Carnegie Mellon

CVPR 2013
Portland, Oregon
June 23-28

Jointly Aligning and Segmenting Multiple Web Photo Streams for the Inference of Collective Photo Storylines

Gunhee Kim

Eric P. Xing

School of Computer Science,
Carnegie Mellon University

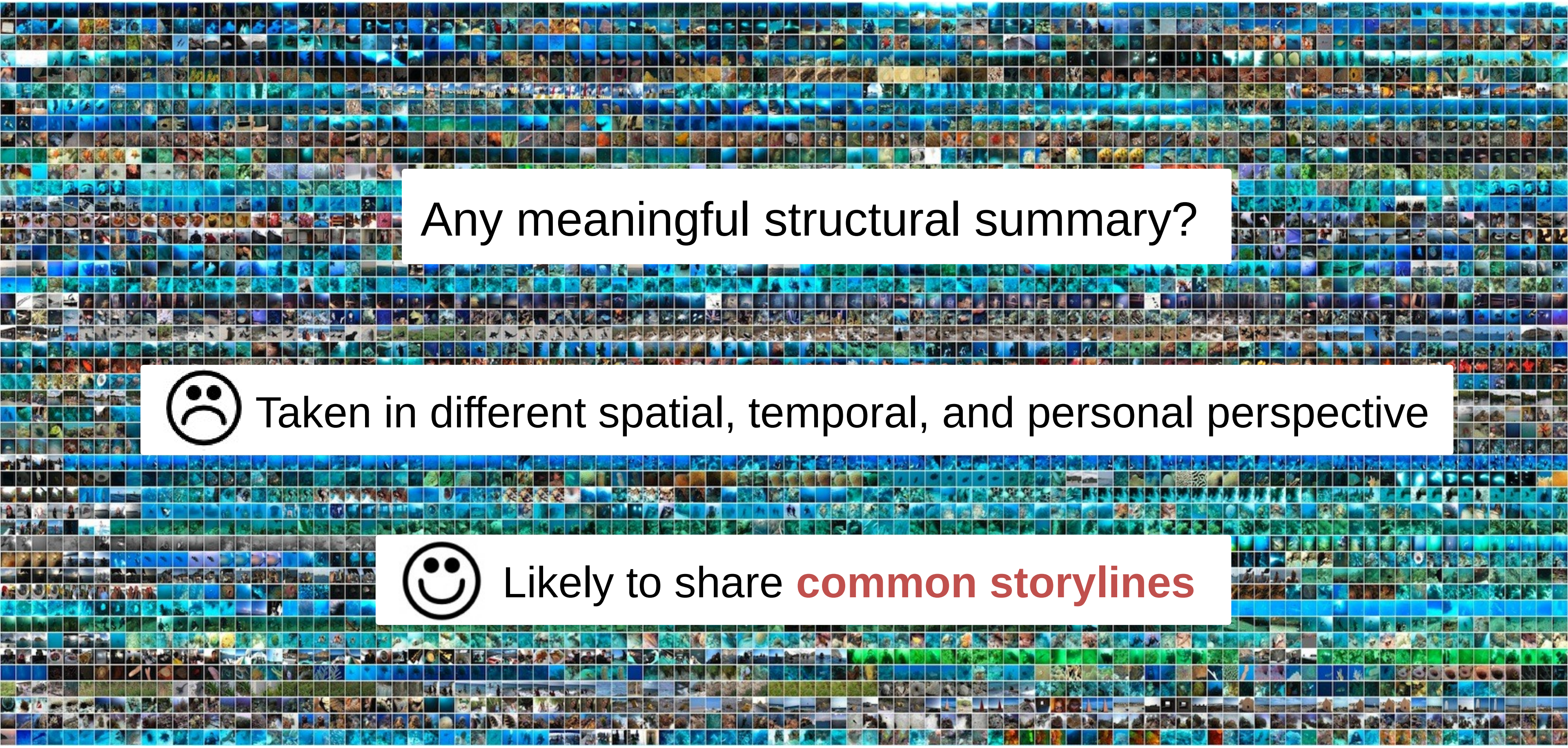
June 19, 2013

Outline

- Problem Statement
- Algorithm
  - Dataset and preprocessing
  - Alignment of Multiple Photo Streams
  - Large-scale Cosegmentation
- Experiments
- Conclusion

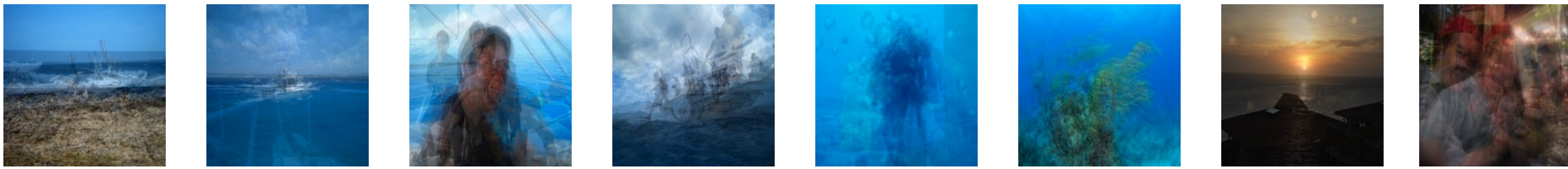
Background

Query *scuba+diving* from Flickr



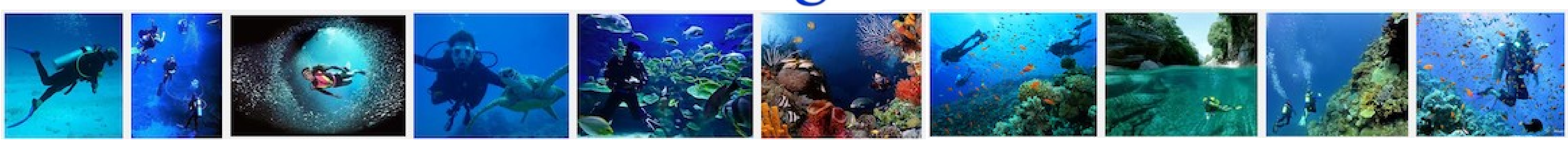
Our Ultimate Goal

An example of *scuba+diving* storyline



beach | boat | on boat | diving | underwater | coral | sunset | dinner

cf) ranking and retrieval by Google



Narrative structural summary vs. independently retrieved images

Reconstructing **photo storylines** from large-scale online images

Objective of This Paper

As a first technical step, jointly perform two crucial tasks...

