# Accurate Object Localization with Shape Masks

Marcin Marszałek  Cordelia Schmid

LEAR, INRIA / LJK, Grenoble, France

CVPR 2007

# Outline

# Outline

# Object class localization

**Given an unseen image and a known object class. . .**

. . . decide where in the image the object of this class is

Open questions:

- How should we answer the question “where”?
  - Center of the object
  - Bounding box
  - Object outline
- What if there is no object or if there are several of them?
  - The concept of “object” is crucial

# Hough space voting with fragment backprojection

## Leibe, Seemann and Schiele [CVPR'05], Opelt, Pinz and Zisserman [CVPR'06]

- Hough space implies low-dimensional localization hypotheses, so parametrized shapes have to be used
- Articulated objects and multiple viewpoints may be confused, backprojection suffers from global consistency problems
- We replace the Hough space with a high-dimensional hypothesis space based on shape masks

Leibe et al.

# Pixel annotation and object segmentation

## Shotton, Winn, Rother and Criminisi [ECCV'06], Todorovic and Ahuja [CVPR'06]

- The notion of object concept is necessary to separate multiple instances
- Segmentation does not include occluded object parts, but in fact the object is there
- We aim to separate object instances and to determine approximate object outlines

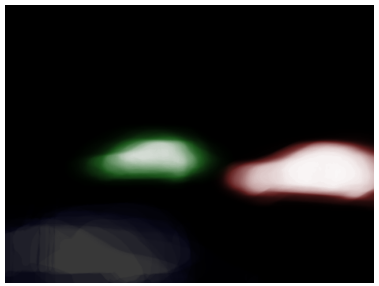

Shotton at al. [ECCV'06]

Todorovic and Ahuja [CVPR'06]

# Our approach: Using shape masks as hypotheses

- Local features and shape masks can be used to cast localization hypotheses [CVPR'06]


- We propose to evaluate the hypotheses when cast to clean the hypothesis space before looking for maxima
- We show how to cluster the hypotheses to find maxima in the high-dimensional hypothesis space

# Features of our approach

- Object localization with approximate outlines (rich answers)
- Implicit handling of multiple object aspects (detection during training and combination during testing)
- Detection of multiple object instances per image
- Segmentation of occluded object parts

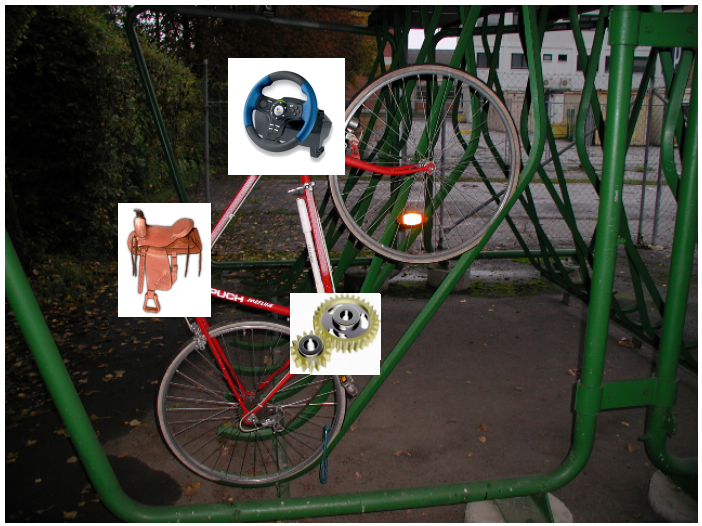
# Outline

# Casting localization hypotheses

- To compute features, Harris-Laplace and Laplacian interest points are detected and described with SIFT
- For each feature *i* the rectification matrix $\theta_i$ is saved and for training features a pointer to the shape mask $\zeta_i$ is kept
- By matching the test features with the training features, localization hypotheses in the form of shape masks can be generated
- The mask $\zeta_i$ can be projected to the reference frame of test feature *j* by composing it with the transformation matrix $P_{ij} = \theta_i^{-1}\theta_j$

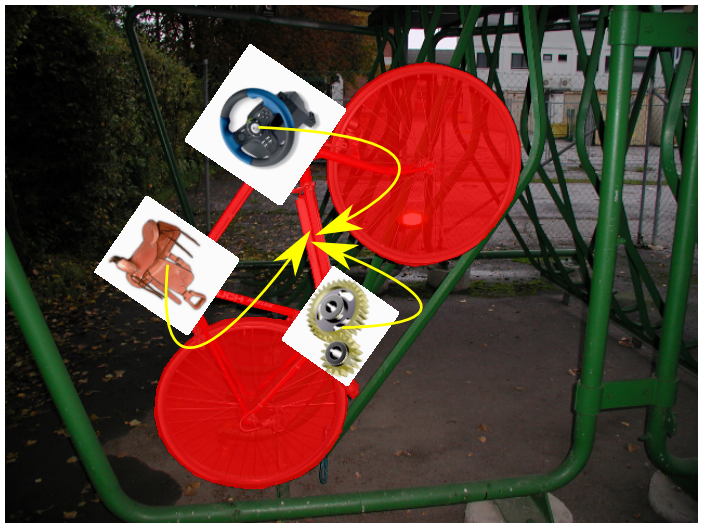
# Casting localization hypotheses

- To compute features, Harris-Laplace and Laplacian interest points are detected and described with SIFT
- For each feature *i* the rectification matrix $\theta_i$ is saved and for training features a pointer to the shape mask $\zeta_i$ is kept
- By matching the test features with the training features, localization hypotheses in the form of shape masks can be generated
- The mask $\zeta_i$ can be projected to the reference frame of test feature *j* by composing it with the transformation matrix $P_{ij} = \theta_i^{-1}\theta_j$



