Coloring Bag-of-Words Based Image Representations

A dissertation submitted by **Fahad Shahbaz Khan** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor en Informàtica**.

Director	**Dr. Joost van de Weijer** Dep. Ciències de la Computació & Centre de Visió per Computador Universitat Autònoma de Barcelona
Co-director	**Dr. Maria Vanrell** Dep. Ciències de la Computació & Centre de Visió per Computador Universitat Autònoma de Barcelona
Thesis committee	**Dr. Gustavo Camps** University of Valencia **Dr. Theo Gevers** University of Amsterdam **Dr. Oriol Pujol** Universitat de Barcelona

This document was typeset by the author using LaTeX 2_ε.

The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

ISBN:

Printed by

Acknowledgements

The work done in this thesis could not be possible without the support and help of many individuals. I would take this opportunity to thank all those people who helped me during my 4 years in Barcelona. Specifically, I would like to single out following people for their wonderful support.

First and foremost I would like to thank Joost van de Weijer, my guru, mentor and advisor, for his utmost support, advice and invaluable guidance since my first arrival in Barcelona 4 years ago. This thesis would not have been possible without the encouragement, support and guidance of him. Whatever very little I know today about computer vision is owing to his support and dedication. I learnt how to do research, write papers and give talks from him. I really hope that some of his skills have rubbed off on me and will stay there forever. In short, I cannot thank him enough for his guidance and support.

I would also like to thank my Co-supervisor, Maria Vanrell for her great support and guidance during last 4 years. Words are not enough to thank her for supporting me during some of the very tough times in my life. Many a thanks to all the CVC personnel for their kind support during my 4 years stay. I have enjoyed collaborating with Pep Gonfaus, Marco Pedersoli, David Rojas, Xavi Bosch, Noha Elfiky, Jordi Gonzalez and Andrew Bagdanov. I thank them for teaching me different amazing things and helping me in building my knowledge about computer vision. Special thanks to Andrew Bagdanov for teaching me all the amazing things and tricks during our time for the PASCAL challenge competitions. His guidance and support played a pivotal role in the success of CVC in all those PASCAL competitions. My years in CVC and Barcelona have been pricesless due to all those amazing people.

My stay in Barcelona was made amazing and exciting by numerous friends and colleagues who I would like to thank for their support. Many a thanks to Shida, Javier, Sadiq, Paresh, Naveen, Bhaskar, Pep, Marco, David, Naila, Wenjuan, Noha, David, Ramon, Jordi, Eduard, Xavi, Alejandro , Robert, Bilal, Vikas and other CVC friends. I would like to pay gratitude to my best friend Anwer Rao for all his support during my tough times. Also thanks to zohaib, Ashfaq, Shahid, Shakeeb, Nauman and Kashif for their support and friendship.

Lastly, I would like to thank my family for their endless support and care. I would like to thank to the most important person in my life, my mother, for all her support, affection and love. Thanks alot for making me the person I am today. Losing you in 2010 was the most painful moment of my life. The immense support from my father and brother during these years means alot to me. Thanks to all my friends who supported me during that time.

Abstract

Put succinctly, the bag-of-words based image representation is the most successful approach for object and scene recognition. Within the bag-of-words framework the optimal fusion of multiple cues, such as shape, texture and color, still remains an active research domain. There exist two main approaches to combine color and shape information within the bag-of-words framework. The first approach called, early fusion, fuses color and shape at the feature level as a result of which a joint color-shape vocabulary is produced. The second approach, called late fusion, concatenates histogram representation of both color and shape, obtained independently.

In the first part of this thesis, we analyze the theoretical implications of both early and late feature fusion. We demonstrate that both these approaches are sub-optimal for a subset of object categories. Consequently, we propose a novel method for recognizing object categories when using multiple cues by separately processing the shape and color cues and combining them by modulating the shape features by category specific color attention. Color is used to compute bottom-up and top-down attention maps. Subsequently, the color attention maps are used to modulate the weights of the shape features. Shape features are given more weight in regions with higher attention and vice versa. The approach is tested on several benchmark object recognition data sets and the results clearly demonstrate the effectiveness of our proposed method.

In the second part of the thesis, we investigate the problem of obtaining compact spatial pyramid representations for object and scene recognition. Spatial pyramids have been successfully applied to incorporate spatial information into bag-of-words based image representation. However, a major drawback of spatial pyramids is that it leads to high dimensional image representations. We present a novel framework for obtaining compact pyramid representation. The approach reduces the size of a high dimensional pyramid representation upto an order of magnitude without any significant reduction in accuracy. Moreover, we also investigate the optimal combination of multiple features such as color and shape within the context of our compact pyramid representation.

Finally, we describe a novel technique to build discriminative visual words from multiple cues learned independently from training images. To this end, we use an information theoretic vocabulary compression technique to find discriminative combinations of visual cues and the resulting visual vocabulary is compact, has the cue binding property, and supports individual weighting of cues in the final image representation. The approach is tested on standard object recognition data sets. The results obtained clearly demonstrate the effectiveness of our approach.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Images play an integral part in our daily communication. The advent of new technologies together with widespread access to internet services have provided a dominant platform for photo sharing. Online photo sharing websites such as Flickr are awash in digital photos. This huge amount of pictures digitally available on the internet are difficult to manage manually. An automatic system for managing the photos would significantly reduce the labor. Automatic image concept classification is however a challenging task. To automatically retrieve an image, search engines typically make use of the text associated with it. This reliance on the metadata and associated text, while ignoring the semantics of an image, hampers the retrieval performance.

Contrary to modern search engines, humans have an outstanding ability of classifying images based on their visual content. When asked about the content of an image, a person can tell whether there is a car, a building or a zebra, etc., in a fraction of a second [60, 68]. Visual content based image classification is a long awaited goal of the computer vision community. Fig. 1.1 shows images of different object categories. Is there a train or a bottle in the top left image in Fig. 1.1? The problem of image classification deals with such queries. However, automatic image classification is a challenging task due to large variations between images belonging to the same category. Several factors such as significant fluctuations in viewpoint and scale, illumination, partial occlusions and multiple instances also have a significant influence on the final results and thus make the problem of description of images even more complicated [7, 14, 21].

Although color is an important visual cue in human perception, and also plays a paramount role in the visual search mechanism [24, 41, 51, 100], still many computer vision approaches to object recognition focus on shape features and ignore color [14, 21, 46]. To improve visual search, color should be incorporated to extract extra visual information in cases where shape is not the most discriminative cue. Subsequently, an object recognition system based on the combination of color and shape cues can be expected to improve recognition performance.

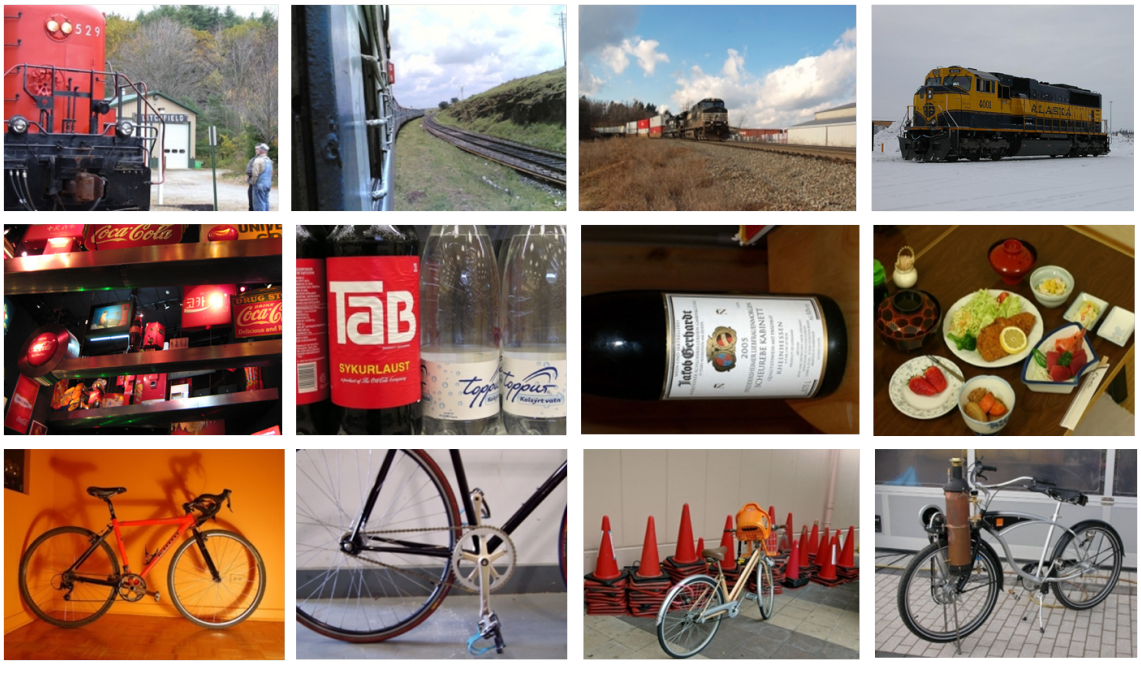


Figure 1.1: Example images of different object categories from the PASCAL VOC data set. Image classification is concerned with assigning one or multiple category labels to each image without localizing the object. In this thesis, we aim at improving the bag-of-words framework by combining color and shape cues for object recognition.

1.1 Bag-of-Words based Object Recognition

In recent years the bag-of-words based framework has been demonstrated to be one of the most successful approaches to object and scene recognition [7, 46, 72]. The first stage in the pipeline, feature detection, involves detecting keypoint regions in an image either by employing dense sampling or through interest point detection. The feature detection step is followed by representing the selected keypoint regions using local descriptors in the feature description stage. Afterwards, a visual vocabulary is constructed by quantizing the local descriptors into a fixed-size visual vocabulary. Finally, in the assignment stage, the image is represented by a histogram over the visual code-book. These histogram representations are then used to train a classifier to recognize different object categories. This relatively simple image representation was found to obtain superior results on image classification tasks even for difficult cases such as those shown in Fig. 1.1.

The success of the bag-of-words approach is dependent on various factors such as the quality of visual vocabularies, the combination of multiple features, efficient sampling strategies, the classification techniques, etc. In this thesis, we focus on three problems related to combining color and shape features within the bag-of-words framework.

Figure 1.2: Example images from the raspberry and foliage categories. Late fusion is better suited to classify the raspberry images since shape is constant and color changes significantly. To classify the foliage category, early fusion is expected to provide better performance.

Combining color and shape features: Several features such as shape, texture and color are used to describe the local patches. Shape is the most commonly used feature within the bag-of-words framework due to its high discriminative power [53]. There are several stages within the bag-of-words framework where color can be incorporated. Firstly, color can be added into the feature detection stage by focusing on the salient color regions in an image. Secondly, color can also be introduced into the feature extraction stage by describing the local patches using color information. Generally, the focus of incorporating color information in the bag-of-words framework is at the feature extraction stage.

Recent approaches used to combine color and shape information often provide below-expected results on a wide range of object categories [56]. The inferior results obtained might be attributed to the way color is incorporated. Traditionally there exist two approaches to combining color and shape features. The first approach, termed early fusion, combines color and shape features locally before the vocabulary construction stage. As a result of this a joint color-shape visual vocabulary is constructed. The second approach, called late fusion, combines the two visual cues after the vocabulary construction stage. In late fusion, separate visual vocabularies are constructed for color and shape and the two representations are then concatenated to construct the image representation.

