Effective Use of Frequent Itemset Mining for Image Classification

Basura Fernando[1], Elisa Fromont[2], and Tinne Tuytelaars[1]

[1] KU Leuven, ESAT-PSI, IBBT (Belgium)
[2] University of Saint-Etienne(France)

Abstract. In this paper we propose a new and effective scheme for applying frequent itemset mining to image classification tasks. We refer to the new set of obtained patterns as *Frequent Local Histograms* or FLHs. During the construction of the FLHs, we pay special attention to keep all the local histogram information during the mining process and to select the most relevant reduced set of FLH patterns for classification. The careful choice of the visual primitives and some proposed extensions to exploit other visual cues such as colour or global spatial information allow us to build powerful *bag-of-FLH*-based image representations. We show that these *bag-of-FLH*s are more discriminative than traditional bag-of-words and yield state-of-the art results on various image classification benchmarks.

1 Introduction

Even though frequent itemset mining techniques (FIM) and variants thereof are well-established in the data-mining community [1, 2], they are, to date, not commonly used in state-of-the-art image classification methods. This is surprising, since it has been shown that these mining methods allow the construction of high-level sets of compound features which can, in many cases, capture more discriminative information [3]. Nevertheless, most attempts so far applying FIM to image classification [4–7] were not able to demonstrate competitive results on standard datasets. In this paper, we propose a new and effective scheme for applying FIM to image classification by adapting the generic FIM tools to the specific context of visual features extracted from images.

In most state-of-the art image classification methods, images are represented with bag-of-visual-words (BOW), i.e. histograms over vector-quantised local features [8]. These can be computed either globally (BOW) or locally in a neighbourhood around an interest point (LBOW, e.g. [6, 7, 9, 10]). In the context of FIM-based image classification, local bag-of-words are usually preferred, since they result in sparse representations, a better signal-to-noise ratio and an increased robustness to image clutter. But before FIM can be applied, these bags (histograms) need to be converted into *sets of items* known as *transactions*. Usually [4, 5, 7, 10, 11], this is done by considering each visual word as an item, ignoring how many times it appears in the bag (i.e. only considering the absence/presence of the visual word). This can lead to a loss of information.
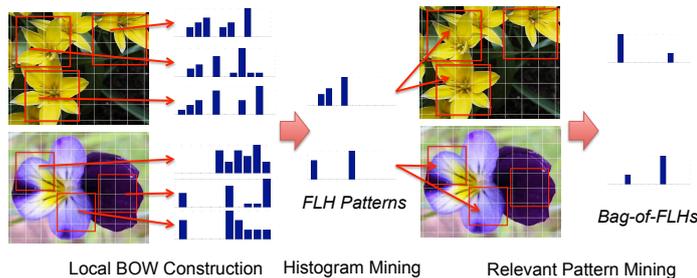
Fig. 1. FLH mining and image representation

Moreover, FIM typically generates a large number of patterns, of which only a subset is relevant for classification. Relevant pattern discovery is studied in the data mining community for the case where each transaction is generated from a different source [3, 12]. However, when using local bag-of-words to create transactions, each source (image) generates multiple transactions and a pattern that is found only in a relatively small number of images can still be frequent. Applying standard relevant pattern discovery methods under these circumstances [5–7, 13] may not be the best strategy.

In this paper we also start from the local bag-of-words representation. Our three major contributions are the following: *Firstly*, we propose a method for histogram mining that discovers a novel set of local patterns called *Frequent Local Histograms* or *FLH*s, avoiding the loss of information during the conversion to transactions. We experimentally show that using less discriminative visual primitives and including histograms extracted from larger neighbourhoods result in larger and more discriminative FLH patterns. *Secondly*, we propose a new relevant pattern mining method to select discriminative, representative and non-redundant FLH patterns taking into account the fact that each image generates multiple transactions. Using this selected set of *FLH* patterns, we build a new image representation called *bag-of-FLHs* and we propose a suitable kernel for classification. The *bag-of-FLHs* creation process is shown in Figure 1. *Thirdly*, we propose a couple of variants to our basic scheme, integrating additional visual cues such as colour or global spatial information. The novel *bag-of-FLHs* and its variants yield powerful image representations, leading to state-of-the-art results on various image classification datasets.

The rest of the paper is organised as follows. First, we review related work in section 2. Section 3 provides details on the construction of relevant FLHs and shows how they can be used for image classification. In Section 4 we show how additional visual cues can be incorporated. Section 5 describes the experimental validation, demonstrating the power of our method for challenging image classification problems. Finally, Section 6 concludes the paper.

2   Related work

Frequent pattern mining techniques have been used to tackle a variety of computer vision problems, including image classification [4, 7, 14, 15], action recognition [16, 13], scene understanding [5], object recognition and object-part recognition [6]. Apart from the application, these methods mostly differ in the image representation used, the way they select relevant or discriminative patterns, and the way they convert the original image representation into a transactional description suitable for FIM techniques.

