Pattern Recognition ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Discriminative compact pyramids for object and scene recognition

Noha M. Elfiky *, Fahad Shahbaz Khan, Joost van de Weijer, Jordi Gonzàlez

Computer Science Department & Computer Vision Center, Edifici O, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Catalonia, Spain

ARTICLE INFO

Article history:
Received 28 September 2010
Received in revised form
7 July 2011
Accepted 24 September 2011

Keywords:
Object and scene recognition
Bag of features
Pyramid representation
AIB
DITC

ABSTRACT

Spatial pyramids have been successfully applied to incorporating spatial information into bag-of-words based image representation. However, a major drawback is that it leads to high dimensional image representations. In this paper, we present a novel framework for obtaining compact pyramid representation. First, we investigate the usage of the divisive information theoretic feature clustering (DITC) algorithm in creating a compact pyramid representation. In many cases this method allows us to reduce the size of a high dimensional pyramid representation up to an order of magnitude with little or no loss in accuracy. Furthermore, comparison to clustering based on agglomerative information bottleneck (AIB) shows that our method obtains superior results at significantly lower computational costs. Moreover, we investigate the optimal combination of multiple features in the context of our compact pyramid representation. Finally, experiments show that the method can obtain state-of-the-art results on several challenging data sets.

1. Introduction

Bag-of-words based image representation is one of the most successful approaches for object and scene recognition [1–10]. The first stage in the method involves selecting key points or regions followed by a suitable representation of these key points using robust local descriptors, like SIFT [11]. The descriptors are then vector quantized into a visual vocabulary, after which an image is represented as a histogram over visual words. The final representation lacks any spatial information since the location of the local features is ignored. This is generally considered as the foremost shortcoming of the standard bag-of-words representation.

Including spatial information into bag-of-words has therefore received considerable attention. The spatial pyramid scheme proposed by [12] is a simple and computationally efficient extension of an order-less bag-of-words image representation, as it captures the spatial information in such a way that traditional histogram-based image representations do not. This technique works by representing an image using multi-resolution histograms, which are obtained by repeatedly sub-dividing an image into increasingly finer sub-regions. The final representation is a concatenation of the histograms of all the regions. Many applications, such as classification and detection, [13–17] benefit from the spatial pyramid representation.

However, spatial pyramids have a major drawback due to the high dimensionality of the generated histograms while going towards the finest level of representation. This drawback is especially apparent for challenging data sets such as Pascal VOC where it is found that large size visual vocabularies generally improve the overall results. The combination of large vocabularies with spatial pyramids can easily lead to image representations as big as 4194K words (e.g. [18]). If these large pyramid representations could be optimized for discrimination between different categories, a more compact representation would be sufficient. This will lead to compact yet efficient pyramid representations that have the advantages of the original pyramid representation [12] while avoiding their computational burden. This is precisely what we aim at, keeping in mind the constraint of reducing the size of the spatial pyramids while maintaining or even improving the performance.

Many recent works addressed the problem of compact vocabulary construction [19–21]. One popular strategy starts with a large vocabulary (e.g. generated by hierarchical k-means) and subsequently clusters these words together while intending to maintain the discriminative power of the original vocabulary [22,23]. Slonim and Tishby [22] proposed a compression technique, denoted as Agglomerative Information Bottleneck (AIB), that constructs small and informative dictionaries by compressing larger vocabularies following the information bottleneck principle. Interestingly, Fulkerson et al. [20] proposed a fast implementation of the AIB algorithm and showed good performance for the construction of visual vocabularies. Following these trends, we will apply the theory and algorithms developed in these works, for the construction of compact discriminative spatial pyramids.

* Corresponding author.
E-mail addresses: noha@cvc.uab.es (N. M. Elfiky), fahad@cvc.uab.es (F. Shahbaz Khan), joost@cvc.uab.es (J. van de Weijer), poal@cvc.uab.es (J. Gonzàlez).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.09.020

These methods are especially appropriate due to the high dimensionality of the pyramid representation.

An additional advantage of compact pyramid representations is that it allows us to combine more features at the same memory usage for image representation. Combining multiple features especially color and shape has recently shown to provide excellent results [3,4,10,24–26] on standard image classification data sets. The two main most common approaches to combine multiple features are early and late fusion. Early fusion based schemes combine features before the vocabulary construction phase. In case of late fusion separate visual vocabularies are constructed for each feature. Subsequently, the bag-of-word representations (histograms) over the different vocabularies are concatenated. Both fusion approaches have been investigated within the context of standard bag-of-words. However, in the context of spatial pyramids, it is still uncertain which of the two fusion approaches is more beneficial. Therefore, in this paper we investigate which fusion approach is more appropriate within the spatial pyramids framework.

In summary, the objective of this paper is twofold. First, we show that the AIB approach used to compress the vocabulary size significantly degrades accuracy when applied at spatial pyramids. To overcome this problem, we propose to use the divisive information theoretic feature clustering (DITC) technique [23] that preserves the overall accuracy while reducing the dimensionality of the pyramid histogram significantly. Our results clearly suggest that pyramid compression based on the DITC approach provides superior results. Furthermore, DITC is computationally superior to AIB. Second, we evaluate the two existing fusion approaches for combining multiple features at the spatial pyramids level. We conclude that late fusion significantly outperforms early fusion based approaches in spatial pyramids. Finally, we combine both proposed contributions and obtain promising results on challenging data sets.

This paper is organized as follows: Section 2 describes the data sets used in the experiments. Section 3 discusses how AIB and DITC can be used for building compact pyramids. Subsequently, Section 4 proposes both an early and a late fusion strategies for combining multiple features in the context of spatial pyramids. Section 5 compares our results with current state-of-the-art performance results. Finally, Section 6 concludes this paper and describes the most important lines of future research.

2. Data sets and implementation details

In this section we provide details about the data sets which will be used throughout the paper, followed by the experimental setup employed to validate the two main contributions of our approach, namely the use of DITC for vocabulary compression and the use of early and late fusion in spatial pyramids. Fig. 1 shows some example images from the five data sets.

2.1. Data sets

For scene classification, the experiments are performed on Sports Events data set and 15 category Scenes data set. The Sports Events data set [27] contains 8 Sports Events categories collected from the Internet namely: bocce, croquet, polo, rowing, snowboarding, badminton, sailing, and rock climbing. The number of images in each category varies from 137 (bocce) to 250 (rowing). For each event class, 70 randomly selected images are used for training and 60 are chosen for testing.

The 15 class Scenes recognition data set [12] is composed of 15 scene categories. Each category has 200–400 images. The major sources of the pictures in the data set include the COREL collection, personal photographs, and Google image search.

