Heterogeneous Visual Codebook Integration via Consensus Clustering for Visual Categorization

Roberto. J. López-Sastre, *Member, IEEE,* Javier Renes-Olalla, Pedro Gil-Jiménez, Saturnino Maldonado-Bascón, *Member, IEEE,* Sergio Lafuente-Arroyo

***Abstract*—Most recent category-level object and activity recognition systems work with visual words, *i.e.* vector-quantized local descriptors. These visual vocabularies are usually built by using a local feature, such as SIFT, and a single clustering algorithm, such as K-means. However, very different clusterings algorithms are at our disposal, each of them discovering different structures in the data. In this paper, we explore how to combine these heterogeneous codebooks and introduce a novel approach for their integration via consensus clustering. Considering each visual vocabulary as one modal, we propose the Visual Word Aggregation (VWA) methodology, to learn a common codebook, where: the stability of the visual vocabulary construction process is increased, the size of the codebook is determined in an unsupervised integration, and more discriminative representations are obtained. With the aim of obtaining contextual visual words, we also incorporate the spatial neighboring relation between the local descriptors into the VWA process: the Contextual-VWA (C-VWA) approach. We integrate over-segmentation algorithms and spatial grids into the aggregation process to obtain a visual vocabulary that narrows the semantic gap between visual words and visual concepts. We show how the proposed codebooks perform in recognizing objects and scenes on very challenging datasets. Compared with unimodal visual codebook construction approaches, our multi-modal approach always achieves superior performances.**

***Index Terms*—consensus clustering, clustering aggregation, visual words, object recognition, scene recognition**

I. Introduction

THE *Bag-of-Words* (BoW) [1], [2] is a popular strategy for representing images within the context of image categorization (*e.g.* [3], [4], [5]) and activity recognition (*e.g.* [6], [7]). The essential idea behind this type of representation is to characterize an image by the histogram of its visual words, *i.e.* vector-quantized local features (see Figure 1). Popular candidates for these local features are local descriptors [8], such as SIFT [9] or SURF [10], that can be extracted at specific interest points (*e.g.* [1]), densely sampled over the image (*e.g.* [11]), or via a hybrid scheme called *dense interest points* [12].

The local descriptors have to be quantized, and there are very different clustering methods that can be used. K-means or variants thereof, such as approximate K-means [13], and mean-shift based approaches (*e.g.* [11]) are currently the most common.

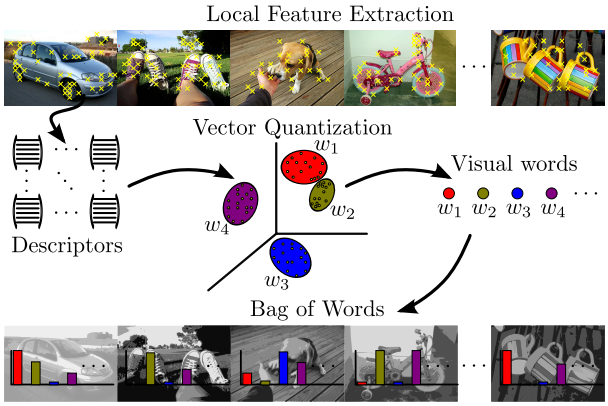


Fig. 1. BoW approach overview. It starts with the extraction of local features followed by a robust description of the features, *e.g.* using SIFT [9]. The following step consists in vector quantizing the high dimensional space of local image descriptors so as to obtain a visual vocabulary. A BoW representation is then built as a histogram of visual word occurrences.

Subsequently, each local feature in an image is mapped to a cluster so as to represent any image as a histogram over the clusters. This BoW representation has been shown to characterize the images and objects within them in a robust yet descriptive manner, in spite of the fact that it ignores the spatial configuration between visual words. Moreover, this approach has inspired a lot of research efforts (obtaining impressive results, *e.g.* [5], [14]), being the basic recipe for most of the methods submitted to the PASCAL VOC Challenge [15].

These visual vocabularies are usually constructed by using a *single* clustering algorithm (normally K-means). Because different clustering algorithms (or even the same clustering but with a random initialization) discover different structures in data, it is true that one particular quantization approach shall obtain a better solution than the others. If we were able to integrate different clustering algorithms, we could build a generally more robust and more discriminative visual codebook. How to combine clustering solutions is becoming a challenging problem nowadays. The consensus clustering solution has been recently proposed [16], [17]. It is defined as the optimization problem where, given a set of m clusterings, the objective is to find the clustering that minimizes the total number of disagreements with the m input clusterings. In other words, consensus clustering can be considered as a

R. J. López-Sastre, J. Renes-Olalla, P. Gil-Jiménez, S. Maldonado-Bascón, and S. Lafuente-Arroyo are with the GRAM research group, Department of Signal Theory and Communications, University of Alcalá. Corresponding author: R. J. López-Sastre, e-mail (robertoj.lopez@uah.es), phone (+34 918856720), fax (+34 918856699). Office S-342, Polytechnic School. University of Alcalá. Campus Universitario, A2 Km 33.600. 28805, Alcalá de Henares (SPAIN)

Manuscript received December 18, 2012; revised xx.

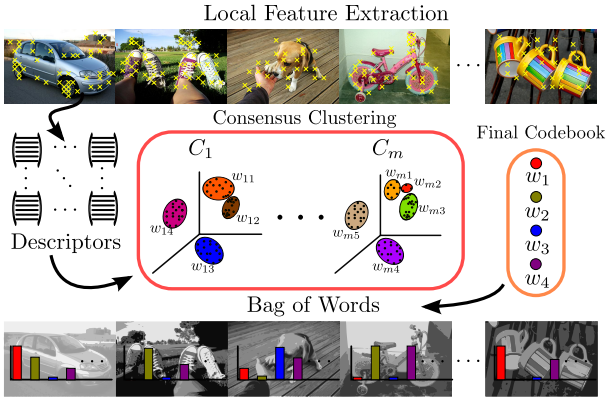


Fig. 2. VWA approach overview. The final codebook is obtained as the output of the consensus clustering approach, where we combine m heterogeneous visual codebooks $\{C_1, \ldots, C_m\}$ obtained from the local features. A BoW is finally built as a histogram of visual word occurrences of the final codebook.

metaclustering method to improve stability and robustness of clustering by combining the results of many clusterers.

In this paper, we propose a novel approach to combine heterogeneous visual codebooks via consensus clustering. We introduce a multi-modal approach (considering each visual vocabulary as one modal), the Visual Word Aggregation (VWA), which: increases the stability of the visual vocabulary construction process, automatically determines the codebook size, and obtains more discriminative solutions.

