Multi-Class Video Co-Segmentation with a Generative Multi-Video Model

Wei-Chen Chiu and Mario Fritz
Max Planck Institute for Informatics, Saarbrücken, Germany
{walon, mfritz}@mpi-inf.mpg.de

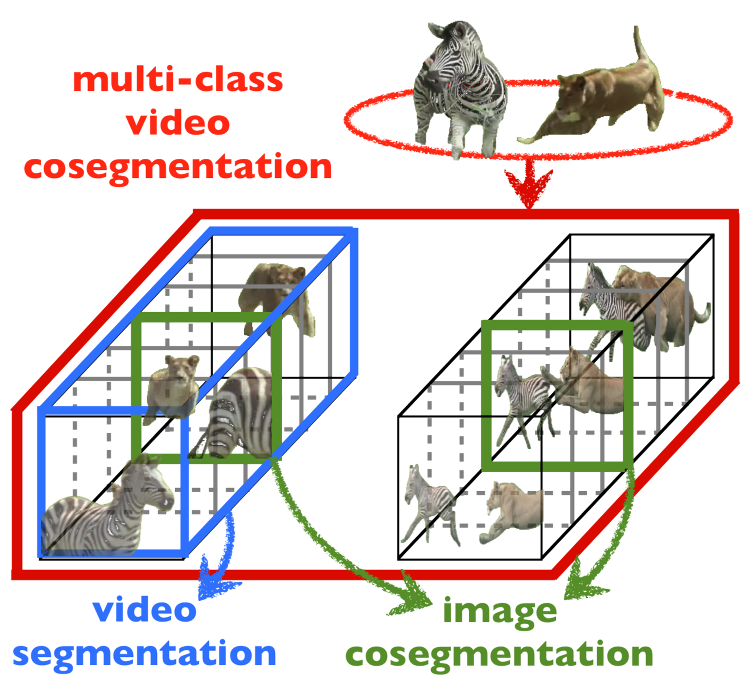
max planck institut informatik

International Max Planck Research School for Computer Science

Motivations

Infer segmentation from multiple videos to extract their semantic structures:

- Video data is a one of the fastest growing resource of publicly available data on the web.
- Reason across videos in order to reveal object class structure and resolve ambiguities caused by observing only a singe video.



Contributions

- First benchmark dataset for multi-class video co-segmentation task
- Video segmentation prior based on non-parametric bayesian spatio-temporal clustering process
- Joint segmentation of videos and learning of shared appearance models across videos
- Improved performance over video segmentation[2] and image co-segmentation[1] baselines

References

[1] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012.

[2] P. Ochs and T. Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011.

Data & Source

Dataset and source code available at:
http://scalable.mpi-inf.mpg.de/

Proposed Method

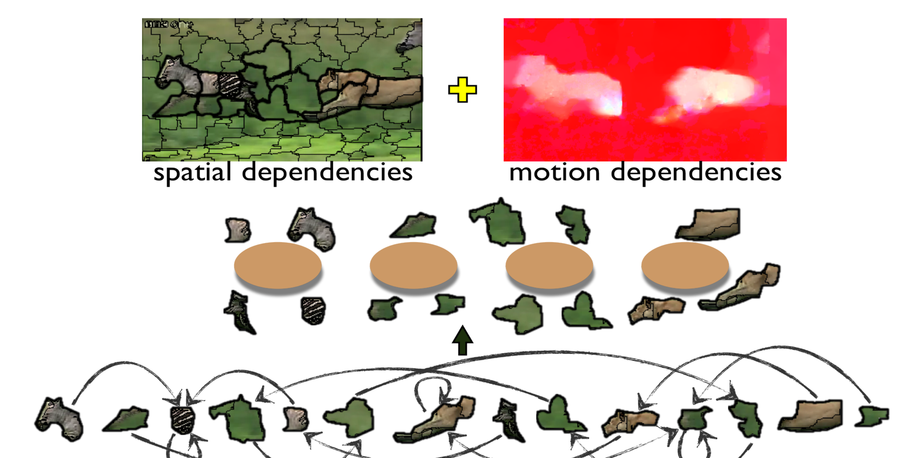
metaphor between HDP ⇔ videos	restaurant franchise	shared menu of dishes	restaurants	tables	customers
	set of videos	object classes	videos	object instances	superpixels

We propose global appearance classes to reason across multiple videos as well as a video segmentation prior to model contiguous segments of coherent motion based on distance dependent Chinese Restaurant Process(CRP).

ddCRP Video Segmentation Prior

Encourage to cluster nearby superpixels with similar motions for contiguous segments in spatio-temporal & motion domains.

