Recognizing Human Activities At Scale

Juan Carlos Niebles
www.niebles.net
@jcniebles
An image is worth one thousand words

**Image:**
- "a tiger attacking a person on a grass field"

**Video:**
- "the tiger is being playful"

**Video:**
- "a man packing a suitcase in a store"
- "the man is unpacking the suitcase"
- "someone unlocking a combination lock"
- "the person is attempting to pick the lock"
Video Generation and Consumption is Huge

Netflix: 100 million hours watched per day

YouTube: 400 hours uploaded per minute

Cisco: ~1 million minutes of video per second by 2020

~200 peta-pixels/second
Recognizing Human Actions
Action Recognition in Videos

KTH dataset [Schuldt, 2004]

HOHA dataset [Laptev, 2008]

UCF101 dataset [Soomro, 2012]

Short, pre-trimmed videos, only containing one action
Action Recognition in Videos

Traditional action classification pipeline

Input Video (pre-trimmed) → Feature extraction (handcrafted/learned) → Feature encoding → Classifier

Yes → Activity
No → Inactivity
Temporal Detection of Actions

Polishing shoes
Temporal Detection of Actions

Long input video

“Polishing shoes” Classifier

- Apply complex classifier at each temporal location frame
- Exhaustive search
- Repeat for all actions we want to detect
- Questionable scalability
Fast Activity Proposals for Action Detection

Long input video

- Runs very quickly (>130 fps)
- Find all temporal intervals that contain “any activity”
- Retrieve action intervals with high recall

[Caba Heilbron, Niebles & Ghanem. CVPR 2016]
[Escorcia, Caba Heilbron, Niebles & Ghamen, ECCV 2016]
Fast Activity Proposals for Action Detection

Long input video

“Polishing shoes” Classifier

“Polishing shoes” Classifier

“Polishing shoes” Classifier

[Caba Heilbron, Niebles & Ghanem. CVPR 2016]
[Escorcia, Caba Heilbron, Niebles & Ghamen, ECCV 2016]
Fast Activity Proposals for Action Detection

[Caba Heilbron, Niebles & Ghanem. CVPR 2016]
[Escorcia, Caba Heilbron, Niebles & Ghamen, ECCV 2016]
Recognizing Human Actions
Temporal Action Labeling

Ground Truth: SIL

Test Video

Learned Model

Prediction

Pan Fry

Put to Plate
Fully Supervised Learning

- Many training videos with per frame action labels
- Costly to annotate!
Weakly-Supervised Learning

- Only use action ordering
- Disambiguate by aggregating across with many training videos

[Pan Fry then Put to Plate]

- Train Video
- Model
- Label

[Huang, Fei-Fei & Niebles, ECCV 2016]
Extended Connectionist Temporal Classification

- Extends the CTC framework
- Explores space of frame-to-labels assignments efficiently
- Incorporates pairwise frame similarities

[Huang, Fei-Fei & Niebles, ECCV 2016]
Our approach starts without label correspondences for the training videos and iteratively improves the localization of the actions.
Weakly Supervised Activity Segmentation Results

[Huang, Fei-Fei & Niebles, ECCV 2016]
Weakly Supervised Action Detection Results

Drive Car

Drive Car: 0.126272

[Huang, Fei-Fei & Niebles, ECCV 2016]
Weakly Supervised Action Detection Results

Drive Car

Drive Car: 0.798997

[Huang, Fei-Fei & Niebles, ECCV 2016]
Hierarchical Modeling of Composable Activities

[Lillo, Soto & Niebles, CVPR 2014]
[Lillo, Niebles & Soto, CVPR 2016]
Recognizing Human Actions

Action Detection

Learning Actions With Weak Supervision
ActivityNet – www.activity-net.org

A Large-Scale Video Benchmark for Human Activity Understanding

Our benchmark aims at covering a wide range of complex human activities that are of interest to people in their daily living. We illustrate three scenarios in which ActivityNet can be used to compare algorithms for human activity understanding: global video classification, trimmed activity classification and activity detection.

[Caba Heilbron, Escorcia, Ghanem & Niebles, CVPR 2015]
Thank you!

Juan Carlos Niebles

www.niebles.net

@jcniebles