For object classification, the experiments are performed on Butterflies [28] and Pascal VOC 2007 and 2009 data sets [15]. The Butterflies data set consists of 619 images of seven classes of butterflies, namely: Admiral, Swallowtail, Machaon, Monarch 1, Monarch 2, Peacock and Zebra. Finally, the experiments are also performed on the Pascal Visual Object Classes Challenge (VOC) data sets: the Pascal VOC 2007 data set consists of 9963 images of 20 different classes with 5011 training images and 4952 test images, while the Pascal VOC 2009 data set contains 13 704 images of 20 different object categories with 7054 training images and 6650 test images.

2.2. Implementation details

We shortly discuss the implementation details we use for the bag-of-words based image classification. We apply a standard

Fig. 1. Example images from the data sets. From top to down: Butterflies, Sports Events, 15 class Scenes and Pascal VOC data sets.

Please cite this article as: N. M. Elfiky, et al., Discriminative compact pyramids for object and scene recognition, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.09.020

multiple-scale grid detector along with interest point detectors (Harris–Laplace and blob detector). In the feature extraction step, we use SIFT descriptor [11] for shape features, Color Names [29] descriptor for color features and the SelfSimilarity descriptor [30] to measure similarity based on matching the internal self-similarity. We use a standard k-means for constructing visual vocabularies. Finally we use a non-linear SVM with intersection kernel for classification as in [31].

2.3. Image representation using spatial pyramids

Spatial pyramid scheme proposed by Lazebnik et al. [12] has recently proven very successful results. These are formed by representing an image using weighted multi-resolution histograms, which are obtained by repeatedly sub-dividing an image into increasingly finer sub-regions by doubling the number of divisions in each axis direction and computing histograms of features over the resulting sub-regions. Resemblances found at finer resolutions are closer to each other in image space and are therefore more heavily weighted. To accomplish this, each level l is weighted to $l/2^{L-l}$, where L is the total number of pyramid levels considered. When histograms for all sub-regions at all levels have been created, these histograms are concatenated to form the final image representation. For example, a level 2 spatial pyramid is constructed by concatenating a total of $1+4+16=21$ histograms.

Although a notable performance gain is achieved by using the spatial pyramid method, the resulting histogram is often a magnitude higher in dimensionality over its standard bag-of-words based counterpart.[1]

3. Compact pyramid representation

As discussed in the Introduction, one of the main drawbacks of the spatial pyramid representation is its memory usage. We will discuss two existing approaches, namely AIB and DITC, which were shown to be successful for compact text document representation [22,23]. Only AIB has been applied for compact image representation [20], and none of them has been studied in the context of spatial pyramids. In this section we will show experimental results on the Sports Events [27] and 15 class Scenes [12] data sets to demonstrate that our proposed compact pyramid representation maintains the performance of their larger counterparts.

In practice the final size of the pyramid is dependent on the application, where users have to balance compactness versus classification accuracy. Depending on the task a smaller representation could be preferred over larger at the cost of performance (e.g. real-time object detection based on ESS [13,33], or large scale image retrieval [34]). In the case that users do not want a drop in accuracy but do want to compress their representation, cross-validation could be used to select the optimal cluster size. Throughout this paper we consider that the final representation size is an input parameter to the compression algorithm.

3.1. Highly informative compact spatial pyramids

Let C be a discrete random variable that takes on values from the set of classes $C=\{c_1, \ldots, c_l\}$ and let W be the random variable that ranges over the set of words $W=\{w_1, \ldots, w_m\}$. It is important to note that we consider the number of words for the spatial pyramid representation to be equal to the number of words used for the visual vocabulary times the number of sub-regions in the spatial pyramid. For a level two pyramid constructed from a 1000 word vocabulary, this will lead to a final representation of $(1+4+16)\times 1000=21\,000$ words. We will consider clustering these 21 000 words into a smaller set where each cluster represents words with a similar discriminative power.

The joint distribution $p(C,W)$ is estimated from the training set by counting the number of occurrences of each visual word in each category. The information about C captured by W can be measured by the mutual information:

$$I(C,W)=\sum_i\sum_t p(c_i,w_t)\log\frac{p(c_i,w_t)}{p(c_i)p(w_t)}, \tag{1}$$

which measures the amount of information that one random variable contains about the other. Ideally, in forming word clusters we aim at preserving the mutual information; however, usually clustering lowers the mutual information. Thus, we aim at finding word clusters that minimize the decrease in the mutual information:

$$I(C,W)-I(C,W^C), \tag{2}$$

where W^C are the word clusters $\{W_1, \ldots, W_k\}$. Note that this is equal to maximizing the mutual information $I(C,W^C)$. Eq. (2) can be rewritten as

$$\sum_i\sum_t \pi_t p(c_i|w_t)\log\frac{p(c_i|w_t)}{p(c_i)}-\sum_i\sum_j\sum_{w_t\in W_j}\pi_t p(c_i|w_t)\log\frac{p(c_i|W_j)}{p(c_i)}, \tag{3}$$

where π_t is the prior of word, and is given by $\pi_t=p(w_t)$.

In the seminal work [23], Dhillon et al. prove that this is equal to

$$I(C,W)-I(C,W^C)=\sum_j\sum_{w_t\in W_j}\pi_t KL((p(C|w_t)),(p(C|W_j))), \tag{4}$$

where the Kullback–Leibler (KL) divergence is defined by

$$KL(p_1,p_2)=\sum_{x\in X}p_1(x)\log\frac{p_1(x)}{p_2(x)}. \tag{5}$$

Eq. (4) is a global objective function that can be applied to measure the quality of word clustering. This object function states that we should group words w_t into clusters W_j, in such a way that the summed KL-divergence between the word distributions $p(C|w_t)$ and their cluster distributions $p(C|W_j)$ is as low as possible. Since the KL-divergence is a measure of similarity between distributions, we are clustering words together which contain a similar information with respect to the classes as described in $p(C|w_t)$. Next we discuss two existing algorithms which aim to find the optimal clusters W_j as defined by Eq. (4).

AIB compression [22]: AIB iteratively compresses the dictionary W by merging the visual words w_i and w_j that cause the smallest decrease in the mutual information given by Eq. (1). The decrease in the mutual information is monotonically reduced after each merging. Merging is iterated until one obtains the desired number of words. AIB is greedy in nature as it optimizes the merging of just two word clusters at every step (a local optimization) and thus the resulting algorithm does not directly optimize the global criteria defined in Eq. (4).

DITC compression [23]: Other than AIB which iteratively reduces the number of words until the desired number of clusters is reached, DITC immediately clusters the words into the desired number of clusters (during initialization) after which it iteratively improves the quality of these clusters. Each iteration monotonically reduces the

[1] The winners of Pascal VOC 2007 [32] showed that dividing an image horizontally 3×1 yields better performance than a conventional 4×4 structure. The resulting histogram is therefore reduced from vocabulary size $\times 21$ to vocabulary size $\times 8$.

