Towards Multi-View Object Class Detection

Alexander Thomas[1], Vittorio Ferrari[2], Bastian Leibe[3], Tinne Tuytelaars[1], Bernt Schiele[4] and Luc Van Gool[1,3]

[1]KU LEUVEN

VISICS

Kasteelpark Arenberg 10,
3001 Heverlee
BELGIUM

http://www.esat.kuleuven.be/psi/visics/

[2]INRIA Rhône-Alpes

LEAR

GRAVIR-INRIA
655 avenue de l'Europe
38330 Montbonnot
FRANCE

http://lear.inrialpes.fr/

[3]ETH Zürich

BIWI

Sternwartstrasse 7
ETH Zürich
8092 Zürich
SWITZERLAND

http://www.vision.ee.ethz.ch/

[4]TU Darmstadt

Multimodal Interactive Systems

Hochschulstrasse 10
D-64289 Darmstadt
GERMANY

http://www.mis.informatik.tu-darmstadt.de/

Basis techniques

Image exploration
(Ferrari *et al., ECCV04*)

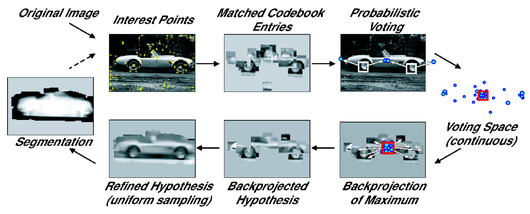
- Find dense two-view correspondences
- Connect two-view correspondences into *region tracks* = **set of dense multi-view correspondences**

Implicit Shape Model (ISM)
(Leibe & Schiele, *DAGM 2004*)

- Build a codebook and record *occurrences* where entries match on training images, each occurrence is stored together with a local segmentation
- During recognition, cast votes for the object center, based on the occurrences
- Local peaks in the voting space are *object hypotheses*.
- Contribution of each patch to the hypothesis score:

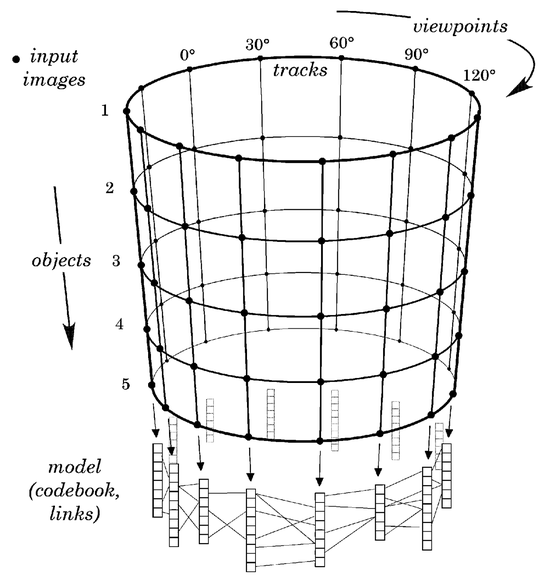
$$P(o_n, \lambda | e, l) = \sum_i P(o_n, \lambda | c_i, l)\, p(c_i | e)$$

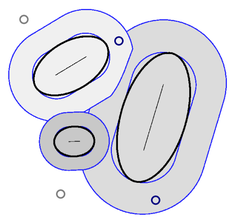
- Using the local segmentations, derive figure/ground probability
- Perform MDL selection procedure between hypotheses, using P(figure/ground) to remove overlapping hypotheses



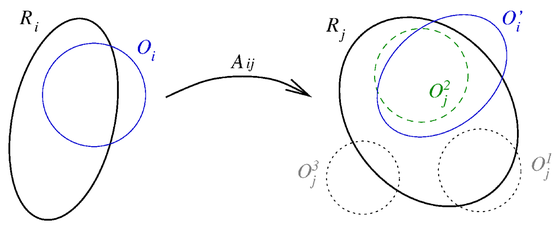
Building a multi-view ISM model

- Training data = (sparse) *M*x*N* matrix (*M* object instances, *N* viewpoints)
- Find region tracks across viewpoints, train ISMs across object instances
- Establish **'activation links'** between the occurrences of different codebooks, using the region tracks:
  - Find closest region to each occurrence, and project it into other views using affine transform between regions
  - Determine which occurrences in the other view match with the transformed occurrence
  - Establish links between matching occurrences and the original occurrence