To accommodate multiple cues within the bag-of-words framework, two properties are especially desirable for the final image representation: feature binding and feature compactness. Feature binding involves combining color and shape information at the local level and not at the image level. This property is essential for distinguishing images with red squares and green circles from images with green squares and red circles. Feature compactness involves having a separate visual vocabulary for each of the different features. The feature compactness property prevents the

different cues from becoming diluted, which happens in the case of early fusion where a combined shape-color vocabulary is constructed. Existing approaches used to combine color and shape information have not succeeded so far in combining both these properties in a single image representation.

Early fusion binds the color and shape features locally at the feature level. The binding of color and shape cues is also performed by the humans to select and combine the individual features in an accurate manner for classifying object categories [82]. This binding ability is especially desireable to correctly classify object categories that exhibit constancy over both color and shape features. Contrary to early fusion, late fusion does not possess the feature binding property. However, late fusion provides compact image representations since separate visual vocabularies are constructed for both color and shape cues. This compact feature representation is very useful for object categories where one of the two visual cues varies significantly. Fig 1.2 shows a two class problem of classifying raspberries and foliage. The foliage category will be better represented by early fusion as both color and shape are constant whereas late fusion will provide better recognition performance for raspberries category since shape is constant while color changes a lot.

Fig 1.2 also shows that both early and late fusion based approaches are suboptimal for a number of object categories. Since both approaches are suited for different sets of object categories, this is especially problematic for complex and challenging object recognition data sets that contains a variety of object categories possessing varying degrees of importance of color and shape cues. Therefore, finding an approach combining the strengths of the two fusion approaches for such challenging and complex data sets is expected to further improve recognition performance.

Spatial information within bag-of-words: Image representations obtained using the standard bag-of-words approach lacks spatial information. The spatial pyramid matching approach [47] provides a simple way of introducing spatial information within the bag-of-words framework. The technique works by dividing an image into increasingly finer sub-regions and constructing histograms for each region. This results in a multi-resolution histogram that captures the spatial layout of an object or scene. Although spatial pyramid schemes yield excellent performance, the resulting histogram has very high dimensionality. Fig 1.3 shows an image with a spatial pyramid scheme. The final dimensionality of the pyramid histogram depends on the size of visual vocabulary and increases by going deeper into the pyramid levels. This is especially problematic when combining spatial pyramids of multiple cues such as color and shape. Therefore, a compact pyramid representation is expected to allow combining multiple visual cues such as color and shape efficiently.

There exist several approaches [22, 45, 101] to compress the size of visual vocabularies. These approaches aim at reducing the size of the visual vocabulary at the standard bag-of-words level. However, compressing the size of spatial pyramid representations is still an open problem. A compact pyramid representation is expected to yield several advantages. Firstly, a compact pyramid representations will

Figure 1.3: An example image with spatial pyramid scheme of [47]. An image is divided into finer regions and a histogram is constructed for each region. Consequently, histograms from all the regions are concatenated into a single representation. The dimensionality of the final histogram is equal to the number of regions times the size of the visual vocabulary.

decrease the classification time and memory consumption. Secondly, reducing the histogram dimensionality will allow the incorporation of more features resulting in better recognition.

Multi-cue visual vocabularies: Conventionally, local color and shape features are concatenated within the bag-of-words framework to construct a single joint color-shape vocabulary. To obtain good performance, relative weighting of the two visual cues is performed since different object categories within a data set possess varied importance of color and shape features. The weights are learned through cross-validation on a validation set. Although early fusion based visual vocabularies do possess the feature binding property discussed above, this comes at an expensive cost of constructing visual vocabularies iteratively for a given set of weights.

The problem of constructing efficient visual vocabularies have been investigated in the past [36, 77, 84, 101]. There exist several approaches [22, 69] to introduce top-down information using the category labels to improve visual vocabularies. However, none of these approaches handle the problem of multi-cue visual vocabularies for a large number of object categories. Moreover, the approach of [69] scales with the number of object categories in the data set. This is especially cumbersome for data sets such as caltech-101, flower-102 and bird-200 etc. where there exists more than 100 object categories. In summary, a new approach to construct multi-cue vocabularies which does not scale with the number of object categories while allowing efficient weighting of the visual cues is highly desired for large object recognition data sets.

A simple way of ensuring feature binding is to construct a color-shape visual vocabulary that contains one visual word for each combination of original shape and color features. This may lead to a visual vocabulary of millions of visual words originating from a few initial color and shape visual words. Constructing such visual vocabularies is infeasible due to the difficulty of sampling from limited training data. Furthermore, with several parameters to tune the resulting classifier is subjected

to overfitting. These limitations prohibited further investigation in this direction. Recently, a number of visual vocabulary compression techniques have been proposed that derive compact and discriminative vocabularies from very large ones. The most successful methods are based on information theoretic clustering algorithms that are based on robust estimation of category-conditional visual word probabilities. The success of vocabulary compression techniques [13, 22, 78] allow us to reconsider the direct, Cartesian product approach to building multi-cue visual vocabularies which have the feature binding property while making it easy to weight the relative contribution of color and shape cues

1.2 Objectives and Approach

Above we discussed three aspects of combining color and shape cues within the bag-of-words framework. This analysis has led us to the following three objectives of the thesis research.

- **Combining feature binding and compactness:** The analysis of early and late feature fusion suggests that both approaches possess different properties, each suitable for only a subset of object categories. The two desired properties: feature binding and feature compactness, found in early and late fusion respectively, should be combined in a single image representation. Therefore a new image representation is required to counter the shortcomings of both early and late fusion. This prompts us to propose a new approach that exploits the advantages of both early and late fusion.

As mentioned above, the advantages of both early and late fusion should be combined in a single image representation. This motivates us to look into the human visual attention literature for an alternative approach of combining color and shape cues. Differently than most computer vision approaches [7, 10, 86, 87], the human visual system processes the basic visual features such as color and shape separately in a parallel way [81]. To obtain the binding of these visual features into a recognizable object, visual attention plays a pivotal role [81,103,104]. This attention mechanism is employed by the visual system to reduce the computational cost of visual search. The two distinct ways by which information can be used to direct attention are, bottom-up attention (memory-free), where the attention is directed rapidly to the salient and potentially important regions and, top-down attention (memory-dependent), which enables goal directed task demanded visual search [102].

The above analysis shows that, differently than state-of-the-art computer vision methods, the human visual system processes basic visual features such as color and shape separately. Subsequently, they are combined in the presence of visual attention. These observations inspire us to propose a new approach to solve the problem raised mentioned above. The proposed approach presented in chapter 3, modulating

shape features by color attention, processes color and shape cues separately. The two visual cues are then combined using bottom-up and top-down mechanisms of attention. The top-down information is introduced by using learned, class-specific color information. This color information is then used to construct category-specific color attention maps of the object categories. Subsequently, top-down color attention maps are used to modulate the weights of the bottom-up shape features. Finally, a class-specific color attention histogram is constructed for each category.

- **Discriminative compact spatial pyramid representations:** The conventional spatial pyramid scheme has been demonstrated to significantly improve the performance over standard bag-of-words approach. However, this performance gain is obtained at a high computational and memory cost due to the high dimensionality of spatial pyramids. Therefore, a compact pyramid representation should reduce the computational cost without deteriorating classification performance. Moreover, such a compact pyramid representation is also expected to allow the combination of multiple visual cues such as color and shape efficiently.

In chapter 4, the problem of constructing compact and discriminative spatial pyramids for object and scene recognition is investigated. Including spatial information using spatial pyramids has been shown to significantly improve recognition results over standard bag-of-words approach [47]. The spatial pyramid scheme works by dividing an image into increasingly finer regions. A histogram is then constructed for each region. Although spatial pyramids improve the results, the resulting histograms are of high dimensionality increasing the classification time and memory usage significantly. This problem is more apparent for difficult data sets such as the PASCAL VOC where the higher performance is achieved by using very large visual vocabulary. Furthermore, the dimensionality problem also prevents combing multiple visual cues due to high computational cost.

To make use of spatial pyramid more efficiently, a lower dimensional representation is highly desirable without significant loss of accuracy. Another advantage of compact spatial pyramid representation is that it allows the combination of visual cues without increasing the classification time. Therefore, an approach is proposed that preserves the overall accuracy while reducing the dimensionality of the pyramid histogram significantly and counters the problem posed above. A divisive information theoretic feature clustering algorithm [13] is used to construct compact pyramid representation. Moreover, an evaluation of combining color and shape cues at the spatial pyramid level is performed. The experiments clearly demonstrate the effectiveness of the proposed approach at a significantly lower computational cost.

- **Compact and discriminative multi-cue visual vocabularies:** Early fusion based visual vocabularies possess the feature binding property. These visual vocabularies are typically constructed by leveraging the contribution of

color and shape. Without weighting the color and shape cues, the vocabularies provide inferior classification results. Typically, the weighting parameter is learned through cross-validation. This is an extremely time consuming procedure. On the other hand, the late fusion approach allows efficient weighting of color and shape cues but lacks feature binding. Therefore, we aim at constructing compact multi-cue visual vocabularies that possess the feature binding property while allowing the weighting of different visual cues efficiently.

As mentioned earlier, early fusion has the property of feature binding. Late fusion lacks feature binding but allows efficient weighting of the visual cues to balance the contribution of shape and color features. Cue weighting in early fusion is problematic, implying that visual vocabularies and histogram construction must be performed for each weighting factor, thus making the cross-validation procedure extremely expensive.

A direct way to construct a multi-cue visual vocabulary is by having a visual word for each combination of original color and shape cues. Although such multi-cue vocabulary ensures feature binding nevertheless this can result in a vocabulary of millions of visual words. Moreover, limited amounts of training samples together with tuning of multiple parameters further make it infeasible to construct such vocabularies. Recent advances in information-theoretic clustering techniques [13,22, 78] permit us to revisit the problem of constructing multi-cue visual vocabularies. The information-theoretic clustering algorithms are based on robust estimation of class-conditional visual word probabilities.

To this end, an approach is presented in chapter 5 that constructs a multi-cue visual vocabulary. We show that for the task of image classification, modeling joint-cue distributions independently is more statistically robust than empirically estimating the dependent, joint-cue distribution directly. A divisive information theoretic feature clustering algorithm [13] is employed to construct a multi-cue visual vocabulary by compressing the cartesian product of primitive features. The resulting visual words are compact, have the feature binding property, and supports individual weighting of visual cues in the final image representation. Experiments demonstrate the effectiveness of the proposed approach.

Chapter 2

Bag-of-Words Based Object Recognition

Object recognition is the problem of determining whether an image contains an object instance or not. Typically a predefined list of object categories is provided and the task is to correctly assign a category label to an image. Visual categorization is a difficult task, interesting in its own right, due to large variations between images belonging to the same class. Several other constraints such as view point changes, variations in illumination, object residing in wide range of context also makes object recognition an extremely difficult task to accomplish. There exist a variety of object categories ranging from man-made object categories such as car, bus, boat, aeroplane, piano etc. to natural object categories such as plants, sheep, dolphins etc. Such diversity further increases the complexity of the problem.

