Bag of Visual Words revisited - An exploratory study on robust image retrieval exploiting fuzzy codebooks

Marian Kogler
University of Klagenfurt
Institute of Information Technology
9020 Klagenfurt, Austria
mkogler@itec.uni-klu.ac.at

Mathias Lux
University of Klagenfurt
Institute of Information Technology
9020 Klagenfurt, Austria
mlux@itec.uni-klu.ac.at

ABSTRACT
Visual information retrieval systems have gained importance due to the increasing amount of available digital multimedia data. Local features employing a bag of words approach from text retrieval have outperformed global features and have enhanced retrieval performance in large data sets. In this paper we conduct an exploratory study revisiting the bag of visual words approach for content based image retrieval. We apply a fuzzy clustering technique for visual words creation and visual words assignment and show in a first attempt that fuzzy clustering leads to more robust results in terms of retrieval performance.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

General Terms
Algorithms, Experimentation

Keywords
CBIR, Bag of Visual Words, Fuzzy Codebooks

1. INTRODUCTION
Nowadays a huge amount of cheap multimedia capturing hardware is available, which accounts for an astonishing growth rate of digital multimedia data. Flickr (flickr.com), a popular web platform, where millions of people store, organize and share their photos, records 6,000 photos uploaded each minute1. In order to handle this growth rate, semantic and meaningful multimedia content organization and retrieval needs to be pursued intelligently and automatically. This is still an open topic in visual information retrieval. Content based image retrieval, as described in [3] and [12], aims to tackle theses issues either with global features or with local feature like Scale Invariant Feature Transform (SIFT) [9] or Speeded Up Robust Features (SURF) [1]. Whereas global features describe a picture in a holistic way (e.g. one histogram represents a fingerprint for the whole image) local features describe several salient patches around key points within images, which leads to a set of feature vectors. The cardinality of the set of local feature vectors depends on the amount of detected key points, which can be huge. To cope with the mass of generated data by local feature analysis the bag of words approach was adopted from text retrieval. Under the assumption that local features are words describing the image content the approach is called bag of visual words. For these visual words inverted lists are employed for indexing and fast search.

When employing the bag of visual words approach for content based image retrieval, local features are extracted from an image collection and clustered typically using k-means [7]. The computed cluster centers are called visual words and form a so called codebook (depicted in Figure 1), which serves as the basis for indexing newly added images. If a new image is added to the existing image collection, local features will have to be extracted and assigned (e.g. with a nearest neighbor search) to the best fitting visual word(s). The resulting local feature histogram representing the distribution of local features over the former computed clusters serves as a fingerprint for the new image (depicted in figure 1). While the bag of words approach works very well for many application scenarios and domains, parameters of the methods have to be determined for each and every application. The resulting local feature histograms heavily depend on the predefined size of the created codebook (visual word vocabulary determined by the k-means clustering step). A unified amount of visual words for different image collections can hardly be defined beforehand. In literature codebook sizes ranging from hundreds to thousands of visual words are reported. Small codebooks might not be discriminative enough for retrieval tasks in one image collection, whereas they fit really well for another one. On the contrary large vocabularies may contain redundant visual words, which could have been merged or left out, in order to achieve better results. This leads to necessary evaluations to determine the appropriate number of visual words for a specific domain.

In this paper we outline a revisited bag of words approach by applying a fuzzy clustering technique for codebook generation and a fuzzy assignment to visual words. So instead of assigning a local feature to a single visual word we determine the degree of membership of a local feature to a visual

---


Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. MDMKDD’10, July 25th, 2010 Washington, DC, USA
Copyright 2010 ACM 978-1-4503-0220-3 ...$10.00.
Our work is based on the hypothesis that fuzziness in the codebook creation step as well as in the histogram creation process leads to more robust behavior of the bag of visual words approach in terms of codebook size. We show that our approach leads to similar, but more robust results in retrieval performance when the number of visual words is varied. The remainder of this paper is organized as follows: Section 2 gives an overview on related work considering discussions and adoptions of the bag of visual words approach and discusses our adaptations in context of the current state of the art. In Section 3 we outline our approach and the novel adaptations compared to the typical approach. The experimental setup, the evaluation and the results are presented in Section 4. Finally Section 5 discusses the results and their impact, concludes our paper and outlines future work.