# Casting localization hypotheses

- To compute features, Harris-Laplace and Laplacian interest points are detected and described with SIFT
- For each feature *i* the rectification matrix $\theta_i$ is saved and for training features a pointer to the shape mask $\zeta_i$ is kept
- By matching the test features with the training features, localization hypotheses in the form of shape masks can be generated
- The mask $\zeta_i$ can be projected to the reference frame of test feature *j* by composing it with the transformation matrix $P_{ij} = \theta_i^{-1}\theta_j$

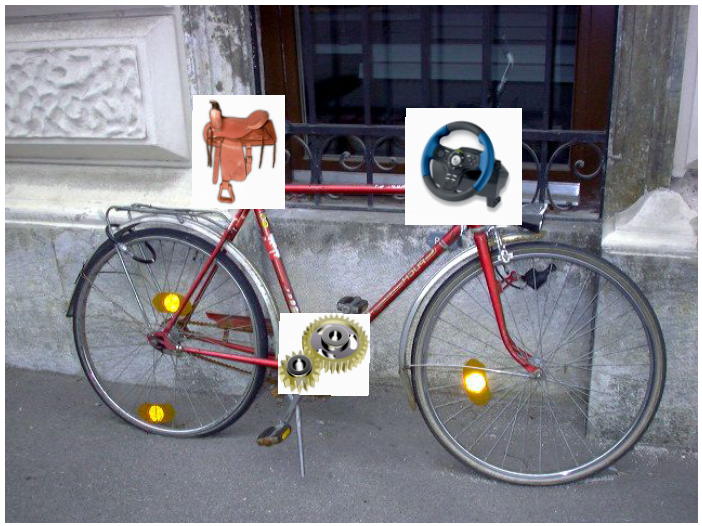


# Casting localization hypotheses

- To compute features, Harris-Laplace and Laplacian interest points are detected and described with SIFT
- For each feature *i* the rectification matrix $\theta_i$ is saved and for training features a pointer to the shape mask $\zeta_i$ is kept
- By matching the test features with the training features, localization hypotheses in the form of shape masks can be generated
- The mask $\zeta_i$ can be projected to the reference frame of test feature *j* by composing it with the transformation matrix $P_{ij} = \theta_i^{-1}\theta_j$

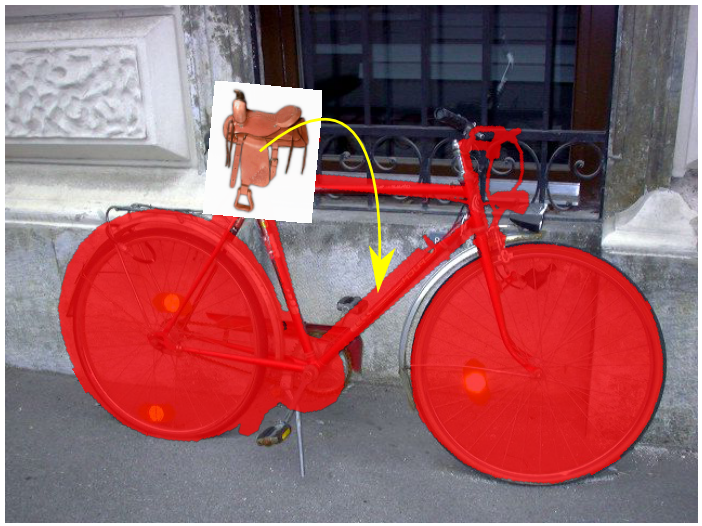
# Casting localization hypotheses

- To compute features, Harris-Laplace and Laplacian interest points are detected and described with SIFT
- For each feature *i* the rectification matrix $\theta_i$ is saved and for training features a pointer to the shape mask $\zeta_i$ is kept
- By matching the test features with the training features, localization hypotheses in the form of shape masks can be generated
- The mask $\zeta_i$ can be projected to the reference frame of test feature *j* by composing it with the transformation matrix $P_{ij} = \theta_i^{-1}\theta_j$