Although such ideas appear to be quite exciting, there is still a main challenge that need to be overcome. Since the clustering is unsupervised, such a representation does not group semantically meaningful object parts (*e.g.* wheels or eyes). That is, visual words tend to be much more ambiguous than texts. In practice, if the data set is sufficiently coherent (*e.g.* images of only one particular class), only a reduced number of visual words actually represent semantic object parts [18]. Moreover, when an unsupervised quantization is applied to a more diverse data set, synonyms and polysemies are the norm rather than the exception [18]. Typically, the spatial context of the local features is lost during the visual vocabulary construction, *i.e.* the clustering algorithms ignore the semantic relationship between local features that normally co-occur. There are some exceptions that model the spatial and semantic cues of the local patches (*e.g.* [3], [19], [20]), but ours is the first that proposes to use a consensus clustering methodology. So, with the aim of obtaining contextual visual words, we also incorporate the spatial neighboring relation between the local descriptors into the VWA process: the Contextual-VWA (C-VWA) approach. In short, our approach consists of the following steps. We first quantize the local descriptors following traditional clustering algorithms. Additionally, we integrate *codebooks* obtained with over-segmentation algorithms and spatial grids into the aggregation process to group neighboring local patches. By considering each local group as a new cluster, we are able to incorporate these quantizations to the input of the consensus clustering approach as new clustering solutions that encode contextual information.

Via our new heterogeneous visual codebook integration approaches, we can always achieve a superior image classification performance than the performance of traditional BoW approaches that just use a single clustering algorithm.

The rest of this paper is organized as follows. Section II contains a review of the related work. In Section III, we briefly review the consensus clustering theory. Section IV gives a detailed description of the novel approaches we propose to adapt the clustering aggregation techniques to the visual vocabulary construction process. Experiments in image categorization and scene recognition are described in Section V, and Section VI concludes the paper.

II. Related work

Traditionally, following a BoW approach consists of extracting local descriptors and applying a K-means clustering for the vocabulary construction. That is, single-feature and single-clustering based approaches. These vector-quantized local features have been referred to as 'object parts' [21], 'visual words' [2] or 'codebooks' [22].

This visual codebook construction step is critical within a BoW scheme. It has significant impact on recognition accuracy, training and test efficiency, and system complexity. In the literature, many codebook generation methods have been proposed, mainly including clustering-based methods [2], [1], mean-shift [11], latent space models [18], [23], information theoretic approaches [24], randomized trees [25], and the recently developed sparse coding methodology [26].

Within the context of vector quantization based approaches, we identify two main problems. On the one hand, the visual codebooks are obtained in an unsupervised way, which ignores the semantic and spatial context between local features during the grouping. On the other hand, there exist the limitations of the clustering algorithms themselves. In general, data clustering usually has associated the stability problem: it is not possible to use cross validation for tuning the clustering parameters because of the absence of ground truth; the dependence on the initialization is a common problem for most of the iterative methods; the objectives pursued by each clustering algorithm are different and different structures in data may be discovered.

Different attempts have been made to overcome the first problem. First, there are several works based on frequent itemset mining [27], [28], [19]. Typically, finding representative visual words boils down to finding frequent co-occurring groups of descriptors in a transaction database obtained from the training images. In [29] Leibe *et al.* presented how to learn semantic object parts for object categorization. They use what they call co-location and co-activation to learn a visual vocabulary that generalizes beyond the appearance of single objects, and often obtains semantic object parts. In [30], Wang *et al.* propose a novel regularized K-means clustering for discovering the spatial co-occurrence of local features, and the feature co-occurrence patterns of different features types. Such co-occurrence patterns can be used to handle the ambiguities of visual primitives.

There are also supervised approaches that use image annotation to guide the semantic visual vocabulary construction (*e.g.* [25], [31], [32]). Specifically, Moosmann *et al.* [25]

utilize extremely randomized clustering forest to organize the vocabulary. The discriminative power of the vocabulary is increased by incorporating image class labels to guide the tree construction. In image classification tasks, this representation provides more accurate results than the conventional K-means based BoW. In [31] mutual information is used between the features and class labels to create the semantic vocabulary from an initial and relatively larger vocabulary quantized by the K-means algorithm. Yang *et al.* [32] unify the vocabulary construction with classifier training. Other interesting works use: diffusion maps to learn a semantic visual vocabulary from abundant quantized mid-level features using K-means [33]; semantic-aware distance metric learning to efficiently cluster the local patches [20]; a method for capturing the spatial contextual information between visual words by counting the occurrence of meaningful visual word pairs [3]; or the well known spatial pyramid matching kernel [34] which considers the spatial layout relation of local features by partitioning an image into increasingly fine grids and computing the BoW inside each grid cell.

The second problem, *i.e.* the limitations of the clustering algorithms themselves, has not been fully studied in the context of visual word generation. K-means has become *de facto* standard. However, it has well known limitations: its output depends on the initialization as the procedure only undertakes the search of a local optimum, the number of clusters must be specified by hand, and it is computationally expensive for big values of K. Other approaches propose to use efficient hierarchical clustering schemes [35] or mean-shift based algorithms [11].

Recently, the use of sparse coding instead of traditional vector quantization algorithms has been proposed too. These approaches are naturally derived by relaxing the restrictive cardinality of vector quantization, *i.e.* vectors with membership in multiple *clusters*. Then, any image characterized by a set of descriptors, can be represented by computing a single feature vector based on some statistics of the codes of the descriptors. In [26] a spatial pyramid image representation [34] based on sparse codes of local features is proposed. Furthermore, in [36], Wang *et al.* present a novel coding scheme called Locality-constrained Linear Coding (LLC) in place of the traditional vector quantization coding, which can work with linear classifiers, performing better than the traditional nonlinear approaches [34]. Compared with the sparse coding strategies [26], the objective function used by LLC has an analytical solution. In contrast to clustering based approaches (like ours), in the sparse coding schemes it is computationally challenging to learn a set of highly over-complete dictionary bases and to encode the test data with the learned bases, although some efficient techniques have been recently proposed (*e.g.* [37]). In Section V-B3, we offer a direct comparison of our models with some of the described sparse coding approaches, as well as with state-of-the-art methods.

Our approach significantly differs from the previous revised works. In this paper, we start proposing the VWA approach in order to overcome the second problem. We explore how the consensus clustering algorithms [17] can be efficiently used for building visual vocabularies from large-scale data sets of high dimensional local descriptors. A preliminary version of this approach was described in [38]. Although multiple vocabularies are also used in [39], the work of Aly *et al.* [39] differs considerably from our approach. First, the dictionaries are generated independently, and no consensus clustering techniques are applied for building the final codebook. Second, the final histogram for characterizing an image is the concatenation of the individual histograms obtained from each codebook. As an alternative approach, they propose to build a unified dictionary from the concatenation of the visual words from all the independent dictionaries.

Moreover, in order to overcome the first problem, we also present the novel C-VWA approach, where we incorporate into the vocabulary construction process the spatial neighboring relation between the local descriptors. This information is captured using over-segmentation algorithms and spatial grids, and is integrated in a heterogeneous consensus clustering pipeline to obtain a more discriminative visual vocabulary that narrows the semantic gap between visual words and visual concepts.