Please cite this article as: N. M. Elfiky, et al., Discriminative compact pyramids for object and scene recognition, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.09.020

decline in the mutual information as given by Eq. (4), therefore the algorithm is guaranteed to terminate at a local minimum in a finite number of iterations.

To optimize the global objective function of Eq. (4), DITC iteratively performs the following steps:

1. Compute the cluster distribution $p(C|W_j)$ according to:

$$p(C|W_j) = \sum_{w_t \in W_j} \frac{\pi_t}{\pi(W_j)} p(C|w_t), \quad (6)$$

where $\pi(W_j) = \sum_{w_t \in W_j} \pi_t$.

2. Reassign the words w_t to the clusters W_j based on their closeness in KL-divergence:

$$j^*(w_t) = \arg\min_j KL(p(C|w_t), p(C|W_j)), \quad (7)$$

where $j^*(w_t)$ is new cluster index of the word w_t.

The initialization of the k clusters is obtained by first clustering the words into l clusters, where l is the number of classes. Every word w_t is then assigned to cluster W_j such that $p(c_j|w_t) = \max_i p(c_i|w_t)$. This strategy guarantees that every word w_t is part of one of the clusters W_j. Subsequently we split each cluster arbitrarily into $\lfloor k/l \rfloor$ clusters. In the case that $l > k$ we further merge the l clusters to obtain k final clusters. The above algorithm is only an approximation of the minimum but it was found to yield accurate results [23].

The basic implementation of the DITC algorithm can result in a large number of empty clusters, especially for large vocabularies. To overcome this problem we propose a modified version of the basic DITC algorithm. At each iteration our algorithm retrieves the index e of the empty word clusters c_e, where $e \subset j$. Subsequently we assign at least one word w_t to each c_e. This is done using Eq. (7) by first assigning each word w_t to its closest word cluster c_j. Based on this assignment, we select that w_t with the maximum KL value returned by Eq. (7), i.e. that w_t found at the furthest distance from its currently assigned word cluster c_j. Then we reassign this w_t to c_e and remove it from c_j.

Comparing the computational cost of the two algorithms shows one of the advantages of DITC: AIB results in a high computational cost of $O(m^3c)$ operations as it runs an agglomerative algorithm until k clusters are obtained. Here m is the total number of words and c is the number of classes in the data set. The fast implementation of the AIB costs $O(m^2c)$. On the other hand, the DITC algorithm requires Eq. (7) to be computed for every pair, $P(C|w_t)$ and $p(C|W_j)$ at a cost of $O(mkc\tau)$, where generally $k \ll m$. The number of required iterations τ to obtain convergence is typically around 15. We found DITC in practice to be computationally superior to AIB, obtaining a speedup between one or two orders of magnitude. On a typical run for obtaining 100 clusters from 20 000 words on a data set with 15 classes, AIB (using [20]) took 14 460 seconds while DITC converged in 234 seconds using a standard PC.

3.2. Experimental results

In this section, we compare the two algorithms discussed above on the task of constructing compact spatial pyramids. To the best of our knowledge we are the first to apply DITC to visual word vocabulary construction. Lazebnik and Raginsky [21] propose a method for discriminative vocabulary construction which uses ideas of the theory of DITC [23]. However, the word clusters were restricted to lie in Voronoi cells, whereas in the original algorithm words are clustered without restrictions on their location in feature space, and thus allowing for multi-model distributions. We show that the pyramid compression based on DITC has a lower loss of discriminative power, and is computationally more efficient compared to compression based on AIB [20].

Table 1 shows numerical results obtained by applying AIB on both the Sports Events and 15 Scenes data sets for different sizes. We started by using vocabulary of size 1000 for constructing a three level pyramid of 21 000 dimensionality, after which we compress this vocabulary to a dimensionality of 5000, 1000 and 500. We can notice that by applying AIB compression on the pyramids the performance drops significantly, especially when we are going towards lower dimensionality. We attribute this to the fact that the information bottleneck technique is agglomerative in nature and results in a sub-optimal word cluster because it greedily merges just two word clusters at every step and it does not directly optimize the global objective function of Eq. (4).

Table 2 shows the results obtained using DITC. The main observation is that the DITC approach succeeds in conserving the discriminative power while reducing dimensionality of the image representation. Furthermore, for both sets reducing the dimensionality leads to an improvement of the classification score, and even at the smallest dimensionality of 500 similar results are obtained as with the total 21 000 word vocabulary.

Classification accuracies of both compression approaches are shown in Fig. 2 which supports the two main conclusions: first, using DITC compression mechanism leads to a compact pyramid representation that reduces the dimensionality of the original pyramid yet preserves or even improves its performance. Second, compact pyramid representation based on DITC achieves better results than those based on AIB approaches at all the vocabulary sizes. Moreover the performance gain is more significant for smaller vocabularies.

We also perform experiments comparing the performance of DITC compression with Principle Component Analysis (PCA) and Partial Least Square (PLS) techniques. Fig. 3 shows the comparison on two data sets. We only show the performance for very compact pyramid representations, since PLS is known to obtain better results for compact representation and quickly deteriorates for larger representation. Moreover, the number of dimensions of PCA is bounded by the number of observations. DITC based pyramid compression consistently outperforms the other two compression techniques. It is worthy to mention that DITC also provides better performance compared to both PCA and PLS with a very small compact pyramid representation (50 bins).

The performance difference between DITC and AIB becomes especially apparent for high compression. An initial pyramid

Table 1
Classification score (percentage) on both the Sports Events and 15 class Scenes data sets. The results demonstrate that by applying the AIB compression [20] a considerable loss in performance occurred for compact vocabularies.

Method	Level	Size	Sports Events	15 class Scenes
Pyramid	2	21 000	83.8	84.1
$Pyramid_{AIB}$	2	5000	81.5	81.7
$Pyramid_{AIB}$	2	1000	79.8	80.4
$Pyramid_{AIB}$	2	500	78.8	78.3

Table 2
Classification score (percentage) on both the Sports Events and 15 class Scenes data sets. The results demonstrate that DITC successfully compresses the vocabularies while preserving their discriminative power.

