# Leopar: Online Learning of Patch Perspective Rectification for Efficient Object Detection

Stefan Hinterstoißer[1], Selim Benhimane[1], Nassir Navab[1], Pascal Fua[2], Vincent Lepetit[2]
[1] Chair for Computer Aided Medical Procedures (CAMP) - TUM, Germany
[2] Computer Vision Laboratory (CVLAB) - EPFL, Switzerland

## Overview

### Problem

How can we efficiently estimate the 3D pose of a poorly textured object in an input image, given a reference image of the object? A single patch provides enough 3D information, however affine region detectors are not accurate enough.

The transformations retrieved with our method are very accurate.

By contrast, affine region detectors retrieve comparatively inaccurate transformations.

How can we extract this rectification information efficiently and accurately?

### Proposed Solution

- Rough estimate of the rectification through trained classifiers
- Patch rectification refinement through Hyperplane Template Matching [Jurie 2002]
- Outlier removal with Normalized Cross Correlation
- Global refinement with e.g. ESM [Benhimane 2004]

Patch Rectification → Coarse Pose Estimation → Global Refinement

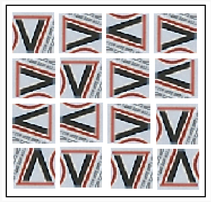
### Results

- Fast (~10-20fps) and accurate tracking by detection
- Retrieval of full perspective transformations (not only affine)
- Robust to large perspective distortions and to some amount of deformations
- Only little texture and few feature points necessary to estimate the pose
- Online learning possible

**Fast, robust and very accurate tracking by detection**
**!!! Pose estimation possible with only ONE feature point !!!**

## Rough Estimation of Patch Rectification

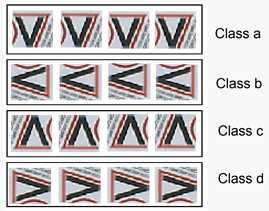
**Ferns [Ozuysal 2007]:**
1 class ID per feature point




$$id = \arg\max_{id} P(ID = id \mid patch)$$

**We learn for each feature point (not only its class but also) its rough discretized pose**

**Leopar:**
1 class ID per (feature point + pose)

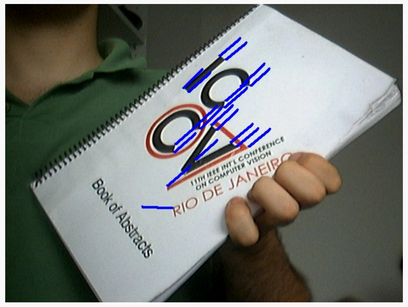


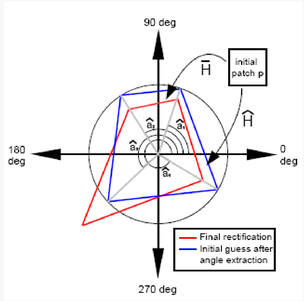

$$\begin{cases} id = \arg\max_{id} P(ID = id \mid patch) \\ p = \arg\max_{p} P(P = p \mid ID = id; patch) \end{cases}$$

## Rough Pose Parametrization

### Proposed Parametrization:

In practice, we are only able to retrieve a rough orientation defined by the angles of the four corner points which parametrize a restricted homography

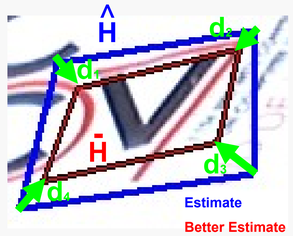





## Iterative Refinement with Linear Predictors

$$\begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{pmatrix} = A(I(\hat{H}) - I^*)$$

[Jurie 2002]




where $\hat{H}$ is the initial homography estimate, $I(\hat{H})$ the normalized intensity vector of the patch under matrix $\hat{H}$ and $I^*$ the normalized intensity vector of the reference patch. This refinement has to be applied iteratively to converge to the right solution.

Matrix $A$ can simply be learned by warping the patch of interest by random pose changes and computing the corresponding normalized intensity differences.

$$A = XD^T(DD^T)^{-1}$$

$X$: matrix of training pose changes
$D$: matrix of corresponding intensity differences

Matrix $A$ can also be learned online with constant memory (see paper for details).

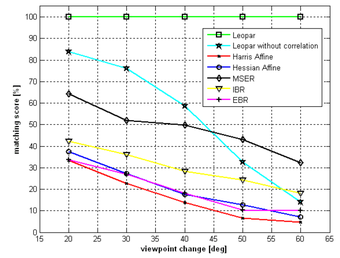
## Robust Outlier Removal with Correlation

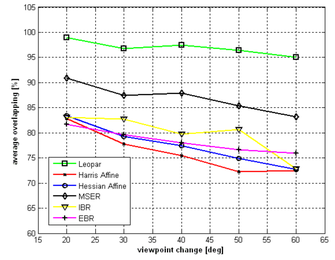
Thanks to the very accurate outcome of the Linear Predictors, outliers are easily removed by simply computing:

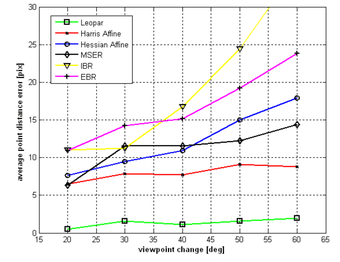
$$I(H_{final})^T \cdot I^* \geq \tau_{ncc}$$

where $H_{final}$ is the final transformation obtained with the linear predictor. In practice, we use normalized intensity vectors and a threshold $\tau_{ncc} = 0.9$.

## Experimental Results




Leopar is compared to *ad hoc* affine region detectors (Harris/Hessian Affine, MSER, IBR, EBR) with respect to the matching score, the average overlap and the average point distance error. Leopar clearly outperforms all other methods.

## Robust Real-Time Tracking by Detection

Leopar is applied on the English grammar book, on the ICCV and on the ISMAR booklet. The pose of the objects is retrieved with one extracted patch only. The outcome of Leopar (yellow) is then refined with the ESM algorithm (green) [Benhimane 2004]. The pose estimation is robust to high perspective distortions, to scale changes, to occlusion and even to deformations.