Impact of a Codebook Filtering Step on a Galois Lattice Structure for Graphics Recognition

Ameni Boumaiza, Salvatore Tabbone

Impact of a Codebook Filtering Step on a Galois Lattice Structure for Graphics Recognition

Ameni Boumaiza and Salvatore Tabbone
Université de Lorraine- LORIA UMR 7503, 54506 Vandœuvre-lès-Nancy, France
{boumaiza, tabbone}@loria.fr

Abstract

In this paper, we propose an evaluation of the impact of a codebook filtering step on the recognition rate of a Galois lattice classifier. Unlike the usual approach which only considers a whole visual dictionary and is likely to over-fitting, we boost the Galois lattice using a filtered dictionary by assigning a probability of appearance to each visual word in a symbol model. The retrieval performance and behavior of the method have been compared to state-of-the art and proved that is suitable to the recognition process. Experimental results show that the Galois Lattice classifier combined with a filtered codebook outperforms classic classifiers. Interestingly, due to the high selection of features from the dictionary, the accuracy improvement is obtained with a considerable computational cost reduction.

1 Introduction

We address the task of graphics classification, which aims to assign a category for a queried symbol. Motivated by the great success of the Galois lattice classifier in the text mining field [4], several techniques in the symbol recognition field have been proposed [2, 5, 7, 1] which have shown very good performances. These state-of-the-art methods require an intensive learning/training stage due to the use of a big number of features either by the use of the bag of words [2] or by the discretization step for features vectors [5]. We argue that these two practices commonly used in symbol classification methods, have led to good performance but still also have some limits. For the purpose of better discrimination and less computational complexity, bag of words features representation should be reduced and filtered out which is indispensable and crucial to the development of a reliable graphics recognition system. The present work[1] is part of an ongoing efforts [2, 3] to improve the performance of a symbol recognition system based on the Galois lattice classifier by studying the impact of the codebook on the Galois lattice performances. We will see that the filtering of the codebook into frequent visual words will boost the performance of the lattice.

2 Proposed approach

2.1 Codebook Generation

In the codebook generation, a clustering algorithm (e.g. k-means) takes the training image descriptors as an input and quantizes the descriptors into n clusters. Then, the center points of each cluster are used to define the codebook and to give a visual word representation of features.

2.2 Codebook Filtering

When the size of the codebook is large and the amount of training data is small, empty or almost empty clusters could be generated in the codebook. These small clusters might affect the performance of the system, and it might be possible to improve the performance by pruning empty and small bins out from the codebook. This study aims to evaluate each part of the dictionary because rare visual words can be of major interest as frequent are. The smaller the pruning factor, the fewer words in the codebook after pruning. Let $|S|$ be the number of symbols in the database, $|\Omega|$ the size of the dictionary, $|S_k|$ the number of symbols in a class k, $|w_k^i|$ the number of visual word $i\,(=1..\Omega)$ describing the symbols belonging to the class k. Then, $Freq[w_k^i, S_k] = \frac{|w_k^i|}{|S_k|}$ is the frequency of appearance of a visual word i into a class k. The higher the value of

[1] This work was supported by the European project Eureka 4462 SCANPLAN

$Freq$ the more frequent is the visual word for the class i. For instance $Freq = 1$ means that the visual word i appears in each symbol of the class k and reciprocally if $Freq = 0$ the class k is not encoded by the word i but this word could encode another classes of symbols.

