J. Vis. Commun. Image R. 28 (2015) 120–129

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

A quantitative evaluation of the conceptual consistency of visual words and visual vocabularies ☆

M. Stommel [a,*], O. Herzog [b,c], W.L. Xu [d]

[a] School of Engineering, Auckland University of Technology, New Zealand
[b] Artificial Intelligence Group, University of Bremen, Germany
[c] Visual Information Technologies, Jacobs University Bremen, Germany
[d] Mechatronics Group, The University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 6 April 2014
Accepted 24 November 2014
Available online 2 December 2014

Keywords:
Computer vision
Pattern recognition
Bag of visual words
Codebook
SIFT
SURF
Image classification
Text image analogy

ABSTRACT

Codebooks are a widely accepted technique to recognise objects by sets of local features. The method has been applied to many classes of objects, even very abstract ones. But although state of the art recognition rates have been reported, the method is still far away from being reliable in any sense that is related to human vision. The literature on this topic emphasises detailed descriptions of statistical estimators over a basic analysis of the data. A deeper understanding of the data is however needed to achieve a further development of the field. In this paper, we therefore present a set of quantitative experiments on codebooks of the popular SIFT descriptors. The results discourage the use of illustrative but overly simplifying descriptions of the *visual words* approach. It is in particular demonstrated that (1) there are more visually distinct patterns than can be listed in a codebook, (2) one element of a codebook represents a set of many, visually distinct patterns, and (3) there are no single, selective SIFT descriptors to serve as codebook elements. This makes us wonder why the method works after all. We discuss several options.

1. Introduction

Visual vocabularies or codebooks of *visual words* are among the most popular image representations for object recognition. They have been used for image classification [1–3] and matching [4], video retrieval and indexing [5], and activity recognition [6–9].

Grauman and Leibe [10] summarise the method to create a visual vocabulary as a two step process of "(1) collecting a large sample of features from a representative corpus of images, and (2) quantising the feature space according to their statistics". The partitioning of the feature space forms the visual vocabulary. Single images can be represented with respect to the visual vocabulary by performing a feature extraction and assigning each feature to the nearest bin in the feature space. In its most simple form, the popular *bag-of-visual-words* approach, images are represented by histograms over the visual vocabulary and compared by histogram-based distance measures, e.g. histogram intersection.

By using histograms, the method allows to classify images that are represented by a set of feature descriptors instead of only one descriptor. In particular, visual vocabularies are often used to work with local feature descriptors like SIFT [11] or related methods [12–15]. By a simple histogram concatenation, it is possible to fuse different feature spaces. This makes the method very attractive for multimedia processing, where heterogeneous sets of video and audio features must be fused [5,16,17]. Compared to other multimodal approaches to video processing (e.g. [18]) visual vocabularies can represent single video fragments by feature sets of differing cardinality and descriptor length. The idea behind this approach is to gain a maximum amount of information by applying multiple, complementary feature extractors. The approach is popular in semantic video classification. Principle Component Analysis can be used to reduce dimensionality [17].

The partitioning of the feature space is usually done by clustering a sufficiently large set of samples [19,10], so the resulting bins represent similar densities. The actual presence of separate clusters in the data is neither presupposed nor questioned [19,2,16]. Vector quantisation has been recognised to not affect the matching of SIFT descriptors [19]. Moreover, a maximally coarse quantisation in form of a thresholding has been found to improve both speed and matching accuracy [20,21]. By using a hierarchical clustering, so called *vocabulary trees* can be created [22–25]. The tree structure can be used both to encode descriptors compactly by a path in the tree and to match features in a coarse-to-fine strategy.

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail address: mstommel@aut.ac.nz (M. Stommel).

http://dx.doi.org/10.1016/j.jvcir.2014.11.015
1047-3203/

An adaptive clustering [24,25] is advantageous over a strict pyramidal approach both in terms of speed and accuracy.

The exact nature of the information represented by visual vocabularies remains surprisingly unclear, even after a decade of research. Obvious dependencies on the type of feature descriptor, image material, and parameterisation of the method make it difficult to give a general answer. Grauman and Leibe state that "in general, patches assigned to the same visual word should have similar low-level appearance" [10]. For SIFT, the most popular descriptor, examples of a few codebook entries have been published [19] and they appear to be homogeneous. This is however no proof for SIFT clusters in general. Because of the density based clustering, it can be conjectured that some clusters extend over large, visually heterogeneous regions in feature space. Since not much is known about the connection between visual appearance and the structure of visual vocabularies, the parameters of the clustering are mostly tuned to the recognition rate [19]. This results in vastly differing vocabulary sizes spanning several orders of magnitude (from 10^3 [17] over 10^4 [19] to 10^7 [25]). More integrated approaches consider class membership during the creation of the vocabularies [26]. When combining different feature spaces, there will also be task-specific and class-specific trade-offs between the properties of different features [27].

