Seeing through bag-of-visual-word glasses: towards understanding quantization effects in feature extraction methods

Alexander Freytag[*], Johannes Rühle[*], Paul Bodesheim[*], Erik Rodner[*], Joachim Denzler[*]

[*]Computer Vision Group, Friedrich Schiller University Jena

{*firstname.lastname*}@*uni-jena.de*

Problem formulation Vector-quantized local features frequently used in bag-of-visual-words approaches are the backbone of popular visual recognition systems due to both their simplicity and their performance. Despite their success, standard bag-of-words-histograms basically contain low-level image statistics (e.g., number of edges of different orientations). The question remains how much visual information is lost in quantization when mapping visual features to visual “words” and elements of a codebook?

Summary To answer this question, *we present an in-depth analysis of the effect of local feature quantization on human recognition performance*. Our analysis is based on recovering the visual information by inverting quantized local features and presenting these visualizations with different codebook sizes to human observers (Fig. 2). Although feature inversion techniques are around for quite a while, to the best of our knowledge, our technique is the first visualizing especially the effect of feature quantization. Thereby, we are now able to compare single steps in common image classification pipelines to human counterparts. Our results show that (i) humans perform significantly worse than machine learning approaches when being restricted to the visual information present in quantized local features rather than having access to the original input images, and (ii) that *early stages of low level local feature extraction seem to be most crucial* with respect to achieving human performance on original images. Finally, we demonstrate (iii) that large codebook sizes in the order of thousands of prototypes are essential not only for good machine learning performance, but more interestingly, also for human image understanding.

Method Our technique is simple and in line with current trends in image reconstruction from local features. For an unseen image, we extract local features on a dense grid and follow the bag-of-words paradigm by quantizing them using a pre-computed codebook. Based on inversion techniques for local features, we can compute the most probable image patch for every prototype, *i.e.*, we can visually inspect the quantization quality for a given codebook. Thus, for any local feature, we vector-quantize it with a codebook and draw the inverted prototype into the reconstruction image with position and size according to the support of the extracted local feature (Fig. 1). In contrast to previous works for feature inversion, we aim at explicitly visualizing quantization effects. For the simplicity of demonstration, we use HOG features, where code for the inversion technique is publicly available[1].

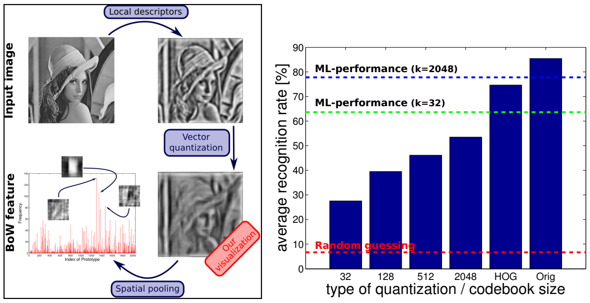

Fig. 1: Inversion strategy and human classification result in comparison to machine performance

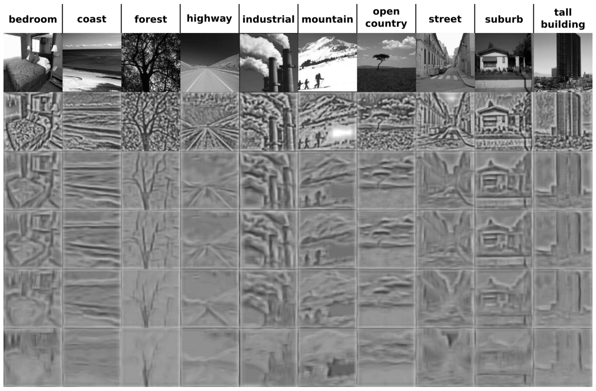

Fig. 2: Overview of the images presented to human observers during the experiment (15 Scenes dataset)

Evaluation *We performed human studies, where more than* 3,000 *images were classified by several individuals with different levels of quantization* (Fig. 2). The results are given in the right plot of Fig. 1 and will be presented in detail at the poster[2].

[1] http://web.mit.edu/vondrick/ihog/

[2] Source code, detailed results, and a web interface to our evaluation server are available at http://www.inf-cv.uni-jena.de/en/image_representation .