Image representations In [5], Yao *et al.* present a structured image representation called *group-lets*. To find discriminative group-lets, they mine for frequent class association rules. Most other methods [6, 7, 10, 15, 16] start from local bag-of-words as image representations. Spatial configuration mining based on LBOW was shown by Quack *et al.* [6], but they did not use these configurations for classification. More structured patterns such as sequences and graphs capturing the spatial distribution of visual words have been used by [4], while [11] uses boosting on top of binary sets of visual words discovered by FIM. Gilbert *et al.* [13] have applied itemset mining to action recognition using rather primitive features like corners, while in [14] high level features such as attributes [17] are successfully used with mining techniques.

Mining relevant patterns With an appropriate selection criterion, frequent patterns can be more discriminative than individual features (e.g. visual words), since a pattern is a combination of several primitives and therefore likely to capture more of the underlying semantics of the data. However, many works applying FIM in computer vision ignore issues such as redundancy and (lack of) repeatability across images of the mined patterns. They use simple class-based association rules [5, 6, 13, 16] or other supervised methods [4, 11]. In [7], Yuan *et al.* present a semantically meaningful visual pattern mining algorithm based on a likelihood ratio test to find relevant patterns in an unsupervised manner. However, none of these works considers the issue of repetitive structures in images, causing frequent yet not representative patterns.

Transforming bags to transactions As indicated earlier, most methods simply use individual visual word as an item in a transaction. Transactions are created in such a way that if a visual word is present in the histogram, then it is also present in the transaction (i.e. an itemset). In [15], Kim *et al.* use a new representation called *Bag-to-set* (B2S) to transform a histogram into a transactional form without loosing information. The B2S representation is, to our knowledge, the only unsupervised effort to explicitly avoid information loss in the conversion to transactions. However it can generate artificial visual patterns that do not exist in the image database. Another possibility could be to give each visual word a weight depending on how many times it appears in the histogram and apply weighted itemset mining [18]. However, this method only postpones the loss of information instead of really solving it, as it then simply sums all the itemset weights together to discover useful patterns.

Spatial configurations without mining Apart from pattern mining techniques, other methods have been proposed to exploit local spatial information as well. However, most of them are limited to the use of pairs or triplets of features. Only a few have used higher-order statistics (co-occurrence of visual words), either on a single image [19] or pairs of images [20]. Unlike FIM, they do not exploit database-wide statistics. FLH patterns are also somewhat similar to the mid-level features learnt by [21]. However, *FLH* patterns are loosely coupled to the local geometry and robust to spatial deformations, occlusions and image clutter.

Also related is the work on hyper-features of Agarwal and Triggs [9]. Like us, they start from local-bag-of-words but *cluster* them recursively (in a hierarchical fashion) to find a new set of spatial features called hyperfeatures. Then they represent each image as a bag-of-hyperfeatures. While this also captures larger patterns, it does not have the same flexibility in local geometry as our scheme.

3   FLH-based Image Representation and Classification

After introducing some notations, we explain how we mine frequent local histograms (FLHs) (section 3.1). We then show how we select the most relevant set of FLHs for image classification (section 3.2) and present a suitable kernel for frequent relevant pattern-based image classification (section 3.3).

Each image I is described by a set of key points $\{f_i | i = 1 \ldots n_I\}$ and a class label c, $c \in \{1 \ldots C\}$. We assume that all the descriptors have been clustered to obtain a set of so-called visual words. Then, each key point f_i is given a label $w_i \in W$ known as the visual word index. $|W|$ is the visual word dictionary size. In our approach, for each key point f_i we compute a *local histogram* also called a *local bag-of-words* (LBOW), $\mathbf{x_i} \in \mathbb{N}^{|W|}$ using the K nearest neighbours of f_i (based on the distance between image coordinates and also including f_i itself as a neighbour). The set of all the local histograms $\mathbf{x_i}$ created from all images is denoted by Ω.

3.1   Frequent local histogram mining

Items, Transactions and Frequencies : In order to avoid loss of information during the transaction creation process we propose the following new definition of an *item*. An item is defined as a pair (w, s), $w \in W$ and $s \in \mathbb{N}$, with s being the frequency of the visual word w in the local histogram. Note that $0 < s \leq K$. We define Γ as the set of all possible items, so $|\Gamma| = K \cdot |W|$. We create the set of *transactions* X from the set of local histograms Ω. For each $\mathbf{x} \in \Omega$ there is one transaction x (i.e. a set of items). This transaction x contains all the items (w_j, s_j) such that the bin corresponding to w_j in $\mathbf{x}$ has the nonzero value s_j. A *local histogram pattern* is an itemset $t \subseteq \Gamma$. For any local histogram pattern t, let $X(t) = \{x \in X | t \subseteq x\}$ be the set of transactions that include the pattern t. The *frequency* of t is $|X(t)|$ also known as the *support* of the pattern t or *supp(t)*.