***Mutually rewarding*!**

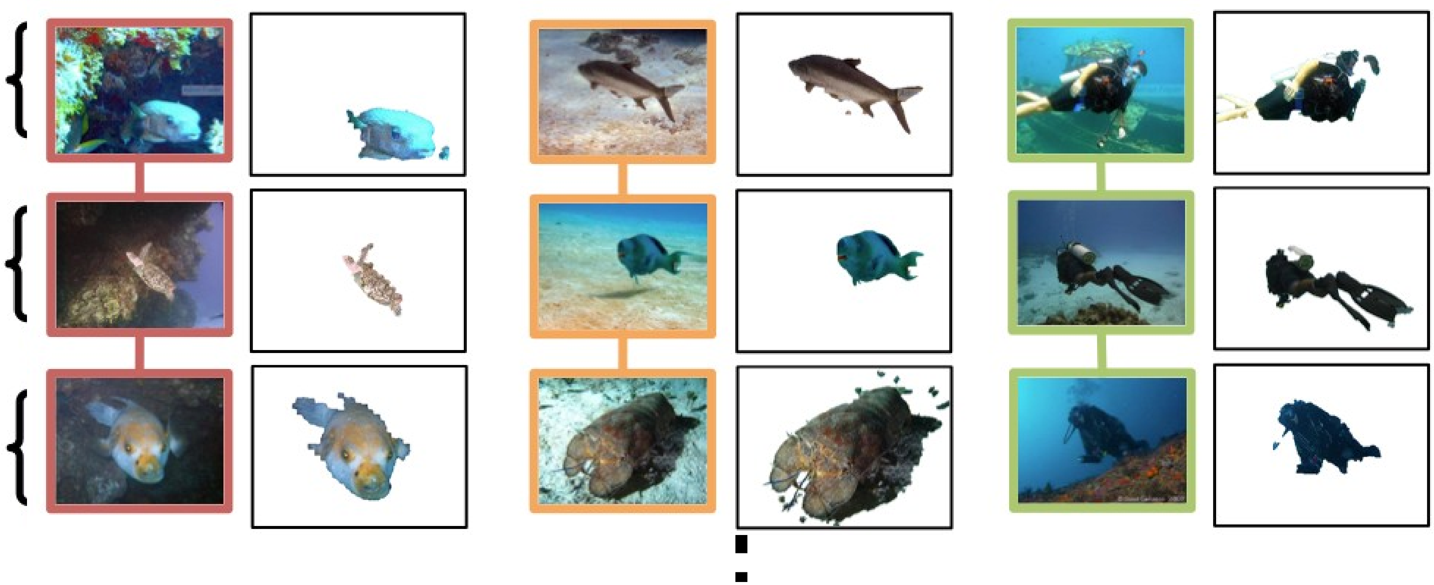
Alignment

Match images from
- different photo streams

Cosegmentation

Segment K common regions
- from aligned M images

PS2 User 1 at 10/19/2008 (Cayman Islands)



Objective of This Paper

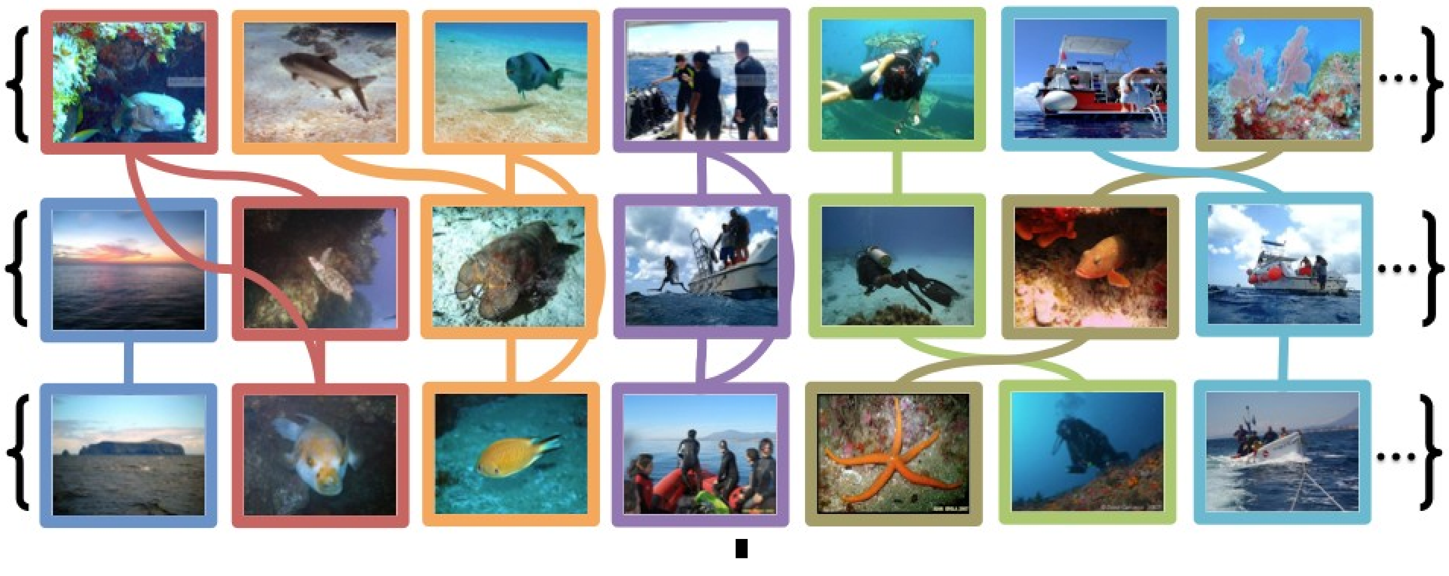
As a first technical step, jointly perform two crucial tasks...

***Mutually rewarding*!**

Alignment → **Cosegmentation**

- Online images are too diverse to segment together at once
  - The alignment discovers the images that share common regions

PS2 User 1 at 10/19/2008 (Cayman Islands)



Objective of This Paper

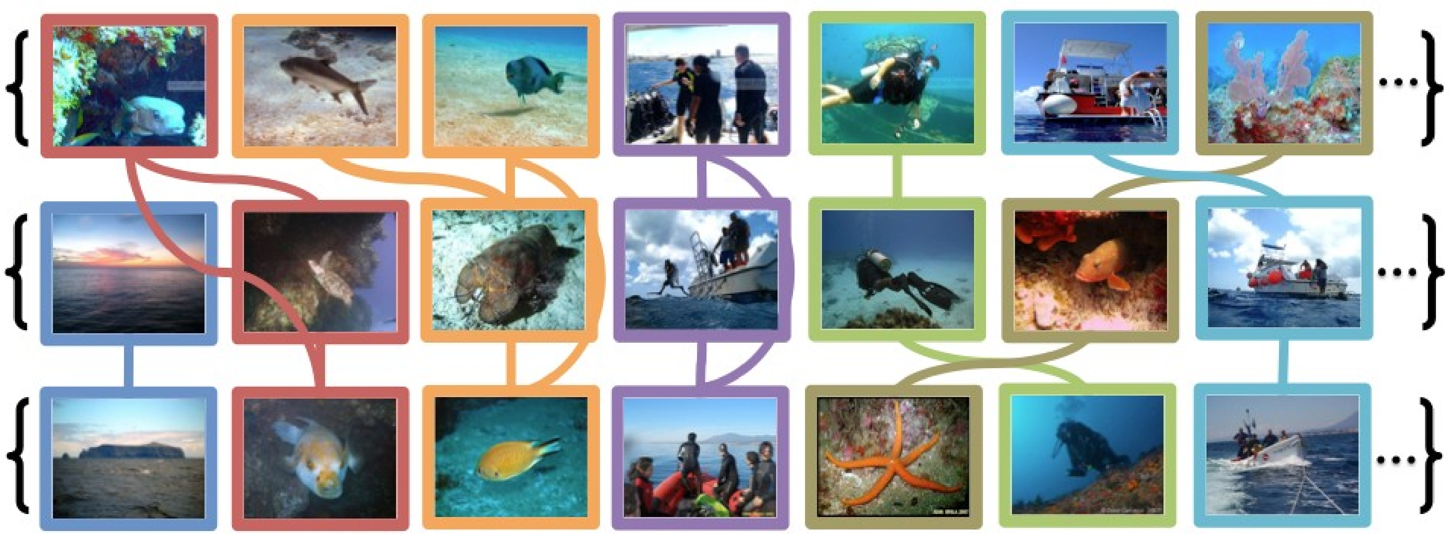
As a first technical step, jointly perform two crucial tasks...

