Three Challenges for

Concept Pair Detection

Cees Snoek

University of Amsterdam

The Netherlands

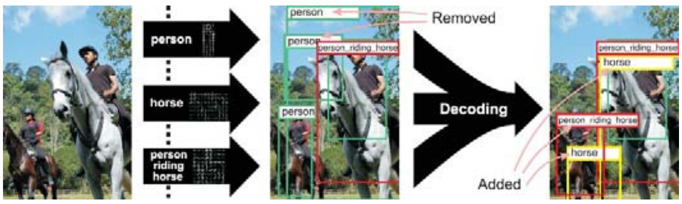
Task

- **Use case**
  - Searching for the co-occurrence of two visual concepts in unlabeled images is an important step towards answering complex user queries.

- **System task**
  - Given the test collection, master shot reference, and concept definitions, return for each concept-pair a list of at most 2,000 shot IDs from the test collection ranked according to their likeliness of containing the concept-pair.

Approaches from the literature

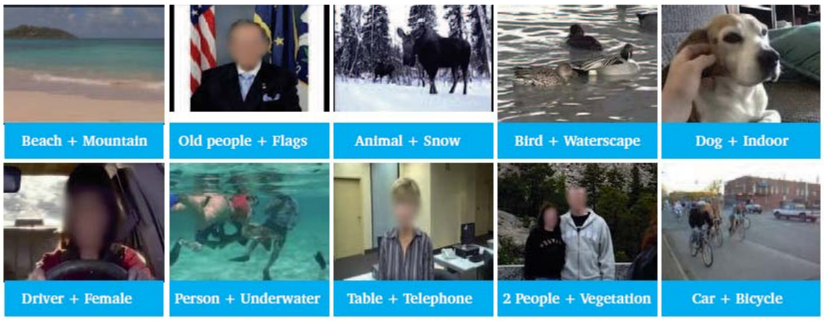
1. Combine individual concepts TRECVID 2005-present
2. Directly learning from training data Li, TMM 2012
3. Combine localized objects Farhadi, CVPR 2011



Data

- Same as regular Semantic Indexing Task
- No additional annotations provided
  - As the number of possible concept-pairs is gigantic, manually collecting training examples seems infeasible in practice.

2012 Concept Pairs



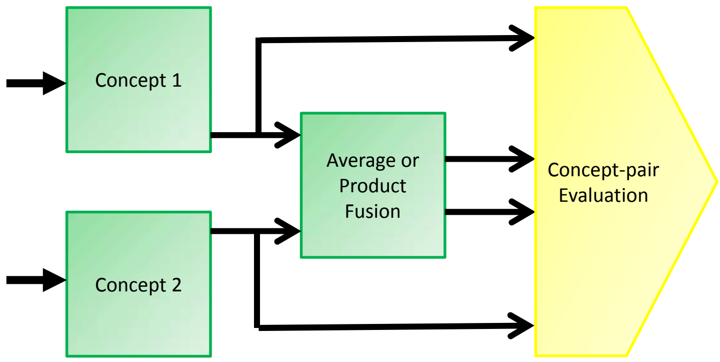
Slide credit: Silvia-Laura Pintea

Finishers

CMU	Carnegie Mellon University - Informedia
FTRDBJ	The France Telecom Orange Labs Beijing
FudaSys	Fuzhou University
ITI_CERTH	Centre for Research and Technology Hellas
TokyoTechCanon	Tokyo Institute of Technology & Canon
UvA	University of Amsterdam – MediaMill

