Learning Local Patch Orientation with a Cascade of Sparse Regressors

Alain Pagani
http://av.dfki.de/~pagani
Didier Stricker
http://av.dfki.de

DFKI
Augmented Vision Lab
Technical University of Kaiserslautern
Kaiserslautern, Germany

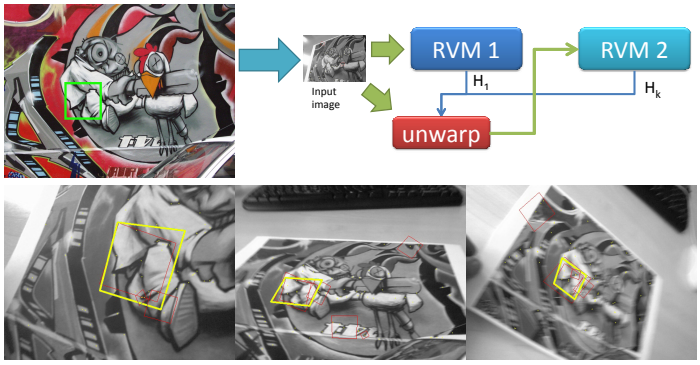

Figure 1: Overview of our approach. From a single reference image (top left), we train a cascade of regressors (top right) to infer the patch orientation from its appearance (bottom).

The problem of identifying good points of interest in an image has a long history in computer vision. The efforts made to improve the repeatability of points detectors culminated with the advent of scale and rotation invariant detectors [2] and affine invariant detectors [3]. However, the retrieved transformations are limited to similarities or affine transformations, and the full perspective transformation (a homography in case of a locally planar surface) cannot be found. In this paper, we present a method that accurately retrieves the local 3D transformation of a patch from its appearance.

Recently, a learning-based method for retrieving the local perspective warping of a patch has been presented [1]. This method uses a classifier which classifies patches into pre-defined quantized pose estimates. To achieve this, the pose space has to be quantized in a finite number of classes, which poses the problem of the pose space tesselation: a high number of classes increases the classifier complexity while a small number of classes can lead to poor estimation results. In this paper, we advocate the use of a keypoint-specific regressor to learn the pose as a function of the patch appearance. Thanks to the regressor, smooth variations in the patch appearance result in smooth variations of the pose. Our method relies on a learning stage, where examples of randomly warped views of the patch are used to train the regressor. We show that a good choice for the predictor is a set of sparse regressors applied sequentially in a cascade. With sparse regression, we mean using a machine that needs to retain only a fraction of the examples in the training set, like *e.g.* Support Vector Machines, leading to an efficient prediction step. With cascading, we mean that the regressor is implemented as a cascade of functions with increasing precision and complexity, which permits an early termination in the spirit of the Attentional Cascade of [5]. In our case, we use a set of parametrized multivariate relevance vector machines (MVRVM) [4]

Our method performs in several steps (see figure 1). In a first step, we apply a first RVM, which has been trained to be fast and infer a rough pose estimate from a local square image patch. The local neighborhood of the keypoint is warped using this first estimate, and only a few point hypotheses are kept, based on the similarity of the patch with the reference one. For each of the selected hypotheses, the warped patch is used as input for a second RVM specialized in small variations. The second RVM yields a better estimate of the patch. This last step is repeated until the changes in the pose are small (usually less than 10 iterations). We call this successive use of RVMs a Cascade of RVMs, and our method Caspar, for CAscade of SPArse Regressors.

Our idea is to design the cascade in such a way that the first regressor is very fast while providing a very coarse estimate of the homography. A similarity test based on the normalized cross correlation rejects most of the candidate already after the first step. The remaining hypotheses are then pushed to more complex regressors with a increasing precision upon several levels. More specifically, let $\mathbf{p}_1$ be a orientation and scale corrected patch of size s_1 found in the image. We start by applying the first, fast and coarse regressor $\mathbf{f}_1$: $\mathbf{h}_1 = \mathbf{f}_1(\mathbf{p}_1)$ and compute the score of the patch $S_1 = ncc(\mathbf{p}^*, u(\mathbf{p}_1, \mathbf{h}_1))$ where $u(\mathbf{p}, \mathbf{h})$ is the function unwarping the patch $\mathbf{p}$ using the homography $\mathbf{h}$, *ncc* is the normalized cross correlation between two patches, and $\mathbf{p}^*$ is a canonical view of the patch. If S_1 is smaller than a threshold τ_1, then the patch is rejected. Otherwise, the patch is processed by the next cascade level. Thus, the entire cascade can be described by repeated applications of tests:

$$\textit{if } ncc(\mathbf{p}^*, u(\mathbf{p}_k, \mathbf{h}_k)) > \tau_k, \textit{ then } \mathbf{h}_{k+1} = \mathbf{f}_{k+1}(u(\mathbf{p}_k, \mathbf{h}_k)), \textit{ else reject patch}$$

Thus, each level is governed by the set of parameters (s_k, σ_k, τ_k). During learning, we automatically adapt the kernel width σ_k and the patch size s_k through training/validation experiments to achieve a given type of regressor: the first level should be fast, so we set a small value for s_1 and adapt σ_1 to keep the number of relevance vectors under a given maximum (about 20). The correlation threshold τ_1 is set to keep the number of hypotheses lower than 5. For the remaining levels ($k = 2$ to 10), the patch size can be higher and the kernel size σ_k is automatically adapted to keep the number of relevance vectors under 100.

Our experiments show that the correct homography is retrieved with high accuracy even under large tilts, and that our method allows for correcting imprecisions from the point detector. Figure 2 show several frames of a video sequence where the 3D orientation of the patch is retrieved at 28 fps.

[1] S. Hinterstoisser, S. Benhimane, N. Navab, P. Fua, and V. Lepetit. Online learning of patch perspective rectification for efficient object detection. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[2] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60:91–110, 2004.

[3] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.V. Gool. A comparison of affine region detectors. *Int. Journal of Computer Vision*, 65:43–72, 2005.

[4] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. S. Torr, and R. Cipolla. Multivariate relevance vector machines for tracking. In *European Conference on Computer Vision (ECCV)*, 2006.

[5] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001.

Figure 2: Estimation of the local patch transformation. Our system detects the learnt point and estimates its pose in real time, even in presence of large scale variations. *A full video is available on the authors' website at* `http://av.dfki.de/~pagani`.