2.3 Structure of the Galois Lattice Classifier

Algorithm 1 Proposed Algorithm

```
1. *Codebook Filtering*
**Input** $w^i$ a visual word, $\Omega$: the visual dictionary,
S: the database, $Freq[w^i, S_k]$ the frequency of appearance of the visual word $i$ in the class $k$
$\theta$ a threshold representing x% of the dictionary of each visual word
For each w$^i \in \Omega$ do
If Freq[w$^i_.$, S$_.$]$\geq \theta$
$\Omega_{Freq}$=Push(w$^i_.$)
else $\Omega_{NonFreq}$=Push(w$^i_.$)
EndIf
2. *Building the Discretized Table of the Galois lattice*
If a symbol S is characterized by w$^i$ then
then R(S, $w^i_.$)=1
else R(S, $w^i_.$)=0
EndIf
3. *Building the Hass Diagram of the Galois lattice based on the Filtered Codebook (For more details, we use a modified Bordat algorithm for the building step[6])*
For k=1 to NumberOfSymbolsPerCategories
For i = 1 to NumberOfSymbolsPerCategory $\mathcal{C}$[k]
For j = 1 to NumberOfConcepts
if ((x$_i$, y$_i$, FIND (x$_i$, y$_i$)=**FALSE**)
For j=1 to NumberOfAttributes
Insert($A${*VisualWords*[j]}) in the Galois lattice=f(C{*VisualWords*[i]}))
$\mathcal{O}_{Ck}(x_i$, y$_i$)=$\mathcal{C}${k}(x$_i$, y$_i$)
else "Concept already exists in the Galois lattice"
Endfor
Endfor
Endfor
**Output** The Galois Lattice:{Symbols, FilteredProperties}
```

Our approach aims to find the frequent word of the dictionary Ω which guarantees the best recognition rate of the Galois Lattice following the frequent and non-frequent sets of words defined respectively as: Ω_{Freq}: w$_{ik}$ →R, where Freq[w$_i$, S$_k$]$\geq \theta$, and $\Omega_{NonFreq} = \overline{\Omega_{Freq}}$.

The Galois lattice (GL) based only on the frequent visual words is named in this paper the Dominant Galois Lattice (DGL) and NDGL (Non Dominant Galois Lattice) for the non-frequent words (see Algorithm 1).

2.4 The traversal of the Hass diagram of the Galois lattice for the classification phase

Given a query symbol, and its visual words, the Hass diagram is traversed in order to recognize the query symbol. A traversal of the Hass diagram often leads to visit all the concepts containing a set of attributes describing the queried symbol. The node containing the primitives of the query is returned as an result of the recognition process.

3 Experimental Results

In order to carry out our experiments we use the GREC2003 symbol database as models. By running experiments with different sizes of the visual dictionary, we study the behavior of the lattice and more precisely how it would perform either with frequent or non-frequent words as shown in Figure 1.

3.1 Impact of the Codebook size on the recognition rate of the Galois Lattice classifier

In this section, we will study the impact of the visual dictionary filtering step on the Galois lattice recognition performances. This evaluation provides an empirical basis for designing visual-words that are likely to produce superior classification performance. In our approach, we opt to a specific criterion for measuring the "informativeness" of each word in order to eliminate the least informative ones before the classifier construction step. As shown in Figure 1, the size of the codebook is crucial for the performance. It affects features that are generated and fed to the Galois lattice classifier. When the codebook size is less than 100 visual words, we remark in Figure 1. a and b that the system looses its ability to classify queried symbols when using frequent or non frequent features separately. This can be explained by the fact that with fewer words in the codebook, features are not distinctive enough, and when using the whole vocabulary Ω the system can recognize symbols better. We conclude that if the codebook is pruned too much then it looses important characteristics especially when its size is smaller. When the size of the vocabulary $\|\Omega\|$ is larger than 300, DGL based on dominant visual words $(\{\Omega\} \backslash \{\Omega_{NonFreq})\}$ shows best results because the most rare words in their corresponding symbol class are removed in the pruning step. This technique improves the classifier performances and when the dictionary is pruned, it performs better than the GL which is not pruned. We remark also that when $\|\Omega\|$ varies from 500 to 800, the performance of the DGL still stable and constant. The NDGL gives bad results in the classification step and many outliers appear when returning concepts from the Galois lattice classifier which are likely to contain the queried symbol.