Many approaches have been used in the past to tackle the problem of object recognition. The bag-of-words approach which represents an image as a histogram of local features is currently the most successful approach for object and scene recognition [7, 21, 46, 47]. The approach works by constructing a visual vocabulary of local features after which a histogram is built by counting the occurrences of each visual word in an image. The histogram is then used as an input to a classifier. A model is trained using a set of training images by projecting the histogram values into a space aiming to optimize the gap between examples of different object categories. Consequently, given a test image the model is used to predict the category label of the image.

In this chapter, we provide a detailed overview of each stage of the bag-of-words pipeline. The bag-of-words framework consists of two main parts namely, image representation and the machine learning. To obtain an image representation, the subsequent stages to follow are feature detection, feature extraction, vocabulary construction and assignment. We provide an overview of each of these stages within the bag-of-words framework. Furthermore, we also provide an overview of existing approaches used to combine multiple cues within this framework. Finally, we present

an overview of our PASCAL VOC 2009 challenge image classification submission which is based on bag-of-words framework. The main novelty in our submission is the introduction of a new approach to combine color and shape cues presented in this thesis.

2.1 Feature Detection

The first stage within the bag-of-words approach involves detecting keypoints or regions in an image. There exist multiple strategies for selecting regions in an image. These strategies can be divided into two broad categories namely: dense sampling and interest point sampling strategies. The dense sampling technique works by scanning the image with either single or multiple scales at fixed locations forming a grid of rectangular windows. Dense sampling scheme is often advantageous for scene classification since all regions in the image provide information for the recognition task.

The second class of sampling strategy employed to find regions is called interest point sampling. Interest point techniques rely on finding salient points (such as corners, blobs etc.) in an image. Interest point strategies are often helpful for object recognition task as they ignore the homogeneous areas and focus on the object and its surroundings in an image. Several interest point strategies have been proposed in the literature [58, 64, 90]. The Harris-Laplace point detector [58] focuses on locating corners that are scale invariant in an image. The Laplacian operater is used to find the scale of the corner. Other then finding corners in an image, there also exists blob like structures in an image. Laplacian-of-gaussian is a commonly used blob detector where an image is convolved using a gaussian kernel at certain scales to obtain a scale space representation. Most of the existing interest point schemes make use of shape saliency as a selection criteria for detection.

Among the color based interest point detectors proposed in the literature, color saliency boosting [90] is the most commonly used approach. The method exploits the saliency of color edges which is computed by applying information theory to the statistics of color image derivatives. The color boosting approach has been successfully applied for object recognition and retrieval tasks [80, 86]. Fig 2.1 shows example of different point sampling strategies. The dense sampling is covering the whole image whereas the interest point sampling strategies target salient regions of an image.

2.2 Feature Extraction

The next stage within the bag-of-words framework involves describing the extracted regions of an image. All the patches extracted in an image are normalized to a

Figure 2.1: Sampling strategies used for selecting regions in an image. The second image from the left showing a dense grid representation followed by two interest point sampling techniques (blob and color-boosted blob detection). Note that the color-boosted detector puts more emphasis on the red beak of the bird.

standard size and descriptors are computed for all regions. Many features such as color, texture, shape have been used to describe visual information for object recognition. In the next paragraphs, we provide an overview of the two most commonly used visual cues namely, shape and color. We will put more emphasis on the color descriptors which play an important role in this thesis.

2.2.1 Shape Feature Extraction

Most of the current approaches within the bag-of-words framework rely on extracting shape features predominantly SIFT [53] to represent an image [7,14,59,87]. The SIFT descriptor works on grey level images ignoring the color contents of an image. SIFT operates by computing gradients within a region of interest. The local appearance of the region is described by edge histograms. The region of interest is first divided into 4x4 grid of cells where each of the four quadrants have its own edge directional histogram computed from the local gradient direction weighed by the magnitude of the gradient. The SIFT descriptor is highly invariant to changes in scale,illumination, and orientation. It is also partially invariant to 3D viewpoint. Each SIFT keypoint has 132 dimensions where 128 are spatial orientation bins,plus the coordinates, rotation and the scale of the keypoint. Fig 2.2 shows computation of a SIFT descriptor based on the gradient and orientation of each image sample point in a region around the feature. The SIFT descriptor was found to outperform other descriptors in an evaluation performed by [59].

2.2.2 Color Feature Extraction

Color descriptors represent the color aspects of an image. The measured color values vary significantly due to large amount of illumination variations. Here we describe three popular color descriptors namely HUE, Color names and ColorSIFT, used extensively in this thesis.

HUE descriptor: The HUE descriptor [87] is based on the hue channel of the

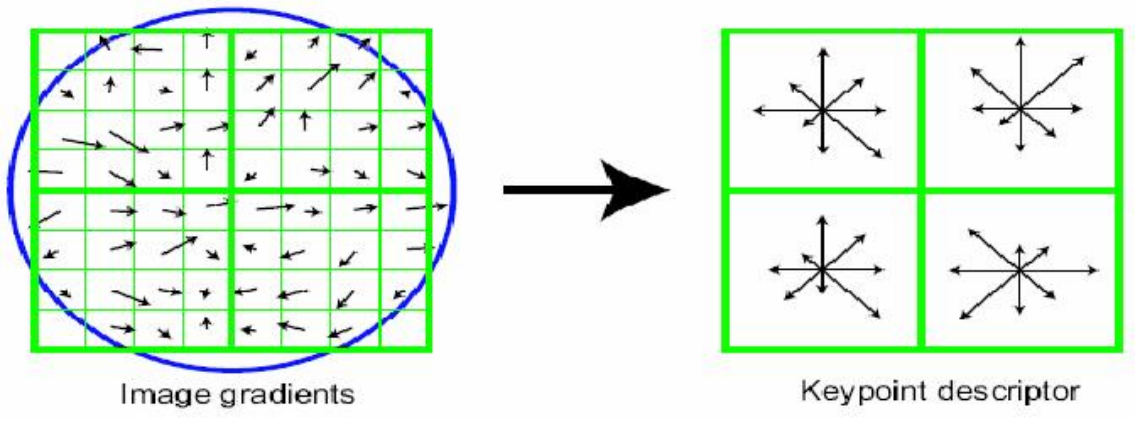


Figure 2.2: An example of SIFT computation. A region in an image is divided into four quadrants where each of the four quadrants contains 16 samples of the image gradient. The direction of the gradient together with magnitude samples are combined into a histogram of 8-bins gradient. Consequently, each of the four quadrants has its own histogram. The figure is taken from [53].

HSV color space. To obtain efficient local color histograms, [87] argues that the color descriptor should be robust to photometric changes such as shadows, shadings and variations in the light sources. Moreover, the descriptor should also be robust to geometric variations and handle photometric stabilities. These events are modeled by the well known Dichromatic Reflection Model [75].

$$\mathbf{f} = m^b \mathbf{C}^b + m^s \mathbf{C}^s \tag{2.1}$$

We use boldface to indicate vectors, for example, $\mathbf{f} = (R, G, B)$ and $\mathbf{C}^b = (C^b_R, C^b_G, C^b_B)$. The reflection of light comprises of two components namely, body reflection part ($\mathbf{C}^b$) and specular reflectance part ($\mathbf{C}^s$). Both terms are multiplied by a geometrical term, m^b and m^s, depending on the scene geometry (viewing and illumination direction, and objects orientation). The body reflectance describes the light reflected after interacting the surface albedo whereas the interface reflectance describes the light portion that is immediately reflected to the surface thereby causing specularities. The dichromatic reflection model can be used to derive the photometric invariance of color features.

The opponent colors can be computed from RGB by:

$$\begin{pmatrix} O1 \\ O2 \\ O3 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.2}$$

It can be shown that the opponent colors O1 and O2 are invariant with respect to specularities in case of white illuminant [27]. For example filling Eq.2.1 into O1

yields the following result.

$$O1 = m^b C_R^b + m^s - \left(m^b C_G^b + m^s\right) = m^b \left(C_R^b - C_G^b\right), \tag{2.3}$$

where we use the fact that white light $\mathbf{C}^s = \{1, 1, 1\}$. As can be seen $O1$ is independent of specularities because m^s drops out of the equation.

The hue equation is given by

$$hue = \arctan\left(\frac{O1}{O2}\right), \tag{2.4}$$

where O1 and O2 are the two opponent channels derived from the RGB space. By substituting the $O1$ and $O2$ using Eq. 2.1 in Eq. 2.4:

$$hue = \arctan\left(\frac{\sqrt{3}\left(C_R^b - C_G^b\right)}{\left(C_R^b + C_G^b - 2C_B^b\right)}\right) \tag{2.5}$$

which only depends on the body color $\mathbf{C}^b$ and is both invariant to m^b and m^s. Therefore, the hue is invariant for shadow and shading variations and specularities.

The hue is unstable around the grey axis. To counter this problem, an error propagation analysis has been applied by [28, 87] to the hue transformation. The error propagation analysis shows that the certainty of the hue is inversely proportional to the saturation. To counter the instability of hue around the grey axis, the hue samples are weighted by its saturation.

Fig 2.3 shows the resemblance between the computation of the SIFT and the HUE descriptor. The SIFT descriptor is based on the local gradients. The local patch is represented as a histogram over the direction of the gradients. Each gradient has a weight in the histogram which is equal to its length, which is the gradient strength. Similarly, the HUE descriptor looks at the chromaticity of each RGB value as is described by O1 and O2. This can be understood as a mapping of an RGB value to a vector. Next, a histogram of the direction of these vector is made. Again each vector has a weight in the histogram which is equal to its length, which in this case is the saturation of the RGB value. Other than SIFT only a single histogram is made to represent the patch.

In conclusion, the HUE descriptor provides a compact color description which is invariant to specularities (under white light assumption), and shadow-shading events.

Color names descriptor: Color names involve the assignment of linguistic color labels to image pixels. The 11 basic color terms of the English language are black, blue, brown, grey, green, orange, pink, purple, red, white and yellow [4]. Color names display a certain amount of photometric invariance because several shades of a color are mapped to the same color name. It also provides an added advantage of allowing the description of the achromatic colors such as black, grey, white etc.

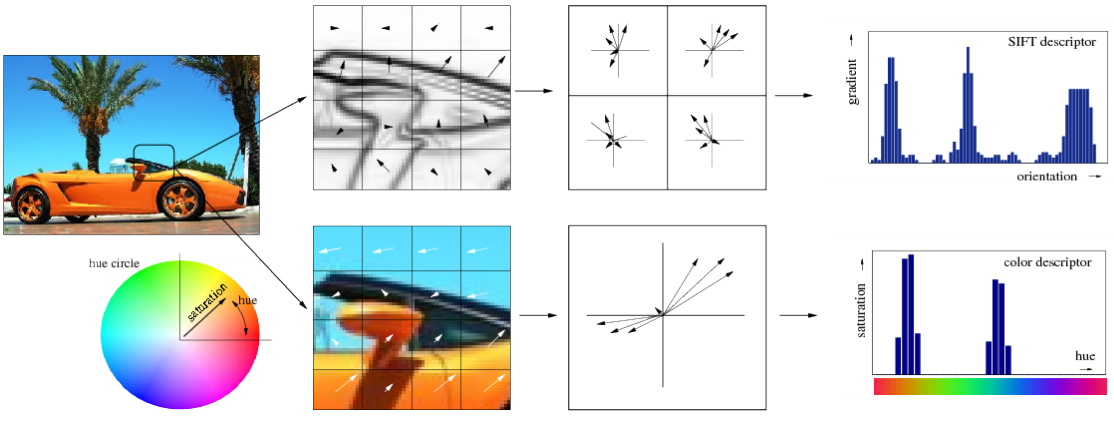


Figure 2.3: An example of Hue description. The top row shows the computation of SIFT descriptor and the bottom row shows the working of HUE descriptor. Note the similarity between the computation of SIFT and HUE.

which are impossible to distinguish from the photometric invariance perspective. The color name descriptor [88] CN is defined as a vector containing the probability of a color name given an image region R.

$$CN = \{p\left(cn_1|R\right), p\left(cn_2|R\right), ..., p\left(cn_{11}|R\right)\} \tag{2.6}$$

with

$$p\left(cn_i|R\right) = \frac{1}{P}\sum_{x\in R} p\left(cn_i|\mathbf{f}\left(x\right)\right) \tag{2.7}$$

where cn_i is the i-th color name, x are the spatial coordinates of the P pixels in region R, $\mathbf{f} = \{L^*, a^*, b^*\}$, and $p\left(cn_i|\mathbf{f}\right)$ is the probability of a color name given a pixel value. The probabilities $p\left(cn_i|\mathbf{f}\right)$ are computed from a set of images collected from Google. To learn color names, 100 images per color name are used. To counter the problem of noisy retrieved images, PLSA approach is used by [89]. Fig 2.4 shows an example image along with the color names description of the pixels.

In conclusion, color names possess some degree of photometric invariance. However, they also allow to encode the achromatic colors such as black, grey and white, leading to higher discriminative power.

ColorSIFT descriptor: The two above descriptors are pure color based. The features discussed here are combined color and shape descriptors. Recently, a performance evaluation of color descriptors has been performed by Van de Sande et al. [86]. The work aims at combining color information with SIFT descriptor. Among several ColorSIFT descriptors evaluated in their study, opponentSIFT is shown to provide superior performance for object recognition task.

The opponentSIFT is based on the opponent color space as shown in Eq. 2.2. The

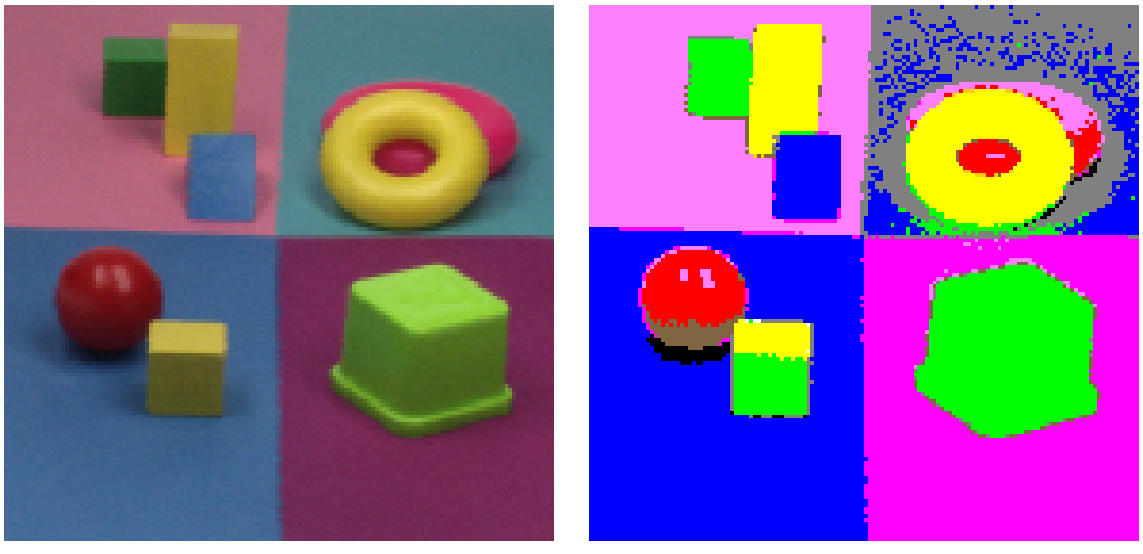


Figure 2.4: An example of color name description. For each pixel the best representative color name is assigned.

O3 channel describes the intensity information whereas the color information is represented in O1 and O2. In OpponentSIFT, SIFT is computed on the three opponent channels respectively. The resulting feature vectors are concatenated into a single representation. It is further shown in [86] that C-SIFT performs best on the PASCAL VOC 2007 data set. The C-SIFT descriptor is derived from the opponent color space as $\frac{O1}{O3}$ and $\frac{O2}{O3}$. Both C-SIFT and opponentSIFT descriptors are invariant to light intensity changes. Fig 2.5 shows an example image along with the three opponent channels.

2.2.3 Visual Vocabulary and Histogram Construction

Feature extraction is followed by visual vocabulary construction stage within the bag-of-words framework. Typically, a visual vocabulary is constructed using K-means algorithm. The algorithm is a simple iterative approach where the number of clusters are predefined. Initially the cluster centers are initialized by randomly selecting descriptor points. The distance is calculated for each sample point to the cluster centers and the point is assigned to the cluster having the closest center. After assigning all the points, the cluster centers are updated by averaging all the points in a cluster. The procedure is repeated for a fixed amount of iterations.

The quality of visual vocabulary depends on the size of the vocabulary. Generally improved results are obtained using larger visual vocabularies. Although larger visual vocabularies improve the performance, yet this improvement comes at the cost of high classification time. One popular strategy is to use information-theoretic clustering techniques [13, 22] to compress the visual vocabularies while maintaining the discriminative power of the original visual vocabulary. The compression of

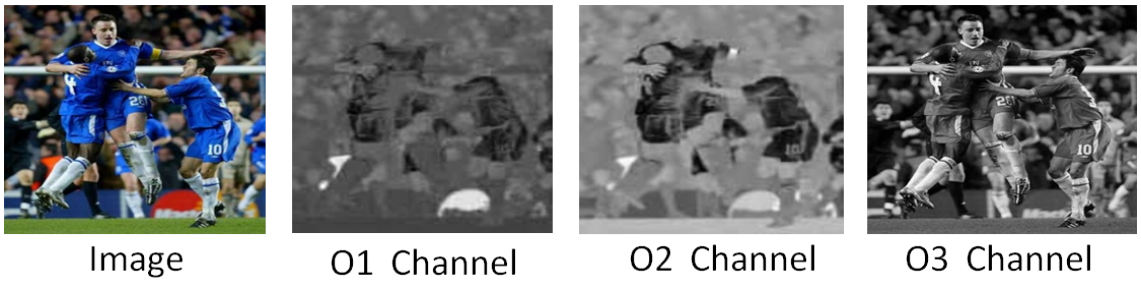


Figure 2.5: An example image with Opponent channel representations. In case of OpponentSIFT, SIFT is computed on each opponent channel respectively.

larger visual vocabularies also allow to incorporate more features to improve the classification performance as we will further investigate in this thesis.

After constructing a visual vocabulary, each descriptor is assigned to a single visual-word in the codebook. Consequently, a histogram is constructed by counting the number of occurrences of each visual-word in an image. Recently, it has been shown that the performance of the conventional histogram approach can be improved by using a soft-assignment through kernel vocabularies [33]. A kernel function is used for smoothing the conventional histogram assignment of image features to visual vocabulary.

2.2.4 Image Classification

The histogram constructed in the previous stage is then input to a machine learning algorithm for classification. The feature vectors and class labels are provided and supervised classification is performed. The goal is to learn a classifier that provides an estimation about previously unseen feature vector of being an instance of a particular class. Generally, classification is performed using support vector machines technique. Support Vector Machines work by finding the hyperplane in the feature space that can best separate the data points. The decision function of support vector machines classifier for a test image with feature vector $\mathbf{F}_t$ is:

$$g\left(\mathbf{F}_t\right) = \sum_{\mathbf{F} \in trainset} \alpha_{\mathbf{F}} cl_{\mathbf{F}} k\left(\mathbf{F}, \mathbf{F}_t\right) - \beta \tag{2.8}$$

where $cl_{\mathbf{F}}$ is the category label of $\mathbf{F}$, β is the threshold learned, $\alpha_{\mathbf{F}}$ is the weight learned from the training example $\mathbf{F}$ and $k\left(\mathbf{F}, \mathbf{F}_t\right)$ is the kernel function based on some distance metric. A variety of kernels for support vector machines have been proposed in literature [29, 113].

Two distance measures are especially useful to compare histograms. The histogram intersection kernel is based on computing the distance between the two

feature vectors as:

$$k\left(\mathbf{F}, \mathbf{F}_t\right) = \min\left(\mathbf{F}, \mathbf{F}_t\right) \tag{2.9}$$

where the minimum is taken for each bin in the histogram. Hence, a smaller value of k indicates that the two feature vectors are different and vice versa. The χ^2 kernel is based on the χ^2 distance computed between the two feature vectors as:

$$k\left(\mathbf{F}, \mathbf{F}_t\right) = e^{-\frac{1}{S} dist\chi^2(\mathbf{F}, \mathbf{F}_t)} \tag{2.10}$$

where S is a scalar normalizing the distance and commonly learned through cross-validation. Among several non-linear kernels, χ^2 kernel is shown to provide excellent performance for image classification task [86, 113]. In this thesis both intersection and χ^2 kernels are used.

2.3 Combining Color and Shape Features for Object Recognition

Generally, the local description is performed by extracting low-level appearance or texture features in an image. However, recently combining color and shape features have shown to provide excellent results on benchmark object recognition data sets [86]. There exists two main approaches to incorporate color information within the bag-of-words framework [73,79]. The first approach, early fusion, combines color and shape cues at the local feature level. This combination at the feature level results in constructing a joint color-shape visual vocabulary. A weight vector β is introduced to tune the relative weight of the color and shape in the combined vocabulary V_{sc} .

$$V_{sc} = (\beta\ V_c, (1-\beta)V_s) \tag{2.11}$$

where V_c are the color features and V_s are the shape features. The weight vector β is learned through cross-validation on the training data.

The second approach, late fusion, fuses color and shape information at the histogram level by concatenating the color and shape histograms obtained independently. Here the different vocabularies are concatenated after quantization. A weight vector α is introduced to obtain a combined histogram $\mathbf{F}(w|I)$ of color and shape vocabularies for an image I.

$$\mathbf{F}(w_{s\&c}|I) = \begin{bmatrix} \alpha\ \mathbf{F}(w_c|I) \\ (1-\alpha)\ \mathbf{F}(w_s|I) \end{bmatrix} \tag{2.12}$$

where w is the number of total vocabulary words, w_c are color words and w_s are shape words. The weight vector α is learned through cross-validation on training data.

The two approaches, early and late fusion, have their own advantages and drawbacks as well. Early fusion provides a more discriminative visual vocabulary since the color and shape words are constructed by quantizing the local color and shape cues combined at the feature level. This helps for recognizing object categories having consistent color and shape features which is commonly the case in many natural categories like plants and lions. On the other hand, late fusion provides a more compact representation of both color and shape as separate visual vocabularies are constructed for individual cues. This is especially important for man made categories such as cars and chairs vary considerably in color. To further elaborate the differences of early and late fusion consider a two class problem of recognizing the sun and balloon categories. Early fusion provides the best representation for the sun category as both the shape (round) and color (yellow) are constant for the category. Late fusion based image representation is problematic for this category because it looses the connection between shape and color. However, such a representation provides better description for the balloon category because only shape is constant and color varies significantly.

2.4 PASCAL VOC 2009 Image Classification Submission

To give an insight into state-of-the-art image classification system, we discuss here the PASCAL VOC 2009 image classification competition. The abovementioned bag-of-words approach has been employed in our PASCAL VOC 2009 image classification competition submission. The PASCAL VOC Challenge 2009 data set consists of 13704 images of 20 different classes with 7054 training images and 6650 test images as shown in Fig 1.1. The test set ground-truth is not available for this data set and the results are submitted to the organizers directly. For this data set the average precision is used as a performance metric in order to determine the accuracy of recognition results. The average precision is proportional to the area under a precision-recall curve. The average precisions of the individual classes are used to get a *mean average precision* (MAP).

The whole pipeline used for our submission is shown in Fig 2.6. In the feature detection step, we use Harris Laplace [58], Color Boosted HarisLaplace, Dense Multiscale Grid, Blob, and Color Boosted Blob detectors. For feature extraction stage, SIFT [53] , Hue [87], Color names [88], Color-SIFT [86], GIST [65] have been used. Spatial information is captured using spatial pyramid histograms [47] by dividing the image into 2×2 (image quarters) and 1×3 (horizontal bars) subdivisions. We compressed the visual vocabularies using the agglomerative information bottleneck approach [22]. The main novelty in the pipeline is the introduction of color attention for combining color and shape information. Finally, the classification scores are combined with object localization scores obtained through HOG pyramids [67] and ESS detector using [30]. Our submission of combined classification and object

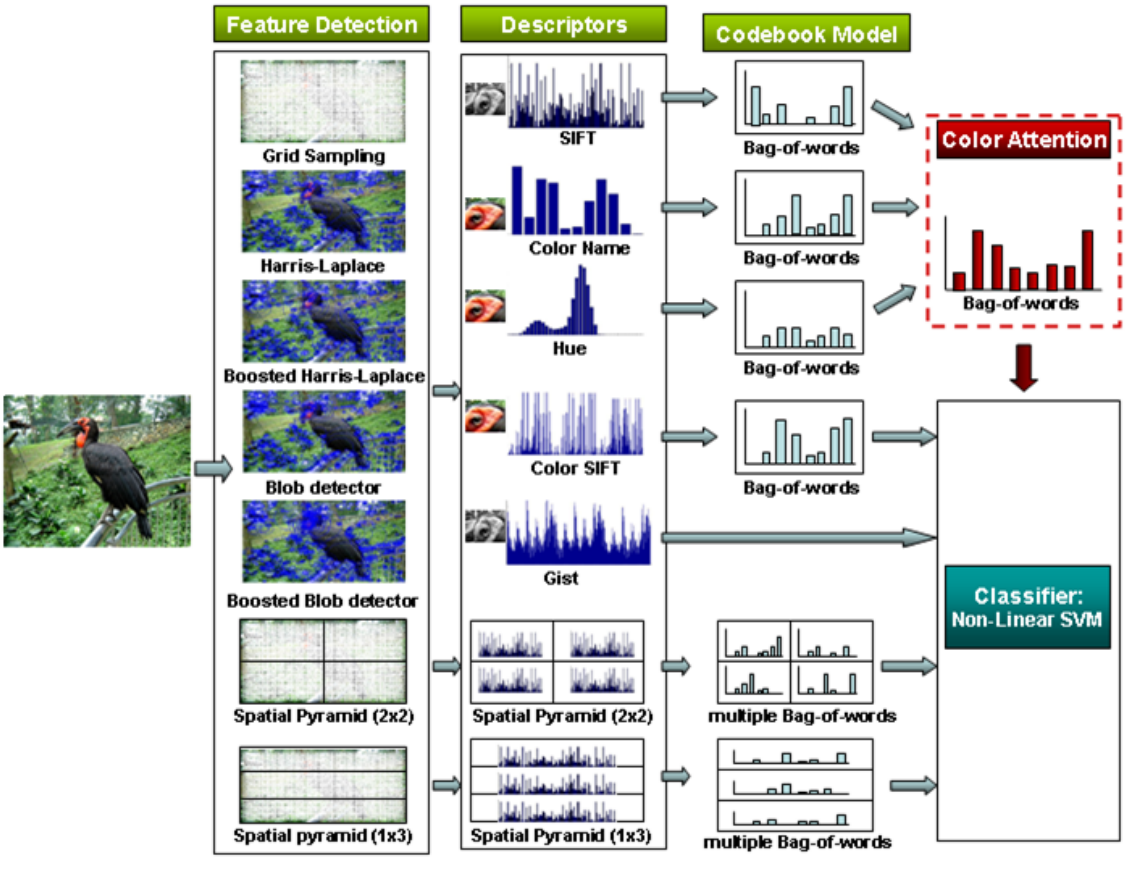


Figure 2.6: An overview of our pipeline used for the VOC 2009 image classification challenge. The main novelty in our whole pipeline is the introduction of color attention proposed in chapter 3 of this thesis.

localization results obtained best results on pottedplant and tvmonitor category [1].

Fig 2.7 shows per category results obtained by the top 3 submissions. The submission from NEC obtained best scores in 18 out of 20 categories. The NEC submission aims at improving the coding scheme within the bag-of-words framework and does not contain any color information. The submission from UVA is based the ColorSIFT descriptors presented in [86] and is more proximal to our approach which also aims at exploiting color information.

2.5 Conclusions

In this chapter, we have provided an overview of the bag-of-words approach for object and scene recognition. The stages within bag-of-words framework namely, feature detection, feature extraction, vocabulary construction and assignment have been

[1]For detailed results: http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/results/

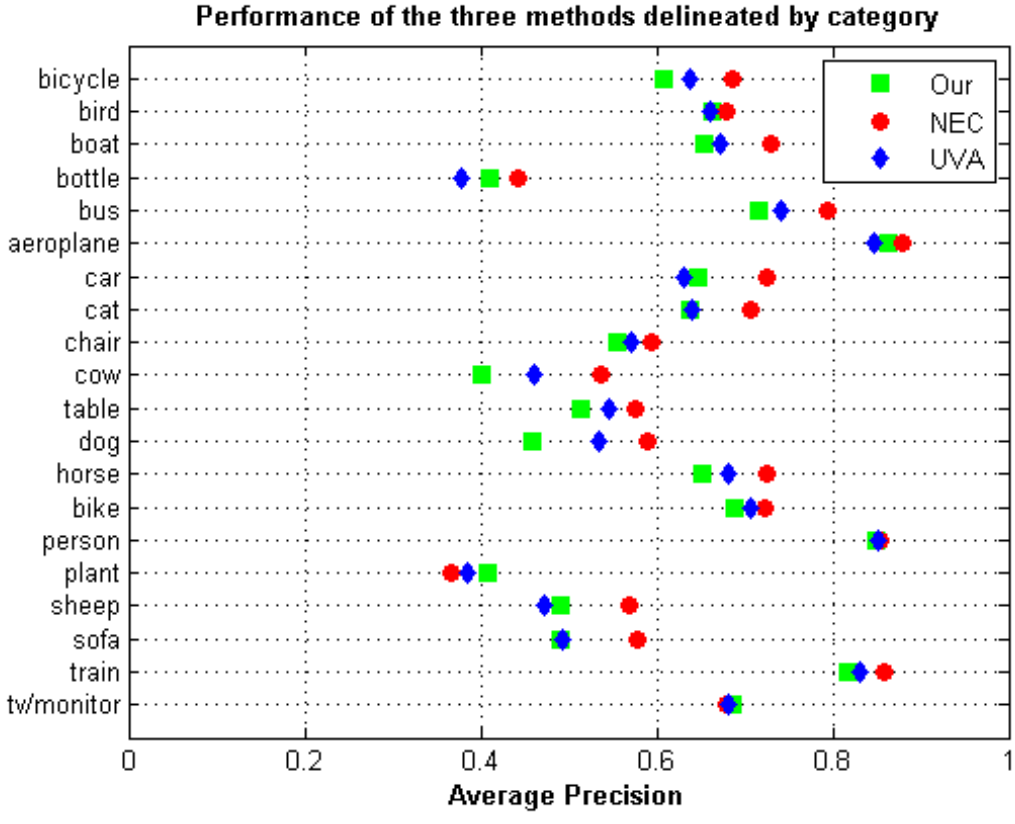


Figure 2.7: Results per category on PASCAL VOC 2009 data set. Only top 3 submissions are shown here. Note that our approach obtains best results on pottedplant and tvmonitor categories..

discussed. The discussion is followed by an overview of the existing approaches, early and late feature fusion, used to combine color and shape cues. Finally, we provide a brief overview of our submission to PASCAL VOC 2009 image classification challenge. In the following chapters, we will investigate several aspects of introducing color information into the bag-of-words framework.

Chapter 3

Modulating Shape Features by Color Attention for Object Recognition[1]

In this chapter we present a novel method for recognizing object categories when using multiple cues by separately processing the shape and color cues and combining them by modulating the shape features by category-specific color attention. Color is used to compute bottom-up and top-down attention maps. Subsequently, these color attention maps are used to modulate the weights of the shape features. In regions with higher attention shape features are given more weight than in regions with low attention.

We compare our approach with existing methods that combine color and shape cues on five data sets containing varied importance of both cues, namely, Soccer (color predominance), Flower (color and shape parity), PASCAL VOC 2007 and 2009 (shape predominance) and Caltech-101 (color co-interference). The experiments clearly demonstrate that in all five data sets our proposed framework significantly outperforms existing methods for combining color and shape information.

3.1 Introduction

Object category recognition is one of the fundamental problems in computer vision. In recent years several effective techniques for recognizing object categories from real-

[1]Accepted for publication by the International Journal of Computer Vision [40]. Part of this chapter appeared in ICCV 2009 [39].

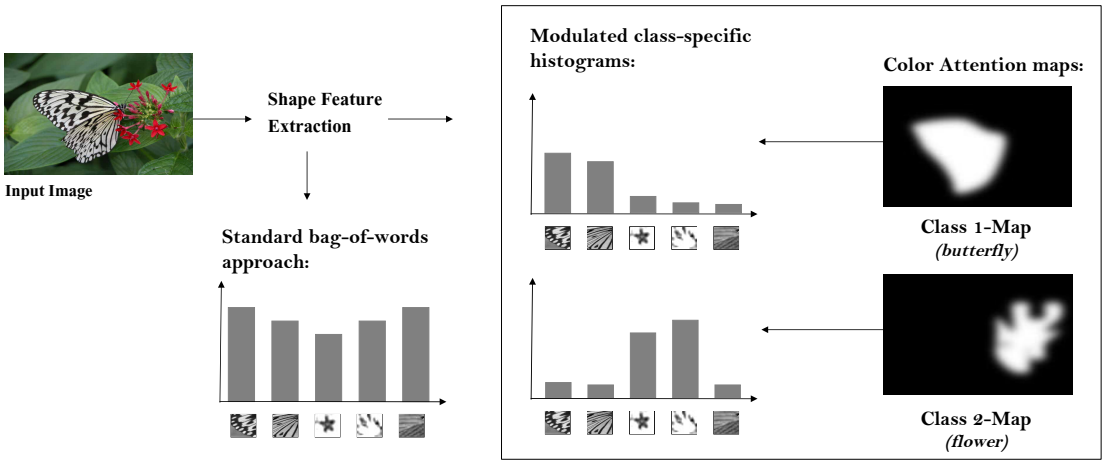


Figure 3.1: Top-down control of visual attention based on color. In standard bag-of-words the image representation, here as distribution over visual shape words, is constructed in a bottom-up fashion. In our approach we use top-down class-specific color attention to modulate the impact of the shape-words in the image on the histogram construction. Consequently, a separate histogram is constructed for the all categories, where the visual words relevant to each category (in this case flowers and butterflies) are accentuated.

world images have been proposed. The bag-of-features framework, where images are represented by a histogram over visual words, is currently one of the most successful approaches to object and scene recognition. Many features such as color, texture, shape, and motion have been used to describe visual information for object recognition. Within the bag-of-words framework the optimal fusion of multiple cues, such as shape, texture and color, still remains an active research domain [10, 26, 86]. Therefore in this chapter, we analyze the problem of object recognition within the bag-of-words framework using multiple cues, in particular, combining shape and color information.

There exist two main approaches to incorporate color information within the bag-of-words framework [73, 79]. The first approach called, *early fusion*, fuses color and shape at the feature level as a result of which a joint color-shape vocabulary is produced. The second approach, called *late fusion*, concatenates histogram representation of both color and shape, obtained independently. Early fusion provides a more discriminative visual vocabulary, but might deteriorate for classes which vary significantly over one of the visual cues. For example, man-made categories such as cars and chairs vary considerably in color. On the other hand, late fusion is expected to perform better for such classes, since it provides a more compact representation of both color and shape as separate visual vocabularies are constructed for individual cues. This prevents the different cues from getting diluted, which happens in case of a combined shape-color vocabulary. However, for classes which are characterized by both cues the visual vocabulary of late fusion will not be optimal. Such classes

include natural categories like cats and trees which are better represented by early fusion based schemes.

Combining color and shape within the bag-of-words, using an early fusion based approach, has recently shown to provide excellent results on standard object recognition data sets [20, 86]. [1] propose to compute the SIFT descriptor in the HSV color space and concatenate the results into one combined color-shape descriptor. Photometrically invariant histograms are combined with SIFT for image classification by [87]. A study into the photometric properties of many color descriptors and an extensive performance evaluation is performed by [85, 86]. In summary, most successful approaches [1, 86, 87] proposed to combine color and shape features are based on early fusion scheme. As discussed before these early fusion methods are all expected to be suboptimal for classes where one of the cues varies significantly, like in the case of man-made objects.

This observation inspires us to propose a new image representation which combines multiple features within the bag-of-words framework. Our approach, *modulating shape features by color attention*, processes color and shape separately and combines them by means of bottom-up and top-down modulation of attention [2] as shown in Fig. 3.1. The top-down information is introduced by using learned class-specific color information to construct category-specific color attention maps of the categories. In Fig. 3.1 two color attention maps are visualized for the butterflies and flowers categories. Subsequently, this top-down color attention maps are used to modulate the weights of the bottom-up shape features. In regions with higher attention shape features are given more weight than in regions with low attention. As a result a class-specific image histogram is constructed for each category. We shall analyze the theoretical implications of our method and compare it to early and late fusion schemes used for combining color and shape features. Experiments will be conducted on standard object recognition data sets to evaluate the performance of our proposed method.

The chapter is organized as follows. In Section 3.2 we discuss related work. In Section 3.3 the two existing approaches namely, early and late fusion, are discussed. Our approach is outlined based on an analysis of the relative merits of early and late fusion techniques in Section 3.4. Section 3.5 starts with an introduction to our experimental setup followed by data sets used for our experiments and finally experimental results are given. Section 3.6 finishes with concluding remarks.

3.2 Related Work

There has been a large amount of success in using the bag-of-visual-words framework for object and scene classification [7, 14, 21, 46, 59, 72, 87] due to its simplicity and

[2] Throughout this chapter we consider information which is dependent on the category-label as top-down, and information which is not as bottom-up.

very good performance. The first stage in the method involves selecting keypoints or regions followed by representation of these keypoints using local descriptors. The descriptors are then vector quantized into a fixed-size vocabulary. Finally, the image is represented by a histogram over the visual code-book. A classifier is then trained to recognize the categories based on these histogram representations of the images.

Initially, many methods only used the shape features, predominantly SIFT [53] to represent an image [14, 21, 46]. However, more recently the possibility of adding color information has been investigated [7,10,86,87]. Previously, both early and late fusion schemes have been evaluated for image classification [73]. The comparison performed in recent studies suggest that combining multiple cues usually improves final classification results. However, within the bag-of-words framework the optimal fusion of different cues, such as shape, texture and color, still remains open to debate.

Several approaches have been proposed recently to combine multiple features at the kernel level. Among these approaches, multiple kernel learning, MKL, is the most well-known approach and significant amount of research has been done to exploit kernel combinations carrying different visual features [3, 6, 74, 91, 92]. Other than MKL, averaging and multiplying are the two straight-forward and earliest approaches to combine different kernel responses in a deterministic way. Surprisingly, in a recent study performed by [26] it has been shown that in some cases the product of different kernel responses provide similar or even better results than MKL. It is noteworthy to mention that our approach is essentially different from MKL because it proposes a new image representation. Like early and late fusion it can further be used as an input to an MKL.

Introducing top-down information into earlier stages of the bag-of-words approach has been pursued in various previous works as well, especially in the vocabulary construction phase. [45] propose to learn discriminative visual vocabularies, which are optimized to separate the class labels. [69] proposes to learn class-specific vocabularies. The image is represented by one universal vocabulary and one adaptation of the universal vocabulary for each of the classes. Both methods showed to improve bag-of-words representations, but they do not handle the issue of multiple cues, and for this reason could be used in complement with the approach presented here. [95] semantically label local features into a number of semantic concepts for the task of scene classification. [110] propose an optimization method to unify the visual vocabulary construction with classifier training phase. [22] propose a method to generate compact visual vocabularies based on agglomerative information bottleneck principle. This method defines the discriminative power of a visual vocabulary as the mutual information between a visual word and a category label.

There have been several approaches proposed in recent years to learn an efficient visual codebook for image classification and retrieval tasks. [77] propose an approach to object matching in videos by using inverted file system and document ranking. [101] propose a method for image categorization by learning appearance-based object models from training images. A large vocabulary is compressed into

a compact visual vocabulary by learning a pairwise merging of visual-words. [36] argue that visual vocabulary based on standard k-means algorithm on densely sampled patches provides inferior performance and propose an acceptance-radius based clustering approach for recognition and detection. [84] propose a data independent approach to construct a visual vocabulary by discritizing the feature space using a regular lattice for image classification. [11] propose an approach for estimating codebook weights especially in scenarios when there are insufficient training samples to construct a large size visual codebook. The abovementioned approaches mainly aim at improving the visual codebook construction stage, whereas the novelty of our proposed method is that we use feature weighting as a mechanism to bind color and shape visual cues.

Humans have an outstanding ability to perform various kinds of visual search tasks constantly. But how is it that the human visual system does this job with little effort and can recognize a large number of object categories with such an apparent ease? Research on the human vision system suggests that basic visual features such as shape and color are processed in parallel, and are not combined in an early fusion manner. For example, in the two-stage architecture of the well known *Feature Integration Theory* by [81], the processing of basic features in an initially parallel way is done in the first stage, also known as the "preattentive stage". These basic visual features processed separately are loosely bundled into objects before they are binded into a recognizable object [103, 104]. It is further asserted that the basic features are initially represented separately before they are integrated at a later stage in the presence of attention. Similarly, we propose a framework where color and shape are processed separately. Other than late fusion, where histograms of individual features are concatenated after processing, we propose to combine color and shape by separately processing both visual cues and then modulating the shape features using color as an attention cue.

Several computational models of visual attention have been proposed previously. The work of [83] uses top-down attention and local winner-take-all networks for tuning model neurons at the attended locations. [32] propose a model for bottom-up selective visual attention. The visual attention mechanism has been based on serial scanning of a saliency map computed from local feature contrasts. The saliency map computed is a two-dimensional topographic representation of conspicuity or saliency for every pixel in the image. The work was further extended by [96] from salient location to salient region-based selection. [57] propose a coherent computational approach to the modeling of bottom-up visual attention where contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions are some of the features implemented in the model. [71] introduce a spatial attention model that can be applied to both static and dynamic image sequences with interactive tasks. [23] propose a top-down visual saliency framework that is intrinsically connected to the recognition problem and closely resembles to various classical principles for the organization of perceptual systems. The method aims at two fundamental problems in discriminant saliency, feature selection and saliency detection. In summary, the visual attention phenomenon has been well studied in

the fields of psychology and neuroscience but still has not been investigated within the bag-of-words framework for combining multiple visual cues.

3.3 Early and Late Feature Fusion

In this section, we analyze the two well-known approaches to incorporate multiple cues within the bag-of-words framework, namely early and late-fusion.

Before discussing early and late fusion in more detail, we introduce some mathematical notations. In the bag-of-words framework a number of local features f_{ij}, j=1...M^i are detected in each image I^i, i=1,2,...,N , where M^i is the total number of features in image i. Examples of commonly used detectors are multi-scale grid sampling and interest point detectors such as Laplace and Harris corner detector. Generally, the local features are represented in visual vocabularies which describe various image cues such as shape, texture, and color. We focus here on shape and color but the theory can easily be extended to include other cues. We assume that visual vocabularies for the cues are available, $\mathrm{W^k} = \{\mathrm{w_1^k}, ..., \mathrm{w_{V^k}^k}\}$, with the visual words $\mathrm{w_n^k}$, n=1,2,...,V^k and $k \in \{s, c, sc\}$ for the two separate cues shape and color and for the combined visual vocabulary of color and shape. The local features f_{ij} are quantized differently for the two approaches: by a pair of visual words $(\mathrm{w_{ij}^s}, \mathrm{w_{ij}^c})$ for late fusion and by single shape-color word $\mathrm{w_{ij}^{sc}}$ in the case of early fusion. Thus, $\mathrm{w_{ij}}^{\mathrm{k}} \in \mathrm{W^k}$ is the j^{th} quantized feature of the i^{th} image for a visual cue k.

For a standard single-cue bag-of-words, images are represented by a frequency distribution over the visual words:

$$h\left(\mathrm{w_n^k}|I^i\right) \propto \sum_{j=1}^{M^i} \delta\left(\mathrm{w_{ij}^k}, \mathrm{w_n^k}\right) \tag{3.1}$$

with

$$\delta\left(x, y\right) = \begin{cases} 0 & \text{for } x \neq y \\ 1 & \text{for } x = y \end{cases} \tag{3.2}$$

For early fusion, thus called because the cues are combined before vocabulary construction, we compute histogram $h\left(\mathrm{w^{sc}}|I^i\right)$. For late fusion we compute histograms $h\left(\mathrm{w^s}|I^i\right)$ and $h\left(\mathrm{w^c}|I^i\right)$ and concatenate the distributions. It is important to introduce a parameter balancing the relative weight between the different cues. For the results of early and late fusion reported in this work we learn the parameter by means of cross-validation on the validation set.