Spatial information is usually lost in the bag-of-words approach. To add geometrical relationships between local features, the hierarchical clustering can be extended to include the image coordinates [28], or the visual vocabulary can be extended in a way that it represents geometrically ordered pairs of local features [29,30]. However, studies on feature constellations indicate that complex geometrical relationships might not always be crucial for object recognition [31], or might not be worth the computational cost [32], and depend on the level of abstraction [33]. In contrast, temporal information seems to be of more general significance for action recognition. It can be modelled as part of the feature detector [9] or separately by relationships between multiple visual words [8].

The notion of visual vocabularies is usually traced back to the study by Sivic and Zisserman [19], who applied a text retrieval approach to video recognition. The basic assumption is that "local features play the role of 'visual words' predictive of a certain 'topic,' or object class. For example, an eye is highly predictive of a face being present in the image" [19]. The paper demonstrates that the text-image analogy carries relatively far (also with respect to recognition rates) and supports even the idea of visual *stop words*. Inspirations from linguistics are nothing new to image processing. Aside from statistical approaches in topic modelling there have been approaches to use formal languages on the level of pixels [34,35] or geometric primitives [36]. However, the text-image analogy is not based on solid theory but serves as inspiration for some approaches [19,3]. The most commonly given justifications of working with visual vocabularies are good results mentioned in the literature [1,37,30,26,38], none [4,5,17,8], a novel technical idea related to the original approach [26,25], popularity of the approach [16,7], and fitness for a purpose [2].

Despite the good benchmark results and high popularity, the approach is still far from being comparable to the quality of human vision; occasionally it is considered even unstable [39]. A deeper understanding of the connection between visual vocabularies and feature based image representations is necessary to develop the approach. In this study, we therefore aim to verify or disprove the very foundations and basic assumptions of the approach. In particular, we address the following questions:

- What does visual similarity mean quantitatively? To answer this question, we consider the Hamming distance of binarised SIFT descriptors for image patches of similar low-level appearance.
- How big are clusters of visually similar features? Here, we are not interested in the density but the diameter of a cluster of similar image patches in feature space.
- How many clusters are there? More precisely, we want to know if it is possible to create a visual vocabulary from a clustering where every cluster represents only patterns of similar low-level appearance.
- What do the most discriminative SIFT features represent?
- Where are features located that represent the class prototypes most closely?

As reviewed above, these data-related questions have been neglected in favour of detailed discussions of machine learning algorithms. But we think that a deeper understanding of the nature of the problem is necessary to draw the right conclusions from the text-image analogy.

2. Materials and methods

2.1. Feature extraction

The Scale Invariant Feature Transform (SIFT) [11] is among the most popular local feature extractors for object recognition and robot perception. The method identifies points of interest (or *keypoints*) as corner like local structures in band pass filtered images at multiple spatial frequencies (called *scales*). The band pass filtering is represented by a scale space pyramid. To compute the pyramid, the input image is iteratively smoothed by a Gaussian filter. The scale space consists of the difference images of the images between before and after Gaussian smoothing at each iteration. The scale of a pyramid level corresponds to the bandwidth of the Gaussian smoothing which accumulates over the iterations. The detected keypoints are annotated with the scale of the pyramid level where they are detected. The local structure is modelled by a feature vector that represents the concatenated bins of 16 histograms of the intensity gradient direction. To compute these histograms, a squared region around each keypoint is chosen. The size of the region is adapted to the scale of the keypoint. The square is rotated to the predominant gradient directions around the keypoint (the *canonical orientation*). The region is then subdivided into a regular 4×4 grid. For each grid cell, a histogram of the gradient directions with 8 bins is computed. The canonical orientation is subtracted from the gradient directions to make the descriptor rotationally invariant. The concatenation of all histograms results in a descriptor with $4 \times 4 \times 8 = 128$ elements. In cases of ambiguous gradient orientation, multiple keypoints are generated. The method is widely appreciated for its robustness against changes in illumination, its rotational invariance, its high tolerance against perspective transformations, and the convenient output in vector form.

However, the method is also known for its computational complexity in matching and sharp discontinuities in the descriptor value between adjacent pixels [40]. In order to address the first point, we apply the feature binarisation used in the bitvector machine [20]. SIFT binarisations have been found to reduce the descriptor size and lead to faster and more robust matching [21,20,41]. The used method is much faster than random hyperplanes [41] and does not require learning except from a straightforward median computation. The binarisation does not merge visually heterogeneous patterns, is not affected by dimensionality (as opposed to SIFT), and has statistical properties that can be simulated by a synthetic codebook without the need to perform a