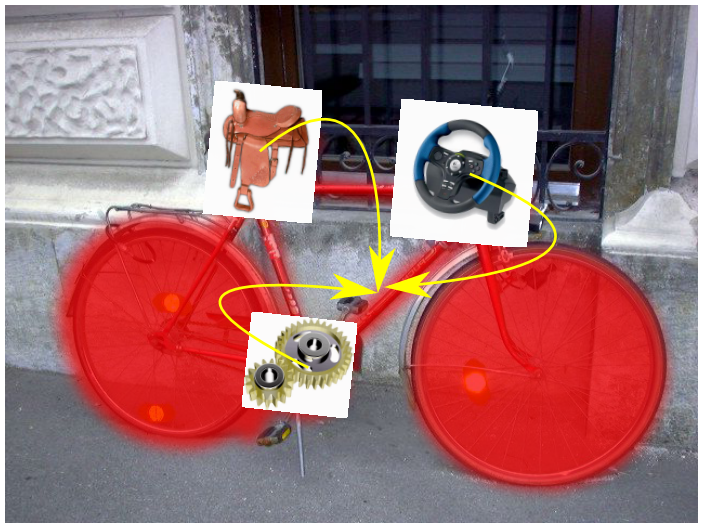
# Casting localization hypotheses

- To compute features, Harris-Laplace and Laplacian interest points are detected and described with SIFT
- For each feature *i* the rectification matrix $\theta_i$ is saved and for training features a pointer to the shape mask $\zeta_i$ is kept
- By matching the test features with the training features, localization hypotheses in the form of shape masks can be generated
- The mask $\zeta_i$ can be projected to the reference frame of test feature *j* by composing it with the transformation matrix $P_{ij} = \theta_i^{-1}\theta_j$



# Casting localization hypotheses

- To compute features, Harris-Laplace and Laplacian interest points are detected and described with SIFT
- For each feature *i* the rectification matrix $\theta_i$ is saved and for training features a pointer to the shape mask $\zeta_i$ is kept
- By matching the test features with the training features, localization hypotheses in the form of shape masks can be generated
- The mask $\zeta_i$ can be projected to the reference frame of test feature *j* by composing it with the transformation matrix $P_{ij} = \theta_i^{-1}\theta_j$



# Similarity between shape masks

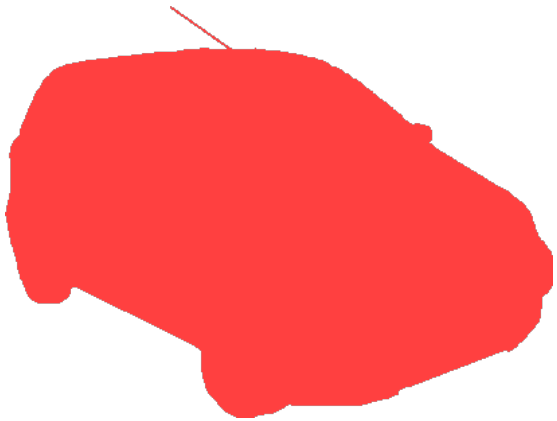
- A shape mask $S : \mathbf{R}^2 \rightarrow \mathbf{R}$ is a natural generalization of the discrete binary segmentation mask $S_b : \mathbf{Z}^2 \rightarrow \{0, 1\}$
- A commonly used overlap area measure

$$o_b(Q_b, R_b) = \frac{|Q_b^1 \cap R_b^1|}{|Q_b^1 \cup R_b^1|} = \frac{\sum \min(Q_b, R_b)}{\sum \max(Q_b, R_b)}$$

  is generalized to a mask overlap similarity measure

$$o_s(Q, R) = \frac{\int \min(Q, R)}{\int \max(Q, R)}$$

- We define a similarity measure between shape masks $\zeta_i$ and $\zeta_j$ associated with features $i$ and $j$ as $o(i, j) = o_s(\zeta_i \circ P_{ij}, \zeta_j)$



# Similarity between shape masks

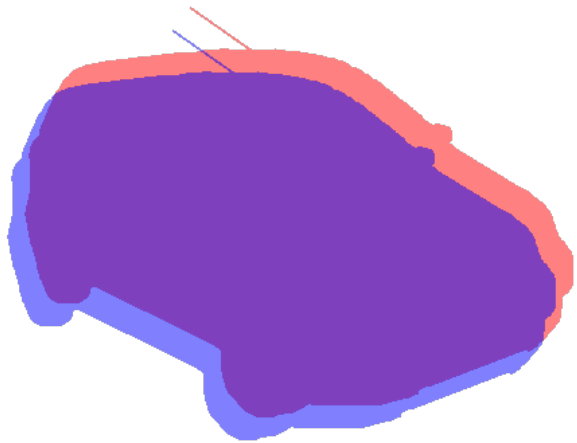
- A shape mask $S : \mathbf{R}^2 \rightarrow \mathbf{R}$ is a natural generalization of the discrete binary segmentation mask $S_b : \mathbf{Z}^2 \rightarrow \{0, 1\}$
- A commonly used overlap area measure

$$o_b(Q_b, R_b) = \frac{|Q_b^1 \cap R_b^1|}{|Q_b^1 \cup R_b^1|} = \frac{\sum \min(Q_b, R_b)}{\sum \max(Q_b, R_b)}$$

  is generalized to a mask overlap similarity measure

$$o_s(Q, R) = \frac{\int \min(Q, R)}{\int \max(Q, R)}$$

- We define a similarity measure between shape masks $\zeta_i$ and $\zeta_j$ associated with features $i$ and $j$ as $o(i, j) = o_s(\zeta_i \circ P_{ij}, \zeta_j)$