III. Background: Consensus Clustering

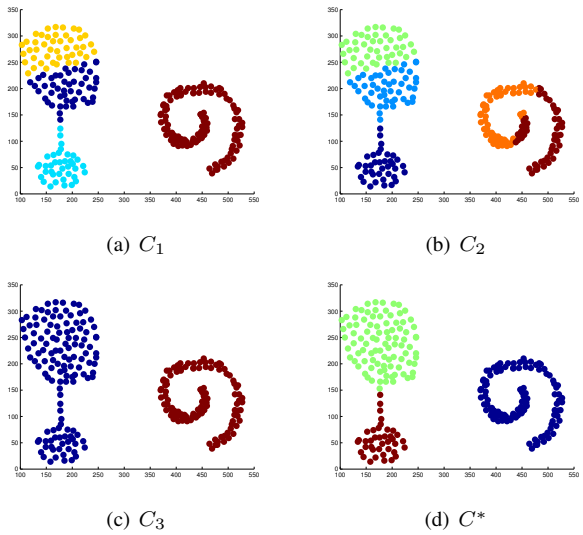
The problem of consensus clustering has been considered under a variety of names: clustering aggregation, clustering combination and cluster ensembles. Many approaches have been proposed, *e.g.* the information theoretic method [17], the graph cut method [40] and the Bayesian method [41].

Consensus clustering is defined as the optimization problem where, given a set of m clusterings, $\{C_1, C_2, \ldots, C_m\}$, the objective is to find the clustering C^* that minimizes the total number of disagreements with the m clusterings. So, consensus clustering can be considered as a metaclustering method to improve stability and robustness of clustering by combining the results of many clusterings. Moreover, it can determine the appropriate number of clusters while detecting outliers. A toy example to illustrate how the clustering aggregation works is depicted in Figure 3.

Consider a set of n objects $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. A clustering C_i of V is a partition of V into k_i disjoint sets $\{S_1, S_2, \ldots, S_{k_i}\}$, *i.e.* $\bigcup_i^{k_i} S_i = V$ and $S_i \cap S_j = \emptyset$ for all $i \neq j$. The clusters of C_i are the k_i sets $\{S_1, S_2, \ldots, S_{k_i}\}$. For each $\mathbf{v}_j \in V$, $j = 1, \ldots, n$, we use $C_i(\mathbf{v}_j)$ to denote the label of the cluster to which the object $\mathbf{v}_j$ belongs to, *i.e.* $C(\mathbf{v}_j) = l$ if and only if $\mathbf{v}_j \in S_l$.

For this paper, we follow the approach of Gionis *et al.* [16], which is based on correlation clustering techniques [42]. We are given a set of m clusterings $\{C_1, C_2, \ldots, C_m\}$. Our objective is to obtain a single clustering C^* that agrees as much as possible with the m input clusterings. It is possible to define a distance $d(\mathbf{v}_i, \mathbf{v}_j)$ between two vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ as the fraction of the m clusterings that place $\mathbf{v}_i$ and $\mathbf{v}_j$ in different clusters. Our objective is to find a clustering C^* that minimizes the function

$$d(C^*) = \sum_{C^*(\mathbf{v}_i)=C^*(\mathbf{v}_j)} d(\mathbf{v}_i, \mathbf{v}_j) + \sum_{C^*(\mathbf{v}_i)\neq C^*(\mathbf{v}_j)} (1 - d(\mathbf{v}_i, \mathbf{v}_j)) \,. \quad (1)$$



(a) C_1 (b) C_2 (c) C_3 (d) C^*

Fig. 3. Toy example. (a)-(c) are 3 different clusterings $\{C_1, C_2, C_3\}$ over a data set of 2D points. (d) depicts the result of the clustering aggregation algorithm, the clustering C^*. Note that the solution C^* improves the clustering robustness and finds the 3 clusters in the data set. We have used different colors to identify different clusters.

For a candidate solution C^*, if C^* places $\mathbf{v}_i$ and $\mathbf{v}_j$ in the same cluster, it will disagree with $m \times d(\mathbf{v}_i, \mathbf{v}_j)$ of the original clusterings, whilst if C^* places $\mathbf{v}_i$ and $\mathbf{v}_j$ in different clusters, it will disagree with the remaining $m \times (1 - d(\mathbf{v}_i, \mathbf{v}_j))$ clusterings.

Several approaches are proposed in [16] to solve this optimization problem. For the particular problem of building visual codebooks, *i.e.* high dimensional vector quantization in large scale sets, we have adapted the the *Balls* and the *Agglomerative* (*Agg*) approaches [16]. Both algorithms take as input a complete graph with all the distances between vectors. The *Balls* algorithm tries to find groups of nodes that are within a ball of fixed radius and far from other nodes. Once such a set is found, the algorithm considers it a new cluster and proceeds with the rest. The *Agg* is a bottom-up algorithm which starts with every node in a cluster. It merges two vertices if the average distance between them is less than a fixed value. In Section IV we describe the novel strategies designed so as to efficiently build visual codebooks from large-scale data set of high dimensional local descriptors.

IV. Heterogeneous Visual Codebook Integration via Consensus Clustering

A. The Visual Word Aggregation

Our aim is to design a novel approach for combining heterogeneous visual codebooks via consensus clustering: the Visual Word Aggregation (VWA). This new strategy has a threefold objective: to increase the stability of the visual vocabulary construction process, to automatically determine the codebook size, and to obtain more discriminative solutions.

Before we go into the details of our approach, some frequently used notations will be introduced. In a BoW framework, a number of local descriptors $\mathbf{v}_{ij}$, $j = 1 \dots m_i$ are extracted in each image I_i, $i = 1 \dots n$. Being V the set of all the local descriptors extracted, a visual vocabulary W is obtained via vector quantizing the local descriptors in V,

$$W = \mathrm{VQ}(V, \mathbf{p}) \ , \tag{2}$$

where $\mathbf{p}$ is a vector containing the specific parameters for the vector quantization algorithm VQ, and $W = \{\mathbf{w}_1, \dots, \mathbf{w}_K\}$ is the visual vocabulary obtained of size K. That is, each $\mathbf{w}_k$ is the centroid of one of the K clusters discovered in the data. Once W has been obtained, a BoW approach describes each image I_i by a frequency distribution over the visual words in W. For each word $\mathbf{w}_k$ in the vocabulary W, the BoW estimates the distribution of visual words in the image I_i by

$$h(\mathbf{w}_k | I_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \begin{cases} 1 & \text{if } \mathbf{w}_k = \underset{\mathbf{w} \in W}{\operatorname{argmin}}(D(\mathbf{w}, \mathbf{v}_{ij})) \\ 0 & \text{otherwise} \end{cases} , \tag{3}$$

where $D(\mathbf{w}, \mathbf{v}_{ij})$ is the distance between a codeword $\mathbf{w}$ and local feature $\mathbf{v}_{ij}$. Finally, the image I_i is represented by the histogram of word frequencies,

$$H(I_i, W) = [h(\mathbf{w}_1 | I_i), \dots, h(\mathbf{w}_K | I_i)] \ . \tag{4}$$

In the VWA approach, we propose to build a visual vocabulary W^* via consensus clustering, combining m heterogeneous visual codebooks $\{W_1, \dots, W_m\}$. We abbreviate the optimization problem of consensus clustering described in Section III, as follows. Given a set of of m codebooks $\{W_1, \dots, W_m\}$, the consensus clustering approach CC finds the visual codebook W^* that minimizes the total number of disagreements with the m clusterings,

$$W^* = \mathrm{CC}(W_1, \dots, W_m) \ . \tag{5}$$

Note that the output of the consensus clustering algorithm, *i.e.* W^*, specifies the centroids that define the clustering of the data. Figure 2 shows the major steps of our proposal. In a first step, images are represented by using local features (*e.g.* SIFT [9]). Then, the vector quantization processes start. We define m as the number of clustering algorithms that are executed. The VWA can reconcile clustering information about the same data set coming from different clustering algorithms and/or from different runs of the same algorithm. We will explore all these combinations in the experiments section. Once the clustering aggregation has finished, each image I_i can be represented using a BoW approach with the novel codebook W^*, *i.e.* computing the histogram of the new visual words $H(I_i, W^*)$.