![Diagram](image)

**Figure 1:** Illustration of the process employed for codebook creation and visual words assignment.

## 2. RELATED WORK

The bag of word approach for content based image retrieval has been revisited and discussed in several publications. A broader context and a general discussion of local and global features as well as an evaluation comparing their performance can be found in [4]. Jiang et al. [8] experiment with different components important for efficient use of the bag of visual words approach, considering different key point detectors, kernels, vocabulary size and introducing a new soft weighting scheme for visual words. Their soft weighting approach takes the distances from the key points to the cluster centers into account, by selecting the top $n$ nearest visual words. The closer a key point is situated to a visual word the higher its significance for that specific visual word. The authors state that their approach outperforms the traditional text retrieval weighting schemes like tf or tf*idf [11] supporting their assumption that a key point doesn’t belong to one specific visual word and that the similarities of key points and visual words must be considered during the codebook creation. This also supports our assumption that fuzzy clustering and fuzzy class membership of interest points lead to more robust results, but fuzzy clustering is not investigated in [8]. Yang et al. [16] experiment with various weighting schemes like binary weighting, tf, idf and tf*idf investigating their effectiveness over different vocabulary sizes ranging from 200 up to 80,000. The large amount of visual words is accompanied by a great computational effort during the codebook generation process and poses an additional burden to classifiers due to the high dimensionality of the composed local feature vectors. A thorough overview on bag of visual words approaches in the domain of concept detection in videos is given by Snoek et al. in [13]. They focus on different methods of codebook construction varying the parameters and components of codebook construction. A lot of different feature detectors and descriptors are investigated, hard and soft assignment to visual words are considered and kernel based learning with support vector machines is addressed too. Despite this thorough study and the statement that soft assignment with kernel codebooks lead to better retrieval and concept detection results, utilizing fuzzy clustering algorithms for codebook creation, is not studied. However the study outlines that hard assignment can lead to unsatisfying results, especially if a key point is similar to multiple visual words and can be assigned to two or more different clusters. Van Gemert et al. [5] describe a kernel based approach allowing for assignment of a key point ambiguously to two different clusters. Applying this soft assignment approach improves retrieval accuracy in their evaluation. Uijlings et al. [14] describe their contribution to the bag of visual words pipeline starting from feature extraction, going over to descriptor projection and classification. For descriptor projection they focus on k-means clustering and Random Forests, which are faster than the k-means algorithm but lead to a slightly worse mean average precision (MAP) concerning their test data. However, codebook generation is typically done offline and from a user point of view the MAP is much more important. Multiple approaches to enhance the bag of words approach have been proposed in literature. Some publications conclude that soft assignment of key points to visual words is superior to hard assignment. However, fuzziness in actual codebook creation process has not been considered yet.

## 3. FUZZY CODEBOOKS

The bag of word approach for content based concept detection consists of several different sequential steps:

1. **Extraction of key points.** Multiple salient points are extracted from each image. SIFT and SURF are common choices for key point description. In our work we focus on SIFT utilizing the implementation provided in [10].

2. **Codebook creation.** All local features (one for each key point) are clustered. For each cluster a visual word is found. Typically a mean vector (in case of k-means) or a medoid is used as visual word. In our work we employ a different, fuzzy clustering approach and compare it to the traditional approach utilizing k-means.

3. **Local feature histogram creation.** Each local feature of an image is assigned to the visual word most
similar to the local feature. These assignments are quantized in a histogram, where each bin reflects a visual word. While this approach is called hard assignment (a local feature is assigned to one single bin) in our approach we utilize fuzzy (soft) assignments, where each local feature has a degree of membership to one or several bins.

The whole process results in one local feature histogram per image. These local feature histograms can then be used for content-based retrieval or classification. Pairwise distance can be determined by different metrics. Typically content-based retrieval or classification. Pairwise distance image. These local feature histograms can then be used for the bag of visual words. The sum of membership values of an element \( x \) to all visual words is one: \( \sum_{A \in C} \mu_{A_i}(x) = 1 \).