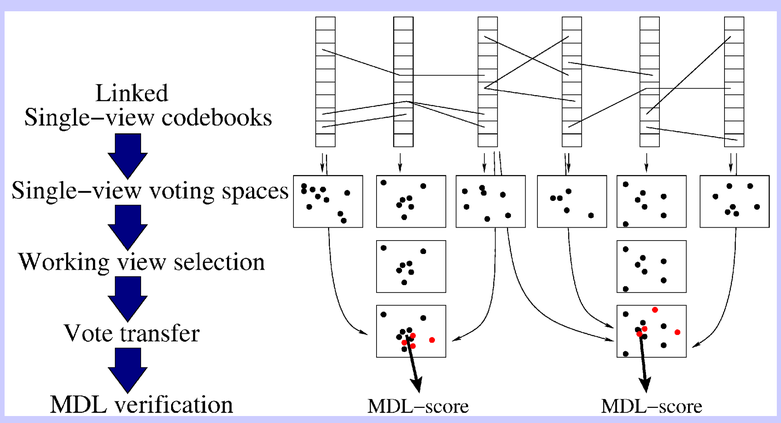
"attraction zones" around regions



Establishing a link between O_i and O_j^2

Multi-view recognition

System overview

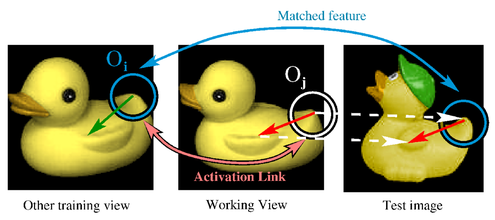


Selecting *'Working Views'*

- Determine which views are likely to correspond with the pose(s) of object(s) in the test image
- A correct hypothesis will cause a strong hypothesis in its view's voting space, and quite strong hypotheses at a similar position in neighboring views.
  → Cluster hypotheses across viewpoints and choose strongest clusters

Transferring votes across views

- Each working view's voting space is *augmented* with votes, transferred from other views:
- If there is an activation link from a matching codebook entry in another view towards this WV, it will cause an additional vote in this WV.



- After transferring votes, hypotheses are re-detected in the augmented voting spaces with a mean-shift procedure, and the MDL procedure is performed on these hypotheses

- After vote transfer, the contribution of a patch e to an object hypothesis is:

$$P(o_n, \lambda | e, l) = \sum_k P(o_n, \lambda | c_k^j, l)\, p(c_k^j | e) + \sum_k \sum_l P(o_n, \lambda | c_k^j, c_l^i, l)\, p(c_l^i | e)$$

Benefits of vote transfer

- A local object part can vary its (apparent) pose, relative to the entire object, and still contribute to the detection (movement of such parts is restricted, due to the way the position of transferred votes is calculated)
- Example: the front wheel of a motorbike can be turned independently from the body of the motorbike
- When the viewpoint falls in-between training viewpoints, vote transfer allows to accumulate *("interpolate")* evidence from both neighboring views, relaxing the constraints on the training images

Experiments and results

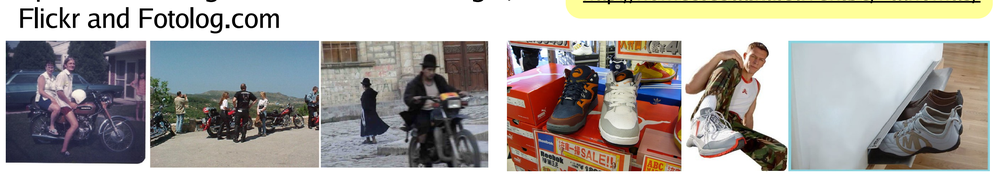
Training images

- 30 motorbikes (16 viewpoints, on average 11 views per bike)
- 42 sports shoes (16 viewpoints, all shoes have all views)

