Discover and Learn New Objects from Documentaries

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

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 HONOLULU 2017

Introduction

- We develop a novel approach to learning object detectors from documentary videos and subtitles in an 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)
- >50 categories
- >4000 annotated tracklets

The elephant are about to march through them. The spiders themselves have a span as wide as a human hand.

But the love serenade is over once a dog arrives.

Australian camels appear sick and emaciated.

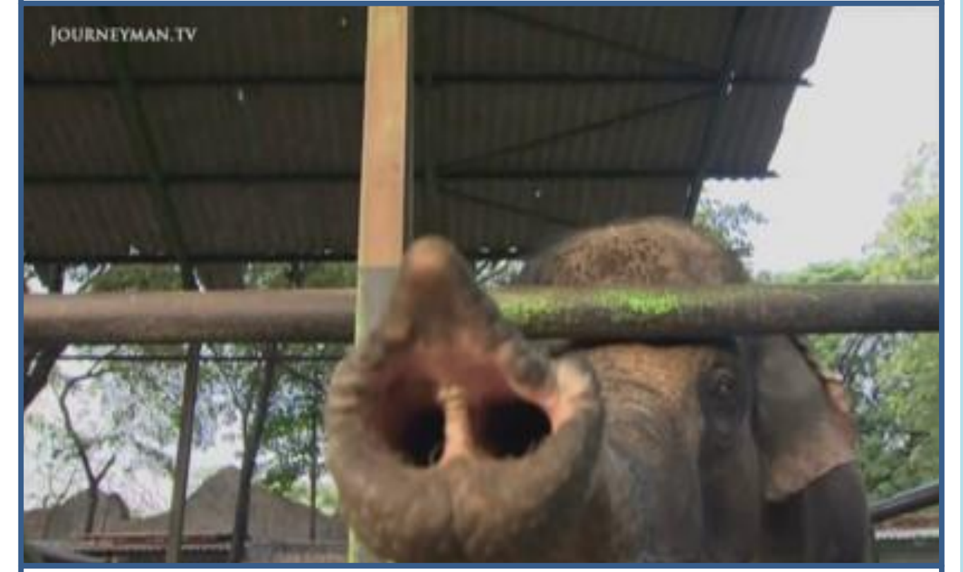
Tigers are one of the few cats that actually enjoy swimming.

Male koalas play no role in parenting.

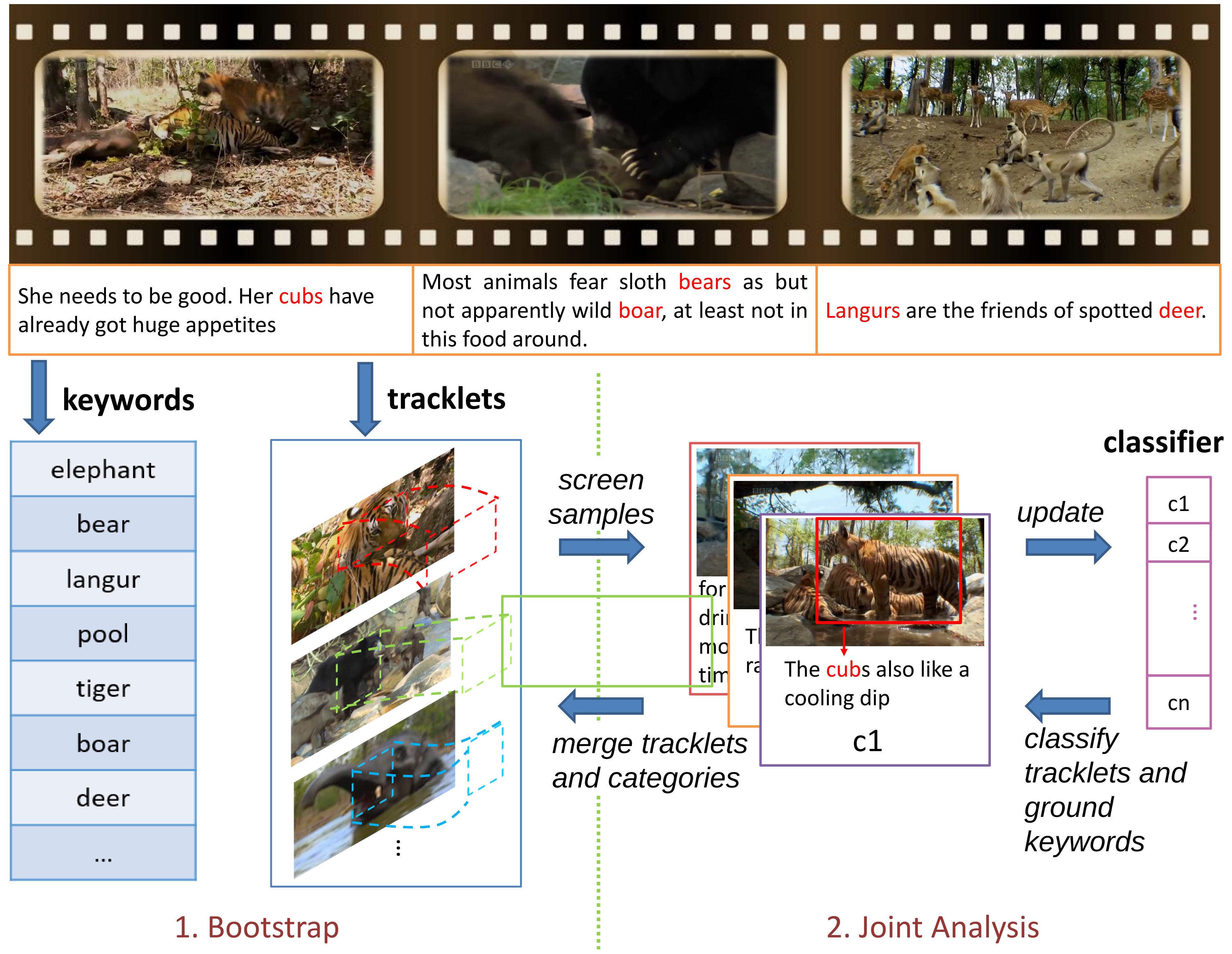
About 50 animals have died in just three months, including this adult orangutan on the day we arrived.