However, a direct application of consensus clustering algorithms is not feasible. When working with visual codebooks, typically, we have to deal with large sets of high dimensional vectors. For instance, if SIFT descriptors are used, each visual vocabulary has to organize the local descriptors in a 128 dimensional space. Furthermore, hundreds of descriptors are extracted from each image, and normally the size of the codebooks used is high too. In this scenario, a standard clustering aggregation algorithm becomes inapplicable: the quadratic complexity is inherit in the correlation clustering problem since a complete graph (with the distance matrix) is

the input to the problem. Some sampling strategies have been introduced to overcome this problem [16]. In a preprocessing step, the algorithm samples a set of nodes uniformly at random from the data set. The sampled set is the input for the clustering aggregation algorithm. In the post-processing step, the algorithm goes through the nodes not in the set and decides whether to place it on one of the existing clusters or to create a singleton. Nonetheless, we observed experimentally that the time complexity of this approach is high within our context, *i.e.* when both the number of clusters and the dimensionality of vectors are high.

In order to reduce the run-time of the visual vocabulary construction, we design a double sampling strategy. Let V be the set of local descriptors extracted from all the images, defined as $V = \{\mathbf{v}_{11}, \ldots, \mathbf{v}_{1m_1}, \ldots, \mathbf{v}_{n1}, \ldots, \mathbf{v}_{nm_n}\}$. Let p be the size of the set V. We start with a uniform and random sampling $R \subset V$ of size $r = \beta p$, where $\beta \in [0, 1]$ is the sampling factor. The set R is sampled again, with a factor β to obtain the subset $S \subset R$. Only the set S is given as input to the clustering aggregation algorithm which builds a clustering $W^* = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$. Note that with this double sampling scheme, the post-processing step of [16] only needs to evaluate the elements in R and not in S, which significantly reduces the run-time of the original approach. Finally, we inspect the vectors in V and not in R and assign them to the nearest centroid. Using this double sampling strategy we can handle large data sets letting VWA converge into a final codebook.

Such a sampling strategy does not jeopardize the object recognition performance of the vocabulary obtained. The visual codebooks are obtained from local descriptors in high dimensional spaces. If these high dimensional descriptors have been densely extracted, then there do not exist separate clusters in the data [43]. Furthermore, the norms used in any vector quantization algorithm, *e.g.* the euclidean norm, tend to concentrate with high dimensional vectors [44]. As a consequence, all pairwise distances in a high-dimensional data set seem to be equal or at least very similar. So, a uniform sampling of the data, will be able to describe the distribution in the feature space. Furthermore, we have experimentally observed that a reduction in the number of vectors (like the proposed with this double sampling strategy) used for building the vocabulary, does not significantly affect the quality of the final codebook obtained. Obviously, we have observed that the sampling factor influences the results obtained. We found that an adequate sampling factor should be greater than 0.5. A thorough evaluation of the influence of this sampling factor on the performance with quantitative results is detailed in Section V-A.

Our experiments reveal that the VWA increases the stability of the visual vocabulary construction process and the performance in image categorization. Furthermore, it can combine the properties from different clustering algorithms, which is something desirable in such high dimensional spaces where the local descriptors reside.

B. Incorporating contextual information

Contextual information plays a fundamental role to recognize visual categories from their appearance (*e.g.* [34], [3]). Via consensus clustering, we propose the Contextual-VWA (C-VWA) to incorporate the spatial coherency among the local descriptors into the visual vocabulary construction. In the VWA, described in Section IV-A, we used as inputs for the consensus clustering process a set of codebooks that organize the local descriptors in a high dimensional space, *i.e.* in the descriptor space. Because the norms used by the vector quantization algorithms tend to concentrate in high dimensional spaces [44], any clustering algorithm which just uses this similarity measure will be limited. Our purpose is then to incorporate the contextual information in the image to guide the quantization algorithm to find a more semantic and discriminative solution. For doing so, we proceed to group neighboring local features in the image space. By considering each local group as a new *cluster*, we are able to incorporate these quantizations into the input of the consensus clustering approach as clusterings that encode the contextual information.

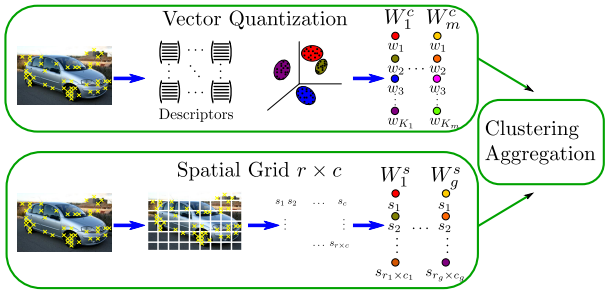


Fig. 4. Using regular grids to add spatial information to the vocabulary construction process. Note that a grid of size $r \times c$ defines a *codebook* of size $r \times c$.

Two approaches have been designed: using spatial grids and over-segmentation algorithms.

1) Via Spatial Grids: This first approach starts with a dense sampling of image patches. We proceed with the local descriptors as it was described in Section IV-A, *i.e.* obtaining a set of visual codebooks $\{W_1^c, \ldots, W_m^c\}$ using traditional clustering algorithms. See Figure 4 upper box. The next step consists in quantizing the same local features but in the image domain. We project a grid over all the images in the database (all the images in the database have been previously scaled to the same size). Each cell of the grid can be considered a spatial cluster. So, a grid of $r \times c$ cells defines a clustering of size $r \times c$, *i.e.* the size of the codebook is $r \times c$, where the local features are quantized. Furthermore, it is possible to use grids of either random or fixed cell sizes. Within this context, we assign to each feature the label of the cell where it falls. If g different grids are used, we obtain a set of g different spatial clusterings $\{W_1^s, \ldots, W_g^s\}$. We use the consensus clustering process to obtain the final codebook $W^* = \mathrm{CC}(W_1^c, \ldots, W_m^c, W_1^s, \ldots, W_g^s)$. Figure 4 depicts the whole process.

2) Via Over-segmentation: Unfortunately, the spatial grids do not often capture image regions of homogeneous appearances, so the local features are not in clusters with semantic coherency. In order to address this issue, we perform an over-segmentation of an image by partitioning it into mul-

(a) original image (b) over-segmented image

Fig. 5. Image over-segmented using [45]. Note that over-segmented regions can group local features which belong to semantic object parts.