Frequent Local Histogram : For a given constant T, also known as the minimum support threshold, a local histogram pattern t is *frequent* if *supp(t) $\geq$ T*. Two patterns t and t' are said to be *equivalent* if $X(t) = X(t')$. This implies that *supp(t) = supp(t')*. Each collection of equivalent patterns forms an equivalent class. The largest element (i.e. the one with the highest number of items) of an equivalent class is called a *closed* pattern. The set of frequent closed patterns is a compact representation of the frequent patterns (i.e we can derive all the frequent patterns from the closed frequent ones). In this work we refer to a frequent and closed local histogram pattern as a *Frequent Local Histogram* or **FLH**. Υ is the set of all FLHs.

FLH Mining: We can use any existing frequent mining algorithm to find the set of FLHs Υ. What is specific to our method is that i) the input of our algorithm is a set of local histograms Ω, and ii) a preprocessing step is performed building the set of transactions X from the local histograms $\mathbf{x_i}$ as described above. Items (w_k, s_k) in a transaction $x \in X$ can then be regarded as standard items in itemset mining. To directly mine closed frequent histogram patterns we use the optimised *LCM* algorithm [2].

Encoding local-bag-of-words using FLH: Given a key point, we compute a LBOW around it, considering its K nearest neighbours. Given this LBOW $\mathbf{x}$, we convert it into a transaction x and check for each FLH pattern $t \in \Upsilon$ whether $t \subseteq x$. If $t \subseteq x$ is true, then $\mathbf{x}$ is an *instance* of the FLH pattern t. The frequency of a pattern t in a given image I_j (i.e., the number of instances of t in I_j) is denoted as $F(t|I_j)$.

3.2   Finding the best FLHs for image classification

We want to use the FLH set Υ as a new set of mid-level features to represent an image. However, we first need to select the most useful FLH patterns from Υ because i) the number of generated FLH patterns is huge (several millions) and ii) not all discovered FLH patterns are equally relevant for the image classification task. Usually, relevant pattern mining selects those patterns that are *discriminative* and *not redundant*. On top of that, we introduce a new selection criterion, *representativity*, that takes into account that, when using LBOW, a single image generates multiple transactions. As a result, some patterns may be frequent and considered discriminative but they may occur in very few images (e.g. due to repetitive structures). We believe that such features are not representative and therefore not the best choice for image classification. A good FLH pattern should be at the same time discriminative, representative and non-redundant. In this section we discuss how we select such patterns.

Relevance criterion: We use two criteria for pattern relevance: a *discriminativity score* $D(t)$ [3] and a new *representativity score* $O(t)$. The overall relevance of a pattern t is denoted $S(t)$ where $S(t) = D(t) \times O(t)$. We claim that if a pattern t has a high relevance score $S(t)$, it is likely to be discriminative and repeatable across images, hence suitable for classification.

Discriminativity score: To find discriminative patterns, we follow the entropy-based approach of [3], where a *discriminativity score* $D(t)$ $(0 \leq D(t) \leq 1)$ for pattern t is defined as :

$$D(t) = 1 + \frac{\sum_c p(c|t) \cdot \log p(c|t)}{\log C}, \tag{1}$$

with $p(c|t)$ the probability of class c given the pattern t, computed as follows :

$$p(c|t) = \frac{\sum_{j=1}^{N} F(t|I_j) \cdot p(c|I_j)}{\sum_{i=1}^{N} F(t|I_i)}. \tag{2}$$

Here, I_j is the jth image and N is the total number of images in the dataset. $p(c|I)$ is 1 if the class label of I_j is c and 0 otherwise. A high value of $D(t)$ implies that the pattern t occurs only in very few classes. Note that in Eq. 1, the term $\log C$ is used to make sure that $0 \leq D(t) \leq 1$.

Representativity score: We compare the distribution of the patterns over all the images with the optimal distribution with respect to a class c. This optimal distribution is such that i) the pattern occurs only in images of class c, i.e. $P(c|t^*) = 1$ (giving also a discriminativity score of 1), and ii) the pattern instances are equally distributed among all the images of class c, i.e. $\forall I_j, I_k$ in class c, $p(I_j|t^*) = p(I_k|t^*) = (1/N_c)$ where N_c is the number of images of class c. To find patterns with distributions close to the optimal one, we define the *representativity score* of a pattern t denoted by $O(t)$. It considers the divergence between the optimal distribution $P(I|t_c^*)$ and $P(I|t)$, and then takes the best match over all classes:

$$O(t) = \max_c (\exp -[D_{\mathrm{KL}}(P(I|t_c^*)||P(I|t))]) \tag{3}$$

where $D_{KL}(.||.)$ is the Kullback-Leibler divergence between two distributions, $P(I|t_c^*)$ is the optimal distribution for the class c and $P(I|t)$ is the distribution for pattern t. $P(I|t)$ is computed empirically from the frequencies $F(t|I_j)$ of the pattern t.