# Similarity between shape masks

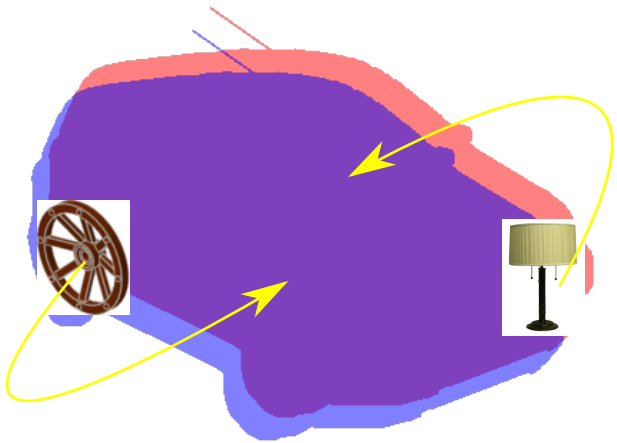
- A shape mask $S : \mathbf{R}^2 \rightarrow \mathbf{R}$ is a natural generalization of the discrete binary segmentation mask $S_b : \mathbf{Z}^2 \rightarrow \{0, 1\}$
- A commonly used overlap area measure

$$o_b(Q_b, R_b) = \frac{|Q_b^1 \cap R_b^1|}{|Q_b^1 \cup R_b^1|} = \frac{\sum \min(Q_b, R_b)}{\sum \max(Q_b, R_b)}$$

  is generalized to a mask overlap similarity measure

$$o_s(Q, R) = \frac{\int \min(Q, R)}{\int \max(Q, R)}$$

- We define a similarity measure between shape masks $\zeta_i$ and $\zeta_j$ associated with features $i$ and $j$ as $o(i, j) = o_s(\zeta_i \circ P_{ij}, \zeta_j)$

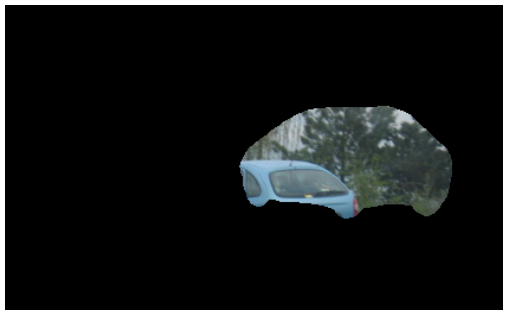


# Evaluation of shape masks

- A bag-of-features representation can be computed for the image part covered by a shape mask
- A non-linear SVM classifier with $\chi^2$ kernel is trained to distinguish between object and background

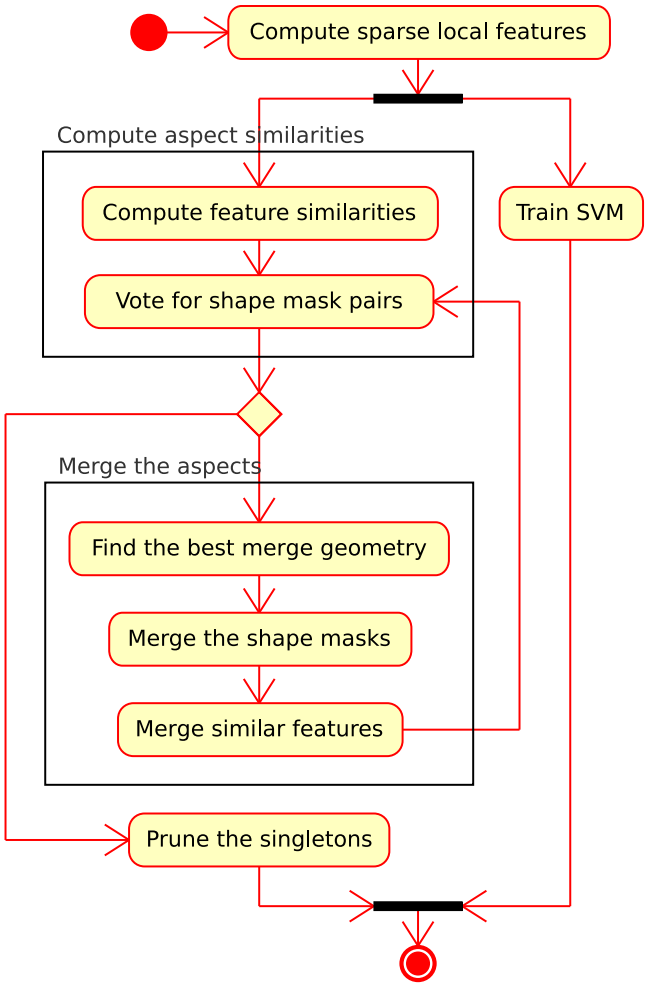
image

positive



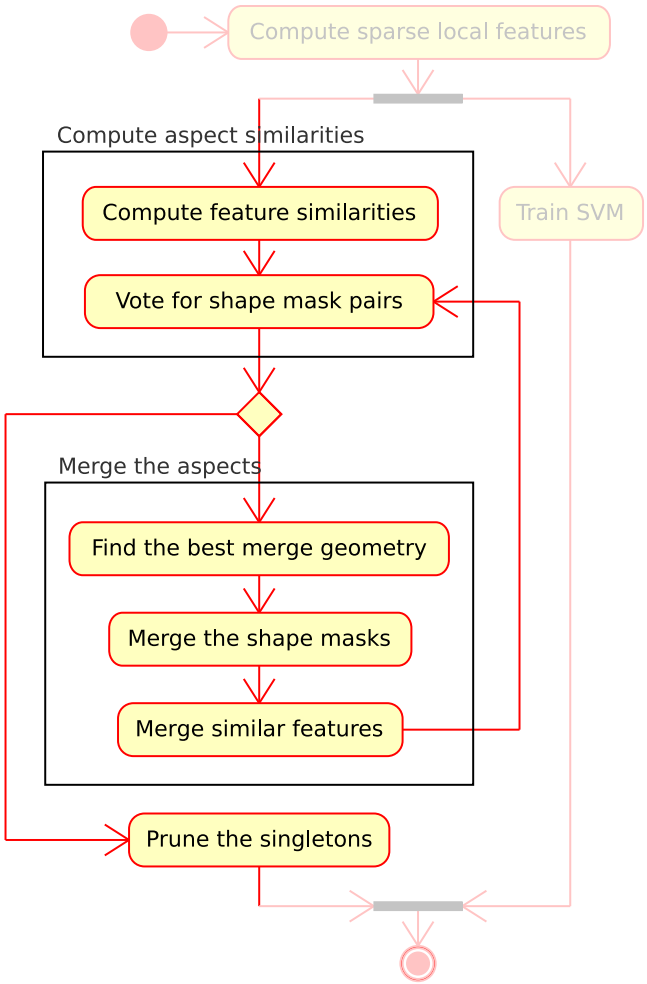
negative

# The training procedure



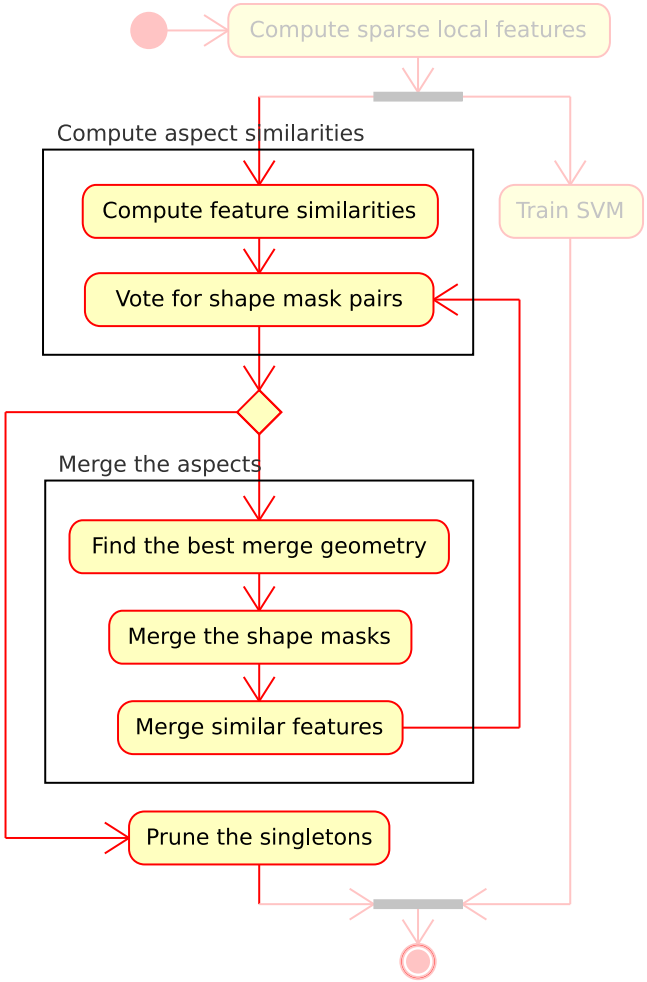

1. Sparse local features are computed for the training images
2. Then
   - Object aspects are learned by agglomerative clustering of object shape masks
   - An object classifier is trained with segmented objects (positive samples) and background (negative samples)

# Agglomerative aspect clustering - main loop



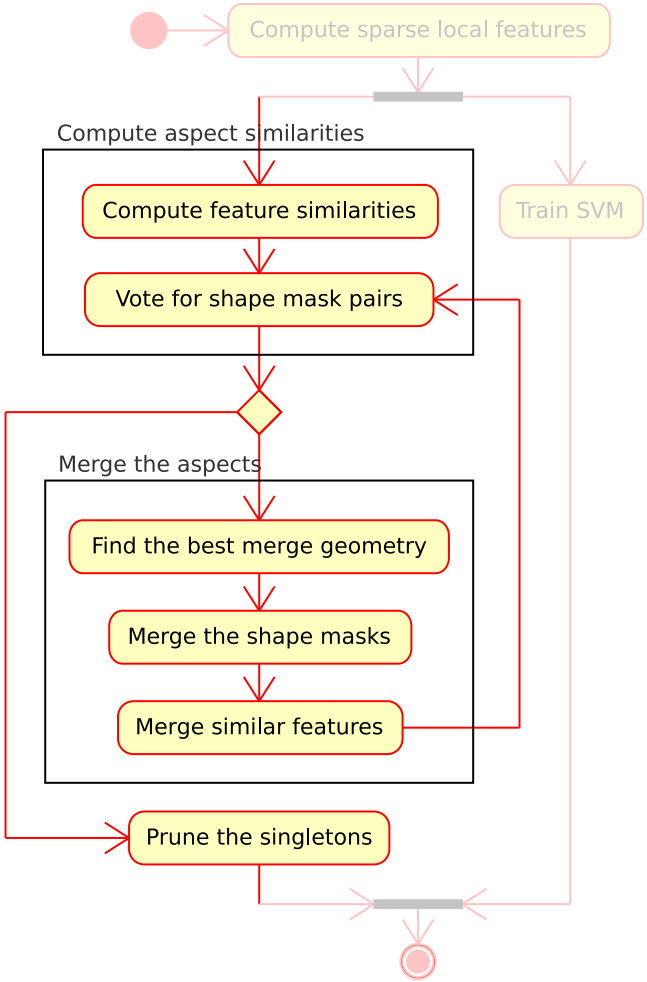

1. For each pair of similar features, the similarity between the associated shape masks is computed
2. Feature pairs with similar shape masks (similarity above $T = 0.85$) vote for shape mask pairs to get merged
3. The shape mask pair with the highest number of votes is merged according to the best feature pair match
4. The features associated with the masks are combined