Test images

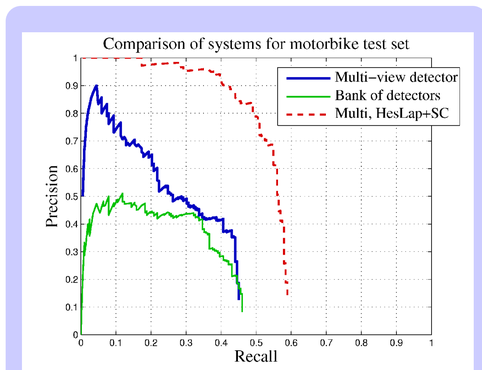
- VOC 2005 challenge, motorbikes-test2 subset
- Sports shoe images collected from Google, Flickr and Fotolog.com

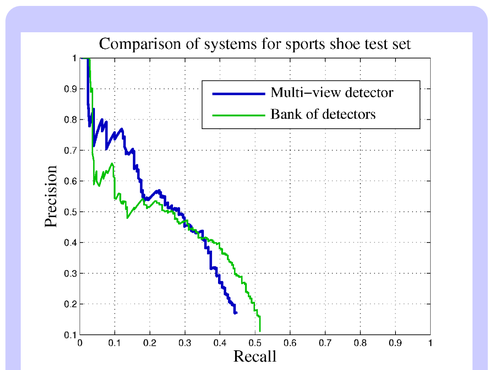
Data available at http://homes.esat.kuleuven.be/~athomas/



Results

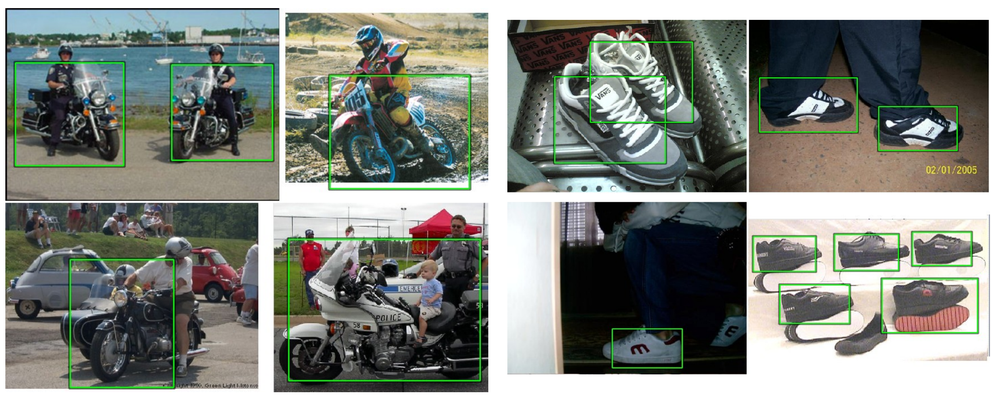
- Our system vs. a battery of independent detectors with identical parameters
- Next to performance improvement, also significant speed-up (2x to 3x)





The multi-view system vs. the battery

Some correct detections

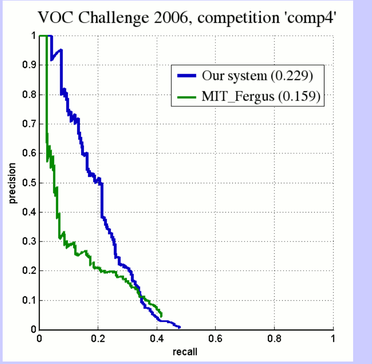


Some missing and false detections

- Either due to too large difference with training instances, or poor quality/contrast

Results on Pascal VOC 2006

- Motorbikes localization task with own training data



Future work

- Extend the MDL stage to work across different viewpoints (currently only per view)
- Remove the need for training views to be ordered according to viewpoint
- Experiment on other object classes