Presentation #5

...

Data Collection

85 Aerial Videos

67 Ground Videos

(Warsaw, Kigali, Dakar, Victoria, Bratislava, Ljubljana, Stockholm, Bern, Dushanbe, Hanoi, Canberra, Tokyo)

THUMOS Challenge

- THUMOS
  - Defined as a “spirited contest” with two challenges:
    - Classification
      - Goal is to determine whether or not a video contains a particular action
    - Temporal Detection
      - Goal is to classify an action and detect its temporal locations in each video
- Temporarily segmented clips do not reflect the real world
  - Actions generally have causal/spatial relations between people objects
    - THUMOS’14 Challenge introduced thousands of untrimmed videos for 101 action classes to provide a dataset for action recognition and temporal detection in realistic settings
- The THUMOS action classes are from UCF101 and can be divided into five categories (videos are annotated and available on Youtube):
  - Human-Object Interaction
  - Body-Motion Only
  - Human-Human Interaction
  - Playing Musical Instruments
  - Sports

THUMOS Challenge

- Objective(s) of THUMOS Challenge:
  - 1) Serve as a benchmark and enable a comparison of different approaches on the tasks of action classification and temporal detection in large-scale realistic video settings
  - 2) Advance the state of the art/accuracy
    - Ex) The accuracy of UCF101 increased from 45% in 2012 to 90% in THUMOS'13
- Early datasets on action recognition in videos would utilize employed actors performing scripted actions under controlled environments
  - The next level of datasets consisted of scripted actors performing under dynamic environments
    - The level that followed (over a decade ago) used more realistic video footage from Hollywood movies and television channels
      - A lot of datasets used spatiotemporal annotations for action instances in short trimmed videos
        - The level of annotation became unrealistic as larger datasets are needed now
          - Most modern datasets use more classes and have more temporal clutter

THUMOS Challenge

Fig. 2. The figure shows the sample frames of the actions from UCF101 dataset (Soomro et al., 2012). The color of frame borders specifies the action type to which they belong: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, Sports (c.f. Appendix A). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

THUMOS Challenge

- Annotation/Verification Procedure
  - If a positive video was found, it was marked as either “Positive” or “Irrelevant” based on the following factors:
    - Slow Motion
    - Sped Up
    - Occlusions/Partial Visibility
    - Motion Blur
    - Clutter/Incorrect Background
    - Unrealistic Instances
    - Animation
- Temporal Annotation
  - Action Boundaries are more abstract than concrete
    - To solve the problem, the 101 action classes were divided into two categories:
      - Instantaneous actions which have a short time span
        - Ex(s): Basketball Dunk
      - Cyclic Actions that are repetitive
        - Biking
        - Playing Guitar

THUMOS Challenge

- There are also other measures to ensure that the evaluation for the task is objective (temporal annotations)
  - 1) Annotate action intervals consistently with the temporal segmentation of corresponding actions in the UCF101 dataset
  - 2) Marked some action instances as ambiguous in cases of partial visibility, incomplete execution or that had strong deviation in style

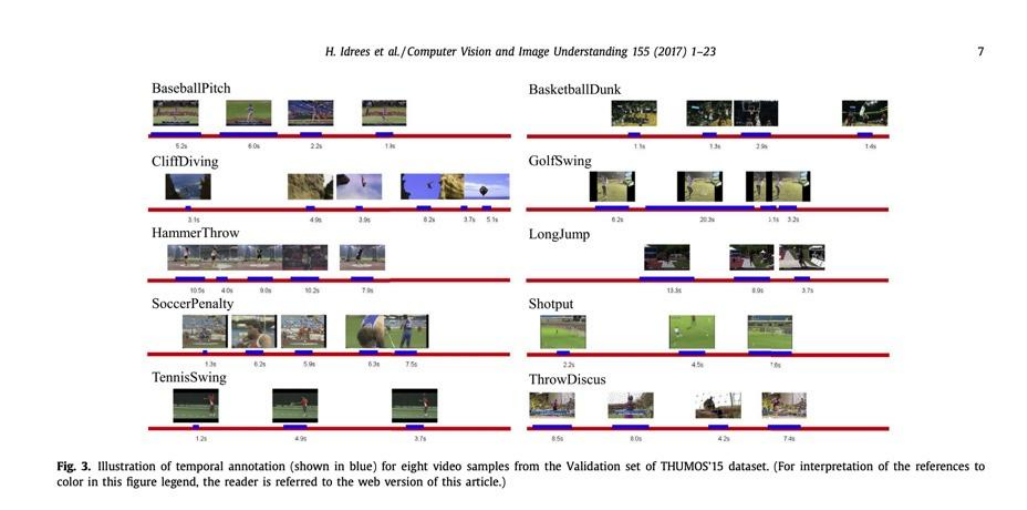

Fig. 3. Illustration of temporal annotation (shown in blue) for eight video samples from the Validation set of THUMOS'15 dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)