3.2 Impact of the probability criterion θ on the Galois Lattice recognition rate

The experiment shown in Figure 2 suggests to search for the optimal value of the parameter θ to be used as a threshold to prune the dictionary and build a competitive classifier. We examine the optimal parameter on the best vocabulary size. We found that if the probability of appearance of a word is greater than θ, the recognition rate of the Dominant Galois lattice is stable and substantially improved when using the frequent words of the vocabulary. The Frequent words extracted from the visual dictionary and which have the most important probability of appearance in the symbol database are clearly more effective since they are used by our approach based on the Dominant Galois lattice. We can remark that following the size of the codebook, we got the best results when we remove at least half ($\theta \simeq 0.5$) of the visual words less frequent in each class. When the size of the visual vocabulary is small (50) the non frequent words are highly correlated to the frequent words and participate on the good recognition rate returned by the classifier. The rare words are informative enough and are complementary to frequent visual words that 's why it will be better to keep the whole codebook to give satisfiable information about a symbol category. When the size of the codebook is doubled from 100 to 200, the performance of the system improves with pruned codebooks. However, the difference in the performances between a non-pruned codebook and the best pruned codebook is minor. Although the difference between non-pruned and the best pruned codebook is relatively small, it is significant especially when the size of the codebook is 200 or larger. Indeed, with the 500 words codebook, the performance is high when the codebook is pruned with a factor greater than 0.5 and the best recognition rate is obtained when we keep only the representative visual words (i.e. $\theta = 1$ means that Freq=1) in each class. We remark that there is a large gap between the performance of the Galois lattice based on frequent attributes and the Galois lattice based on the non frequent attributes which can be explained by the fact that frequent words are widely-spread in the symbol database, and therefore they bring a sufficient information to classify a queried symbol using our classifier. In this case, we can remove all the non frequent words without loss of symbol classification accuracy. These words are not very relevant to the category of the graphical symbol. As shown in this experiment, the recognition rate rises significantly when using frequent words Ω_{Freq} for building the classifier, whereas, when using the non frequent words $\Omega_{NonFreq}$ the level drops in comparison with Ω_{Freq}.

This experiment shows that when the size of the codebook increases the symbol are better described but we add noise in the representation. In this case we improve the performance of the classifier if we keep the more informative words in each class. On the contrary, if the size of the codebook is small we reduce the noise, all the words are useful but the description is poor. Therefore, the performance of the classifier is lower.

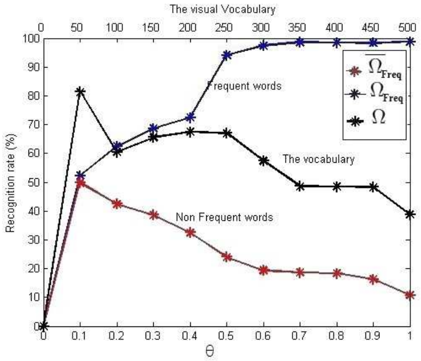

Figure 2. Comparison between the recognition rate of the DGL$_{\Omega_{Freq}}$, NDGL$_{\Omega_{NonFreq}}$ and GL$_{\Omega}$ as a function of θ. In this experiment the size of the vocabulary $\|\Omega\|$ is set between 50 and 500 visual words.

3.3 Comparison with the State of the Art

In Table 1 (column 2) we show the obtained recognition rates when considering just the topmost visual words of the vocabulary, in (column 3) we show the performances of the NDGL based on rare visual words. This experiment indicates that the performance is also highly dependent on the filtering step of the codebook. With our approach which is based on the highest level of frequency of appearance (dominance criterion) reached by visual words, the classifier attains good recognition rates whereas in the lowest level of frequency of appearance some symbol designs are badly recognized provoking some outliers. For instance, as shown in (row4, column2) and (row4, column3), we can remark that the recognition rate obtained using the DGL (98.4%) is sig-