Unlike mechanics, langurs are the friends of spotted deer.

There's a turf war going on and the koalas are losing.


The mayor has declined offers of assistance and expert advice from animal welfare groups.

Framework


She needs to be good. Her cubs have already got huge appetites

Most animals fear sloth bears as but not apparently wild boar, at least not in this food around.

Langurs are the friends of spotted deer.

keywords

elephant
bear
langur
pool
tiger
boar
deer
...

tracklets

screen samples

merge tracklets and categories

The cubs also like a cooling dip

c1

update

classifier

c1
c2
cn

classify tracklets and ground keywords

1. Bootstrap

2. Joint Analysis

CRF formulation

$$p(z,a,r|o;\Theta) = \frac{1}{Z(\Theta)} \exp\left(\boxed{\Psi_{\mathrm{ap}}(\mathrm{z}|\mathrm{o};\theta)} + \boxed{\Phi_{kt}(z,a|o;\eta)} + \boxed{\Phi_{st}(r,z|o)}\right)$$

Appearance potential

Keyword-tracklet potential

Geometric potential

$$\Psi_{\mathrm{ap}}(\mathrm{z}|\mathrm{o};\theta) = \sum_{i=1}^{n} \psi_{ap}(z_i|v_i;\theta)$$

$$\Phi_{\mathrm{kt}}(\mathrm{z},\mathrm{a}|\mathrm{o};\eta) = \sum_{(i,j)\in G} \phi_{kt}(z_i,a_{ij}|\eta)$$