Late and early fusion methods lead to different image representations and therefore favor different object categories. To better understand their strengths we perform an experiment on the PASCAL VOC 2007 data set which contains a wide variety of categories. Both early and late fusion results are obtained using SIFT and Color Names descriptors. The results are presented in Fig 3.2. The axis shows the difference between the average precision (AP) scores of early and late fusion

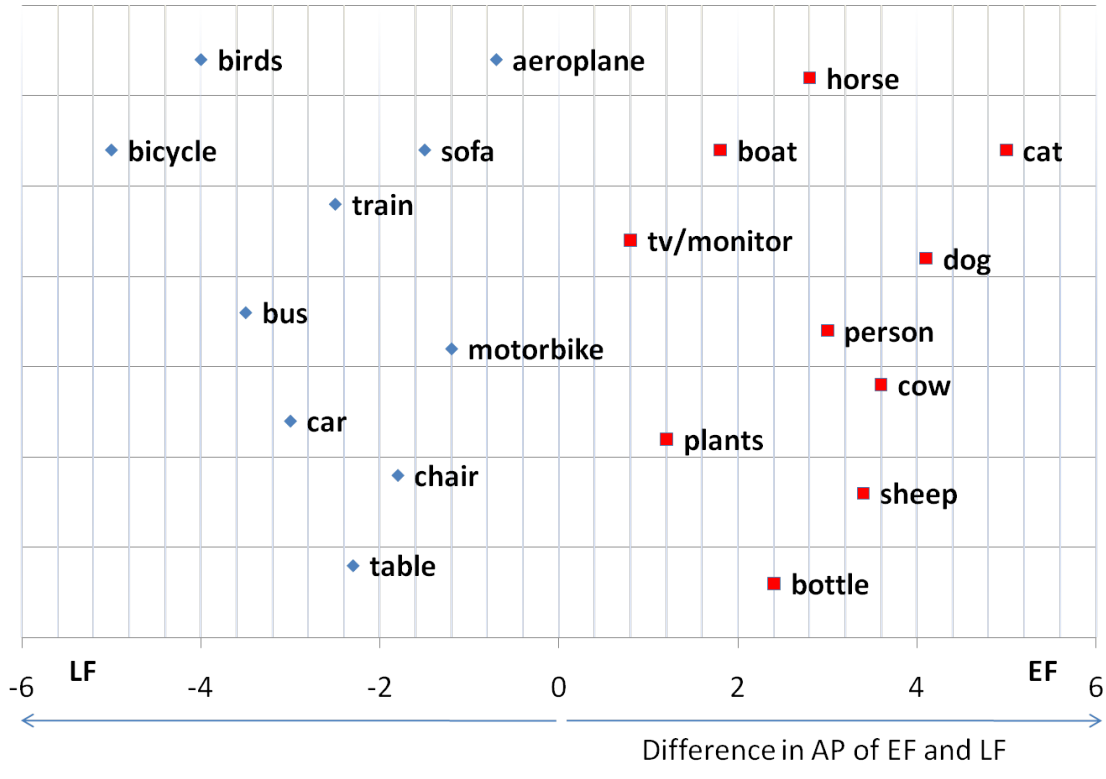


Figure 3.2: Difference in average precision (AP) scores of early and late fusion schemes for the 20 categories of PASCAL VOC 2007 data set. Vertical axis does not contain information. Half of the categories are better represented by early fusion (red) and half by late fusion(blue).

schemes (e.g. bicycle has a 5% higher score when represented by late fusion than by early fusion, and for airplane both representation yield similar results). The results clearly show that neither of the two fusion approaches perform well for all object categories.

Most man-made categories namely, bicycle, train, car and buses performs better with late fusion over its early fusion counterpart. The only exception in this case is the boat category which is better represented by early fusion. On the other hand, natural categories such as cow, sheep, dog, cat, horse etc. are better represented by early fusion. The bird category is the only outlier among natural categories which provides superior performance with late fusion instead of early fusion. Better than the distinction between man-made and natural categories is the distinction between color-shape dependency and color-shape independency of categories. This explains the location of most of the categories along the axes, including the birds class which is represented by a large variety of bird species with widely divergent colors, and the boat class which contains mainly white boats. The difference in strength of early and late fusion on different object categories is illustrated in Fig 3.3.

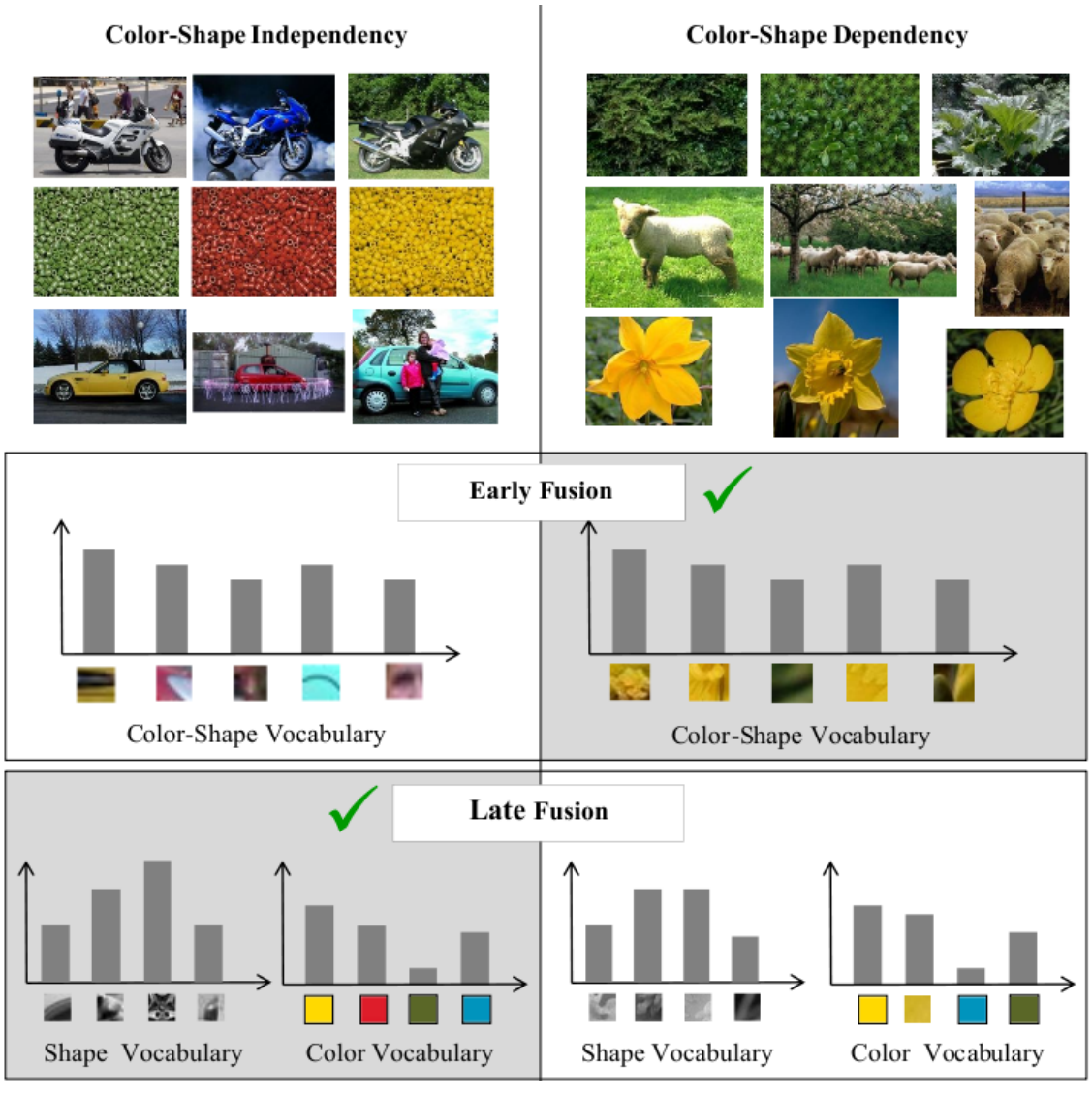


Figure 3.3: Graphical explanation of early and late fusion approaches. Note that for some classes early fusion scheme performs better where as for some categories, late fusion outperforms early fusion methods.

Based on the above analysis of early and late fusion we conclude that, to combine multiple cues, two properties are especially desired. The first property is *feature compactness*. Having this property implies constructing a separate visual vocabulary for both color and shape. This is especially important for classes which have color-shape independency. Learning these classes from a combined shape-color vocabulary only complicates the task of the classifier. Late fusion possesses the property of feature compactness, whereas early fusion lacks it. The second property is *feature binding*. This property refers to methods which combine color and shape information at the local feature level (as desired for categories with color-shape dependency). This allows for the description of blue corners, red blobs, etc. Early fusion has this

property since it describes the joined shape-color feature for each local feature. Late fusion, which separates the two cues, only to combine them again at an image-wide level, lacks this property.

3.4 Color Attention for Object Recognition

In the previous section we elaborated two approaches to combine color and shape features. In this section, we propose an attention-based image representation. Feature binding and feature compactness will be achieved by modulating shape features with bottom-up and top-down components of color attention.

3.4.1 Attention-based Bag-of-Words

We define a generalization of the bag-of-words as given by Eq. 3.3, called *attention-based bag-of-words*:

$$h\left(\mathrm{w}_n^{\mathrm{k}}|I^i\right) \propto \sum_{j=1}^{M^i} a_{ij}\delta\left(\mathrm{w}_{\mathrm{ij}}^{\mathrm{k}}, \mathrm{w}_{\mathrm{n}}^{\mathrm{k}}\right), \tag{3.3}$$

where a_{ij} are the attention-weights which modulate feature $\mathrm{w}_{\mathrm{ij}}^{\mathrm{k}}$. Choosing the a_{ij} weights to be equal to one reduces the equation to standard bag-of-words. The weights can be interpreted as attention maps, essentially determining which features w^{k} are relevant.

Next, we apply attention-based bag-of-words to combine color and shape. For this purpose we separate the functionality of the two visual cues. The shape cue will function as *descriptor cue*, and is used similar as in the traditional bag-of-words. The color cue is used as an *attention cue*, and determines the impact of the local features on the image representation. To obtain our image representation, color attention is used to modulate the shape features according to:

$$h\left(\mathrm{w}_n^s|I^i, class\right) \propto \sum_{j=1}^{M^i} a\left(\mathbf{x}_{ij}, class\right)\delta\left(\mathrm{w}_{\mathrm{ij}}^{\mathrm{s}}, \mathrm{w}_{\mathrm{n}}^{\mathrm{s}}\right), \tag{3.4}$$

where $a\left(\mathbf{x}_{ij}, class\right)$ denotes the color attention of the j^{th} local feature of the i^{th} image and is dependent on both the location $\mathbf{x}_{ij}$ and the *class*. The difference to standard bag-of-words is that in regions with high attention, shape-features are given more weight than in regions with low attention. This is illustrated in the two attention-based bag-of-words histograms in Fig. 3.1 where the attention map of the butterfly results in a bag-of-words representation with an increased count for the visual words relevant to butterfly (and similarly for the flower representation). Note that all histograms are based on the same set of detected shape features and only the weighting varies for each *class*. As a consequence a different distribution over the same shape words is obtained for each *class*.