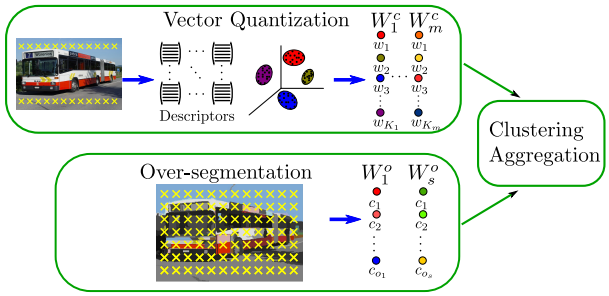


Fig. 6. Using over-segmentation for adding spatial information to the vocabulary construction process.

tiple homogeneous regions. Our approach searches for over-segmented regions that can group local features which belong to semantic object parts. For the experiments, we use the segmentation algorithm described in [45]. Figure 5 shows one example of the over-segmentation process. Note that our method is not tied to a specific segmentation algorithm. We let the segmentation discover semantically meaningful segments. The local features are quantized by these segments, *i.e.* two features in the same segment receive the same label, and the corresponding clusterings are incorporated as inputs in the clustering aggregation algorithm.

As in the approach of Section IV-B1, the first step is to compute the set of visual codebooks $\{W_1^c, \ldots, W_m^c\}$ using traditional clustering algorithms. The second step implies to over-segment the images to establish regions of neighboring appearances. Each segmented region can be considered a spatial cluster, and a different label is assigned to each of them. We assign to each local descriptor the label of the segment where it falls. If s different over-segmentations are processed (using different parameters), we obtain a set of s different spatial clusterings $\{W_1^o, \ldots, W_s^o\}$. Then, we use the consensus clustering process to obtain the final codebook $W^* = \mathrm{CC}(W_1^c, \ldots, W_m^c, W_1^o, \ldots, W_s^o)$. Figure 6 depicts the whole process.

A mixture approach is also viable. With the formulation introduced, we can combine into the same consensus clustering approach both the spatial grid codebooks and the visual vocabularies obtained with the over-segmentation. The final codebook obtained with such a mixture approach is then obtained as $W^* = \mathrm{CC}(W_1^c, \ldots, W_m^c, W_1^s, \ldots, W_g^s, W_1^o, \ldots, W_s^o)$.

V. Results

A. Experimental Setup

Our aim is to evaluate the performance of the proposed approaches within two contexts: visual categorization and scene recognition. For the former, we use the PASCAL VOC Challenge 2007 database [46]. We emphasize that this challenge is widely acknowledged as a difficult testbed for both object detection and image categorization. The dataset contains $9,963$ annotated images, with the number of annotated objects being $24,640$. In the experiments we select the *trainval* and *test* sets for training and testing the classifier respectively. For further details we refer to [46], [15].

We also evaluate our approaches on the scene recognition problem. We use the New York University Depth video data set (NYU-Depth) [47]. It is a new and challenging indoor video scene dataset, which is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. For the scene classification benchmark, the dataset offers $20,000$ images ($10,000$ for training and testing) distributed across 5 different scene-level classes: bathroom, bedroom, kitchen, living room and office. In our experiments, we simply use the RGB images provided in the NYU-Depth.

For image representation, in both datasets, we use SIFT [9] descriptors of 16×16 pixel patches computed over a grid with spacing of 8 pixels. When using the PASCAL VOC 2007, we force the images to be not bigger than 375×500.

We perform the clustering algorithms of a random subset of features from the training set to form the visual codebooks. With these descriptors we obtain all the different codebooks that will be subsequently combined via consensus clustering. As described in Section IV, there are very different clustering methods than can be used. K-means is a popular algorithm for its simplicity. Unfortunately, centroids tend towards denser regions, with the result that they tend to be tightly clustered near dense regions and sparsely spread in sparse ones. Mean shift based approaches (*e.g.* the J&T, proposed by Jurie and Triggs [11]) can be used to overcome some of the limitations of K-means. So, for the experiments, we use both K-means and the J&T.

As it was detailed in Section IV, we have integrated our novel double sampling methodology with the *Balls* (with $\alpha = 0.25$) and the *Agg* consensus clustering algorithms detailed in [16]. Typical values of the sampling factor β for our experiments are $\beta = 0.25$, $\beta = 0.33$ or $\beta = 0.5$.

Support Vector Machines (SVMs) are used for classification. We experiment with the Histogram Intersection Kernel (HIK) which has shown good results in visual categorization. The HIK applied to two-feature vectors $\mathbf{x}$ and $\mathbf{x}'$ of dimension d is defined as $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} \min(\mathbf{x}(i), \mathbf{x}'(i))$. Specifically, we use libSVM [48]. A 10-fold cross-validation on the training sets to tune SVM parameters is conducted.

For image classification, we follow the evaluation procedure proposed by the PASCAL VOC Challenge [46] using the Mean Average Precision (MAP), which is computed by taking the mean of the average precisions for the 20 classes for each method. For the scene recognition problem, we follow the

TABLE I
CODEBOOKS FOR VISUAL CATEGORIZATION

	Codebook description
C1	K-means ($K = 200$)
C2	3 K-means ($K = 200$) and *Balls* ($\alpha = 0.25$) + Sampling ($\beta = 0.5$)
C3	3 K-means ($K = 200$) and *Agg* + Sampling ($\beta = 0.33$)
C4	J&T ($r = 0.83, N = 3000$)
C5	3 J&T ($r = 0.83, N = 3000$) and *Balls* ($\alpha = 0.25$) + Sampling ($\beta = 0.25$)
C6	J&T ($r = 0.92, N = 3000$)
C7	2 K-means ($K = 200$) + J&T ($r = 0.92, N = 3000$) and *Balls* ($\alpha = 0.25$) + Sampling ($\beta = 0.5$)
C8	J&T ($r = 0.8, N = 3000$) + K-means ($K = 2000$) and *Balls* ($\alpha = 0.25$) + Sampling ($\beta = 0.5$)

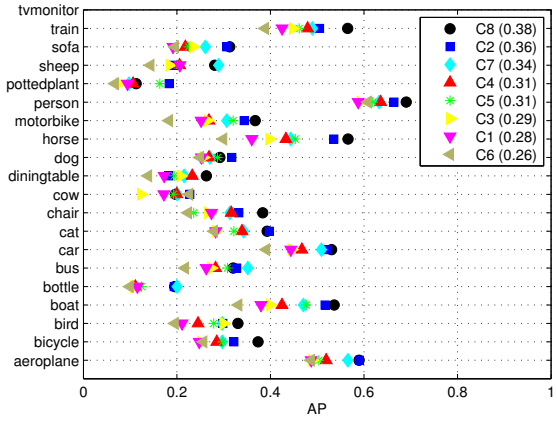


Fig. 7. Evaluation of codebooks on image categorization with the PASCAL VOC 2007 Challenge. Average precision per class for each method is shown. The *legend* indicates the MAP obtained by the corresponding method.

experimental setup described in [47]: the Mean Diagonal of Confusion Matrix (MDCM).

B. Results on image categorization

1) Aggregated Codebooks Performance: Before adding contextual information to our framework, we only combine visual codebooks, coming from different clustering algorithms, following the VWA approach.

We evaluate the MAP in image categorization for the codebook combinations described in Table I. Note that codebooks C1, C4 and C6 have been obtained without using the VWA approach, *i.e.* using a single clustering and a traditional BoW approach. Results per object category are shown in Figure 7. The aggregation of one K-means and one J&T, *i.e.* codebook C8, obtains the best MAP (0.38). Furthermore, all the codebooks generated via VWA using the *Balls* algorithm and our sampling approach (*i.e.* vocabularies C2, C5, C7 and C8), obtain better results than when a traditional BoW is used (C1, C4 and C6). Comparing C2 and C3 we also have observed that the *Balls* algorithm performs better than the *Agg*. Moreover, for the *Balls* algorithm, we have found that $\alpha \leq 0.25$ leads to better results in image categorization.