Redundant patterns: We propose to remove redundant patterns in order to obtain a compact representative set of FLHs. We take a similar approach as in [22] to find affinity between patterns. Two patterns t and $s \in \Upsilon$ are redundant if they follow similar document distributions, i.e if $P(I|t) \approx P(I|s) \approx P(I|\{t,s\})$ where $P(I|\{t,s\})$ gives the document distribution given both patterns $\{t,s\}$. We define the redundancy $R(s,t)$ between two patterns s,t as follows :

$$R(s,t) = \exp -[p(t) \cdot D_{KL}(P(I|t)||P(I|\{t,s\})) + p(s) \cdot D_{KL}(P(I|s)||P(I|\{t,s\}))] \tag{4}$$

Note that $0 \leq R(s,t) \leq 1$ and $R(s,t) = R(t,s)$. For redundant patterns, $D_{KL}(P(I|t)||P(I|t,s)) \approx D_{KL}(P(I|s)||P(I|t,s)) \approx 0$ which increases the value of $R(s,t)$.

Finding the most suitable patterns for classification: We are interested in finding the most suitable pattern subset χ where $\chi \subset \Upsilon$ for classification. To do this we define the *gain* of a pattern t denoted by $G(t)$ s.t. $t \notin \chi$ and $t \in \Upsilon$ as follows:

$$G(t) = S(t) - max_{s \in \chi}\{R(s,t) \cdot min(S(t), S(s))\} \qquad (5)$$

In Eq. 5, a pattern t has a higher gain $G(t)$ if it has a higher relevance $S(t)$ (*i.e. it is discriminative and representative*) and if the pattern t is non redundant with any pattern s in set χ (*i.e. R(s,t) is small*). To find the best k number of patterns we use the following greedy process. First we add the most relevant pattern to the relevant pattern set χ. Then we search for a pattern with the highest gain (non redundant but relevant) and add this pattern into the set χ until k number of patterns are added (or until no more relevant patterns can be found).

3.3   Kernel function for effective pattern classification

After computing the k most relevant and non-redundant FLHs, we can represent each image using a new representation called *bag-of-FLHs* by counting the occurrences of such FLHs in the image. Let L be such a *bag-of-FLHs* for the image I_L and M be the *bag-of-FLHs* for the image I_M. We propose to use the kernel function $K(L, M) = \sum_i min(\sqrt{L(i)}, \sqrt{M(i)})$ to find the similarities between the *bag-of-FLHs* of L and M. This kernel provides good classification accuracies for frequent pattern-based image representation. It is a standard histogram intersection kernel but with non-linear weighting. This reduces the importance of highly frequent patterns and is necessary since there is a large variability in pattern frequencies. Similar power-low normalization methods are used in improved Fisher-Vector based methods [23].

4   Extending FLH

Exploiting both local and global spatial information using FLH: In order to take advantage of both global and local spatial information, we extend the FLH mining process. We build on the spatial pyramid idea [24] and apply it in our FLH mining framework. First we create LBOW for all features in the image. Then we discover grid-specific relevant FLHs by employing the process described in Section 3.2. Then, for each image, we concatenate these grid-specific *bag-of-FLH* representations to create a new representation called *GRID-FLH*. The *GRID-FLH* is a more structured local-global representation with a lot of flexibility.

Combining multiple cues using FLH: The influence of color and shape is different for different object classes. We show two ways to fuse multiple cues such as shape and color using our FLH-based method. One way is to fuse shape-based *bag-of-FLHs* (denoted by FLH_S) with a color-based *bag-of-FLHs* (denoted by FLH_C) using a simple average Kernel. We denote this approach "$FLH_C + FLH_S$". In
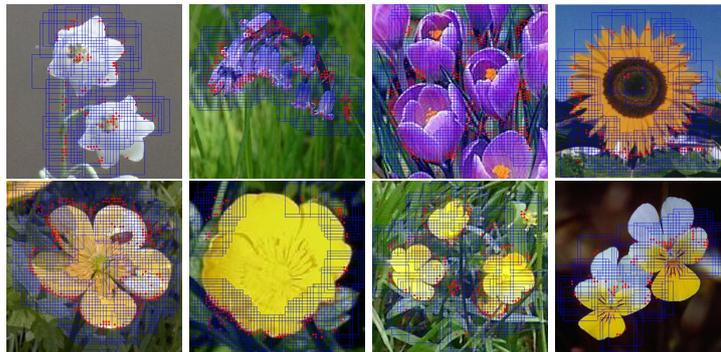
Fig. 2. FLH patterns using shape information where each red dot represents the central location and the blue square represents the size of the FLH pattern

Table 1. Datasets

Dataset	# Classes	# Train. Imgs per class	# Test. Imgs per class	Evaluation Criterion
GRAZ-01	2	same setup as in [24]	as in [24]	ROC Equal Error Rate
Oxford-Flower	17	60	20	Classification accuracy
15-Scenes	15	100	Rest of the images	Mean class-based accuracy
VOC2007	20	see [27]	see [27]	Mean average precision

the second approach, we concatenate a local color histogram with the corresponding local shape BOW and then mine the local color-shape BOW. After that we represent an image using a bag-of *frequent local color-shape histograms*. We call this approach FLH_{CS}. The advantage of FLH_{CS} is that both color and shape information are fused spatially in local neighbourhoods. This allows us to discover useful combinations of local color-shape patterns with a high degree of flexibility.

