Discover and Learn New Objects from Documentaries

Kai Chen, Hang Song, Chen Change Loy, Dahua Lin
The Chinese University of Hong Kong

Introduction
- We develop a novel approach to learning object detectors from documentary videos and subtitles in a weakly supervised way.
- We propose a framework that can effectively integrate visual and linguistic cues.

WildLife Documentary(WLD) Dataset
- Video frames + subtitles
- 15 documentary videos
- >700k frames (7.4h)
- >4000 eight wild animals and humans
- >50 categories
- >6000 annotated tracklets

Examples
- Male tigers play no role in parenting.
- Australian camels appear sick and emaciated.
- Most animals have died in just three months, including the adult orangutan on the day we arrived.
- More tigers are on the move as the tigers are losing.

CRF formulation
\[
p(z, a, r | \theta) = \frac{1}{Z(\theta)} \exp \left( \Psi_{ap}(z | \theta) + \Phi_{kt}(z, a | \eta) + \Phi_{st}(r, z | \phi) \right)
\]
- Appearance potential
- Keyword-tracklet potential
- Geometric potential
- object category
- whether tracklet \( t_i \) associates with keyword \( w_j \)
- whether two tracklets should be merged

Results

Timeline
- Iter 0
  - 417.542 Background
  - 414.4218 Bear
  - 6165.6232 Background
  - 6232-6270 Background
  - 6268-6291 Background
- Iter 1
  - 417.542 Background
  - 414.4218 Bear
  - 6165-6232 Background
  - 6232-6270 (merged)
  - 6268-6291 Background
- Iter 2
  - 417.542 Cub
  - 414.4218 Bear
  - 6165-6231 (merged)
  - 6268-6291 Cow

Ground truth
- Tiger/Cub
- Bear
- Deer

Welcome to visit our project homepage or scan the barcode: