Towards Multi-View Object Class Detection

Alexander Thomas, Vittorio Ferrari, Bastian Leibe, Tinne Tuytelaars, Bernt Schiele & Luc Van Gool
The Goal

Detection of specific objects

Recognizing classes of objects – typically limited to a single viewpoint

Multi-view object class detection

Figure credit: David Lowe
Multi-view specific object recognition system  

Implicit Shape Model for object class detection  
Image Exploration

Ferrari et al.

- Recognizing specific objects from different views
- Creating correspondences among different model views
Region Tracks

A region track is composed of the image region of a single physical surface patch along the views.

A set of region tracks is produced for each specific training object.
Image Exploration

• Dense two-view matches are produced between each model image and all other images within a limited change of the view-point

• All the pairs of matches are integrated into a single multi-view model
Implicit Shape Model
Leibe & Schiele

- Recognizing object categories
- Codebook of local structures:
  - Clustered image features sampled at interest point locations
- Occurrences - map sampled image features from the test image to the codebook entries

Image credit: Grauman & Leibe
Integrating the Systems

- Integrate the two systems to achieve a multi view object class detection, not only by running a collection of single view detectors.
- The single view codebooks (ISM) will communicate using activation link (image exploration)
Training

• M object instances, from N viewpoints.
• The viewpoints should be approximately the same, but each instance does not need to have all of them.
• A set of ISMs is trained independently for each viewpoint.
• The image exploration algorithm is run for every object and create sets of region tracks.
The Data Set

- Region Tracks
- Region Tracks
- Region Tracks
Region Tracks

• Region tracks – contain regions corresponding across the object’s views.
• Region – described by ellipse – affine transformation of a circle.
• The affine transformation between the regions approximates the affine transformation between the image patches they cover.
How to find the closest region to the occurrence?

- There is no one to one correspondence between regions and occurrences.
- Finding the closest ellipse (a region) to a point (the center of the occurrence)
- There is an analytical solution, but it is computationally expansive.
- Approximation – the distance to a line aligned with the major axis of the ellipse, of length $||l|| - ||s||$
- Only if the distance is $< 2\cdot||s||$
Linking Algorithm

Iterate over all occurrences $O_i$ in all training viewpoints of a specific object. For each $O_i$:

- Find the nearest **region** $R_i$ (approximate way)
- For every other view $V_j$ in $R_j$’s track:
  - Transform the circular region $O_i$ with affine transformation $A_{ij}$ (between $R_i$ and $R_j$) to $O_i'$
  - Look for occurrences $O_j^k$ in view $V_j$ that are sufficiently similar to $O_i'$
  - Store all $O_i \rightarrow O_j^k$ as activation links
Matching Occurrences

• Looking for all circles sufficiently similar to the ellipse Oi’.

• Using the heuristics:
  (circle: pc – center, R – radius,
   ellipse: pe – center, || l || || s || - major/minor axis)

\[
\begin{align*}
  || p_c - p_e || & < a \cdot R \\
  \left| 1 - \left(\frac{||s|| \cdot ||l||}{R^2}\right) \right| & < b \\
  || s || / R & > 1/c \\
  || l || / R & < d
\end{align*}
\]

\[
a = 0.35 \quad b = 0.25 \quad c = d = 3.0
\]
Identification Algorithm

Handling a test image:
First stages – similar to ISM:
  – Extracting features and matching to the codebooks of the different ISMs.
  – Casting votes in the Hough spaces of each ISM separately.
  – Detecting initial hypotheses as local maxima.
Identification Algorithm – Selecting Working Views

A trivial criterion – choose the views containing the strongest initial hypothesis but…

– Image clutter can lead to strong hypotheses
– Correct strong hypothesis tend to create maxima in neighboring views while clutter doesn’t.
– The pose of the object mostly falls between two training views.
Identification Algorithm – Clustering the Hypothesis

- Pick the strongest hypothesis.
- Search in the neighboring views for hypothesis in approximately the same locations.
- Extend the cluster as possible to all directions.
- Take the next strong hypothesis etc. till all the hypotheses are clustered.
- The score of the cluster – the sum over all the hypotheses scores.
- Keep clusters > $T \times$ maximum score ($T = 0.7$).
- Choose the working views – the strongest hypothesis of each remaining cluster.
Identification Algorithm – Transferring Votes

Augmenting the Hough transform:

A feature matches to a codebook entry in view \( V_i \) + An activation link between entry’s occurrences in \( V_i \) and \( V_j \) → Cast additional vote in \( V_j \)

- Assume that the part will be in approximately the same location in view \( V_j \) (for estimating the object center).
- If a part was detected in the codebook of \( V_i \), but \( V_j \) is more likely the pose of the object, transfer the evidence of the part to \( V_j \).
Identification Algorithm – Wight of Transferred votes

Expresses the contribution of a patch \( e \) in location \( l \) to an object hypothesis \((o_n, \lambda)\) (\( \lambda = x y s \) – location and scale). \( V_j \) is the current working view.

\[
p(o_n, \lambda|e, \ell) = \sum_k P(o_n, \lambda|c_k^j, \ell)p(c_k^j|e) + \sum_k \sum_l P(o_n, \lambda|c_k^j, c_l^i, \ell)p(c_l^i|e)
\]

- Iterates over all the entries for view \( V_j \)
- The probability of the hypothesis given the codebook entry
- The probability that the entry \( C_k \) in view \( V_j \) is a correct interpretation of the patch
- Iterates over all the entries of the other codebooks – \( V_i \neq V_j \)
- Non-zero only if there is an activation link between \( c_l^i \) and \( c_k^j \)
- The probability that the patch matches the codebook entry \( C_i \) in view \( V_i \)
Testing

- Motorbikes from PASCAL Visual Object Classes (VOC) Challenge and sport shoes.
- Motorbikes - training set - 30 objects, segmented by a bounding box, 16 training views taken on a circle around the object. Average of 11 views per motorbike.
- Average of 22 objects per view point, which is only a small number for training the ISM.
- Sport shoes - training set - 16 views, taken at 2 different elevations.

Testing

- Baseline: a bank of 16 ISM models ran separately.
- All the detections are collected and output together.
- Evaluation protocol like in the PASCAL challenge – detection is correct if its bounding box overlaps more than 50% with the ground truth.
Testing

Comparison of systems for motorbike test set

Comparison of systems for sports shoe test set
Testing

• Comparison versus PASCAL VOC challenge – second using DoG+Patches and first using the new HesLap+SC features.

• After training the ISMs from much fewer motorbike instances.

• Not a perfect comparison:
  – Trained on different instances
  – Used multiple training views per instance
Results

Multi View System

Bank of Detectors