Method	Level	Size	Sports Events	15 class Scenes
Pyramid	2	21 000	83.8	84.1
$Pyramid_{DITC}$	2	5000	84.2	85.4
$Pyramid_{DITC}$	2	1000	85.6	84.4
$Pyramid_{DITC}$	2	500	84.6	84.2

Please cite this article as: N. M. Elfiky, et al., Discriminative compact pyramids for object and scene recognition, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.09.020

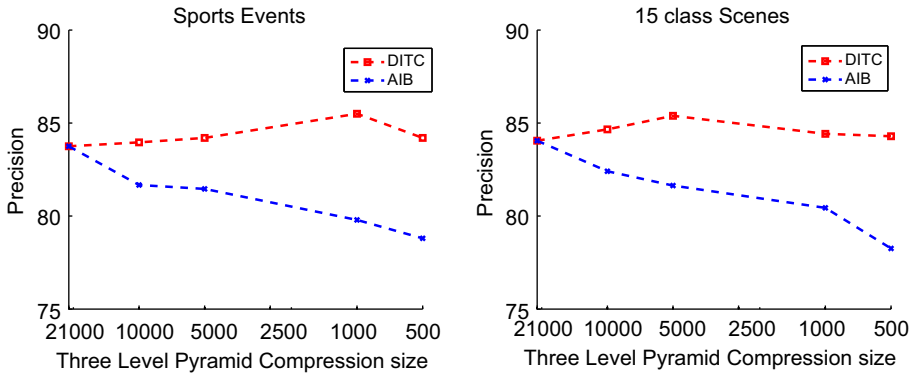

Fig. 2. Sports Events data set (left) and 15 class Scenes data set (right) classification accuracy for compressing the whole pyramid representation leading to a more compact pyramid representation using the two compression approaches considered namely: DITC vs. AIB.

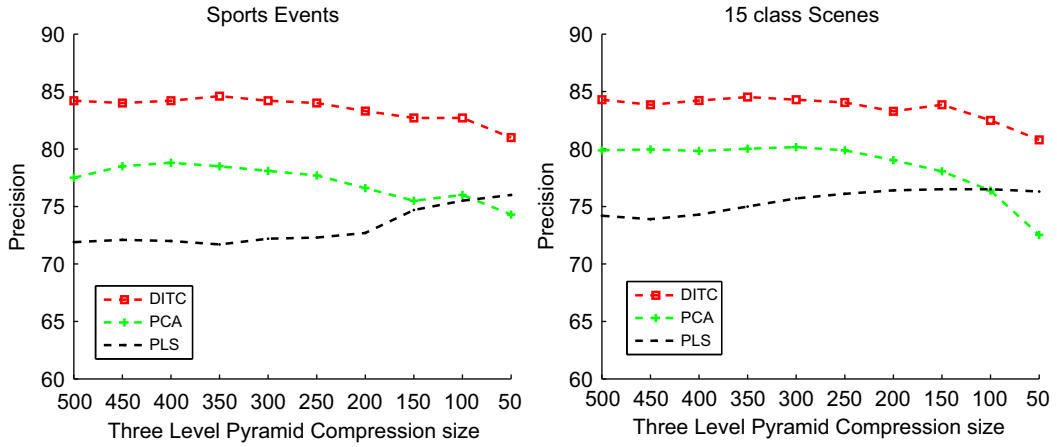

Fig. 3. Sports Events data set (left) and 15 class Scenes data set (right) classification accuracy for compressing the whole pyramid to a compact representation using approaches namely: DITC, PLS and PCA. Note that DITC based compression also provides superior performance for very compact pyramid representations.

Table 3
Average-precision results for all classes of the Pascal VOC 2007 database. Comparison on the average accuracy of the original four level pyramid representation of size 25 500 compressed to size 200. The second row shows the compression results using the AIB [20] and the third row shows the results using DITC [23].

	Plane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table
Pyramid	72.1	54.9	41.9	62.6	23.9	46.3	71.4	51.4	48.8	37.4	46.8
AIB	53.2	28.3	24.6	43.2	11.4	27.5	54.2	29.9	35.6	11.1	13.9
DITC	61.4	50.6	36.5	49.1	20.3	43.9	68.2	44.1	47.1	29.7	38.8

	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	Train	TV	Mean
Pyramid	38.9	72.1	58.1	80.3	25.4	32.4	41	70.5	43.6	50.9
AIB	21.1	41.3	32.3	73.3	10.4	13.9	27.9	40.2	27.8	31.1
DITC	33.4	69.5	53.6	78.9	23.6	22.9	37.6	64.3	42.3	45.8

representation of the Pascal data set of 25 500 words is compressed to 200 clusters. Table 3 shows a 14% higher mean average-precision for having compact pyramid representations based on DITC compared to those obtained using AIB on object recognition.

3.3. Compact pyramid designs

As demonstrated in the last section, we can significantly reduce the dimensionality while preserving or even improving the performance of the original pyramid representation that we started with. We next evaluate and compare two different design strategies for building our final compact pyramid representations. The main aim is to find an optimal design for obtaining compact yet efficient pyramids based on the DITC compression algorithm. The two proposed designs are the following:

1. Compute a vocabulary, compress it using DITC and subsequently build a compact pyramid representation based on the compressed compact vocabulary (the traditionally used schema, denoted as *CompPyr* hereafter).
2. Construct the pyramid representation for an image and subsequently compress the vocabulary of the whole pyramid directly using DITC (strategy presented in Section 3.1 and denoted as *PyrComp* hereafter).

Table 4 shows the results obtained using both of the considered proposed designs on 15 class Scenes and the Sports Events data sets. To compare the classification scores obtained from the

Please cite this article as: N. M. Elfiky, et al., Discriminative compact pyramids for object and scene recognition, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.09.020

two designs, we consider the same dimensionality of size 1000. For the 15 class Scenes data set, using *CompPyr* we got a score of 82.1%, while *PyrComp* gives us a performance of 84.4%. For the Sports Events data set, we observe a similar gain in the obtained results.

These quantitative results suggest how optimal compact pyramid representations can be built: although both designs preserve the accuracy of the original pyramid representation, the best results are obtained following the *PyrComp* strategy, since it does not only preserve the original pyramid performance, but also slightly improves the performance. Additionally Fig. 4 illustrates another interesting conclusion: the gain in performance using *PyrComp* is obtained throughout all sizes, and this gain is more significant at lower sizes.

The *CompPyr* compresses the vocabulary while ignoring the spatial pyramid image representation to which it will later be applied. This strategy is used by most existing methods for compact vocabulary construction [21,35,36]. Our experiment shows that compressing the vocabulary within the spatial pyramid, significantly improves the results. Compression with *PyrComp* has the same freedom as *CompPyr* to merge words within a sub-window. Additionally, it can also merge words of different sub-windows, something which is impossible within the *CompPyr* strategy.

4. Combining multiple features in spatial pyramids

In the previous section, we have provided an efficient method for the construction of compact pyramid representations. The gained compactness allows us to combine more features at the same memory usage of the image representation. Here we analyze how to optimally combine multiple features in a pyramid representation.

We will look at the particular case of combining color and shape, which was shown to provide excellent results for object and scene recognition [14]. In particular we investigate two approaches to combine multiple features, namely the early and late fusion schemes. In the next section we provide results from combining visual cues other than color and shape.

Table 4
Classification score on the Sports Events and 15 class Scenes data sets using the DITC approach comparing the two proposed designs: *CompPyr* (compute a vocabulary, compress it, and then build a compact pyramid representation using this compressed compact vocabulary) and *PyrComp* (i.e. construct a pyramid representation for an image, then compress the words of the whole pyramid afterwards).

