience sees what they see.
 - More concretely, there are parts of the cameraman that are shown on screen (eg legs, arms, helmet, etc.) or it can be inferred from the movement of the camera that the cameraman is performing the main action (eg riding a bike, running)
 - Prefer not selecting this if you are not sure

Foreground Entities

Which of these entities are shown in the video and/or described in the caption?

- People: Woman, Man, Girl, Boy, Kid, Women, Kids, etc
- Animals: Dog, Cat, Shark, ...
- Vehicles: Car, airplane, boat, bicycles, ...
- Food/drinks: ...
- Containers: Wine glass, bucket, boxes, ... - it is used as a container in the video
- Tools: toys, utensils, kitchen appliances, weapons, sports equipment, music instruments, toys, electronic devices, etc. That is, everything else that people actually grab, hold or indirectly use.

Foreground Actions

Which of these actions happen in the video foreground and/or are described in the caption?

- Humans moving (changing position in space): People walking, ...
- Animals moving (changing position in space): Dogs running, ...
- Objects moving (changing position in space): Fall, Roll, Slide, Being pushed, Being pulled, Being thrown, Being transported (eg by conveyor belts, cranes, or elevators), etc
- Humans using objects: eg handling objects, such as tools; manipulate objects, such as opening, closing; interacting with technological objects, such as computers, phones; etc

- By physical touch or it can be inferred that the objects are affected by the humans' actions
- Animals using objects: ...
 - By physical touch or it can be inferred that the objects are affected by the animals' actions
- Static actions (for both humans and animals): talking, standing, ...

Background Actions

Which of these actions happen in the video background and/or are described in the caption?

- Animate entities moving: People walking, dogs running, ...
- Objects moving: Fall, Roll, Slide, Being pushed, Being pulled, Being thrown, Being transported (eg by conveyor belts, cranes, or elevators), etc
- Animate entities using objects: handling objects, such as tools; manipulate objects, such as opening, closing; interacting with technological objects, such as computers, phones; etc
- Static actions: talking, standing, ...
- Dynamic background: Fireworks, explosions, sun setting, waves moving, clouds moving, ...

Foreground Interactions (between two or more animate entities)

Which of these entity interactions are shown in the video and/or described in the caption?

- Dialogue: two or more entities converse with each other
- Direct: two animate entities directly interact with each other (eg there is physical touch)
- Indirect: two animate entities interact with each other with no physical interaction
- Object-based: two animate entities interact with each other through objects (eg one object is given from one character to the other)

Foreground Transitions (happening any time in the video after the first second of the first video segment)

Which of these transitions are shown in the video and/or described in the caption?

- New entities: One or more animate entities are shown in the video for the first time
- New objects: One or more objects are shown in the video for the first time
- Entities vanish: One or more animate entities disappear from the video (eg person hides)
- Objects vanish: One or more objects disappear from the video (eg obfuscated)
- Entities re-enter: One or more animate entities re-enter the video
- Objects re-enter: One or more objects re-enter the video