5   Experimental Setup and Evaluations

In this section we introduce the datasets used to evaluate our FLH-based method; we compare our method with the most relevant baselines; and we present an analysis on parameter selections and design choices. Finally we compare the extensions proposed in Section 4 with the state-of-the-art.

5.1   Datasets and evaluation criteria

We evaluate the new *bag-of-FLH* (hereafter denoted by just FLH) approach on several challenging natural image datasets: *GRAZ-01* [25], *Oxford-Flowers 17* [26], *15-Scenes* [24] and the *PASCAL-VOC2007* dataset [27]. Details of these datasets and evaluation criteria are summarised in Table 1.

For all experiments (and all the datasets), we start from SIFT descriptors [28] densely sampled over the image with 16×16 patches and a grid spacing of 8 pixels. For the Oxford-Flower dataset we experiment not only with SIFT but also with the ColorName descriptors of [29]. We use the K-means algorithm to

Table 2. Comparison with baseline methods. Classification accuracies for GRAZ-01 and Oxford-Flower for a local neighbourhood size of K=5.

	Dict.	Baselines		Mining Methods			
	size	BOW	SPM	FIM	B2S	FLH	FLH + BOW
GRAZ-Person	200	79.4 ± 1.1	79.7 ± 1.4	80.5 ± 2.2	81.8 ± 2.1	**83.5 ± 1.8**	**84.0 ± 1.6**
GRAZ-Bike	200	76.8 ± 2.5	79.6 ± 2.1	78.0 ± 2.1	78.4 ± 1.9	**81.3 ± 2.4**	**82.5 ± 2.2**
Flower-(SIFT)	200	56.4 ± 2.5	57.3 ± 2.3	54.7 ± 2.6	55.3 ± 2.9	**59.0 ± 3.4**	**71.1 ± 2.4**
Flower-(CN)	100	59.4 ± 1.7	61.5 ± 1.6	61.5 ± 2.7	62.4 ± 2.4	**69.5 ± 2.5**	**72.9 ± 2.4**

create visual dictionaries and LIBSVM[3] to train an SVM using a constant cost parameter (C = 1). For our methods, the kernel used is the one presented in Section 3.3.

5.2   Comparison with baseline methods

We compare our *FLH*-based method using some default settings with BOW-based image classification, spatial pyramid matching *(SPM)* [24], visual word-based standard frequent itemset mining called *FIM* (with binarized local bag-of-words) and the *B2S* [15] representation. We also report results for the *FLH*-based method combined with *BOW* using an average Kernel (BOW+FLH). For the other baseline methods, we use an intersection kernel. The baseline results are shown in Table 2.

FIM is comparable with SPM while B2S (which is a lossless histogram transformation approach) slightly outperforms FIM. The *FLH*-based method outperforms all the baseline methods. This result shows the importance of not loosing information during the transaction creation time. We believe that our FLH-based method improves the results over B2S because it does not generate artificial visual patterns (i.e. patterns not actually present in the data set) while B2S does. The combination of *BOWs* and *bag-of-FLHs* gives better results compared to all other methods and is an indication of the complementary nature of both representations. Especially the increase for the case of SIFT features on the Flowers dataset is remarkable.

5.3   Parameter selection and optimization

In this set of experiments we analyse the effect of several parameters of our method: dictionary size, SIFT feature size, and local neighbourhood size. We use a three-fold cross-validation on training data to optimize our parameters. In the remaining experiments (sections 5.4 and 5.5), we then use the found optimal parameters to test our FLH-based methods on the test sets.

Smaller dictionaries: We first evaluate the effect of the dictionary size on our *FLH*-based method using the *Oxford-Flower* dataset. We report results for *FLH* and *FLH+BOW* with different dictionary sizes (Table 3). Note that when combining *FLH* and *BOW* we do not reduce the dictionary size for *BOW*. For the

---

[3] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Table 3. The effect of dictionary size on FLH-based methods. Classification accuracy on Oxford-Flowers training data using cross-validation.

Descriptor	Dict. Size	FLH	FLH + BOW	Descriptor	Dict. Size	FLH	FLH + BOW
SIFT	800	**64.4 $\pm$ 2.9**	64.8 $\pm$ 2.9	CN	300	69.1 $\pm$ 2.4	70.4 $\pm$ 2.2
SIFT	200	56.2 $\pm$ 2.7	68.9 $\pm$ 2.7	CN	100	69.9 $\pm$ 2.1	71.1 $\pm$ 2.4
SIFT	100	54.7 $\pm$ 2.6	**70.3 $\pm$ 2.8**	CN	20	**72.0 $\pm$ 1.9**	**76.2 $\pm$ 2.3**

Table 4. Effect of SIFT feature size on FLH. Classification accuracy on Oxford-Flowers training data using cross-validation.

SIFT-D	Dict. Size	FLH	FLH + BOW
128	100	54.7 $\pm$ 2.6	70.3 $\pm$ 2.8
32	100	**72.7 $\pm$ 1.8**	**80.7 $\pm$ 1.4**

Table 5. Effect of neighbourhood size(K), dictionary size ($|W|$) and SIFT descriptor size (B) on classification accuracy for the Oxford-Flower training data using cross-validation.
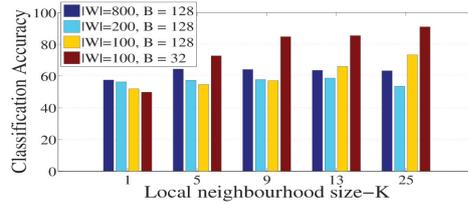
CN descriptor, results improve when the dictionary size is reduced. For SIFT, we see the opposite trend. However, in both cases, the results improve with reduced dictionaries for *FLH+BOW*, indicating that with a smaller dictionary for *FLH*, the complementarity to *BOW* increases. Smaller dictionaries reduce the discriminativity of the visual primitives. However, this does not seem to affect the discriminativity of the FLH patterns, which may even become more robust and stable. Therefore, smaller dictionaries created using even less discriminative features might be better suited for *FLH*. This is tested in the next experiment.

Less discriminative features: We evaluate the effect of less discriminative features on *FLH* using the *Oxford-Flower* dataset with a small dictionary of 100 words and a local neighbourhood size of 5 (Table 4). For this we use $SIFT32$ features that are extracted like the standard SIFT (referred to as $SIFT128$) but performing spatial binning of (2×2) instead of (4×4). Surprisingly, *FLH* using the less discriminative $SIFT32$ features and a smaller dictionary performs as good as the ColorName descriptor in the previous experiment (72.0% for CN and 72.7% for $SIFT32$). When combined with *BOW* (using SIFT128 and a 800 dimensional vocabulary for the BOW) we obtain a classification accuracy of 80.7%, outperforming the ColorName descriptor.

Larger local neighbourhoods We believe that the use of smaller dictionaries and less discriminative SIFT features allows the *FLH*-based method to exploit larger neighbourhoods. To evaluate this relation, we run some further experiments on the *Oxford-Flower* dataset – see Fig. 5. The best results are obtained when reducing both SIFT feature size and dictionary size while increasing the neighbourhood size of the local spatial histograms. The best classification accuracy we obtain with SIFT features is 91.0% for a dictionary size of 100 words, $SIFT32$

Table 6. Effect of relevant pattern mining using SIFT-32

Criterion	Frq.	Rps.	Disc.	Rel.
GRAZ-01	83.1	90.3	90.9	**91.6**
Flower	65.6	84.2	90.9	**92.5**

Table 7. Effect of the choice of the kernel in pattern classification using SIFT-32

$K(\mathbf{x},\mathbf{y}) =$	$\mathbf{x}.\mathbf{y}^t$	$\sqrt{\mathbf{x}}.\sqrt{\mathbf{y}}^t$	$\sum_i min(\mathbf{x_i},\mathbf{y_i})$	$\sum_i min(\sqrt{\mathbf{x_i}},\sqrt{\mathbf{y_i}})$
15-Scenes	80.4	81.3	85.9	**86.2**
Flower	89.5	92.0	91.2	**92.5**

features and a neighbourhood size of 25. Note that this is a larger neighbourhood size (covering up to 48×48 pixels) than what is typically used in the literature [6, 7, 13].

We can conclude that the *FLH*-based method obtains its best results when exploiting larger local patterns with smaller dictionaries and less discriminative features. From now on, for all *FLH*-based experiments we propose to use the SIFT32 descriptor and a neighborhood size of 25 neighbours. For *Oxford-Flower* and *PASCAL-VOC2007* datasets a SIFT32 dictionary of 100 words is used. For all other datasets we propose to use a SIFT32 dictionary of 25 words.

5.4   Effect of relevant pattern mining and of the kernel functions

FLH mining algorithms can generate a large number of FLH patterns (in our case, 1-20 million). Some of these patterns are not relevant for classification. Therefore, we select the most relevant-non-redundant ones, as described in Section 3.2. Here we evaluate the importance of this pattern selection step by comparing different criteria: the most frequent (Frq.), the most representative-non-redundant(Rps.)(eq. 3), the most discriminative-non-redundant (Disc.)(eq.1) and the most relevant-non-redundant (Rel.) (*i.e. representative, discriminative and non-redundant*) patterns (see Table 6). We always select the top 10.000 patterns for each criterion which we believe is sufficient to obtain good results. These results clearly show that only selecting the top-k most frequent patterns (as usually done) is not a good choice for classification. Both representativity and discriminativity criteria alone also do not provide the best results. It's the relevant non-redundant FLH patterns that are the most suitable for classification. Some of these relevant and non-redundant FLH patterns are shown in Fig. 2. These selected FLH patterns are semantically meaningful and capture most of the relevant shape information in an image.

In Table 7, we evaluate the effect of the square root intersection kernel using relevant non-redundant *FLH* patterns on *Oxford-Flower* and *15-Scenes*. The proposed square-root weighting increases in both the linear kernel and the non-linear intersection kernel the classification accuracy for both datasets.

5.5   Comparison to state of the art methods

In this section we compare *FLH* using the parameters optimised as above with, to the best of our knowledge, the best results reported in the literature. Here, we also evaluate the extensions proposed in section 4, integrating shape and color information, or local and global spatial information.

Table 8. Equal Error Rate (over 20 runs) for categorization on GRAZ-01 dataset

Class	State-of-the-art			Our results		
	$SPCK_+$ [19]	NBNN[30]	HF[20]	FLH	FLH+BOW	GRID-FLH
Person	87.2	87.0	84.0	94.0 ± 1.8	95.0 ± 1.6	95.8 ± 1.8
Bike	91.0	90.0	94.0	89.2 ± 1.6	90.1 ± 1.8	91.4 ± 1.1
AVG.	89.1	88.5	89.0	**91.6**	**92.6**	**93.8**

Table 9. Classification accuracy (over 20 runs) on the Flower dataset

Our results Using Shape - SIFT			Our results Using Color - ColorName		
FLH	$FLH + BOW$	$GRID - FLH$	FLH	$FLH + BOW$	$GRID - FLH$
92.5 ± 1.6	**92.7 ± 1.2**	**92.9 ± 1.6**	72.1 ± 1.9	74.4 ± 2.4	74.8 ± 1.9

State-of-the-art			Our results Using Color and Shape		
Nilsback [26]	CA[31]	$L_1 - BRD$[32]	$FLH_C + FLH_S$	FLH_{CS}	FLH_{CS}+BOW
88.3 ± 0.3	89.00	89.02 ± 0.60	**93.0 ± 2.0**	**93.4 ± 1.5**	**94.5 ± 1.5**

GRAZ-01: The results reported in Table 8 show that on average all *FLH*-based methods outperform the state-of-the-art. The $GRID - FLH$ representation, combining local and global spatial information, yields the best results.

Oxford-Flower: The results are reported in Table 9. Note that only using shape information (SIFT32 features) we get a classification accuracy of **92.5%**, outperforming the state-of-the-art. Combining color and shape information further improves these results to **94.5 $\pm$ 1.5** (using FLH_{CS} + SIFT-128 dictionary of 800 and ColorName dictionary of 300 words). Adding global spatial information on this dataset only gives an insignificant improvement of 0.4%.

15-Scenes: Results are shown in table 10. This dataset is strongly aligned. *FLH* does not exploit this and therefore by itself cannot obtain state-of-the-art results. However, the *GRID-FLH* method described in section 4 does take the global spatial information into account and achieves close to state-of-the-art results (86.2%). This is only outperformed by [14] who report 87.8% using CENTRIST and SIFT features along with LLC coding. However, our method uses only simple SIFT features. As far as we know the previous best classification accuracy using SIFT features was reported by Tuytelaars *et al.* in [33] combining a NBNN kernel and a SPM method.

Pascal-VOC2007: Results are reported in Table 11. FLH alone gives a mAP of 60.4. In combination with BOW of SIFT-128 and 5000 visual word vocabulary, we obtain a state-of-the-art mAP of 62.8. Note that the score for each individual class often varies a lot between the FLH+BOW and the Fisher Vector[23] method. Our method does especially well on what are known to be 'hard' classes: bottle (+34% improvement), dining table (+11%), potted plant (+16%), tv monitor (+23%). This suggests that both methods are complementary.

6   Conclusion

In this paper we show an effective method for using itemset mining to discover a new set of mid-level features called *Frequent Local Histograms*. Extensive exper-

Table 10. Results on 15-Scenes dataset

State-of-the-art				Our results		
SPM	*SPCK*++[19]	NBNN Kernel+SPM[33]	(AND/OR)[14]	*FLH*	FLH+BOW	GRID-FLH
80.9	82.5	85.0	**87.8**	70.9 ± 0.4	83.0 ± 0.5	**86.2 ± 0.4**

Table 11. Results on PASCAL-VOC 2007 (Mean average precision)

Class	✈	🚲	🐦	⛵	🍾	🚌	🚗	🐱	🪑	🐄	
FV[23, 34]	78.8	67.4	51.9	70.9	30.8	72.2	79.9	61.4	56.0	49.6	
FLH	67.9	70.6	41.0	54.6	64.9	60.9	85.8	56.6	59.6	40.0	
FLH+BOW	69.2	**73.0**	42.7	56.3	**64.9**	60.9	**86.6**	58.9	**63.3**	41.8	