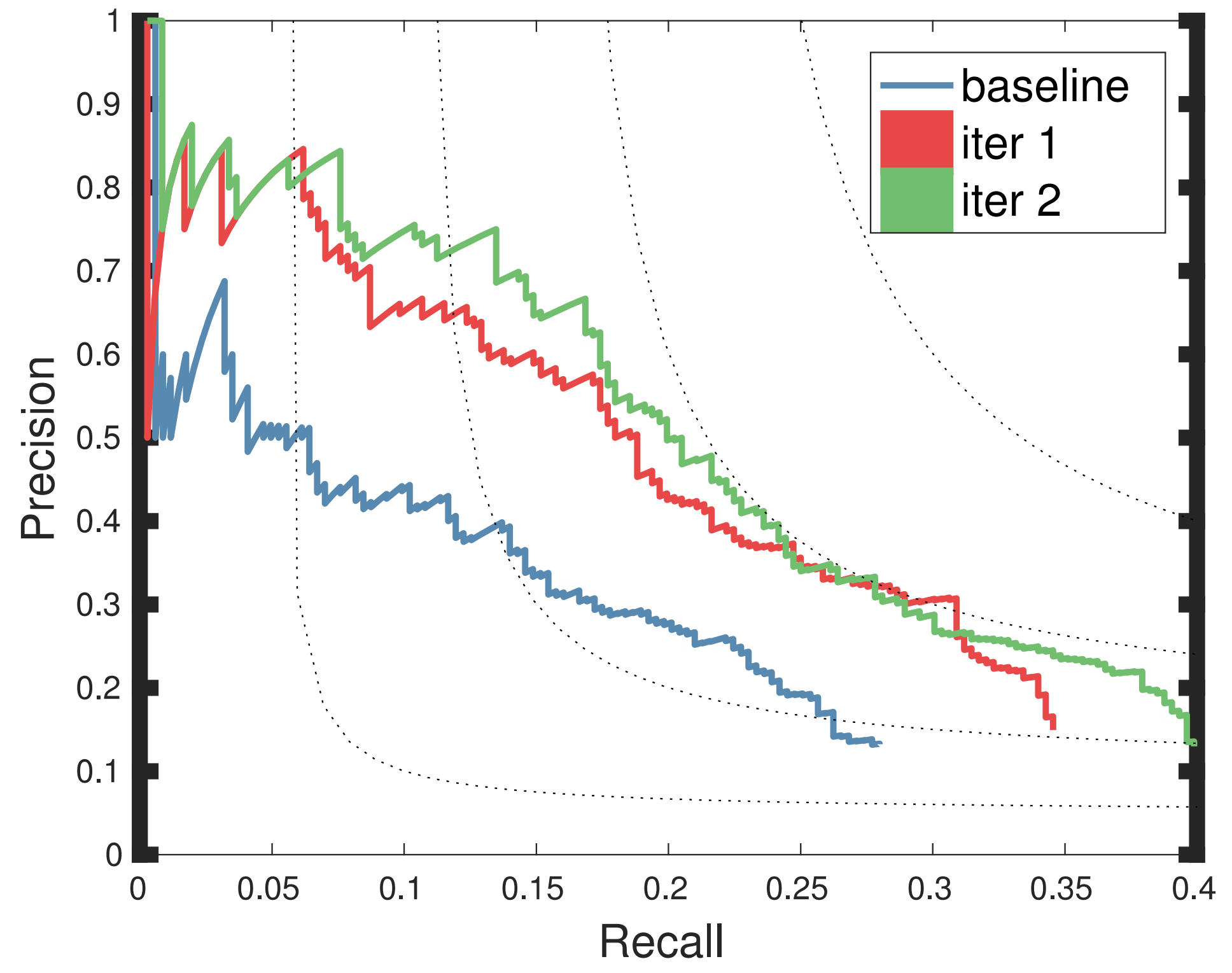
$$\Phi_{\mathrm{st}}(\mathrm{r},\mathrm{z}|\mathrm{o}) = \sum_{(i,i')\in R} \phi_{st}(r_{ii'},z_i,z_{i'}|u_i,u_{i'})$$