Method	Level	Size	Sports Events	15 Class Scenes
Pyramid	2	21 000	83.8	84.1
$Pyramid_{AIB}$	2	1000	79.8	80.4
CompPyr	2	1000	81.9	82.1
PyrComp	2	1000	**85.6**	**84.4**

Note: Bold numbers correspond to the highest classification scores.

4.1. Early and late fusion spatial pyramid matching

In early fusion the local features of color and shape are concatenated into a single feature. Subsequently, the combined color and shape features are quantized into a joint shape–color vocabulary. In general, early fusion results in vocabularies with high discriminative power, since the visual-words describe both color and shape jointly, allowing for the description of blue corners, red blobs, etc. A significant shortcoming of early fusion approach is that it deteriorates for categories which vary significantly over one of the visual cues, for example, man made categories such as cars and chairs which vary considerably in color. In such cases, the visual-words will be contaminated by the "irrelevant" color information. The relevant shape words will be spread over multiple visual-words, thereby complicating the task of the learning algorithm significantly. On the other hand, early fusion is suitable for categories which are constant over both color and shape cues like plants, lions, road-side signs, etc.

The second approach, called late fusion, fuses the two cues, color and shape, by processing the two features independent of each other. Separate visual vocabularies are constructed for color and shape independently, and the image is represented as a distribution over shape-words and color-words. A significant drawback of late fusion is that we can no longer be certain that both visual cues come from the same location in an image. Late fusion is expected to provide better results over early fusion on categories where one cue is constant and the other varies considerably. Examples of such categories are man made objects such as car, buses, chairs, etc.

Typically within the bag-of-words framework a number of local features f^c_{mn}, $m=1,\ldots,M^n$ are extracted from training images I_n. Where $n=1,2,\ldots,N$, and $c\in\{1,2\}$ is an index indicating the different visual features. In case of early fusion, two visual features are concatenated according to

$$f^{1\&2}_{mn}=(\beta f^1_{mn},(1-\beta)f^2_{mn}). \tag{8}$$

Vector quantization of f^1, f^2, $f^{1\&2}$ yields the corresponding vocabularies V_1, V_2, $V_{1\&2}$. We define $h^V(I)$ to be the histogram representation of image I in vocabulary V. Early fusion representation of the image is given by $h^{V_{1\&2}}(I)$ and the late fusion is

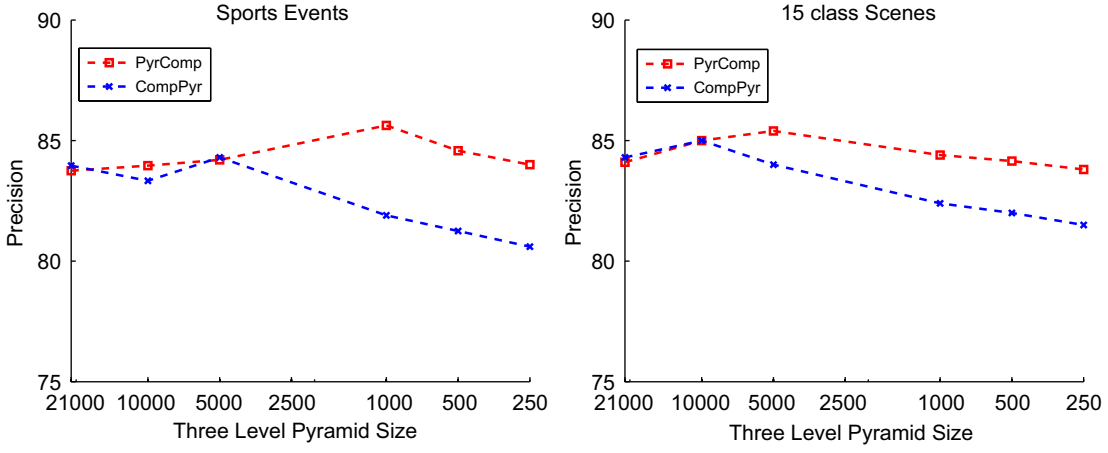

Fig. 4. Classification comparison between *PyrComp* and *CompPyr* strategies for (left) 15 class Scenes and (right) Sports Events data sets.

Please cite this article as: N. M. Elfiky, et al., Discriminative compact pyramids for object and scene recognition, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.09.020

obtained by concatenating the separate histograms:

$$h^{(V_1,V_2)}(I) = [\beta h^{V_1}(I), (1-\beta)h^{V_2}(I)]. \quad (9)$$

Note that we have introduced a weight parameter β for both early and late fusions which allow us to leverage the relative weight of the various cues. In our setting this parameter is learned through cross-validation on the training data. Both fusion schemes can easily be extended to accommodate several visual cues.

Before applying the two schemes on spatial pyramids, we will shortly discuss the relation of existing approaches for the combination of multiple features to early and late fusion. Bosch et al. [3] compute the SIFT descriptor on the H, S, V channels and then concatenate the final descriptor into a single representation. van de Weijer and Schmid [26] compare photometrically invariant representations in combination with SIFT for object recognition. Recently, van de Sande et al. [10] performed a study on the photometric properties of many color descriptors, and did an extensive performance evaluation. In their evaluation OpponentSIFT was shown to be the best choice to combine color and shape features. Since in all these works color and shape are combined before vocabulary construction, they are considered early fusion methods.

Regarding late fusion, several methods explore the combination of multiple features at the classification stage. These approaches, of which multiple kernel learning MKL is the most well-known [37–41], combine kernel combinations of different visual features. A weighted linear combination of kernels is employed, where each feature is represented by multiple kernels. Besides the multiple kernel learning approach, the two conventional approaches that combine different kernels at the classification stage in a specified deterministic way are *averaging* and *multiplying* the different kernel responses. Surprisingly, the product of different kernel responses is shown to provide similar or even better results than MKL in a recent study performed by Gehler and Nowozin [24]. These approaches are considered as late fusion since they perform vocabulary construction separately for the different features. Recently, an alternative method for combining color and shape, called color attention, was proposed by Khan et al. [42]. However, it is unclear how this method can be extended to incorporate spatial pyramids, since the normalization performed in the sub-regions of the pyramid counters the color attention weighting.

For the standard bag-of-features image representation there is no consensus whether early or late fusion is better. Here we investigate the two approaches in the context of spatial pyramids. The common methodology employed in current object recognition frameworks is to build spatial pyramids of early fusion based schemes (such as Opp-SIFT, C-SIFT, HSV-SIFT, etc.) [3,10,26]. We refer to these spatial pyramids that are based on early fusion scheme as *early fusion spatial pyramids* and the spatial pyramids that are based on late fusion as *late fusion spatial pyramids*. Fig. 5 highlights the two spatial pyramid matching approaches.