***Mutually rewarding*!**

Alignment ← Cosegmentation

- Improve image matching by a better image similarity measure

Closing a loop between the two tasks



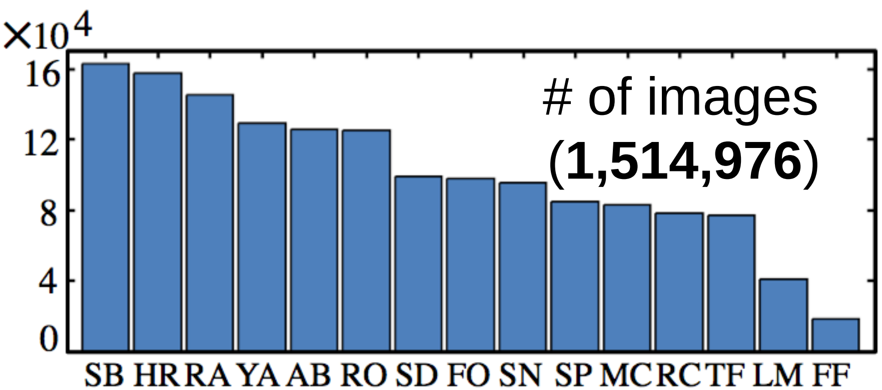
Outline

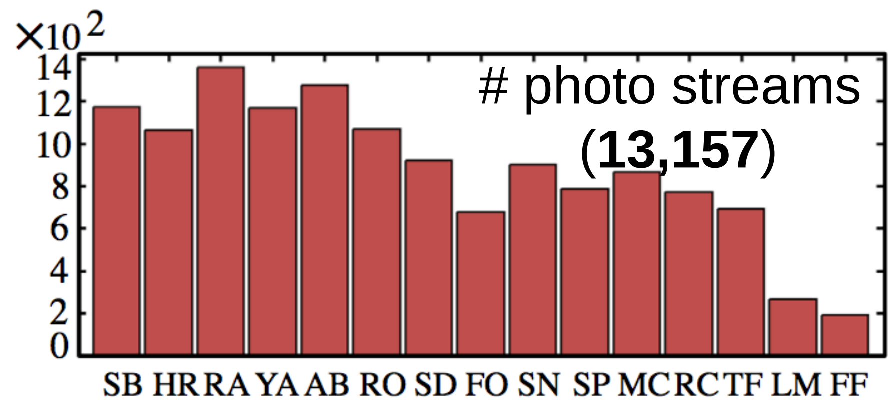
- Problem Statement
- Algorithm
  - Dataset and preprocessing
  - Alignment of Multiple Photo Streams
  - Large-scale Cosegmentation
- Experiments
- Conclusion

Flickr Dataset

Flickr dataset of 15 outdoor recreational activities

- Experiments with more than **100K** images of **1K** photo streams
- Larger than those of previous work by orders of magnitude






***S**urfing **B**each*

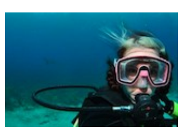

Horse **R**iding


RAfting


YAcht


Air **B**allooning


***RO**wing*


***S**cuba **D**iving*


***F**ormula **O**ne*


***SN**ow boarding*


***S**afari **P**ark*


***M**ountain **C**amping*


***R**ock **C**limbing*


***T**our de **F**rance*


***L**ondon **M**arathon*


***F**ly **F**ishing*

Image Descriptor and Similarity Measure

Image description

- HSV color SIFT and HOG features on regular grid
- L1 normalized spatial pyramid histogram using 300 visual words

Image similarity measure :

- (Our assumption) Segmentation enhances the image alignment.

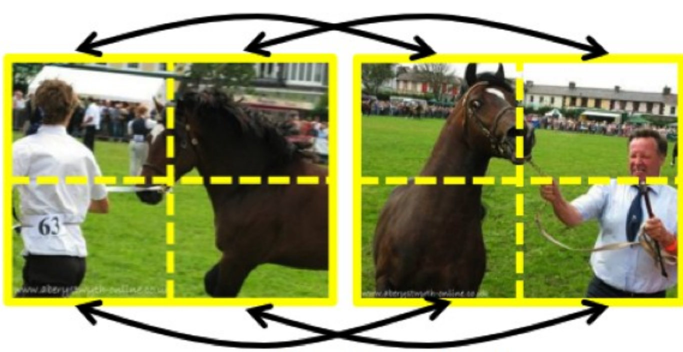
$$\sigma(I_1, I_2) = \max \left(\sum_{s \in \mathcal{F}_1} \sigma_s(s, f_s(s)) \right) / M$$

1. No segmentation available

- Histogram intersection on SPH

☹ Not robust against location/pose

- changes

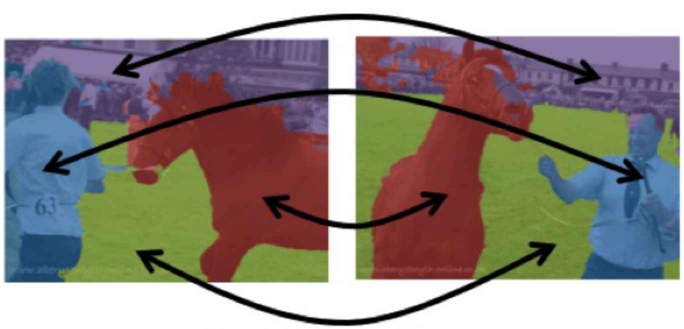


$\sigma(I_1, I_2) = 1.21$

2. Segmentation available

Histogram intersection on the

- best assignment of segments



$\sigma(I_1, I_2) = 1.83$

Outline

- Problem Statement
- Algorithm
  - Dataset and preprocessing
  - Alignment of Multiple Photo Streams
  - Large-scale Cosegmentation
- Experiments
- Conclusion

Alignment of Photo Streams

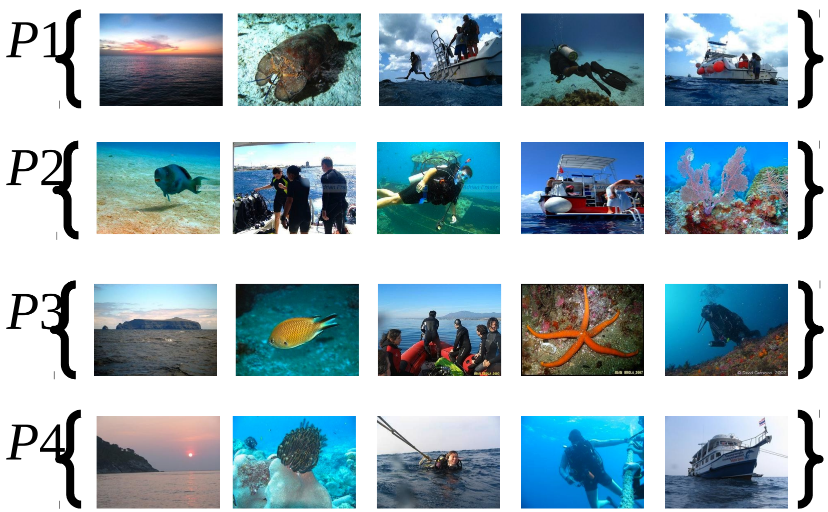
Input: A set of photo streams (PS): $\boldsymbol{P} = \{P1,\ldots,PL\}$