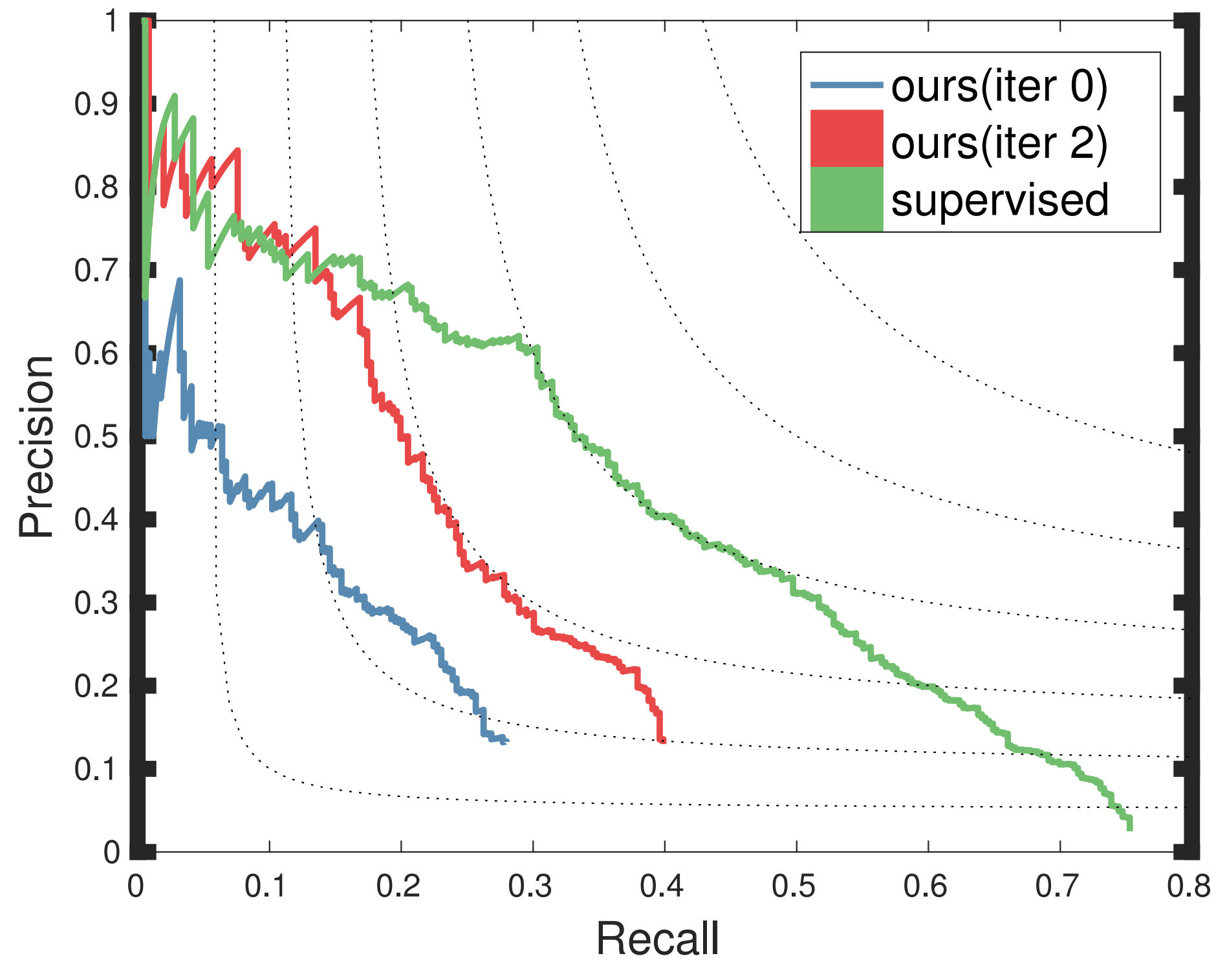
z_i object category

a_{ij} whether tracklet τ_i assiciates with keyword w_j

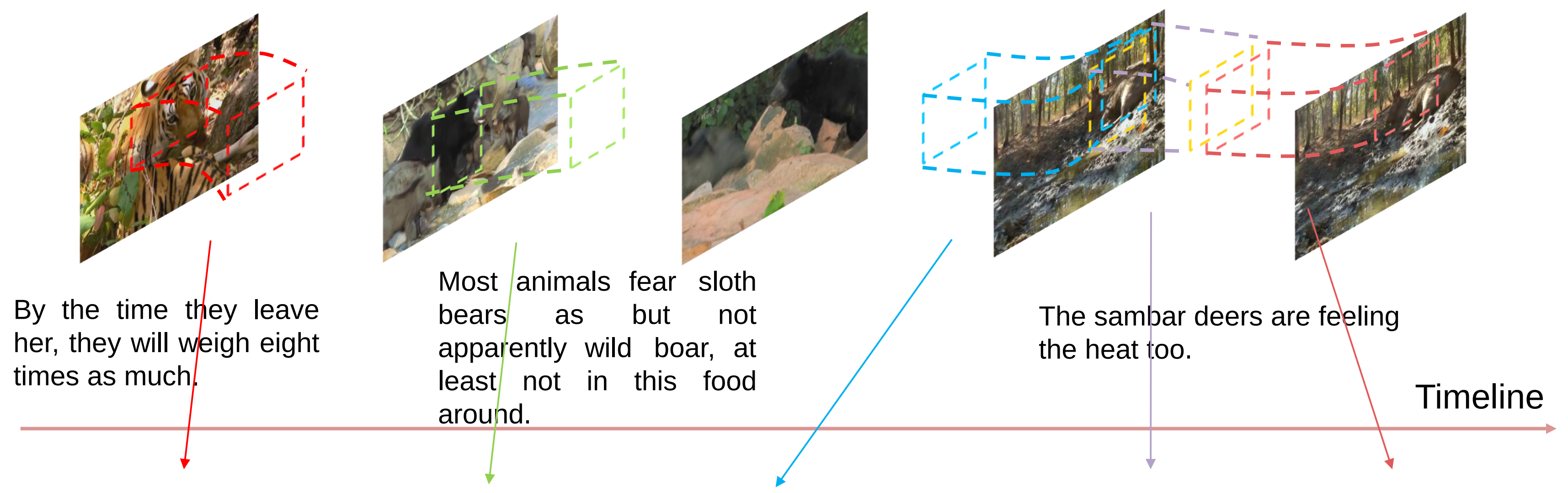
$r_{ii'}$ whether two tracklets should be merged

Results





Examples


By the time they leave her, they will weigh eight times as much.

Most animals fear sloth bears as but not apparently wild boar, at least not in this food around.

The sambar deers are feeling the heat too.

Iter 0	417-542 Background	4149-4218 Boar	6105-6232 Background	6232-6270 Background	6268-6291 Background
Iter 1	417-542 Cub	4149-4218 Bear	6105-6270(merged) Deer		6268-6291 Background
Iter 2	417-542 Cub	4149-4218 Bear	6105-6291 (merged) Deer		
Ground truth	Tiger/Cub	Bear	Deer		

Welcome to visit our project homepage or scan the qrcode.
http://www.chenkai.site/projects/documentary-learning/index.html