# Agglomerative aspect clustering - singleton pruning




- Aspect merging is repeated until no more aspects are found to be similar enough
- After the agglomerative aspect clustering is over, singletons (outliers) can be discarded

# Agglomerative aspect clustering - singleton pruning

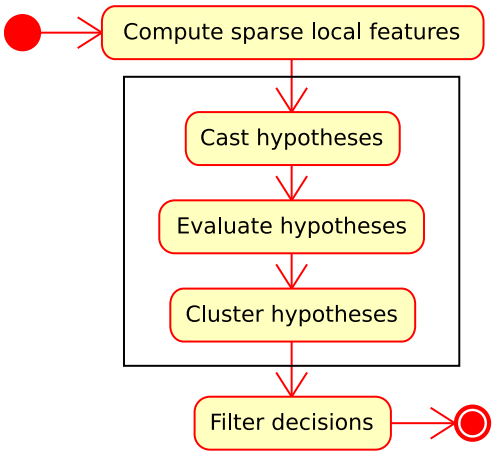



- Aspect merging is repeated until no more aspects are found to be similar enough
- After the agglomerative aspect clustering is over, singletons (outliers) can be discarded

**Demo**

Example for aspect clustering

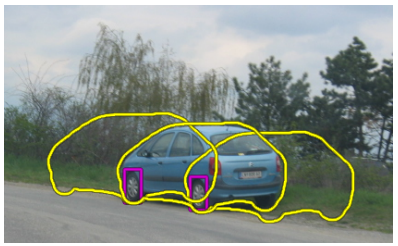
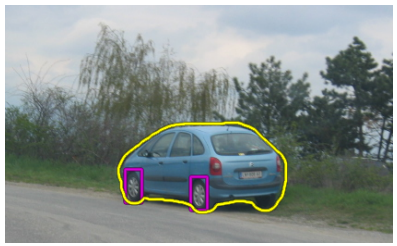
# The recognition procedure



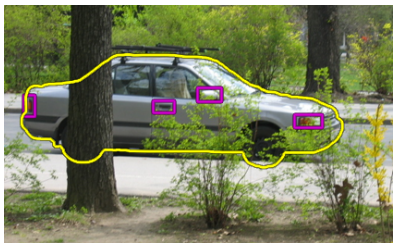

1. Each local feature in a test image is matched with similar training features and a localization hypothesis is generated for each pair
2. Generated hypotheses are evaluated on-line with the object classifier and a score is assigned
3. The hypotheses are clustered on-line (up to $L = 100$ hypotheses are kept). Similar hypotheses are merged (similarity threshold $U = 0.7$) and the scores are added. Non-promising hypotheses (with the lowest score) are dropped.
4. Overlapping hypotheses are removed

# Main points of our framework

- Ambiguities introduced by local features may generate false hypotheses
  - Hypothesis evaluation helps to avoid them in our framework
- Occlusion weakens the discriminative classifier response and the object may be missed
  - This is reduced in our framework by collecting the local evidence provided by consistent features