4.2. Experimental results of early and late fusion based spatial pyramids

To evaluate both early and late fusion spatial pyramids, we perform an experiment for both object and scene recognitions. For scene classification, the experiments are performed on Sports Events data set. We use the Butterflies data set for the object recognition task. To construct a shape vocabulary we use the SIFT descriptor and the Color Names descriptor [29] for creating a color vocabulary. We combine the two cues based on early fusion and late fusion schemes, both at the standard bag-of-words level and at the spatial pyramids level. To obtain a fair comparison between early and late fusions we use the two standard implementations as given by Eqs. (8) and (9). The parameter β in both equations is learned by cross-validation.

We also compare with OpponentSIFT which was shown to be the best color–shape descriptor in a recent evaluation [10]. Table 5 shows the results obtained on Sports Events data set. For this data set, shape is an important cue and color plays a subordinate role. At the standard bag-of-words level, OpponentSIFT provides the best results but as we move into higher levels of spatial pyramids the performance of both early fusion and OpponentSIFT starts to

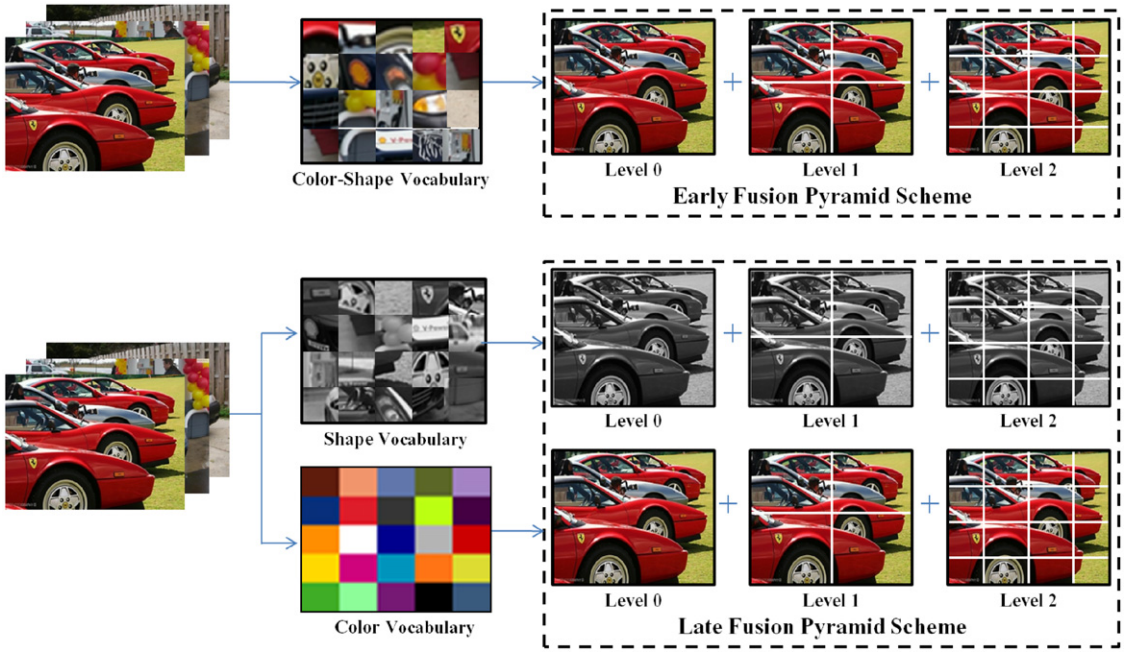

Fig. 5. Early and late fusion pyramid schemes. In the early fusion pyramid scheme a combined color–shape vocabulary is constructed as a result of which a single pyramid representation is obtained. To construct a late fusion pyramid, a separate vocabulary is constructed for color and shape and spatial pyramids are obtained for each cue. We show that late fusion is the recommended approach for combining multiple features. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: N. M. Elfiky, et al., Discriminative compact pyramids for object and scene recognition, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.09.020

Table 5
Classification score (percentage) on Sports Events data set.

Method	Level	Size	Score
Shape	0	800	80.6
Color	0	300	53.9
Opp-SIFT	0	1100	**82.9**
Early fusion	0	1100	80.6
Late fusion	0	1100	81.8
Opp-SIFT	1	5500	82.3
Early fusion	1	5500	80.8
Late fusion	1	5500	**82.7**
Opp-SIFT	2	23 100	80.8
Early fusion	2	23 100	82.7
Late fusion	2	23 100	**84.4**

Note: Bold numbers correspond to the highest classification scores.

Table 6
Classification score (percentage) on Butterflies data set.

Method	Level	Size	Score
Shape	0	1000	79.4
Color	0	300	53.3
Opp-SIFT	0	1500	78.7
Early fusion	0	1500	79.6
Late fusion	0	1300	**81.9**
Opp-SIFT	1	7500	79.6
Early fusion	1	7500	81.7
Late fusion	1	6500	**84.4**
Opp-SIFT	2	31 500	81.0
Early fusion	2	31 500	83.3
Late fusion	2	27 300	**87.9**

Note: Bold numbers correspond to the highest classification scores.

degrade (the performance of OpponentSIFT at the finest pyramid level is below its performance at the standard bag-of-words level). We also combined color and shape at the kernel level with the product rule as advocated by Gehler and Nowozin [24]. However, results were found to be inferior compared to the late fusion spatial pyramid scheme.

Table 6 shows the results obtained on Butterflies data set. Shape plays an important role as depicted from the results of individual visual cues. Late fusion provides better results at the standard bag-of-words level than both early fusion and OpponentSIFT. The performance gain of late fusion is further increasing when more pyramid levels are considered.

In conclusion, in a standard bag-of-words representation both early and late fusions obtain comparative results. However, our experiments show that within a spatial pyramid representation late fusion significantly outperforms early fusion. These results of late fusion could further be improved by applying multi-kernel learning.

5. Comparison to state-of-the-art

In the previous section we have investigated how to optimally compute compact and multi-feature spatial pyramids. We have shown that optimal results are obtained by using DITC algorithm for compression, and using the *PyrComp* strategy for the computation of compact pyramids. Furthermore, as demonstrated in the previous section, late fusion pyramids are shown to be more efficient than early fusion pyramids. In this section, we combine these conclusions to construct compact multi-feature spatial pyramids. First we compute compact spatial pyramids for each feature separately and then combine them in a late fusion manner.

Table 7
Classification score (percentage) on Sports Events, 15 class Scenes, Butterflies, Pascal VOC 2007 and 2009 data sets.

Data sets	Best score		*PS*		PS_C		$PS_C+PC_C+PSS_C$	
	Size	**Score**	Size	**Score**	Size	**Score**	Size	**Score**
Sports	6 K	84.2 [17]	21 K	83.8	1 K	85.6	2 K	**87.1**
15 Scenes	21 K	84.3 [43]	21 K	84.1	1 K	84.4	2 K	**86.7**
Butterflies	2 K	90.6 [28]	21 K	89.5	1 K	89.0	2 K	**91.4**
Pascal 2007	160 K	**60.5** [10]	84 K	57.4	15 K	57.2	25 K	59.5
Pascal 2009	4194 K	**64.6** [18]	84 K	55.7	15 K	55.2	25 K	57.6

Note: Bold numbers correspond to the highest classification scores.