Photo Stream: a set of photos taken in sequence by a single user

- in a single day

Idea: Align all photo streams at once after building K-NN graph

- Naïve-Bayes Nearest Neighbor(NBNN) [Boiman et al. 08] for similarity metric



Alignment of Photo Streams

Input: A set of photo streams (PS): $\boldsymbol{P} = \{P1,\ldots,PL\}$

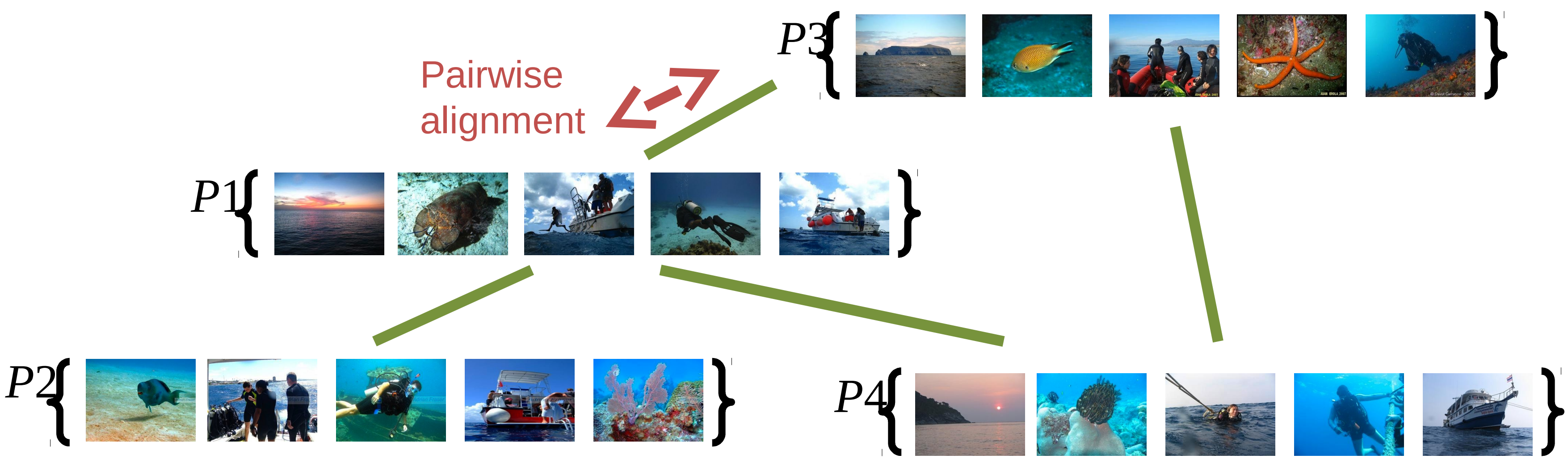
Photo Stream: a set of photos taken in sequence by a single user

- in a single day

Idea: Align all photo streams at once after building K-NN graph

- Naïve-Bayes Nearest Neighbor(NBNN) [Boiman et al. 08] for similarity metric

For simplicity, first consider pairwise alignment of two photo streams



Pairwise Alignment

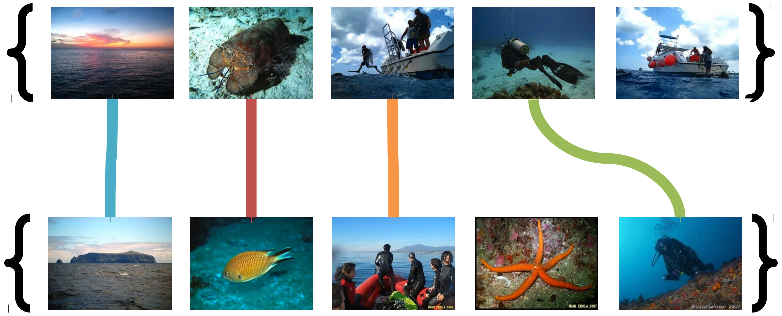
Goal of alignment: find a matching btw a pair of PS $\boxed{f : P^1 \rightarrow P^2 \cup \{\varnothing\}}$

- $\boxed{f(I) = \varnothing}$ means I in $P1$ has no match in $P2$.

Optimization: MRF-based energy minimization

- Flexibility: Various energy terms
- Solved by discrete BP

$$\boxed{E(P^1, P^2) = \sum_{I_i \in P^1} d(I_i, \hat{I}_i) + \sum_{I_i \in P^1} \eta \min\left(\left|t(I_i) - t(\hat{I}_i)\right|, \tau\right) + \sum_{(I_i, I_j) \in \delta} \rho \min\left(\left|t(\hat{I}_i) - t(\hat{I}_j)\right|, \nu\right)}$$



Pairwise Alignment

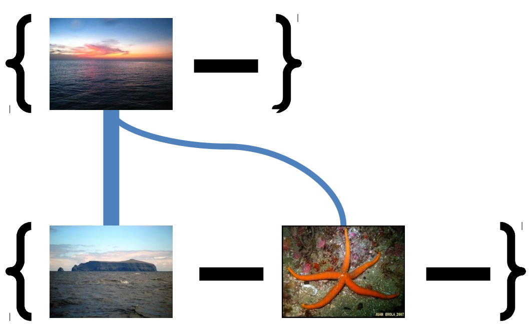
Objective function

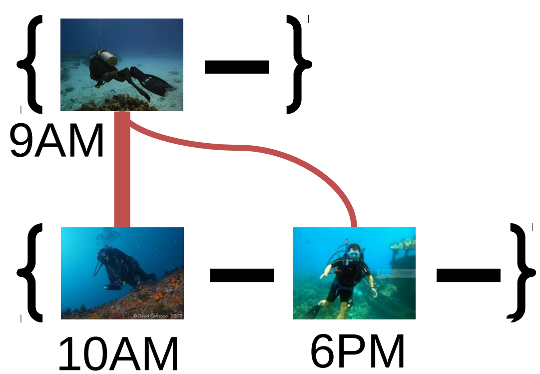
$$E(P^1,P^2)=\sum_{I_i\in P^1} d(I_i,\hat{I}_i)+\sum_{I_i\in P^1}\eta\min\left(\left|t(I_i)-t(\hat{I}_i)\right|,\tau\right)+\sum_{(I_i,I_j)\in\delta}\rho\min\left(\left|t(\hat{I}_i)-t(\hat{I}_j)\right|,\nu\right)$$

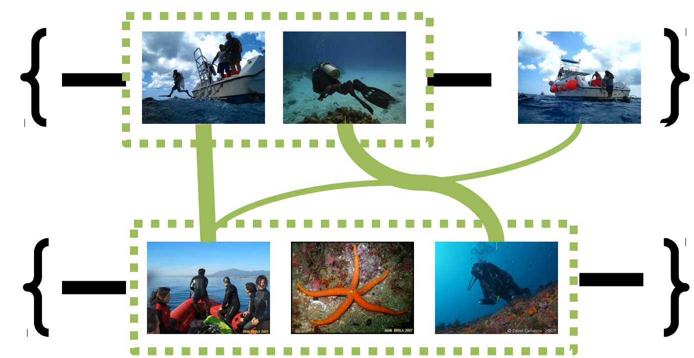
Data term : The matched image pairs should be visually similar.

Time term : The matched image pairs should be temporally similar.

Smoothness term : The matched images to neighbors in *P*1 should be neighbors in *P*2.