We observed experimentally that the sampling factor β directly affects to the classification performance: the best results are obtained for $\beta \geq 0.5$. In our VWA approach, for a combination of three K-means codebooks, and using the *Balls* algorithm, we decreased the sampling factor β from 0.5 to 0.25. The different MAP obtained are detailed in Table II. It is interesting to observe how the performance slightly decreases as soon as β becomes more restrictive, *i.e.* $\beta < 0.5$. Our results show that $\beta = 0.5$ represents a good compromise between recognition performance, runtime and memory footprint.

TABLE II
QUANTITATIVE ANALYSIS OF THE SAMPLING FACTOR β. MAP RESULTS.

Codebook	$\beta = 0.25$	$\beta = 0.33$	$\beta = 0.5$
3 K-means and *Balls* ($\alpha = 0.4$)	0.29	0.29	0.31

Results confirm that the VWA technique can be used to obtain better vocabularies. Our approach also reveals that better results are obtained when combining the properties of different clustering types via consensus clustering. The best result obtained, *i.e.* C8 (MAP of 0.38), is a combination of K-means and J&T clustering. As expected, the characteristics of each kind of clustering algorithm complement each other during the clustering aggregation step. Some values of the algorithm parameters have been tested, and for the C8 combination, we find that the best performance is obtained for $K = 2000$ for K-means, and $r = 0.83$, $N = 3000$ for J&T. If we decide just to use a single clustering algorithm (*e.g.* K-means), the proposed pipeline automatically finds a more discriminative vocabulary by simply combining different executions of the same one (see how the MAP increase from C1 to C2 when we just insert K-means codebooks in the VWA pipeline). In conclusion: the clustering algorithm matters when dealing with BoW systems for visual categorization.

2) Contextual Information Integration: First, we evaluate the approach based on regular grids for incorporating contextual information. We superimpose a spatial grid over all the images and consider each cell as a cluster, *i.e.* a particular grid of $r \times c$ cells defines a clustering of size $r \times c$. For the experiments, grids of different dimensions, from 32×32 to 256×256, have been used.

By simply combining a single J&T clustering, whose MAP is 0.26, with contextual codebooks obtained via different regular grids, the performance in object categorization always increases (see Table III). Similar results are obtained when following the C-VWA with spatial grids and the rest of analyzed clusterings. Table IV shows the most significant results. The MAP is generally optimized when using a grid of 256×256, *i.e.* the denser the grid, the higher the precision. From now on, we select this grid size when combining it with C1, C2, C6, C7 and C8.

Next, we evaluate the performance of our C-VWA approach when we introduce the contextual information via the over-segmentation approach detailed in Section IV-B2. For this experiment, the input images are initially segmented into semantically meaningful regions employing [45]. We have tried with different segmentation parameter combinations (σ and k, see [45] for further details). Table V shows the results obtained when the C-VWA combines C1, *i.e.* a single K-

TABLE III
C-VWA USING SPATIAL GRIDS IN COMBINATION WITH C6. MAP RESULTS.

$r \times c$	c=32	c=64	c=128	c=256
r=32	**0.33**	0.34	0.34	0.34
r=64	0.34	**0.35**	0.35	0.35
r=128	0.35	0.35	**0.36**	0.34
r=256	0.35	0.35	0.35	**0.36**

TABLE IV
C-VWA USING SPATIAL GRIDS IN COMBINATION VARIOUS CLUSTERINGS. MAP RESULTS

	without grid	grid=32 × 32	grid=64 × 64	grid=128 × 128	grid=256 × 256
C1	0.28	0.33	0.35	0.35	**0.36**
C6	0.26	0.33	0.35	0.36	**0.36**
C2	0.36	0.36	0.37	0.37	**0.37**
C7	0.34	0.37	0.37	0.37	**0.37**
C8	0.38	0.37	0.38	0.38	**0.38**

means, with a clustering obtained via the over-segmentation process. We can observe that the C-VWA methodology always improves the MAP of the C1 alone (0.28 – see Figure 7). The segmentation parameters that obtain the best results are $k = 300$ and $\sigma = 0.5$.

So far we have experimented with non-mixture contextual approaches, *i.e.* combining a clustering with either a spatial grid or an over-segmentation process. However, in this experiment, we explore the performance of those mixtures approaches where the contextual information is jointly incorporated by a spatial grid and an over-segmentation. We run all these new experiments using the *Balls* consensus clustering algorithm with $\alpha = 0.25$ and a sampling factor $\beta = 0.25$. We use spatial grids of 256×256, and the over-segmentation parameters are $k = 300$ and $\sigma = 0.5$. We evaluate the MAP in image categorization with the combinations that obtained the highest MAPs: C1, C2, C6, C7 and C8.

Figure 8 compares the result of the original codebook, defined as the *reference*, with the results with the C-VWA approach. Any C-VWA combination always outperforms the results obtained by the reference codebook. This improvement is greater when we start from a single clustering (*e.g.* C1 or C6). The best results are obtained when the C-VWA is used with C8.

TABLE V
C-VWA USING OVER-SEGMENTATION IN COMBINATION WITH CODEBOOK C1. MAP RESULTS.

	$k = 300$	$k = 500$	$k = 1000$	$k = 1500$	$k = 2000$	$k = 2500$
$\sigma = 0$	0.37	**0.38**	0.36	0.37	0.36	0.37
$\sigma = 0.3$	0.36	0.36	0.37	0.37	0.36	0.37
$\sigma = 0.5$	**0.38**	0.36	0.37	0.36	0.36	0.37
$\sigma = 0.8$	**0.38**	0.36	0.36	0.36	0.36	0.36
$\sigma = 1$	0.37	0.36	0.36	0.36	0.36	0.36

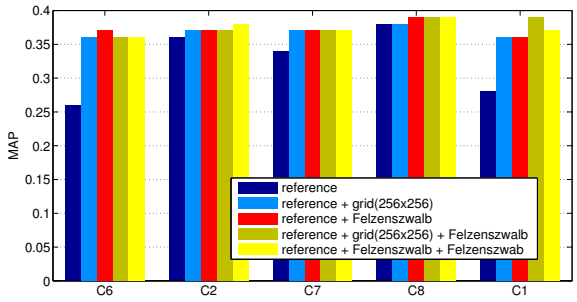


Fig. 8. Evolution of the results by adding contextual information to the reference vocabularies C6, C2, C7, C8 and C1.