In our work we employ the fuzzy c-means clustering algorithm [2]. In fuzzy c-means a membership function is iteratively used to assign data points \( d \in D \) to clusters \( c_i \in C \) with \( \bigcup_{i \in C} c_i = D \) and to compute cluster centers \( m_i \in M \):

\[
\hat{m}_i = \frac{\sum_{d \in D} \mu_{c_i}(d)^m d}{\sum_{d \in D} \mu_{c_i}(d)^m}
\]

\[
\mu_{c_i} = \frac{1}{\sum_{m_k \in M} \left( \frac{L_2(m_i, m_k)}{L_2(m_k, d)} \right)^{\frac{1}{m-1}}}
\]

Parameter \( m \in [1, \infty) \) is called fuzzy. and controls the membership function. The algorithm terminates when a global optimization function

\[
f = \sum_{d \in D} \sum_{m_i \in M} \hat{m}_i^2 \mu_{c_i}(d)^m
\]

reaches a minimum:

1. Randomly select \( n \) cluster centers.
2. Determine membership of each data point to each cluster (using the cluster center).
3. Compute \( f_{\text{last}} \).
4. Recompute cluster centers based on the determined membership values.
5. Determine membership of each data point to each cluster (using the cluster center).
6. Compute \( f_{\text{actual}} \).
7. (a) If \( |f_{\text{actual}} - f_{\text{last}}| < \epsilon \) stop.
(b) Else set \( f_{\text{last}} \) to \( f_{\text{actual}} \) and start over with step 4.

The same membership function \( \mu_A \) is used for soft assignments of local features to visual words and therefore, to create the local feature histograms.

Figure 2 shows an overview on our bag of visual words pipeline for our experiments and the evaluation. Based on SIFT features we either extract a traditional codebook and create the local feature histograms by hard assignments or create fuzzy codebooks and compute local feature histograms with soft assignments. Based on the created local feature histograms a k-nearest neighbour search in our test data set is done and results are evaluated as described in Section 4.

4. OUR EXPERIMENTS

As mentioned in the previous sections, the focus of our work lies on the codebook generation step. We aim at investigating the impact of the use of a fuzzy codebook in the bag of visual words pipeline. Figure 2 shows the available options in our experimental setup. We chose SIFT to extract key points and both, k-means clustered codebooks and fuzzy c-means clustered codebooks, can be computed. For the traditional approach hard assignments are used. In case of fuzzy c-means also the assignment in the local feature histogram creation step is implemented fuzzy. For fuzzy c-means we chose the parameter \( m = 1.7 \) and stop condition \( \epsilon = 0.001 \) (algorithm stops if optimization function value changes less than \( \epsilon \) in absolute numbers between subsequent steps). For evaluation purposes we employ k-nearest neighbour search using the local feature histograms and the \( L_2 \) distance. We use the Simplicity data set also applied in [15] for evaluation. The test data set includes 10 different concepts (topics), each containing 100 images.

The sizes of visual vocabularies ranged from 10 to 400. 30 images per concept were used during the training phase to build the codebooks. The set of those images was selected randomly one time and then used throughout the tests. For evaluation all 1,000 images of the Simplicity data set were used as query images. From the results of the k-nearest neighbor search mean average precision (MAP) and the error rate (ER) were computed to allow objective comparison between both approaches. MAP is based on the average precision (AP), which for a single query \( q \in Q \) is the mean of the precision scores after each retrieved relevant item.

\[
MAP = \frac{1}{|Q|} \sum_{q \in Q} AP(q)
\]

The error rate reflects how often the first hit in the result list is not a correct one.

\[
ER = \frac{1}{|Q|} \sum_{q \in Q} \begin{cases} 
0 & \text{if first element is correct hit} \\
1 & \text{otherwise}
\end{cases}
\]
4.1 Results

As a first step we investigated the MAP of the entire image collection. For each concept each of the 100 members was used as query image and all images assigned to the concept were considered as correct results. Average precision was computed for the full list of 1,000 results. Results are given in Table 1 and Figure 3.