Alignment of Multiple Photo Streams

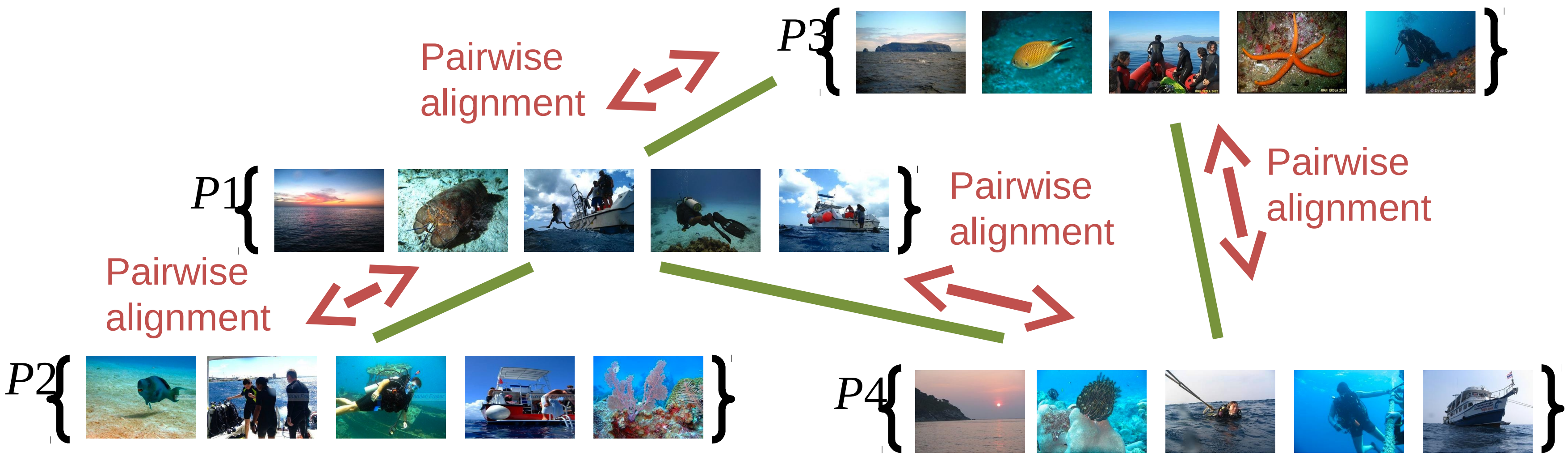
Objective : MRF-based energy minimization

$$\boxed{E_{All} = \sum E(P^i, P^j)}$$

$\boxed{(P^i, P^j) \in \Xi}$: All pairs of NN photo streams

Message-passing based optimization

- until convergence or for fixed iterations



Alignment of Multiple Photo Streams

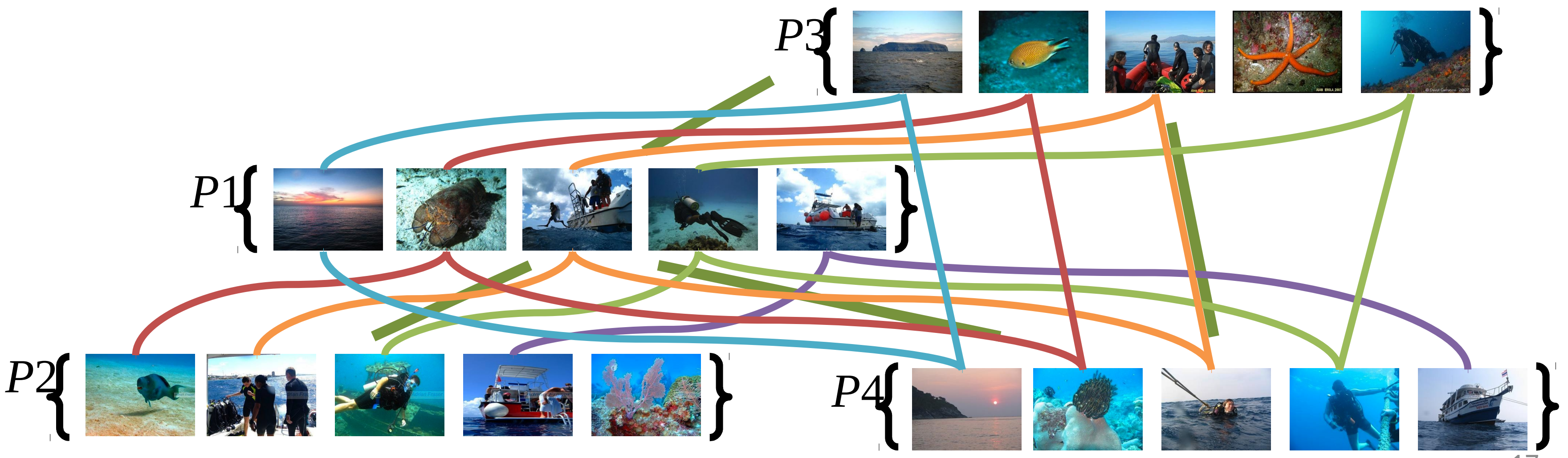
Objective : MRF-based energy minimization

$$E_{All} = \sum E(P^i, P^j)$$

$\boxed{(P^i, P^j) \in \Xi}$: All pairs of NN photo streams

Message-passing based optimization

- until convergence or for fixed iterations



Outline

- Problem Statement
- Algorithm
  - Dataset and preprocessing
  - Alignment of Multiple Photo Streams
  - Large-scale Cosegmentation
- Experiments
- Conclusion

Build an Image Graph

Idea: Connect the images that are similar enough to be cosegmented

Image Graph $G = (\mathbf{I}, \boldsymbol{E})$

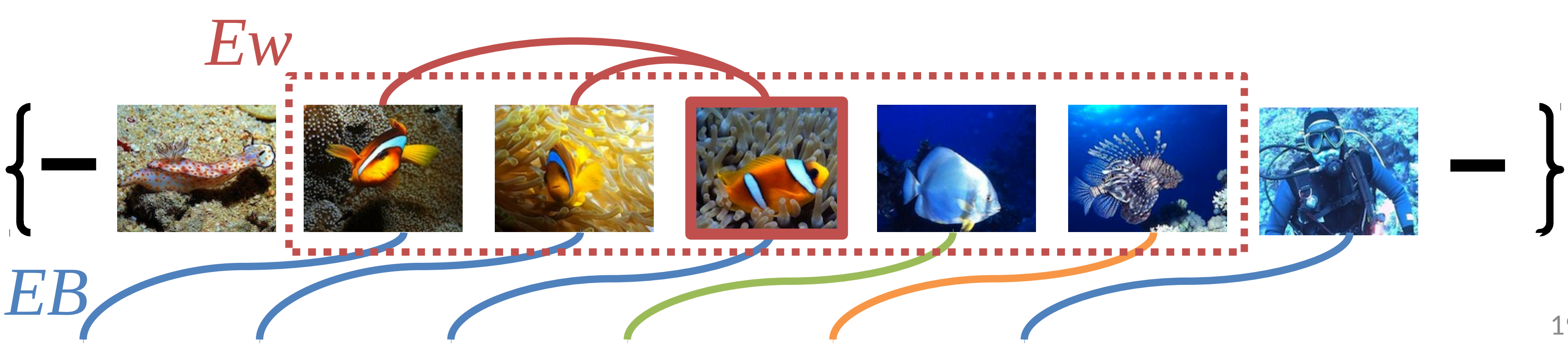
- $\mathbf{I}$: The set of images. $\boldsymbol{E}$: The set of edges.
- $\boldsymbol{E} = \boldsymbol{EB} \cup \boldsymbol{Ew}$

$\boldsymbol{EB}$: Edges between different photo streams (results of alignment)

$\boldsymbol{EW}$: Edges within a photo stream

For each image I, consider the images such that $\boxed{|t(I) - t(I_i)| \leq \delta}$

links I with the K-NN of I (EW).