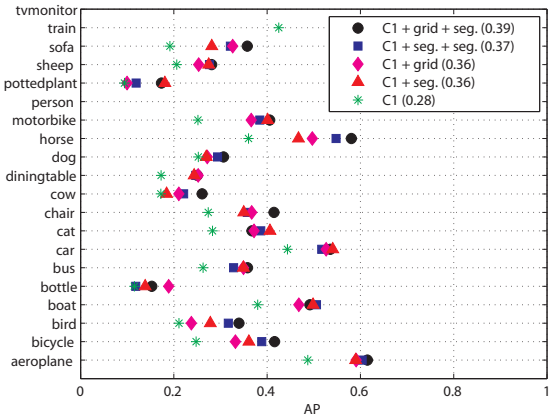
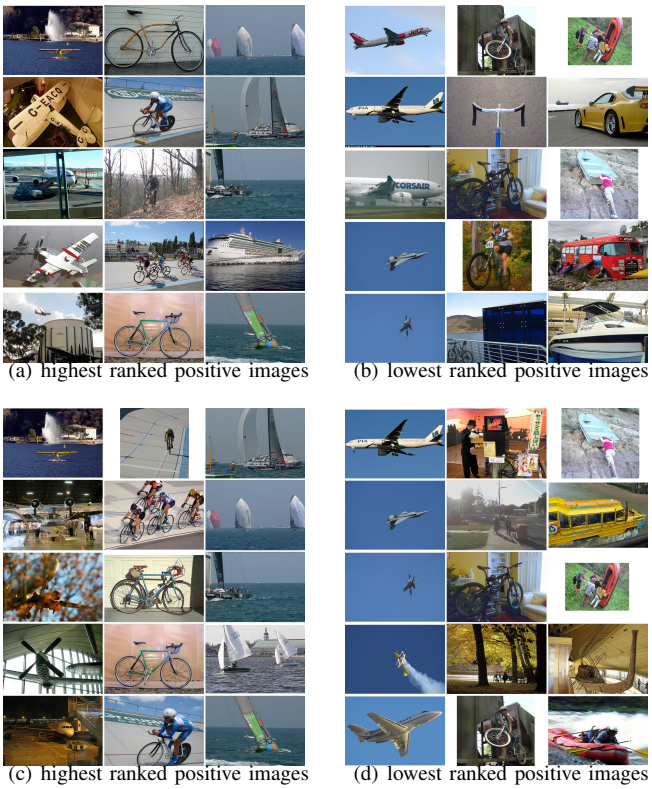


Fig. 9. Evolution of the results by adding contextual information to the reference vocabulary C1. Average precisions per class are shown. The *legend* indicates the MAP obtained by the corresponding method.

It is also interesting to observe how the C-VWA is able to dramatically increase the MAP obtained by a simple single clustering such as C1. We show in Figure 9 how the AP per class evolves for the different combinations starting with the simple C1 as the reference. Furthermore, the MAP obtained by the C-VWA of C1+grid+seg (0.39) is only equal to the MAP obtained by the C-VWA of the more complex combination of clusterings C8 (see Figure 8).

Some qualitative results in visual categorization are shown in Figure 10. Figures 10(a) and 10(b) show some ranked images for 3 different classes when just C1 is used. These can be compared with the same ranking when we make use of contextual information, *i.e.* with the combination C1+grid+seg , illustrated by Figures 10(c) and 10(d).

In light of the results obtained, we can confirm that the addition of contextual information to the process of vocabulary construction, for visual categorization, is always a good practice that contributes to the stability of the final codebook, while increases the performance, and mitigates the dependence on the clustering parameters. Moreover, it is important to mention that with the proposed approaches we are able to avoid expensive cross-validation through the optimization of vocabulary parameters. In a traditional BoW approach, whether a particular clustering is better than another



(a) highest ranked positive images (b) lowest ranked positive images

(c) highest ranked positive images (d) lowest ranked positive images

Fig. 10. Ranked images for the classes aeroplane, bicycle and boat of the PASCAL VOC 2007. (a,b) show the results for the codebook C1. (c,d) show the results for the C1+grid+seg combination.

or not must be evaluated within the context of the entire system, *i.e.* by evaluating its effect on the accuracy of the resulting classifier on a validation set. The main problem with this approach is clear: to validate just the clustering parameters it is necessary to go through the system's entire pipeline. The combinatorics involved result in an explosion of the number of iterations needed. Instead, with the VWA and C-VWA approaches we are able to start with a simple initialization of clustering parameters, *e.g.* C1, and let the process to automatically: determine the final clustering, and increase the performance in visual categorization.

3) Comparison with state-of-the-art results: Our best MAP (39%) in the PASCAL VOC 2007 has been obtained by the C-VWA of C1+grid+seg. For the sparse coding approach in [37], the MAP is of 59.6%. The LLC coding method in [26] and the winner of the PASCAL VOC 2007 challenge [15], report a MAP of 59.3% and 59.4%, respectively. We realize our results are inferior to the winning schemes in the PASCAL VOC 2007 challenge for image categorization. However, to make fair comparison, some noteworthy comments need to be made. The basic recipe for the winning pipelines, using BoW descriptors, consists in incorporating multiple features and nonlinear classifiers [15], and building more complex image representations that include spatial configuration information [34]. In [37], although a sparse coding based representation is proposed, a spatial pyramid structure is also employed to encode the spatial distribution of the features. Instead, our approach builds on a simple and small K-means codebook, using a single feature type (SIFT descriptors), without any spatial pyramid structure, and more important, with a learning approach that is based on a unique kernel. This way, we are able to actually evaluate the performance of the visual vocabularies themselves, *i.e.* to prove that the vocabulary construction step actually matters. Our aim in these experiments is not to compete with these more complex systems described.

Overall, we offer a consistent approach for visual vocabulary construction which systematically outperforms the traditional pipeline where only one vector quantization approach is used. Furthermore, any other BoW based approach can be benefited from incorporating our codebooks.

TABLE VI
CODEBOOKS FOR SCENE RECOGNITION

	Codebook description
C1	K-means ($K = 200$)
C2	3 K-means ($K = 200$)
C1+grid	K-means ($K = 200$) + grid(256×256)
C1+seg.	K-means ($K = 200$) + seg($k = 300$, $\sigma = 0.5$)
C1+grid+seg.	K-means ($K = 200$) + grid(256×256)+ seg($k = 300$, $\sigma = 0.5$)
C2+grid	K-means ($K = 200$) + grid(256×256)
C2+seg.	K-means ($K = 200$) + seg($k = 300$, $\sigma = 0.5$)
C2+grid+seg.	K-means ($K = 200$) + grid(256×256)+ seg($k = 300$, $\sigma = 0.5$)

C. Results on scene recognition

For the scene recognition experiments in the NYU-D dataset, we follow the experimental setup described in Section V-A. After the thorough study of the performance of the different codebook combinations in Section V-B, we propose to evaluate the scene recognition problem using the codebook combinations detailed in Table VI. For all the consensus clustering combinations we use again the *Balls* algorithm (with $\alpha = 0.25$), and a sampling factor $\beta = 0.5$. Note that with this set of codebooks we are going to be able to evaluate the performance of both the VWA and the C-VWA. Codebook C1 has been obtained without using the VWA approach.