We denote our pyramid representation for SIFT with *PS*, and the compact pyramids of SIFT, SelfSimilarity and Color with PS_C, PSS_C and PC_C respectively. We report the final results on all the four challenging data sets obtaining very good classification scores even when reducing the pyramid histograms significantly. In addition, we compare our results with several recent results reported on these data sets in the literature. Table 7 shows our final results and a comparison with the best results reported on the four data sets.

For the *Sports Events data set* experiments are repeated five times by splitting the data set into train and test set and the mean average accuracy is reported. As depicted from the results, each feature's compact representation preserves or even improves the performance over its original pyramid histogram. The original three level pyramid representation of SIFT (PSIFT) with dimensionality 21 000 gives an accuracy of 83.8 while, compressing it to 1000 we improve the score to 85.6. By combining the three compact pyramid representations we obtained a classification score of 87.1 which exceeds the state-of-the-art results obtained on this data set [17,16,43–45]. The final accuracy is obtained with our compact histogram of dimensionality 2000 reduced from the original pyramid histograms of dimensionality 42 000.

For the *15 category Scenes data set*, we followed the standard protocol of splitting the data set into training and testing five times and reported the mean classification score. The results of each feature compact pyramid representation preserves or even improves the performance of its original pyramid representation. The original three level pyramid structure of SIFT (*PS*) with dimensionality 21 000 gives an accuracy of 84.1 while, compressing it to 1000 we improve the score further to 84.4. Since there is no color in this data set, we only combine the compact pyramids obtained from SIFT and SelfSimilarity. Our final compact representation has a histogram of size 2000 reduced from original pyramid histograms having a dimensionality of 42 000. We obtained a classification accuracy of 86.7 which is to the best of our knowledge the best performance on this data set [17,16,43–45].

The *Butterflies data set* shows our approach on a object recognition data set. Our compact pyramid representation of SIFT provides comparable results w.r.t. the original pyramids of SIFT. Our final combination yields a score of 91.4 which outperforms the best reported result in [28].

The results on the *Pascal VOC 2007* show that we reduce the pyramid histogram of SIFT to one-third with a small loss. The final mean average precision of 59.5 is obtained with a histogram size of 25K. Our final results are close to state-of-the-art, but we have significantly reduced the histogram dimension (25K) compared to the approach of van de Sande et al. [10], where SIFT pyramids are combined with 4 ColorSIFT pyramids, leading to higher histogram dimensions of 160K. Lastly, it should be noted that better results (63.5) were reported in [46], where authors include additional information of object bounding boxes from object detection to improve image classification.

Please cite this article as: N. M. Elfiky, et al., Discriminative compact pyramids for object and scene recognition, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.09.020

For the *Pascal VOC 2009*, a similar behavior is noticed. Hence, with an original SIFT pyramid of size 84K a mean average score of 55.7 is obtained. However, we maintained a score of 55.2 using our 15K compact SIFT representation. Finally, the results for multiple features fusion improve the overall mean average score up to 57.6 over the compact SIFT features.

6. Conclusions

A major drawback of spatial pyramids is the high dimensionality of their image representation. In this paper we have proposed a method for the computation of compact discriminative pyramids. The method is based on the divisive information theoretic feature clustering algorithm, which clusters words based on their discriminative power. We show that this method outperforms clustering based on the agglomerative information bottleneck both in accuracy and in computational complexity. Results show that depending on the data set dimensionality reductions up to an order of magnitude are feasible without a drop in performance. The gained compactness leaves more room for the combination of features. We investigate the optimal strategy to combine multiple features in a spatial pyramid setting. Especially for higher level pyramids late fusion was found to significantly outperform early fusion pyramids. We evaluated the proposed framework on both scene and object recognitions, and obtained state-of-the-art results on several benchmark data sets.

For future work we are particularly interested in applying the compact pyramids to the task of bag-of-words based object detection [13,46]. The application of bag-of-words based detection has been particularly advanced due to the efficient subwindow search (ESS) algorithm proposed by Lampert et al. [13]. The usage of compact discriminative pyramids to this application could help obtain faster detection methods without loss in accuracy.

Another line of future research includes investigating the application of DITC to sparse image representation [35,36], which has been shown excellent results in recent works in image restoration and face recognition [47,48]. Although discriminative vocabularies within the context of sparse image representation have been investigated, these methods still ignore the spatial pyramid for the construction of discriminative vocabularies, whereas our work shows that compressing the vocabulary within the spatial pyramid significantly improves results. Therefore, we expect that combining the strengths of both methods will lead to further improvements.

Acknowledgments

This work has been supported by the EU projects ERGTS-VICI-224737 and VIDI-Video IST-045547; the Spanish Research Program Consolider-Ingenio 2010:MIPRCV (CSD200700018); Avanza I+D ViCoMo (TSI-020400-2009-133); and by the Spanish projects TIN2009-14173 and TIN2009-14501-C02-02. Joost van de Weijer acknowledges the support of a Ramon y Cajal fellowship.