Scalable Cosegmentation

Iteratively run the MFC algorithm [Kim and Xing, 2012] on the image graph

Review of MFC algorithm

- Cosegmentation: Jointly segment *M* images into *K+1* regions (*K* foregrounds (FG) + background (BG))

Foreground Modeling

Iterate

Region Assignment

- Learn appearance models of *K* FGs and BG
- Any region classifiers or their combination
- Ex. Gaussian mixture on RGB, linear SVM on SPH

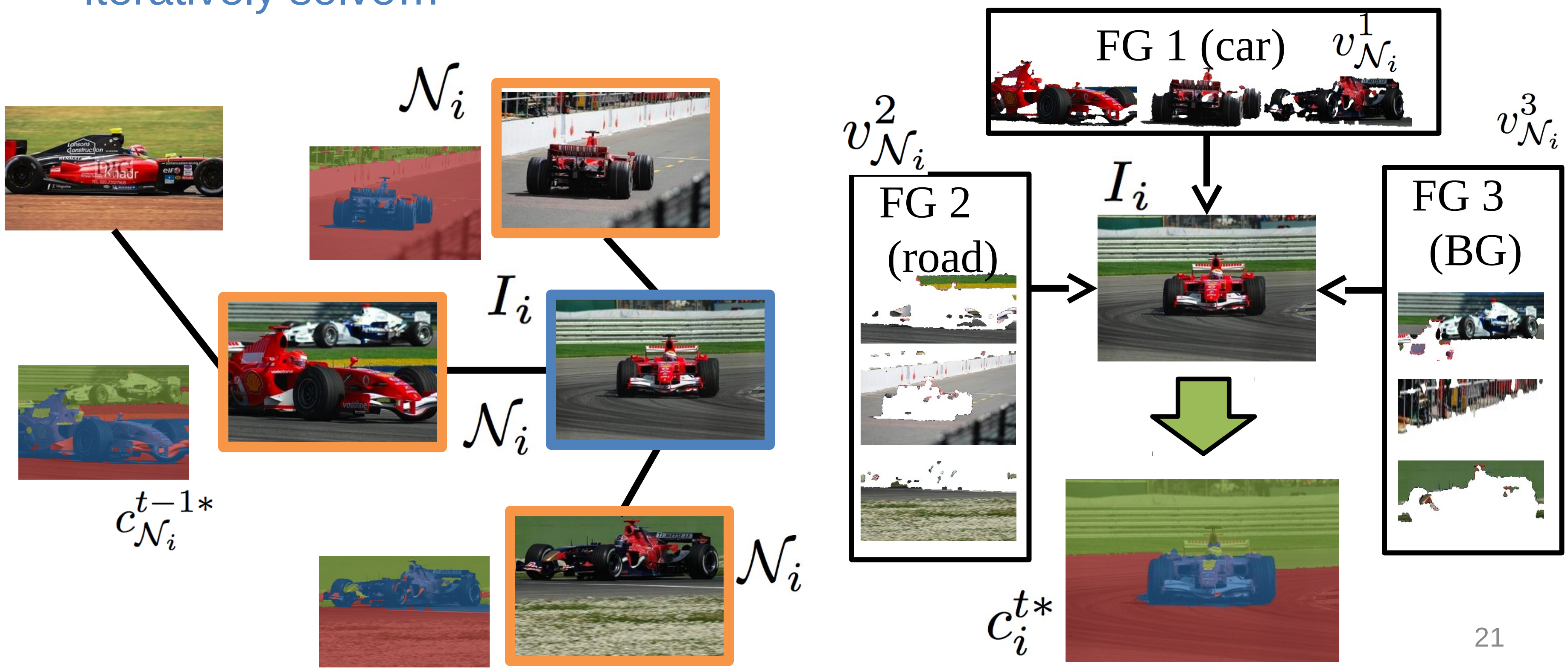
- Allocate the regions of image into one of *K* FGs or BG
- Very efficiently solve using the idea of **combinatorial auction**

Scalable Cosegmentation on Image Graph

Message-passing based optimization

- Learn FG Models from neighbors of Ii.
- Run region assignment on Ii.

Iteratively solve…



Scalable Cosegmentation on Image Graph

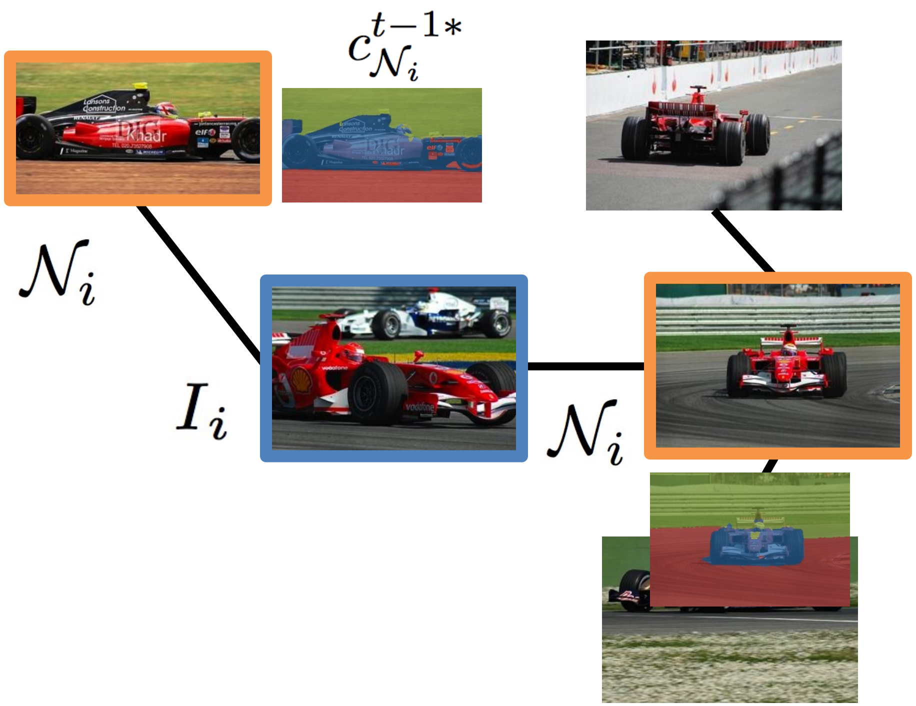
Message-passing based optimization

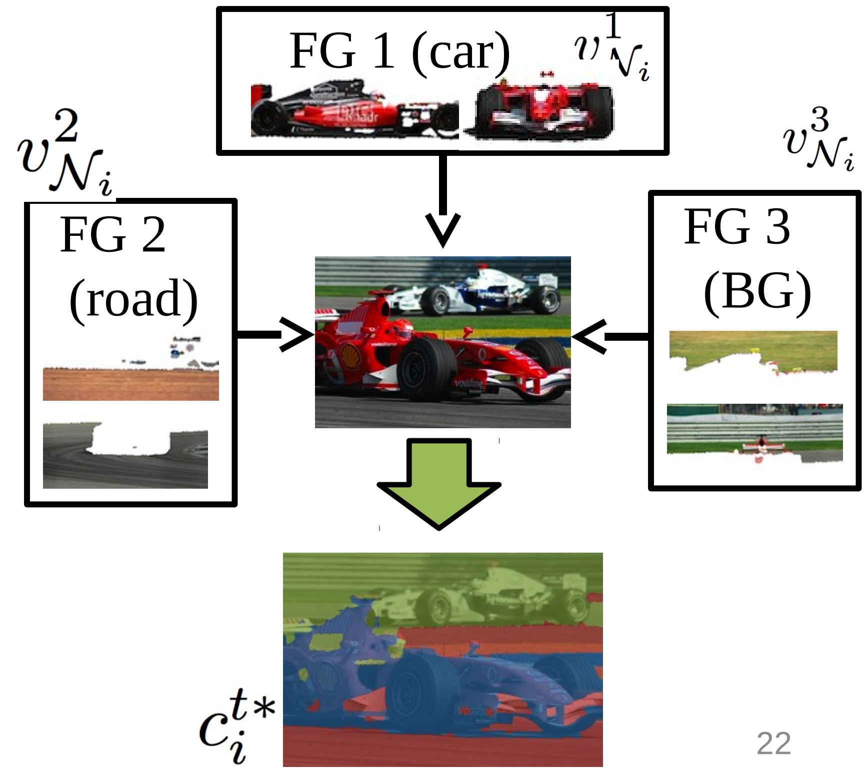
- Learn FG Models from neighbors of I_i.
- Run region assignment on I_i.