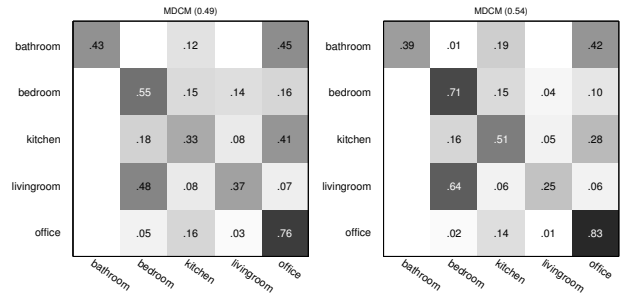
Table VII shows all the results obtained. The baseline codebook C1 obtains the lowest performance. A combination via VWA of three different K-means codebooks, *i.e.* C2, does not immediately improve the results. It is significant that the performance always increases when the contextual information is incorporated. These results demonstrate the convenience of the C-VWA for constructing visual vocabularies. Our best accuracy (54%) is obtained when we combine K-means with the contextual information of a proper segmentation. Furthermore, our results are close to the state-of-the-art reported in [47] when only RGB images are used (56%). Note that in [47], instead of a simple BoW pipeline, a SPMK [34] of four levels is used. We can even improve our results by applying the SPMK [34] with our codebooks too, but this is out of the scope of this paper. Figure 11, shows a comparison of the confusion matrix for the baseline codebook C1, and the combination C1+seg.

VI. CONCLUSION

In this paper we have introduced the VWA methodology, which incorporates the consensus clustering techniques to the

TABLE VII
SCENE CLASSIFICATION RESULTS

Codebook	C1	C2	C1+grid	C1+seg.	C1+grid+seg.	C2+grid	C2+seg.	C2+grid+seg.
MDCM (%)	49	49	52	**54**	50	50	50	52



(a) C1 (b) C1+seg

Fig. 11. Confusion matrices of the scene recognition results in the NYU-D dataset.

visual codebook construction process. Also, a novel sampling strategy has been designed in order to use the VWA approach with large sets of vectors in high dimensional spaces. With the aim of obtaining contextual visual words, we have presented the C-VWA: an approach where we also incorporate the spatial neighboring relation between the local descriptors into the consensus clustering process. We integrate over-segmentation algorithms and spatial grids into the aggregation algorithm in order to capture the contextual relations between the visual words. To the best of our knowledge, this is the first paper to describe such a consensus clustering based methodology within this context. We show the results of the proposed codebooks in visual categorization and scene recognition on very challenging datasets. Results show that the proposed approaches always achieve better performances than traditional BoW approaches.

Most researches simply use a standard clustering algorithm for building the visual vocabularies. In this work, we have investigated how the quality of the visual codebook could be improved, both quantitatively (yielding better classification accuracies) as well as qualitatively (incorporating semantic and contextual information into the visual vocabulary construction process). We have demonstrated that the design choices made in the vector quantization step really matter and have a significant impact on the overall performance of an image categorization system. Exploring other clusterings as well as other data sets is one interesting avenue of future research. Finally, with the aim of making our research reproducible, we release the code[1].

ACKNOWLEDGEMENTS

This work was partially supported by projects TIN2010-20845-C03-03, UAH2011/EXP-030 and IPT-2011-1366-390000.

[1] http://agamenon.tsc.uah.es/Personales/rlopez/data/vwa/

REFERENCES

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV*, 2004.
[2] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.
[3] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *IEEE TCSVT*, vol. 21, no. 4, pp. 381–392, 2011.
[4] D.-X. Li, J.-Y. Jin-ye Peng, Z. Li, and Q. Bu, "LSA based multi-instance learning algorithm for image retrieval," *Signal Processing*, vol. 91, no. 8, pp. 1993–2000, 2011.
[5] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *PAMI*, vol. 32, pp. 1582–1596, 2010.
[6] Y.-G. Jiang, Z. Li, and S.-F. Chang, "Modeling scene and object contexts for human action retrieval with few examples," *IEEE TCSVT*, vol. 21, no. 5, pp. 674–681, 2011.
[7] S. Mukherjee, S. K. Biswas, and D. P. Mukherjee, "Recognizing human action at a distance in video by key poses," *IEEE TCSVT*, vol. 21, pp. 1228–1241, 2011.
[8] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.
[9] D. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999.
[10] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *CVIU*, vol. 110, pp. 346–359, 2008.
[11] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *CVPR*, 2005.
[12] T. Tuytelaars, "Dense interest points," in *CVPR*, 2010.
[13] J. Philbin, O. Chum, J. Isard, M.and Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.
[14] Y. Su and F. Jurie, "Visual word disambiguation by semantic contexts," in *ICCV*, 2011.
[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC)challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.
[16] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 4, 2007.
[17] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions." *Machine Learning Research*, vol. 3, pp. 583–617, 2002.
[18] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling scenes with local descriptors and latent aspects," in *ICCV*, 2005.
[19] J. Yuan and Y. Wu, "Context-aware clustering," in *CVPR*, 2008.
[20] S. Zhang, Q. Tian, G. Hua, W. Zhou, Q. Huang, H. Li, and W. Gao, "Modeling spatial and semantic cues for large-scale near-duplicated image retrieval," *CVIU*, vol. 115, pp. 403–414, 2011.
[21] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *CVPR*, vol. 2, 2003, pp. 264–271.
[22] B. Leibe and B. Schiele, "Interleaved object categorization and segmentation," in *BMVC*, 2003.
[23] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, vol. 2, 2005, pp. 524–531.
[24] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *PAMI*, vol. 31, pp. 1294–1309, 2009.
[25] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE PAMI*, vol. 30, no. 9, pp. 1632–1646, 2008.
[26] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009.
[27] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *ICCV*, 2009.
[28] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool, "Efficient mining of frequent and distinctive feature configurations," in *ICCV*, 2007.
[29] B. Leibe, A. Ettlin, and B. Schiele, "Learning semantic object parts for object categorization," *Image and Vision Computing*, vol. 26, no. 1, pp. 15–26, 2008.
[30] H. Wang, J. Yuan, and Y.-P. Tan, "Combining feature context and spatial context for image pattern discovery," in *IEEE ICDM*, 2011.
[31] J. Winn, A. Criminisi, and A. Minka, "Object categorization by learned universal visual dictionary," in *ICCV*, 2005.

[32] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, "Unifying discriminative visual codebook generation with classifier training for object category recognition," in *CVPR*, 2008.
[33] J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabularies using diffusion distance," in *CVPR*, 2009.
[34] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." in *CVPR*, 2006.
[35] B. Leibe, K. Mikolajczyk, and B. Schiele, "Efficient clustering and matching for object class recognition," in *BMVC*, 2006.
[36] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.
[37] J. Yang, K. Yu, and T. Huang, "Efficient highly over-complete sparse coding using a mixture model," in *ECCV*, 2010.
[38] R. López-Sastre, J. Renes-Olalla, P. Gil-Jiménez, and S. Maldonado-Bascón, "Visual word aggregation," in *Proceedings of the 5th Iberian conference on Pattern Recognition and Image Analysis*, 2011.
[39] M. Aly, M. Munich, and P. Perona, "Multiple dictionaries for bag of words large scale image search," in *ICIP*, 2011.
[40] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *ICML*, 2004.
[41] H. Wang, H. Shan, and A. Banerjee, "Bayesian cluster ensembles," in *SDM*, 2009.
[42] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, pp. 89–113, 2004.
[43] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *ICCV*, 2007.
[44] D. Francois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007.
[45] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, 2004.
[46] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.
[47] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *ICCV 2011 Workshop on 3D Representation and Recognition*, 2011.
[48] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001.