![Figure 3: MAP, full data set, fuzzy and traditional codebooks.](image)

<table>
<thead>
<tr>
<th>clusters</th>
<th>fuzzy</th>
<th>trad.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.082148201</td>
<td>0.099821977</td>
</tr>
<tr>
<td>20</td>
<td>0.082209283</td>
<td>0.104639869</td>
</tr>
<tr>
<td>30</td>
<td>0.082448309</td>
<td>0.10342143</td>
</tr>
<tr>
<td>40</td>
<td>0.082700377</td>
<td>0.100510162</td>
</tr>
<tr>
<td>50</td>
<td>0.082472548</td>
<td>0.101482909</td>
</tr>
<tr>
<td>100</td>
<td>0.082226761</td>
<td>0.092439249</td>
</tr>
<tr>
<td>200</td>
<td>0.082220257</td>
<td>0.079575734</td>
</tr>
<tr>
<td>300</td>
<td>0.082215961</td>
<td>0.07664239</td>
</tr>
<tr>
<td>400</td>
<td>0.082572559</td>
<td>0.075742321</td>
</tr>
</tbody>
</table>

As can be seen in Figure 3 the k-means based approach outperforms the fuzzy codebook approach up to a codebook size of 100 visual words. Most notable fact however is that the MAP value of the fuzzy approach remains stable, also in absolut numbers.

If we take a look at the MAP of specific concepts, we can observe that the dinosaur concept is an outlier compared to the the other concepts (see Figure 4, top right diagram). Note that this topic is composed of artificial images, while the other 9 concepts contain digital photos. With the dinosaur concept traditional codebooks outperform fuzzy codebooks at each choice of vocabulary size having a MAP peak value of 0.58 at a cluster amount of 100 and the worst value for 10 clusters achieving 0.45 compared to fuzzy codebooks with an MAP of 0.34.

Still, k-means clustering shows unstable results over different codebook sizes, a hypothesis supported by the results of the flowers concept, the beach concept and the African people and villages concept (see Figure 4). For all of these concepts the performance in terms of MAP is stable at varying codebook sizes for the fuzzy approach, while the traditional approach leads to peaks at different codebook sizes, typically somewhere between 40 and 200 visual words.

![Figure 4: MAP for single concepts: Africa people and villages, dinosaurs, beach, flowers](image)

Above investigations indicate that fuzzy clustering in contrast to k-means clustering leads to more stable results. However, especially with the dinosaur concept, featuring artificial images, the traditional approach outperforms our fuzzy codebook approach. All other concepts in the investigated image collection solely contain digital photos. Therefore, we conducted a further series of tests investigating only 9 concepts of the Simplicity data set leaving out the dinosaur concept. The overall MAP for 9 concepts is lower than the MAP for 10 concepts for both approaches, due the high impact of the artificial images in terms of MAP as both approaches work best on the dinosaur pictures. What we can see in figure 5 is, that k-means performs better than fuzzy c-means with a vocabulary size ranging from 10 to 40, reaching a peak value of 0.06. The fuzzy approach remains constantly at level of 0.054, performing better than k-means with a codebook size of 50 and more. The performance of of the traditional approach is rapidly declining with growing size of the vocabulary. With 200 visual words for instance the fuzzy codebook approach achieves double MAP compared to the traditional approach.
Table 2: MAP, subset with 9 concepts

<table>
<thead>
<tr>
<th>clusters</th>
<th>fuzzy</th>
<th>trad.</th>
<th>fuzzy</th>
<th>trad.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.053954054</td>
<td>0.059983086</td>
<td>0.054002713</td>
<td>0.057406135</td>
</tr>
<tr>
<td>20</td>
<td>0.054308078</td>
<td>0.059040424</td>
<td>0.054372135</td>
<td>0.054833569</td>
</tr>
<tr>
<td>30</td>
<td>0.054372135</td>
<td>0.059040424</td>
<td>0.054372135</td>
<td>0.054833569</td>
</tr>
<tr>
<td>40</td>
<td>0.054212391</td>
<td>0.052313621</td>
<td>0.054212391</td>
<td>0.052313621</td>
</tr>
<tr>
<td>50</td>
<td>0.054007505</td>
<td>0.037920877</td>
<td>0.054007505</td>
<td>0.037920877</td>
</tr>
<tr>
<td>100</td>
<td>0.054029564</td>
<td>0.026796618</td>
<td>0.054029564</td>
<td>0.026796618</td>
</tr>
<tr>
<td>200</td>
<td>0.054014670</td>
<td>0.022477977</td>
<td>0.054014670</td>
<td>0.022477977</td>
</tr>
<tr>
<td>300</td>
<td>0.054319566</td>
<td>0.020947603</td>
<td>0.054319566</td>
<td>0.020947603</td>
</tr>
<tr>
<td>400</td>
<td>0.054319566</td>
<td>0.020947603</td>
<td>0.054319566</td>
<td>0.020947603</td>
</tr>
</tbody>
</table>