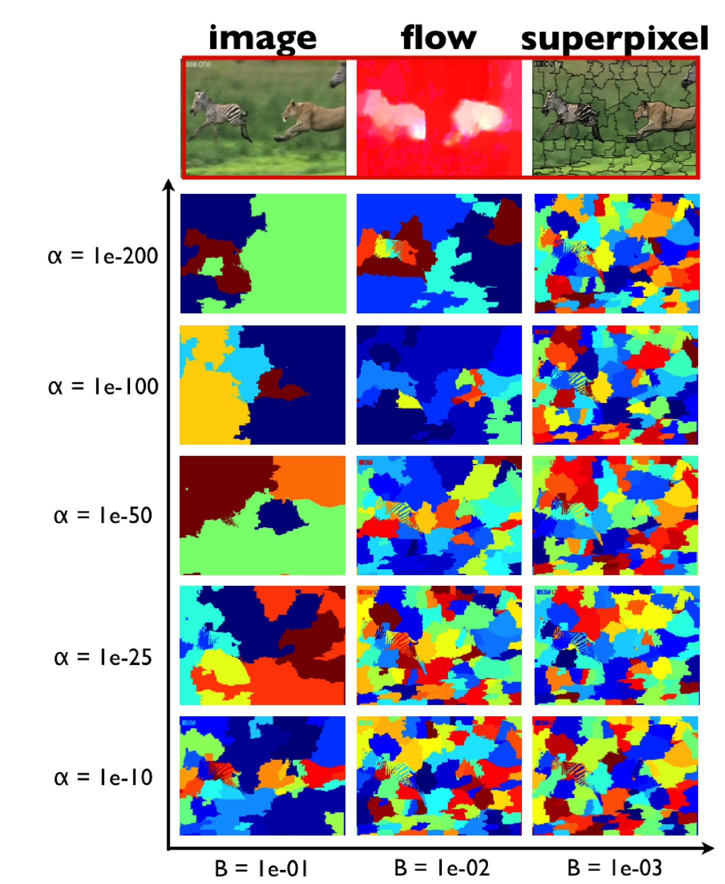
$$p(c_i = j|D, f, \alpha) \propto \begin{cases} f(d_{ij}) & j \neq i \\ \alpha & j = i \end{cases} \quad (1)$$



Generative Multi-Video Model

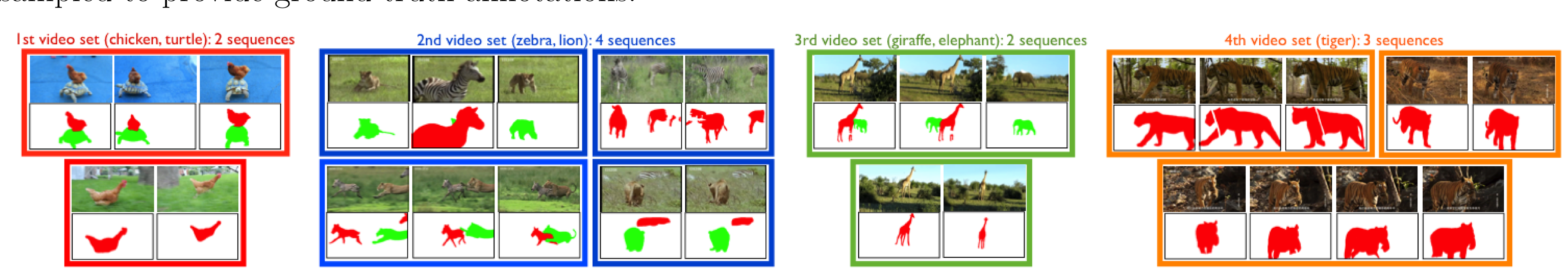
Multiple global object classes with different appearance models shared across videos + unknown number of object instances for each video ⇒ modeled by Hierarchical Dirichlet Process (HDP)

1. For each superpixel i_v in video v, draw assignment $c_{i_v} \sim$ ddCRP(D, f, α) to object instance
2. For each object instance t_v in video v, draw assignment $k_{t_v} \sim$ CRP(γ) to object class
3. For each object class k, draw parameters $\phi_k \sim G_0$
4. For each superpixel i_v in video v, draw observed feature $x_{i_v} \sim P(\cdot|\phi_{z_{i_v}})$, where $z_{i_v} = k_{t_{i_v}}$ the class assignment for i_v.



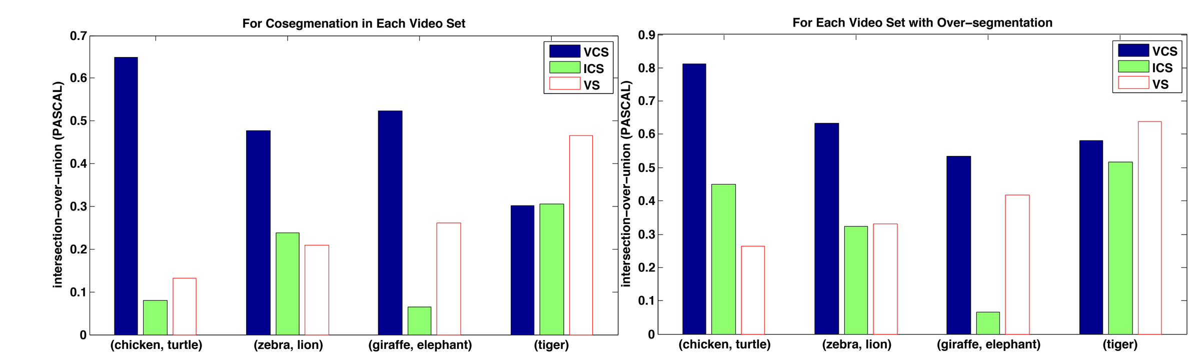
Benchmark

New Multi-Object Video Co-Segmentation (MOViCS) challenge on consumer videos collected from Youtube. The dataset has 4 different video sets including 11 videos with 514 frames in total. 5 frames per video are equidistantly sampled to provide ground truth annotations.



Experimental Results

- Evaluation metric: find for each object class a set of segments that coincide with object instances in video frames.
- Quantified by intersection-over-union metric with/without over-segmentation.



over-segmentation	VCS(Ours)	ICS[1]	VS[2]*
No	**48.75%**	17.25%	26.67%
Yes	**64.1%**	33.91%	41.28%

*VS baseline doesn't do inference across video but benefit from groundtruth

- Outperform recent image co-segmentation (ICS)[1] and video segmentation (VS)[2] approaches
- Analysis of improvement due to joint inference across all videos and learning a global object class model (3.15% better on average)