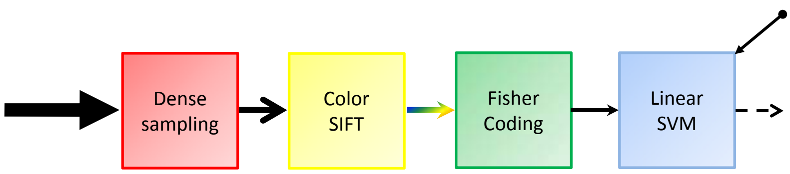
+ 4 Baseline runs

Four baseline runs



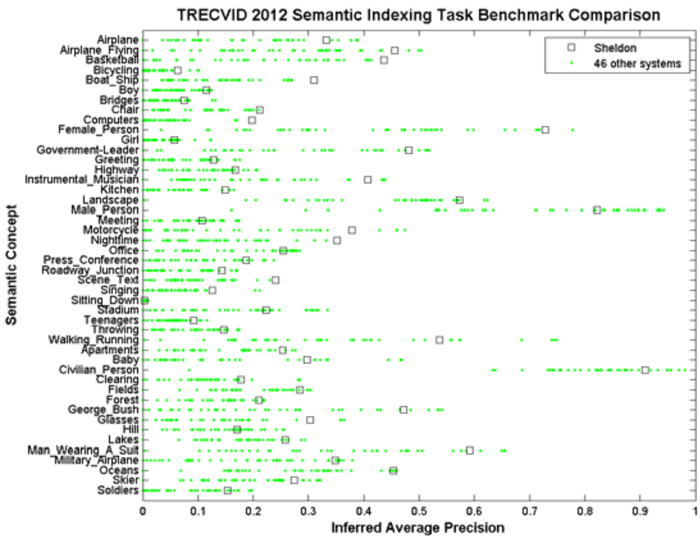
Baseline implementation

Software available for download at http://colordescriptors.com



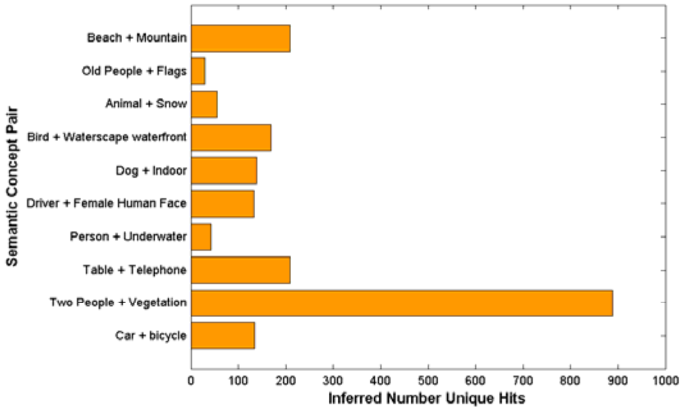
Snoek et al, TRECVID 2004-2012
Perronnin,CVPR 2010
Van de Sande et al, PAMI 2010
Van Gemert et al, PAMI 2010

