ELF: Embedded Localisation of Features in pre-trained CNN

Assia Benbihi, Matthieu Geist, Cédric Pradalier
abbenbihi@georgiatech-metz.fr

Abstract: ELF is a novel feature detector based only on information embedded inside a CNN already trained on a standard learning task (e.g. classification). This information is extracted from the gradient of the feature map with respect to the input image. It provides a saliency map with local maxima on the relevant keypoint locations. We compare our method to hand-crafted and learned feature matching pipelines and reach comparable performances although our method requires neither supervised training nor finetuning.

Method: Feature detection
i) Saliency map \( S(I) = \nabla F(I) \cdot \nabla F(I) \)
ii) Adaptive threshold (Kapur).
iii) Non-Maxima Suppression (NMS).

Feature description
i) Interpolate the feature map on detected keypoints.

State-of-the-Art

Full supervision is the standard training method for recent detector-descriptor. It requires corresponding keypoints generated with either an existing detector or with Structure-from-Motion. Our method is semi-supervised: the CNN may require full supervision when trained on the standard task but it does not require corresponding keypoints.

Results

General performance
We derive ELF on three classification networks as well as SuperPoint's and U-Net's descriptor networks. Overall, VGG provides the best variation; we assume that this is because it has the biggest feature space, hence better discriminative properties. ELF compares with state-of-the-art on HPatches (SuperPoint) and slightly outperforms it on Webcam (TILDE). Lift and U-Net curious underperformance may come from a poor data generalisation from their training data.

Robustness performance
- Scale: Methods that process multiple scale of the same image (e.g., Lift, ELF) can get outperformed by the ones that delegate the multi-scale processing to the network (SuperPoint, ELF).
- Orientation: All methods without explicit orientation estimation degrade (SuperPoint, ELF).
- 3D viewpoint: All methods degrades similarly when the change increases.

Integration performance
ELF detection (dots): When integrated with other descriptors, ELF boosts the matching score.

Simple description (hashes): Even integrating the interpolated descriptors boosts the performance.

These results show that the feature representation and localisation information learnt by a CNN to complete a task are as relevant as when the CNN is trained specifically for feature matching.

Metrics [4]
1. Repeatability: Percentage of keypoints common to both images.
2. Matching Score: Percentage of keypoints that are nearest neighbours in both image space and descriptor space.

Bibliography