References

[1] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: an in-depth study. A comprehensive study, International Journal of Computer Vision 73 (2) (2007) 213–218.
[2] G. Csurka, C. Bray, C. Dance, L. Fan, Visual categorization with bags of key points, in: Workshop on Statistical Learning in Computer Vision, ECCV, 2004.
[3] A. Bosch, A. Zisserman, X. Munoz, Scene classification via PLSA, in: Proceedings of European Conference on Computer Vision, 2006.
[4] A. Bosch, A. Zisserman, X. Munoz, Scene classification using a hybrid generative/discriminative approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (4) (2008) 712–727.
[5] G. Dorko, C. Schmid, Selection of scale-invariant parts for object class recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2003.
[6] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Proceedings of the Computer Vision and Pattern Recognition, 2005.
[7] S. Lazebnik, C. Schmid, J. Ponce, A sparse texture representation using local affine regions, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1265–1278.
[8] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (10) (2005) 1615–1630.
[9] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, L.V. Gool, Modelling scenes with local descriptors and latent aspects, in: Proceedings of the IEEE International Conference on Computer Vision, 2005.
[10] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1582–1596.
[11] D.G. Lowe, Distinctive image features from scale-invariant points, International Journal of Computer Vision 60 (2) (2004) 91–110.
[12] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the Computer Vision and Pattern Recognition, 2006.
[13] C. Lampert, M. Blaschko, T. Hofmann, Beyond sliding windows: object localization by efficient subwindow search, in: Proceedings of the Computer Vision and Pattern Recognition, 2008.
[14] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2008 (voc2008) Results.
[15] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 Results.
[16] J. Wu, A fast dual method for HIK SVM learning, in: Proceedings of the European Conference on Computer Vision, 2010.
[17] J. Wu, J. Rehg, Beyond the Euclidean distance: creating effective visual codebooks using the histogram intersection kernel, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.
[18] J. Yang, K. Yu, T. Huang, Efficient highly over-complete sparse coding using a mixture model, in: Proceedings of the European Conference on Computer Vision, 2010.
[19] J. Winn, A. Criminisi, T. Minka, Object categorization by learned universal visual dictionary, in: Proceedings of the IEEE International Conference on Computer Vision, 2005.
[20] B. Fulkerson, A. Vedaldi, S. Soatto, Localizing objects with smart dictionaries, in: Proceedings of the European Conference on Computer Vision, 2008.
[21] S. Lazebnik, M. Raginsky, Supervised learning of quantizer codebooks by information loss minimization, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (7) (2009) 1294–1309.
[22] N. Slonim, N. Tishby, Agglomerative information bottleneck, in: Advances in Neural Information Processing Systems, 1999.
[23] I.S. Dhillon, S. Mallela, R. Kumar, I. Guyon, A. Elisseeff, A divisive information-theoretic feature clustering algorithm for text classification, Journal of Machine Learning Research 3 (2003) 1265–1287.
[24] P.V. Gehler, S. Nowozin, On feature combination for multiclass object classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.
[25] G.J. Burghouts, J.M. Geusebroek, Performance evaluation of local colour invariants, Computer Vision and Image Understanding 113 (2009) 48–62.
[26] J. van de Weijer, C. Schmid, Coloring local feature extraction, in: Proceedings of the European Conference on Computer Vision, 2006.
[27] L.-J. Li, L. Fei-Fei, What, where and who? classifying events by scene and object recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2007.
[28] D. Larlus, F. Jurie, Latent mixture vocabularies for object categorization and segmentation, Image and Vision Computing 27 (5) (2009) 523–534.
[29] J. van de Weijer, C. Schmid, J.J. Verbeek, D. Larlus, Learning color names for real-world applications, IEEE Transactions on Image Processing 18 (7) (2009) 1512–1524.
[30] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: Proceedings of the Computer Vision and Pattern Recognition, 2007.
[31] S. Maji, A.C. Berg, J. Malik, Classification using intersection kernel support vector machines is efficient, in: Proceedings of the Computer Vision and Pattern Recognition, 2008.
[32] M. Marszalek, C. Schmid, H. Harzallah, J. van de Weijer, Learning object representation for visual object class recognition, in: Visual Recognition Challenge Workshop, in Conjuncture with ICCV, 2007.
[33] C.H. Lampert, M.B. Blaschko, T. Hofmann, Efficient subwindow search: a branch and bound framework for object localization, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (12) (2009) 2129–2142.
[34] F. Perronnin, J. Sánchez, Y. Liu, Large-scale image categorization with explicit data embedding, in: Proceedings of the Computer Vision and Pattern Recognition, 2010.
[35] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of the Computer Vision and Pattern Recognition, 2009.
[36] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the International Conference on Machine Learning, 2009.

[37] F.R. Bach, G.R.G. Lanckriet, M.I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, in: Proceedings of the International Conference on Machine Learning, 2004.
[38] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: ACM International Conference on Image and Video Retrieval, 2007.
[39] P.V. Gehler, S. Nowozin, Let the kernel figure it out: principled learning of pre-processing for kernel classifiers, in: Proceedings of the Computer Vision and Pattern Recognition, 2009.
[40] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, More efficiency in multiple kernel learning, in: Proceedings of the International Conference on Machine Learning, 2007.
[41] M. Varma, D. Ray, Learning the discriminative power invariance trade-off, in: Proceedings of the IEEE International Conference on Computer Vision, 2007.
[42] F.S. Khan, J. van de Weijer, M. Vanrell, Top-down color attention for object recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.
[43] Y.-L. Boureau, F. Bach, Y. LeCun, J. Ponce, Learning mid-level features for recognition, in: Proceedings of the Computer Vision and Pattern Recognition, 2010.
[44] N. Xie, H. Ling, W. Hu, X. Zhang, Use bin-ratio information for category and scene classification, in: Proceedings of the Computer Vision and Pattern Recognition, 2010.
[45] Z. Wang, Y. Hu, L.-T. Chia, Image-to-class distance metric learning for image classification, in: Proceedings of the European Conference on Computer Vision, 2010.
[46] H. Harzallah, F. Jurie, C. Schmid, Combining efficient object localization and image classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2009.
[47] R. Jenatton, J. Mairal, G. Obozinski, F. Bach, Proximal methods for sparse hierarchical dictionary learning, in: Proceedings of the International Conference on Machine Learning, 2010.
[48] M. Yang, L. Zhang, J. Yang, D. Zhang, Robust sparse coding for face recognition, in: Proceedings of the Computer Vision and Pattern Recognition, 2011.

Noha M. Elfiky did her M.Sc. in Computer Vision and Artificial Intelligence from UAB, Barcelona, Spain (2009). Before studying at UAB, she completed her Bachelor's degree in Computer Science and Information Systems from AinShams University, Cairo, Egypt. Since 2007 till October 2008, she worked as a teaching assistant on the German University in Cairo (GUC). She is based in the Computer Vision Center in Barcelona (Spain) since 2008.

Fahad Shahbaz Khan did his M.Sc. in Intelligent Systems Design from Chalmers University of Technology, Gothenburg, Sweden (2007) and he also obtained a Master's degree in Computer Vision and Artificial Intelligence from UAB, Barcelona, Spain (2008). Before studying at Chalmers, he completed his Bachelor's degree in Computer Science from BZU, Multan, Pakistan. He is based in the Computer Vision Center in Barcelona (Spain) since 2007.

Joost van de Weijer received his M.Sc. (Delft University of Technology) in 1998 and his Ph.D. (University of Amsterdam) in 2005. He was a Marie Curie Intra-European fellow in INRIA Rhone-Alpes. He was awarded the Ramon y Cajal Research Fellowships in Computer Science by the Spanish Ministry of Science and Technology. He is based in the Computer Vision Center in Barcelona (Spain) since 2008.

Jordi González completed his Ph.D. in 2004 at Universitat Autónoma de Barcelona (UAB). At present he is an Associate Professor in Computer Engineering and a researcher at a the Computer Vision Center. The topic of his research is the cognitive evaluation of human behaviors in image sequences, or video-hermeneutics. The aim is the generation of both linguistic and visual descriptions, which best explain those observed behaviors.

Please cite this article as: N. M. Elfiky, et al., Discriminative compact pyramids for object and scene recognition, Pattern Recognition (2011), doi:10.1016/j.patcog.2011.09.020