Class	🪑	🐕	🐎	🏍	🧍	🌿	🐑	🛋	🚂	📺	m.AP
FV[23, 34]	58.4	44.8	78.8	70.8	85.0	31.7	51.0	56.4	80.2	57.5	61.7
FLH	64.7	47.3	56.6	65.7	80.7	46.3	41.8	54.6	71.0	77.6	60.4
FLH+BOW	**74.3**	**48.4**	61.8	68.4	81.2	**48.5**	41.8	**60.4**	72.1	**80.8**	**62.8**

iments have proved that the proposed bag-of-FLH representation, the proposed relevant pattern mining method and the chosen kernel all improve the classification results on various datasets. We have also experimentally shown that the best results for FLH methods are obtained when exploiting a large local neighbourhood with a small visual vocabulary and less discriminative descriptors. Finally, we have shown that extending a local approach such as FLH to exploit other cues such as global spatial information or colour information allows us to obtain state-of-the-art results in many datasets. As future work we propose to investigate how to push the relevant and non redundant constraints directly into the local histograms mining process to make it particularly efficient.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB. (1994) 487–499
2. Uno, T., Asai, T., Uchida, Y., Arimura, H.: Lcm: An efficient algorithm for enumerating frequent closed item sets. In: FIMI. (2003)
3. Cheng, H., Yan, X., Han, J., Hsu, C.W.: Discriminative frequent pattern analysis for effective classification. In: ICDE. (2007) 716 –725
4. Nowozin, S., Tsuda, K., Uno, T., Kudo, T., Bakir, G.: Weighted substructure mining for image analysis. In: CVPR. (2007)
5. Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: CVPR. (2010)
6. Quack, T., Ferrari, V., Leibe, B., Van Gool, L.: Efficient mining of frequent and distinctive feature configurations. In: ICCV. (2007)
7. Yuan, J., Wu, Y., Yang, M.: Discovery of collocation patterns: from visual words to visual phrases. In: CVPR. (2007)
8. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Work. on Statistical Learning in CV. (2004) 1–22
9. Agarwal, A., Triggs, B.: Multilevel image coding with hyperfeatures. Int. J. Comput. Vision **78** (2008) 15–27

10. Sivic, J., Zisserman, A.: Video data mining using configurations of viewpoint invariant regions. In: CVPR. (2004)
11. Yuan, J., Luo, J., Wu, Y.: Mining compositional features for boosting. In: CVPR. (2008)
12. Cheng, H., Yan, X., Han, J., Yu, P.S.: Direct discriminative pattern mining for effective classification. In: ICDE. (2008) 169–178
13. Gilbert, A., Illingworth, J., Bowden, R.: Fast realistic multi-action recognition using mined dense spatio-temporal features. In: ICCV. (2009) 925 –931
14. Yuan, J., Yang, M., Wu, Y.: Mining discriminative co-occurrence patterns for visual recognition. In: CVPR. (2011) 2777 –2784
15. Kim, S., Jin, X., Han, J.: Disiclass: discriminative frequent pattern-based image classification. In: Tenth Int. Workshop on Multimedia Data Mining. (2010)
16. Quack, T., Ferrari, V., Gool, L.V.: Video mining with frequent itemset configurations. In: CIVR. (2006) 360–369
17. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009) 1778 –1785
18. Yun, U., Leggett, J.J.: Wfim: Weighted frequent itemset mining with a weight range and a minimum weight. In: SDM'05. (2005)
19. Yang, Y., Newsam, S.: Spatial pyramid co-occurrence for image classification. In: ICCV. (2011)
20. Yimeng Zhang, T.C.: Efficient kernels for identifying unbounded-order spatial features. In: CVPR. (2009)
21. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR. (2010)
22. Yan, X., Cheng, H., Han, J., Xin, D.: Summarizing itemset patterns: a profile-based approach. In: ACM SIGKDD. (2005)
23. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010) 143–156
24. Svetlana Lazebnik, C.S., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178
25. Opelt, A., Fussenegger, M., Pinz, A., Auer, P.: Weak hypotheses and boosting for generic object detection and recognition. In: ECCV. (2004) 71–84
26. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP. (2008) 722 –729
27. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/ challenges/VOC/ voc2007/workshop/ index.html (2007)
28. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV. (1999) 1150–1157
29. van de Weijer, J., Schmid, C.: Applying color names to image description. In: ICIP. (2007) 493–496
30. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR. (2008)
31. Shahbaz Khan, F., van de Weijer, J., Vanrell, M.: Top-down color attention for object recognition. In: ICCV. (2009) 979 –986
32. Xie, N., Ling, H., Hu, W., Zhang, X.: Use bin-ratio information for category and scene classification. In: CVPR. (2010) 2313 –2319
33. Tuytelaars, T., Fritz, M., Saenko, K., Darrell, T.: The nbnn kernel. In: ICCV. (2011) 1824–1831
34. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC. (2011)