Iteratively solve…

Initialization

- Supervised: start from seed labels
- Unsupervised: use the algorithm of CoSand [Kim et al. 2011].





Outline

- Problem Statement
- Algorithm
  - Dataset and preprocessing
  - Alignment of Multiple Photo Streams
  - Large-scale Cosegmentation
- Experiments
- Conclusion

Evaluation – Two Experiments

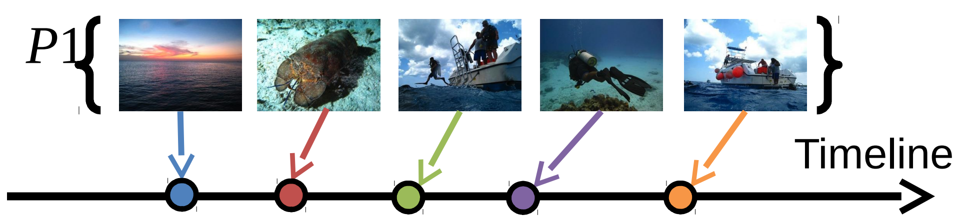
Evaluation for *Alignment*

☹ Very hard to obtain groundtruth!

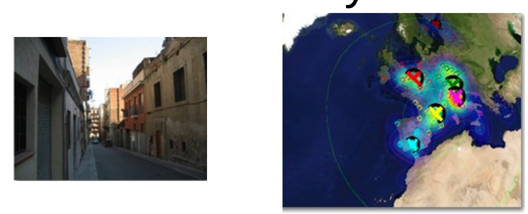
- Correspondences btw two sets of thousands of images?

Task: Temporal localization (inspired by geo-location estimation)

When are they likely to be taken?



Where is it likely to be taken?



[Hays and Efros. 2008]

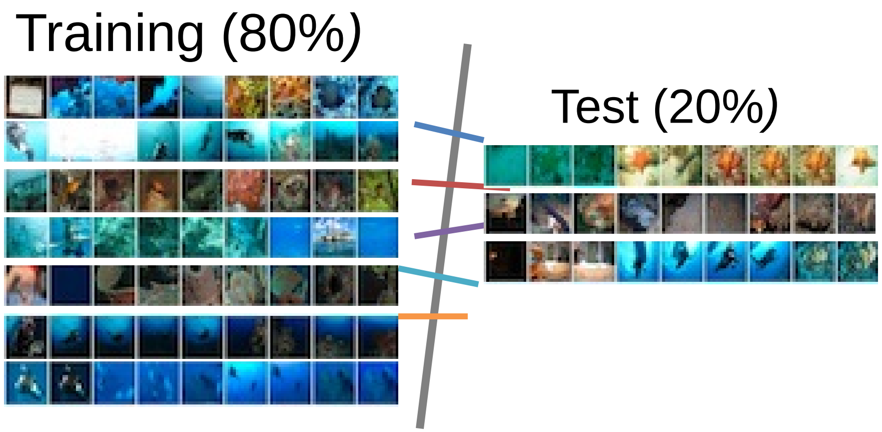
Evaluation for *cosegmentation*

Task: Foreground detection

- We manually annotate 100 images per class
- Accuracy is measured by intersection-over-union $Acc = \frac{GT_i \cap R_i}{GT_i \cup R_i}$

Evaluation of Alignment

Procedures of temporal localization



1. Given a set of photo streams, randomly split training and test sets
2. Run alignment
3. Estimate timestamps of all images in test photo streams
4. Temporal localization is correct if

$$\boxed{\left|t_{gt} - t_{est}\right| \leq \varepsilon}$$

☹ Better temporal localization ≠ Better Alignment

Baselines

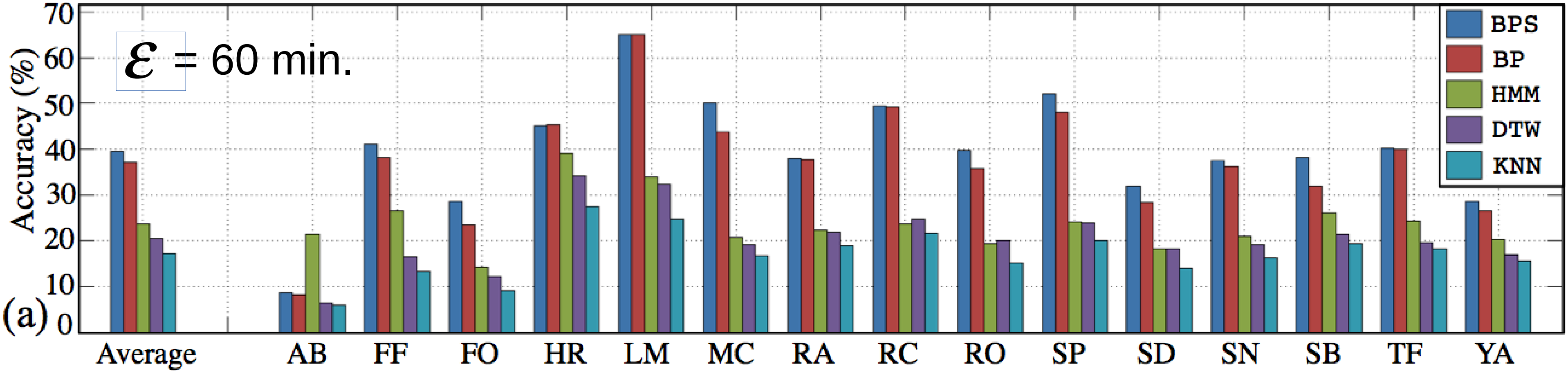
- BPS: Our Alignment + Cosegmentation
- BP: Our alignment only
- KNN: K-nearest neighbors
- HMM: Hidden Markov Models
- DTW: Dynamic Time Windows