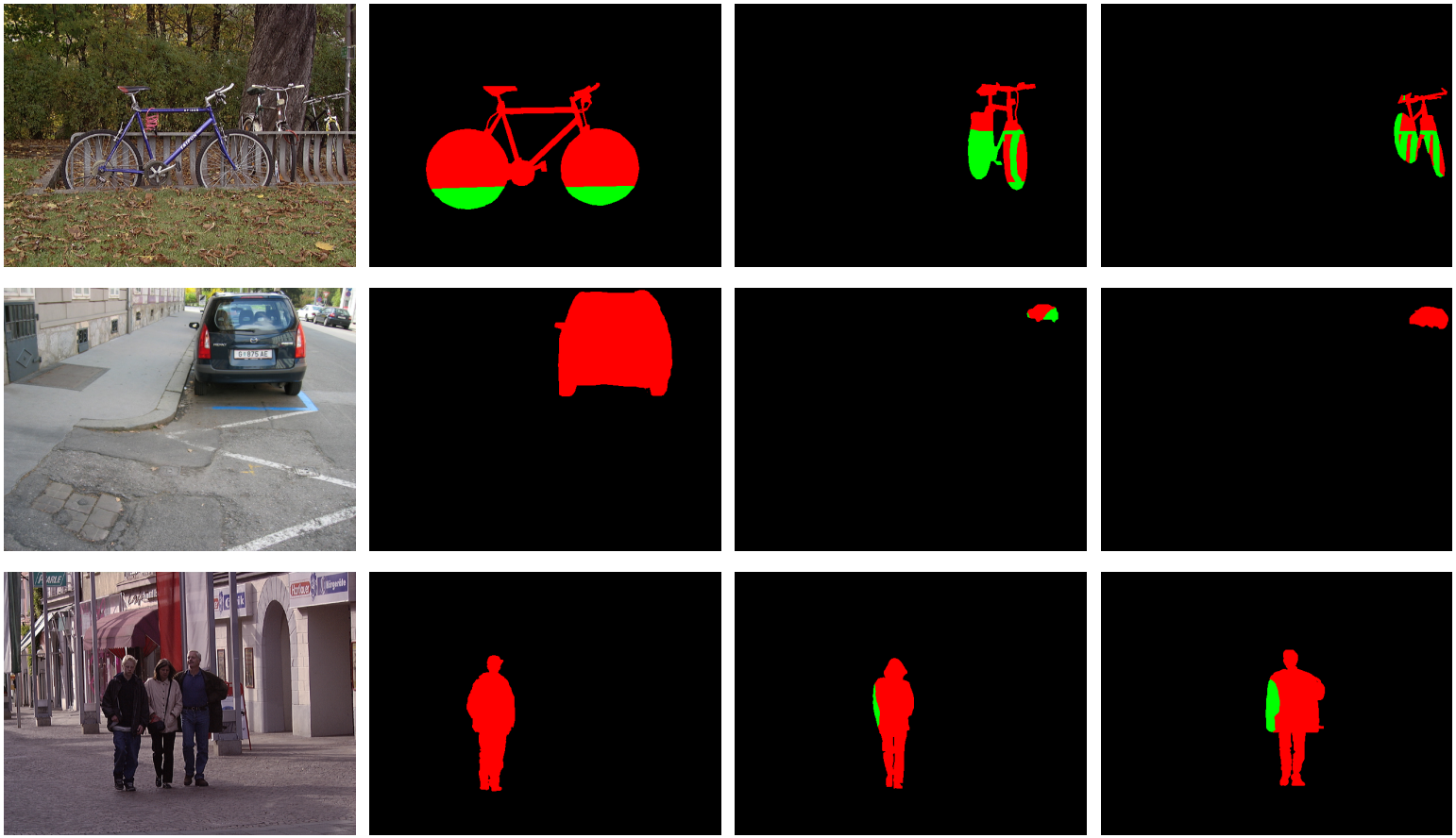
Hypothesis evaluation

Evidence collection

# Outline

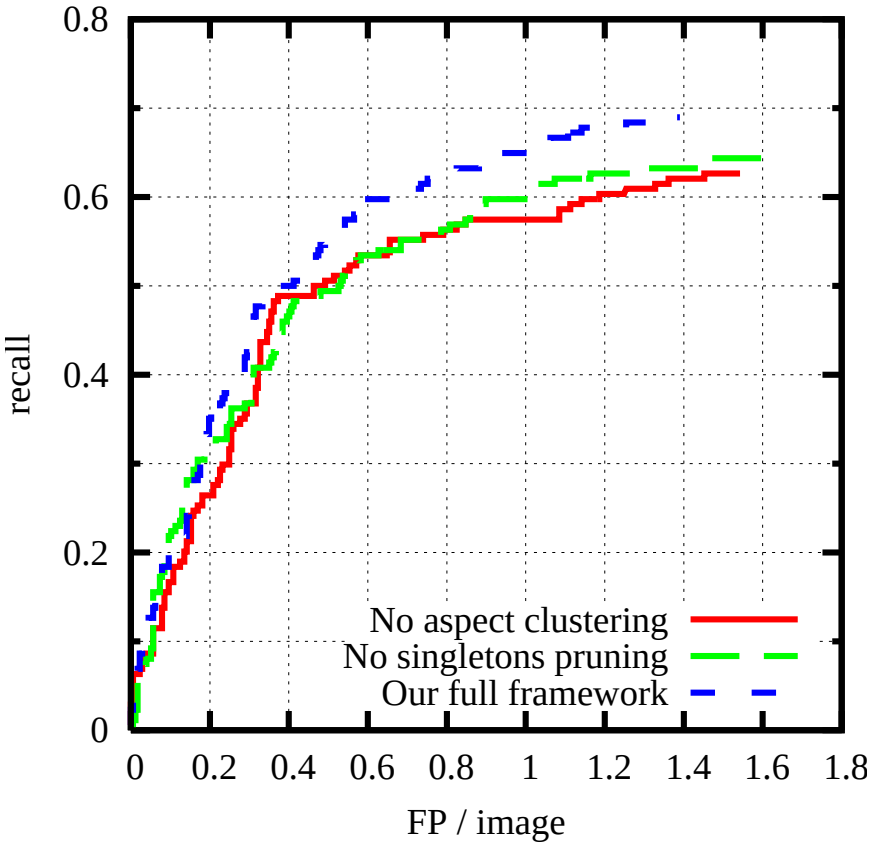
# INRIA Annotations for Graz-02 (IG02)



http://lear.inrialpes.fr/data/

# Impact of aspect clustering

- Aspect clustering improves the recognition results
- Singleton pruning leads to further improvement
- Furthermore, both of them improve runtime complexity