Baseline in SIN task



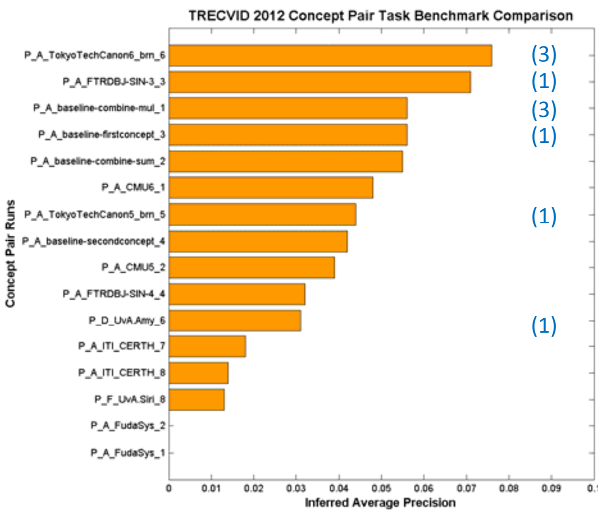
RESULTS

Most pairs are rare

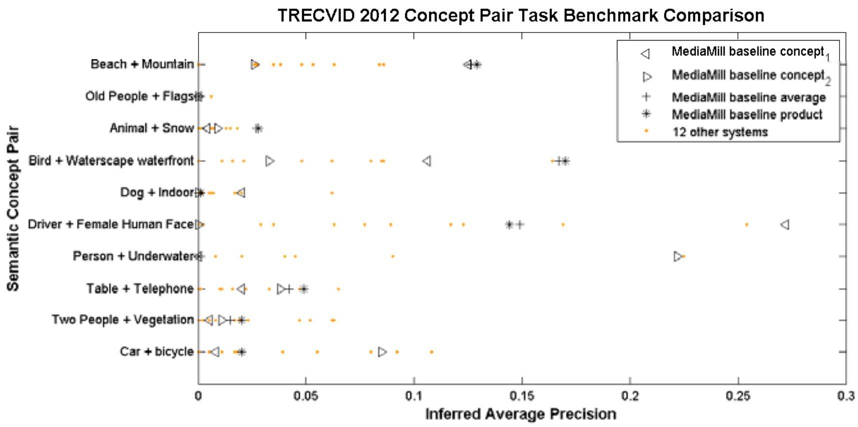


Context captured by bag-of-words no longer informative

Overall results

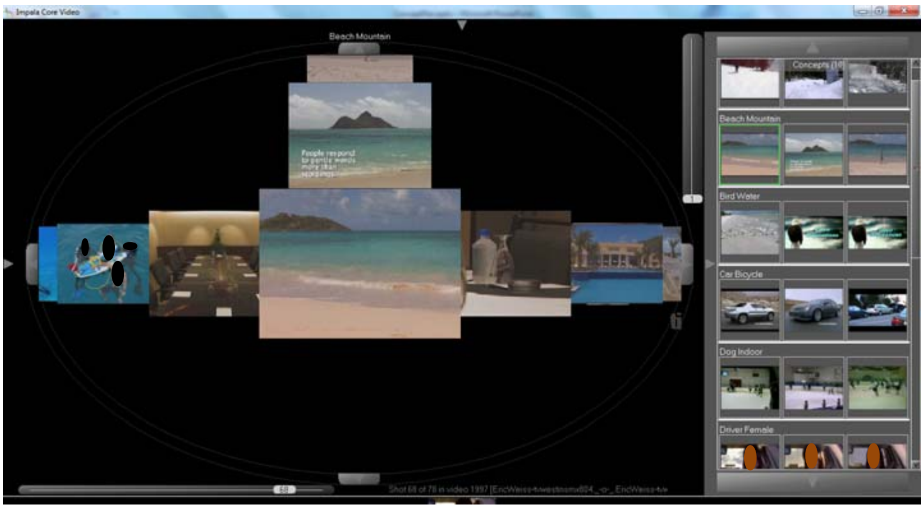


Baselines hard to beat



Euvision Technologies

Demo



PERSONAL OVERVIEW OF FINISHERS

CMU - Informedia

- **Idea:** train individual detectors and then enhance the prediction of pair-concepts using related concepts
  - Beach + Mountain: "Beach", "Mountain“, "Valleys", "Rocky_Ground", "Outdoor", "Lakes", "Islands".
- The difference between the two runs lies in the different weights in combing the final score.
  - **P_A_CMU5_2** employs the average score for each related concepts.
  - **P_A_CMU6_1** applies the score based on the concepts’ prediction accuracy in the development set.

France Telecom Orange Labs - Beijing

- **Idea:** compensate for quality/unbalance of individual detectors
- 7 fusion schemes evaluated in paper
- **P_A_FTRDBJ-SIN-3_3**
  - Fusion by confidence
- **P_A_FTRDBJ-SIN-4_4**
  - Fusion by ordered weighted averaging

FudaSys

- A 45d frequency descriptor with SVM or KNN
- **P_A_FudaSys1**
  - Weighted fusion of KNN and SVM Outputs.
- **P_A_FudaSys2**
  - Concept relation fusion of KNN and SVM outcomes.
  - Score * Prior * Conditional probability

ITI-CERTH

- **P_A_ITI-CERTH-Run 7**
  - Product fusion of concepts from primary SIN run
- **P_A_ITI-CERTH-Run 8**
  - Product fusion of concepts from their SIN run 4.

TokyoTechCanon

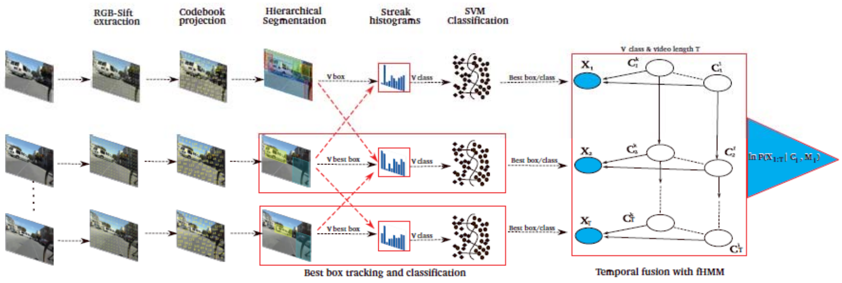
- **P_A_TokyoTechCanon5_brn_5:**
  - Average fusion of their top-performing SIN detectors
- **P_A_TokyoTechCanon6_brn_6**
  - Concept-pair classifier using SIN method. Positive examples based on intersection of individual concept annotations.

UvA - MediaMill

- **P_D_UvA.Amy_6**
  - Spatiotemporal detection for the pairs having concepts that can be localized. [Highlight follows]
- **P_F_UvA.Siri_8**
  - Identify pair-labeled videos on YouTube and learn a joint detector directly.