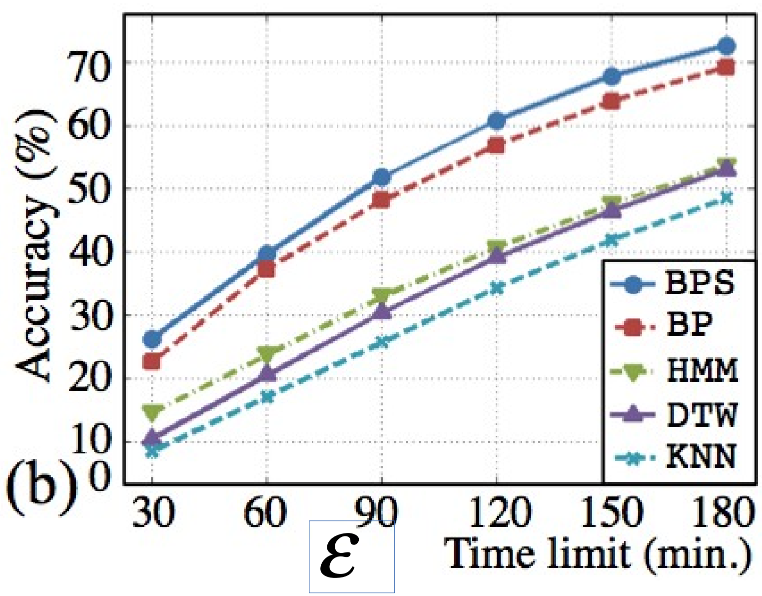
Justify closing a loop

– Image similarity only (the simplest)

Popular multiple sequence alignment

Evaluation of Alignment





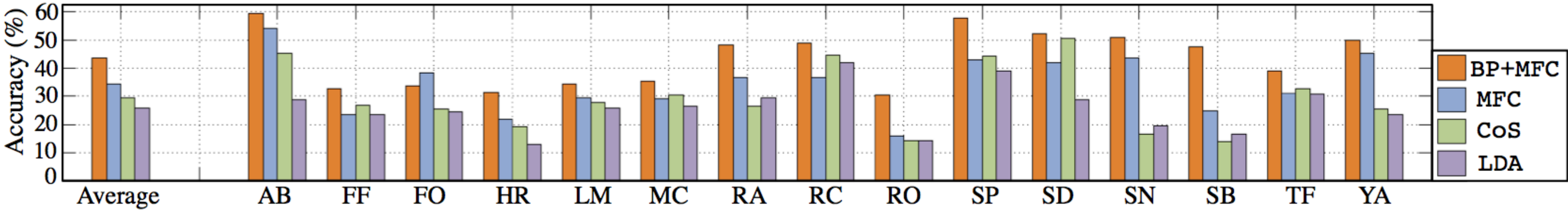
- Temporal localization is correct

if $\left| t_{gt} - t_{est} \right| \leq \varepsilon$

BPS: Our Alignment + Cosegmentation
BP: Our alignment only
KNN: K-nearest neighbors
HMM: Hidden Markov Models
DTW: Dynamic Time Windows

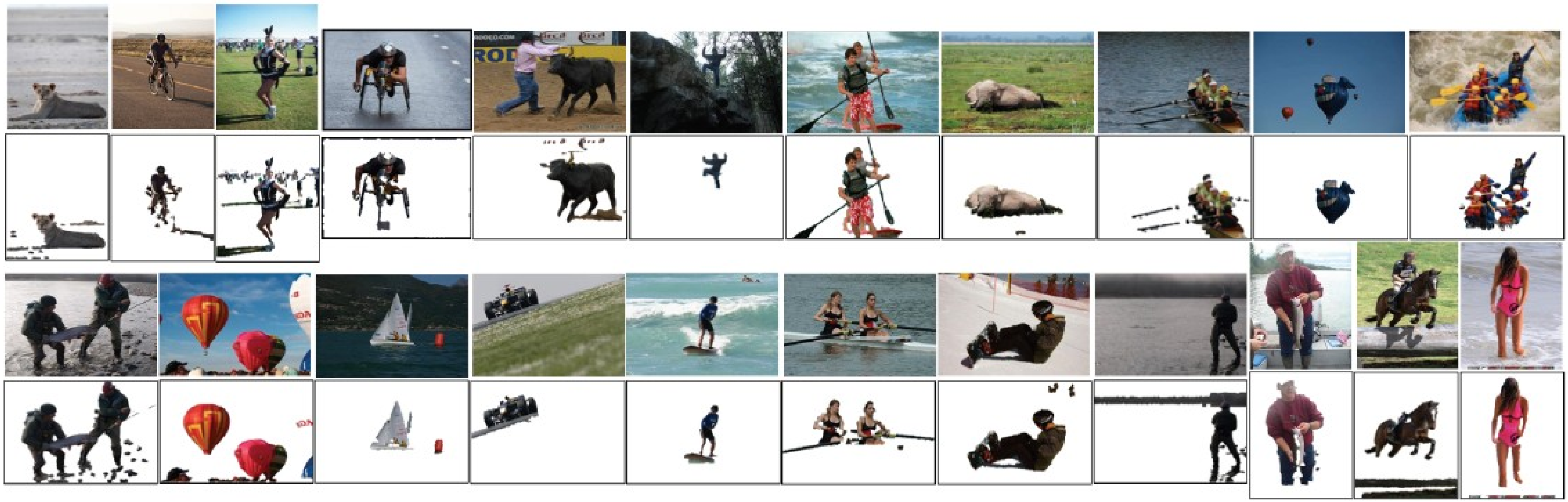
Evaluation of Cosegmentation

Task: Foreground detection



BP+MFC: (Proposed) Alignment + Cosegmentation
MFC: Our cosegmentation without alignment
COS : Submodular optimization [Kim et al. ICCV11]

Examples



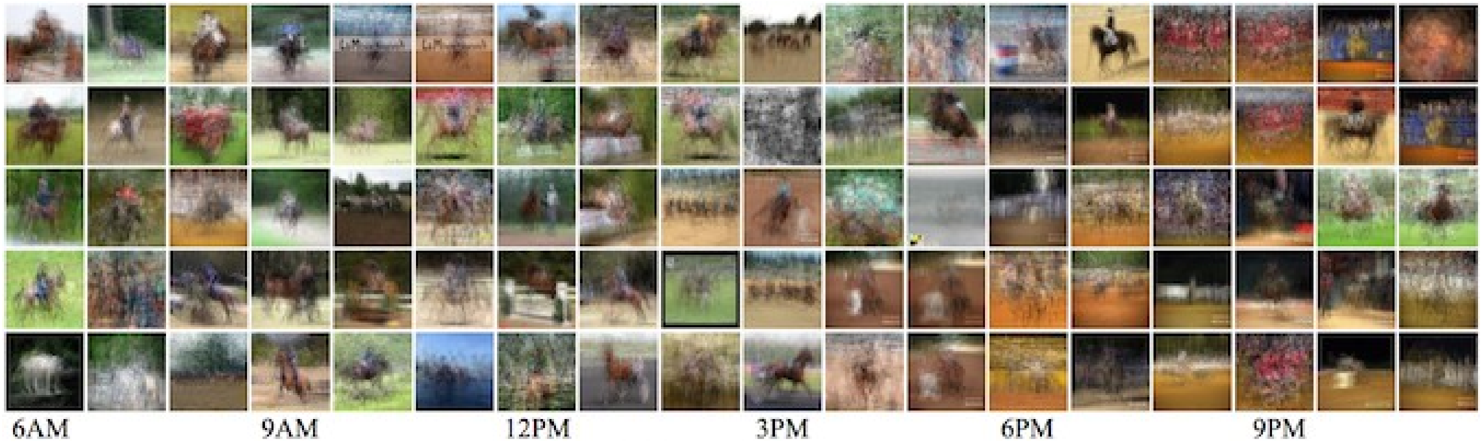
Outline

- Problem Statement
- Algorithm
  - Dataset and preprocessing
  - Alignment of Multiple Photo Streams
  - Large-scale Cosegmentation
- Experiments
- Conclusion

Conclusion

Ultimate goal: building **photo storylines** from large-scale online images

horse+riding



safari+park