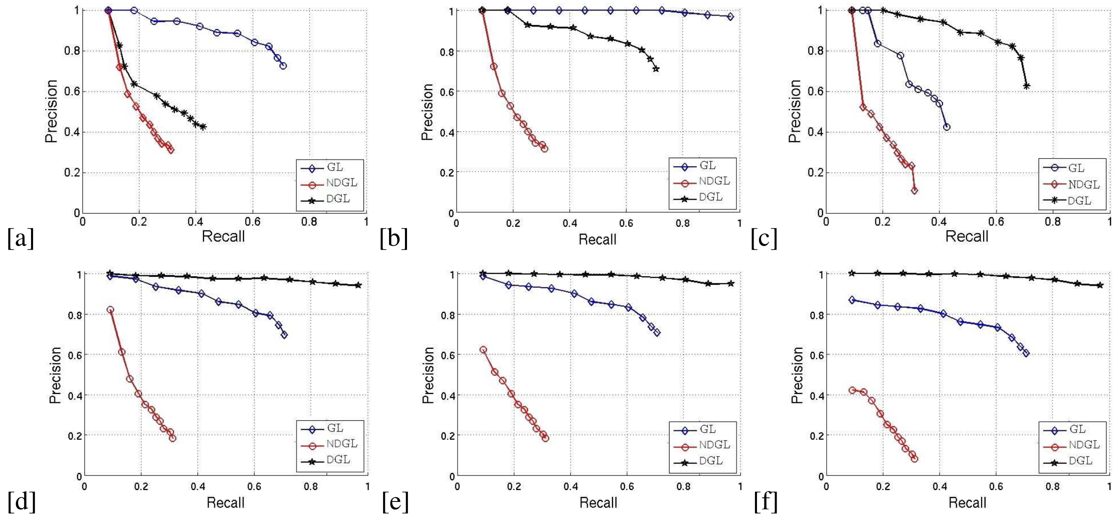

Figure 1. Comparison between the performances of the three classifiers: DGL$_{\Omega_{Freq}}$, NDGL$_{\Omega_{NonFreq}}$ and the GL$_{\Omega}$ as a function of the size of the dictionary (a) $\|\Omega\|$=50, (b): $\|\Omega\|$=100, (c): $\|\Omega\|$=300, (d): $\|\Omega\|$=400, (e): $\|\Omega\|$=500, (f): $\|\Omega\|$=800.

nificantly greater than the other methods based on the Galois lattice too. In addition, our approach guarantees a reduced processing time during the traversal of the Hass diagram (109.7s for our approach instead of 231.74s for [5] and 153.85s for the Galois lattice based on the whole dictionary. The time needed during the classification step is lower because the number of nodes is reduced in the lattice. Due to the high features selection level to build the Galois lattice, there is a larger number of un-informative concepts or nodes eliminated from the graph. The concept which is likely to contain the category of the queried symbol is returned in a reduced time in comparison with the classical Galois lattice. Therefore, the number of concepts is reduced when using our approach (2100 nodes) instead of 3170 nodes for the GL and 4230 nodes for [5].

Table 1. Performance evaluation of GL[2], DGL, NDGL and GL[5] in term of NC: Number of Concepts, PT: Processing Time and RR: Recognition Rate.

Approach	GL[2]	DGL	NDGL	Approach in [5]
NC	3170	2100	1050	4230
PT(s)	153.85	109.7	67.45	231.74
RR(%)	96.08	98.4	43.02	94.62

4 Conclusions

In this paper, we study the impact of the codebook filtering step on classifier performance. From the experiments, we can remark that keeping only dominant words improve the performance of the lattice in terms of recognition rate and time complexity. More precisely, when the size of the codebook increases we improve the description of symbols but we add noise in the representation. In this case we improve the performance of the classifier if we keep the more informative words.

References

[1] S. Barrat and S. Tabbone. A bayesian network for combining descriptors: application to symbol recognition. *International Journal on Document Analysis and Recognition*, 13(1):65–75, 2010.

[2] A. Boumaiza and S. Tabbone. A novel approach for graphics recognition based on galois lattice and bag of words representation. In *ICDAR*, pages 829–833, 2011.

[3] A. Boumaiza and S. Tabbone. Symbol recognition using a galois lattice of frequent graphical patterns. In *10th IAPR Workshop on Document Analysis Systems (DAS)*, 2012.

[4] C. Carpineto and G. Romano. Using concept lattices for text retrieval and mining. In *Formal Concept Analysis*, pages 161–179, 2005.

[5] M. Coustaty, S. Guillas, M. Visani, K. Bertet, and J.-M. Ogier. On the joint use of a structural signature and a galois lattice classifier for symbol recognition. In *GREC*, pages 61–70, 2007.

[6] B. Ganter and R. Wille. *Formal concept analysis - mathematical foundations*. Springer, 1999.

[7] T.-O. Nguyen, S. Tabbone, and O. R. Terrades. Symbol descriptor based on shape context and vector model of information retrieval. In *Document Analysis Systems*, pages 191–197, 2008.