Spatiotemporal detection by tracking

- Selective search for individual object detection
- Foreground-background tracking of identified objects
- Factorial Hidden Markov for spatiotemporal fusion

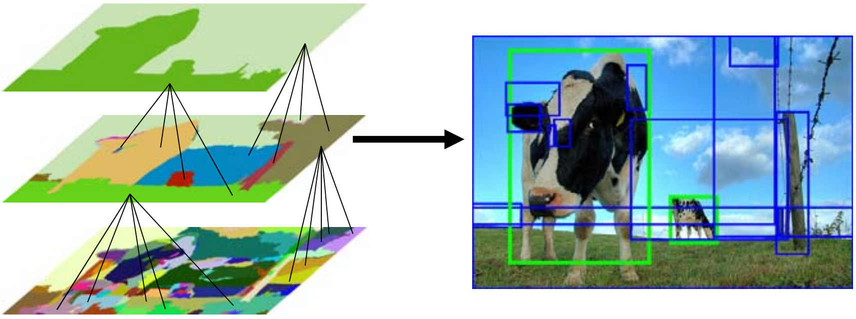


Slide credit: Silvia-Laura Pintea

Van de Sande ICCV 2011

Selective Search

- Object hypotheses based on hierarchical grouping

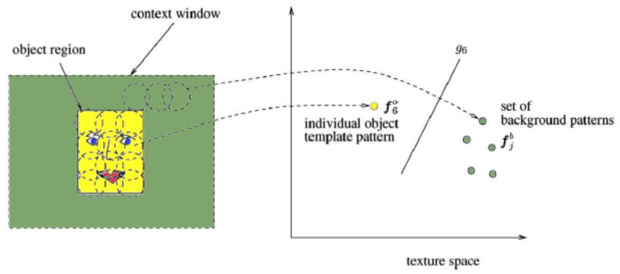


Group adjacent regions on color/texture cues

Selective Search: Example

Nguyen IJCV 2006

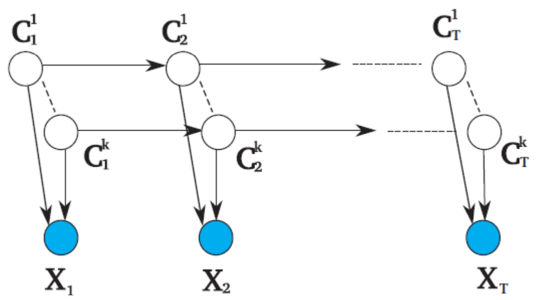
Foreground-Background tracker



- Builds *N* foreground models, 1 background model from the surrounding area
- Train *N* linear discriminants to distinguish between object pixel and background
- No assumptions regarding object appearance or motion

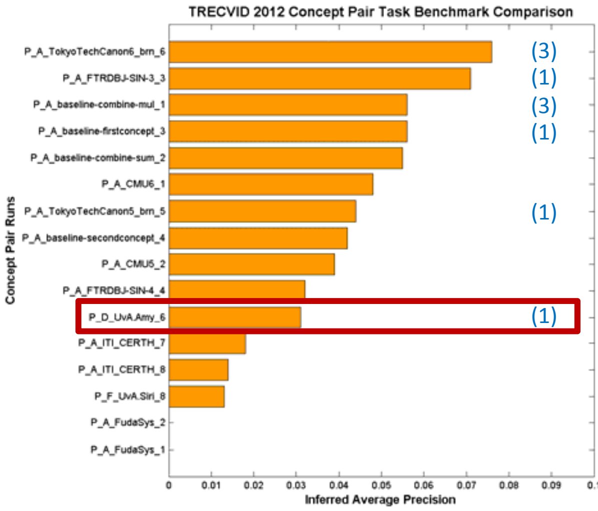
Ghahramani, ML97

Factorial HMM



- Probabilistic graphical model for sequential data
- The observations at each time step *t* depend on multiple non-independent hidden variables

Overall results



Observations

- Several runs similar to baselines
- Novelty wrt fusion, concept context, and spatiotemporal analysis
  - Mostly ‘high-level’, not so much ‘low-level’
- Complaints about lack of training data
  - Not only for pairs but also for localized detectors
  - Training from web video challenging

Conclusion

Reasonable level of participation for first pilot

Three problems waiting to be resolved

1. Manually collecting training examples is infeasible
2. Must outperform simple baselines
3. Need to consider spatiotemporal dependencies

A good challenge

Question for participants

- Shall we do it again next year?
- Should we require each group to submit a baseline?
- Should we adapt the task slightly?
  - Add more pairs?
  - More emphasis on audio concepts?
  - Shall we increase to triples?
  - Alternative evaluation metric, e.g. P@10?
- Anything else?

Contact

- dr. Cees Snoek

www.ceessnoek.info

cgmsnoek@uva.nl

twitter.com/cgmsnoek