Recognition rate for Graz-02 cars

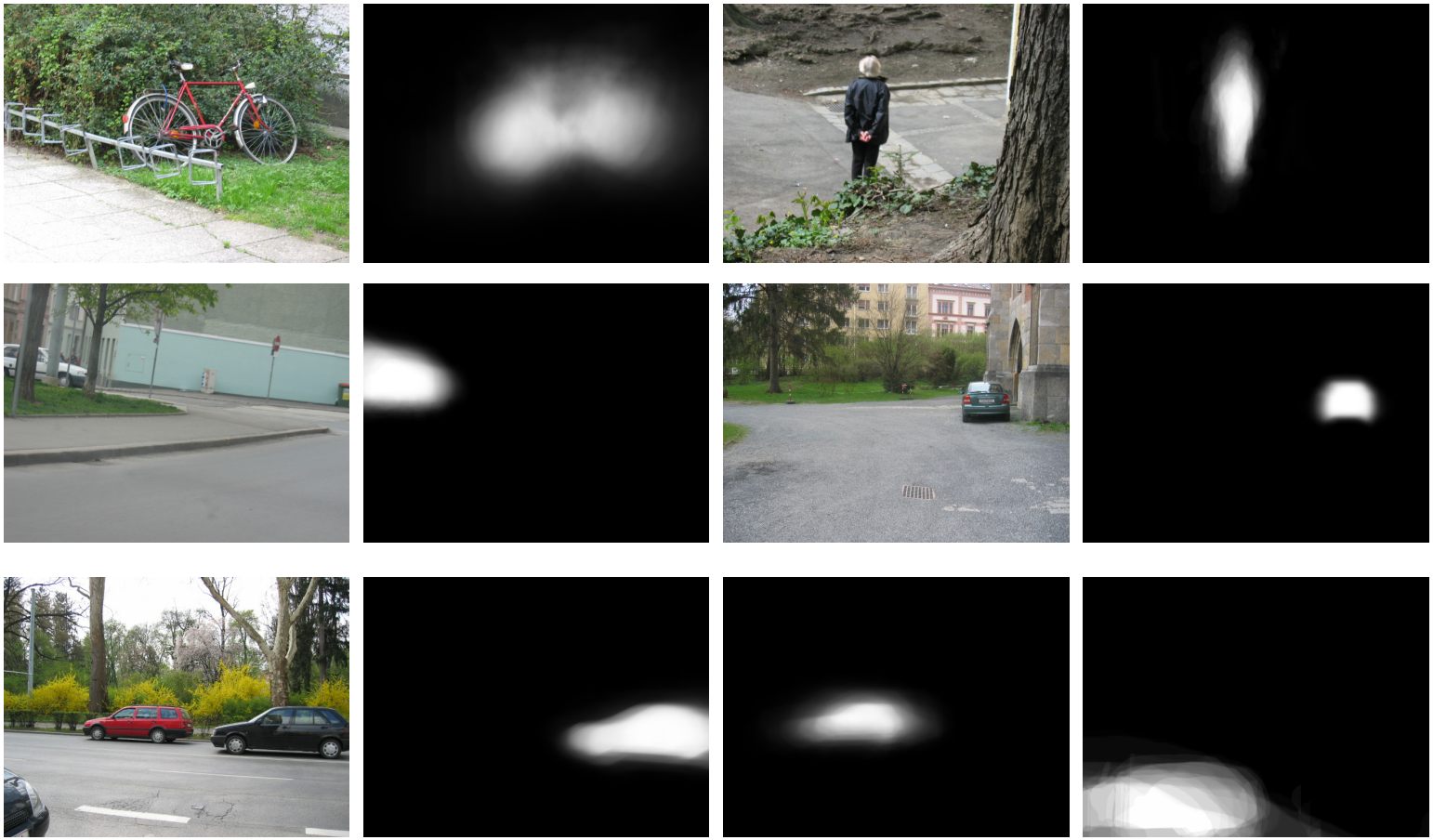
# Importance of recognition components

| object class | cars | people | bicycles |
|---|---|---|---|
| no hypothesis evaluation | 40.4% | 28.4% | 46.6% |
| no evidence collection | 50.3% | 40.3% | 48.9% |
| our full framework | **53.8%** | **44.1%** | **61.8%** |

Table: Pixel-based RPC EER measuring the impact of hypothesis evaluation and evidence collection on Graz-02

- For each class the combined framework shows better performance than hypothesis evaluation or evidence collection separately
- Therefore, both elements are necessary in order to perform precise object class localization

# Results on Graz-02 dataset



| Confidence: | 1103.1 | 561.8 | 4.9 |
|---|---|---|---|

# Comparison to the state-of-the-art

| | |
|---|---|
| Shotton et al. [ICCV'05] | 92.1% |
| Our framework ($T = 0.85$, with singletons) | **94.6%** |
| Our framework ($T = 0.7$, no singletons) | **94.6%** |

Table: RPC EER for Weizmann horse dataset

- We closely follow the experimental setup of Shotton et al.
- Due to large number of articulations, we had to lower the aspect merge threshold or turn off singleton pruning




# Summary

- An object localization framework with shape masks as localization hypotheses was proposed
- The object outline incorporates additional information about viewpoint, articulation, sub-type or state
- At the same time, the experimental results show that the standard localization performance of the method is comparable to the state-of-the-art
- Our method performs well on natural images and handles robustly multiple object aspects, significant intra-class variations, occlusions and background clutter

# Thank you for your attention

I will be glad to answer your questions

INRIA Annotations for Graz-02 (IG02):
http://lear.inrialpes.fr/data/