In addition to the MAP we also investigated the error rate. As with the MAP the error rate remains more or less constant around 0.825 with the fuzzy approach, while the error rate for the traditional approach varies in [0.76, 0.922]. For both experiments, the one with all and the one leaving out the dinosaur concept, the fuzzy approach yields better results with 100 and more visual words. Values from the experiments are given in Table 3. The results are visualized in Figure 6 and Figure 7.

Table 3: Error rate for data sets with 10 and 9 concepts

<table>
<thead>
<tr>
<th>clusters</th>
<th>fuzzy (10)</th>
<th>trad. (10)</th>
<th>fuzzy (9)</th>
<th>trad. (9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.755</td>
<td>0.722</td>
<td>0.833</td>
<td>0.804</td>
</tr>
<tr>
<td>20</td>
<td>0.745</td>
<td>0.700</td>
<td>0.823</td>
<td>0.787</td>
</tr>
<tr>
<td>30</td>
<td>0.743</td>
<td>0.673</td>
<td>0.821</td>
<td>0.760</td>
</tr>
<tr>
<td>40</td>
<td>0.740</td>
<td>0.711</td>
<td>0.821</td>
<td>0.800</td>
</tr>
<tr>
<td>50</td>
<td>0.745</td>
<td>0.679</td>
<td>0.825</td>
<td>0.767</td>
</tr>
<tr>
<td>100</td>
<td>0.748</td>
<td>0.751</td>
<td>0.826</td>
<td>0.848</td>
</tr>
<tr>
<td>200</td>
<td>0.748</td>
<td>0.801</td>
<td>0.826</td>
<td>0.893</td>
</tr>
<tr>
<td>300</td>
<td>0.748</td>
<td>0.830</td>
<td>0.826</td>
<td>0.922</td>
</tr>
<tr>
<td>400</td>
<td>0.741</td>
<td>0.828</td>
<td>0.823</td>
<td>0.921</td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this paper we investigated the consequences of fuzzy clustering for codebook generation in combination with soft assignments and compared it with the traditional codebook approach using hard assignment. Our results indicate that fuzzy codebooks yield more robust results in terms of MAP and error rate with respect to codebook size. Moreover the error rate of the fuzzy approach is considerably lower with codebook sizes of 100 and more visual words compared to the traditional approach. Considering only digital photos from the Simplicity data set by leaving out the dinosaur concept fuzzy codebooks even outperform traditional codebooks in terms of MAP with a codebook size of 50 or more visual words.

While the increase in MAP and error rate seem marginal in terms of absolute numbers, the impact on practical applications of our novel fuzzy approach is still high. Considering applications taking many different domains of images into account, like for instance in a web photo portal where users are allowed to upload arbitrary images, the fuzzy approach allows for stable results over different domains without the need to find the proper vocabulary size for each of the domains. The traditional approach on the contrary appears to be rather sensitive to changes in the vocabulary size. Plainly speaking for the fuzzy approach the codebook size is more or less a one fits all parameter. On the contrary choosing
the wrong codebook size with the traditional approach can go horribly wrong in terms of MAP and error rate.

In the future we aim at examining further fuzzy clustering algorithms (like the Gustafson-Kessel clustering algorithm [6]) and their contribution to the stability and the performance of fuzzy codebooks. We want to compare our fuzzy approach with state of the art soft assignment techniques by using a more significant data set like the PascalVoc and/or the TrecVid dataset. We further want to investigate if the fuzzy codebook approach performs the same way with differing local features, like SURF or image patches as well as the impact of those local features on the stability in general.

6. ACKNOWLEDGMENTS

This work was supported by the Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant 20214/17097/24774.

7. REFERENCES