Similarly as for human vision we distinguish between bottom-up and top-down attention:

$$a\left(\mathbf{x}_{ij}, class\right) = a_b\left(\mathbf{x}_{ij}\right) a_t\left(\mathbf{x}_{ij}, class\right). \tag{3.5}$$

Here $a_b\left(\mathbf{x}_{ij}\right)$ is the bottom-up color attention based on the image statistics and highlights the most salient color locations in an image. The top-down color attention is represented by $a_t\left(\mathbf{x}_{ij}, class\right)$, describing our prior knowledge about the color appearance of the categories we are looking for. The two components will be discussed in detail later.

Two parameters are introduced to tune the relative contribution of the two attention components:

$$a\left(\mathbf{x}_{ij}, class\right) = \left(a_b\left(\mathbf{x}_{ij}\right)^{(1-\beta)} a_t\left(\mathbf{x}_{ij}, class\right)^{\beta}\right)^{\gamma}. \tag{3.6}$$

The parameter, γ, is used to control the influence of color versus shape information. For $\gamma = 0$ we obtain a standard bag-of-words based image representation where a higher value of γ denotes more influence of color attention. The second parameter, β, is employed to vary the contribution of top-down and bottom-up attention, where $\beta = 0$ indicates only bottom-up attention and $\beta = 1$ means only top-down attention. Both γ and β parameters are learned through cross-validation over the validation set.

The image representation proposed in Eq. 3.4 does not explicitly code the color information. However, indirectly color information is hidden in these representations since the shape-words are weighted by the probability of the category given the corresponding color-word. Some color information is expected to be lost in the process, however the information most relevant to the task of classification is expected to be preserved. Furthermore, our image representation does combine the two properties *feature binding* and *feature compactness*. Firstly, *feature compactness* is achieved since we construct separate visual vocabularies for both color and shape cues. Secondly, *feature binding* is achieved by the top-down modulation as follows from Eq. 3.4. Consequently, we expect to obtain better results by combining both these properties into a single image representation.

The attention framework as presented in Eq. 3.3 recalls earlier work on the feature weighting techniques [99]. Replacing $a_{ij} = a_n$ transforms the equation to a classical feature weighting scheme in which separate weights for each feature are introduced, allowing to leverage their relative importance and reduce the impact of noisy features. The main difference with our approach is twofold. Firstly, our weighting is dependent on the position in the image (as indexed by i) which allows for the feature binding. Secondly, we use a different cue, the attention cue, to compute the weight. As a consequence, the final image representation is based on the combination of the two cues, color and shape.

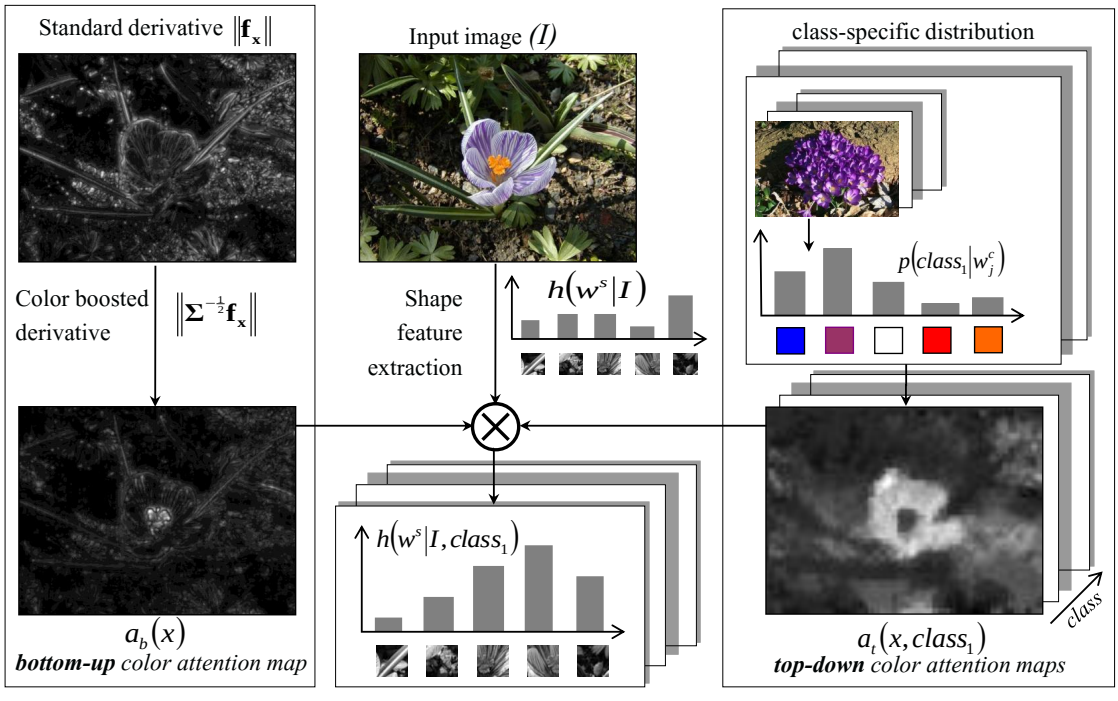


Figure 3.4: An overview of our method. Other than the classical bag-of-words approach, our method modulates the shape features with bottom-up and top-down color attention. Bottom-up attention is based on image statistics to indicate the most salient color regions whereas the top-down attention maps provide class-specific color information. As a result, a class-specific histogram is constructed by giving prominence to those shape visual-words that are considered relevant by the attention maps.

3.4.2 Top-down Color Attention

Here we define the top-down component of color attention of local features to be equal to the probability of a class given its color values and it is defined by:

$$a_t\left(\mathbf{x}_{ij}, class\right) = p\left(class|\mathrm{w}_{\mathrm{ij}}^{\mathrm{c}}\right). \tag{3.7}$$

The local color features at the locations $\mathbf{x}_{ij}$ are vector quantized into a visual vocabulary where $\mathrm{w}_{\mathrm{ij}}^{\mathrm{c}}$ describes a visual word. The probabilities $p\left(class|\mathrm{w}_{\mathrm{ij}}^{\mathrm{c}}\right)$ are computed using Bayes theorem,

$$p\left(class|\mathrm{w}^{\mathrm{c}}\right) \propto p\left(\mathrm{w}^{\mathrm{c}}|class\right) p\left(class\right) \tag{3.8}$$

where $p\left(\mathrm{w}^{\mathrm{c}}|class\right)$ is the empirical distribution,

$$p\left(\mathrm{w}_{\mathrm{n}}^{\mathrm{c}}|class\right) \propto \sum_{I^{class}} \sum_{j=1}^{M^i} \delta\left(w_{ij}^k, \mathrm{w}_{\mathrm{n}}^{\mathrm{c}}\right), \tag{3.9}$$

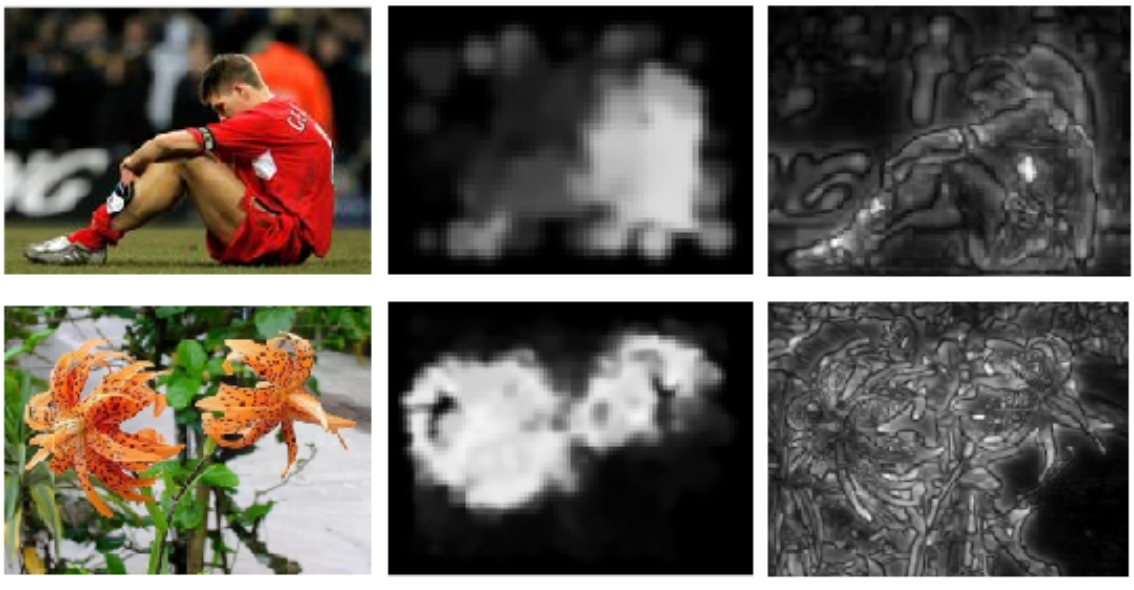


Figure 3.5: Top-down color attention and bottom-up saliency maps. First row: a liverpool class category image from soccer data set, color attention map followed by the saliency map. Second row: a snowdrop flower species image from flower data set, color attention map followed by the saliency map.

obtained by summing over the indexes of the training images for the category I^{class} . The prior over the classes $p\left(class\right)$ is obtained from the training data. For categories where color is irrelevant, $p\left(class|\mathrm{w}^{c}\right)$ is uniform and our model simplifies to the standard bag-of-words representation. If the bounding box information is available it was found that the probabilities computed only from features inside the bounding boxes provide better results. Thus when available we used bounding box knowledge available to obtain the probabilities.

If we compute $p\left(class|\mathrm{w}^{c}\right)$ for all local features in an image we can construct a top-down class-specific color attention map. Several examples are given in Fig. 3.5. The color attention map is used to modulate the local shape features. Each category provides its own attention map, consequently, a different histogram is constructed for each category. The final image representation is constructed by concatenating the category-specific histograms. The image representation is normalized before classification.

3.4.3 Bottom-up Color Attention

Bottom-up attention is employed to determine salient locations obtained from visual features such as color, intensity, orientation etc in an image. Contrary to top-down attention, bottom-up attention is independent of the object categories since it is not task dependent. In this work, we apply the color saliency boosting method [90] to

compute bottom-up attention maps. The color saliency boosting algorithm is based on the application of information theory to the statistics of color image derivatives. It has been successfully applied to image retrieval and image classification [80, 86].

Let $\mathbf{f}_{\mathbf{x}} = (R_{\mathbf{x}}\ G_{\mathbf{x}}\ B_{\mathbf{x}})^T$ be the spatial image derivatives. The information content of first order derivatives in a local neighborhood is given by

$$I(\mathbf{f}_{\mathbf{x}}) = -log(p(\mathbf{f}_{\mathbf{x}})) \tag{3.10}$$

where $p(\mathbf{f}_{\mathbf{x}})$ is the probability of the spatial derivative. The equation states that a derivative has a higher information content if it has a low probability of occurrence. In general, the statistics of color image derivatives are described by a distribution which is dominated by a principal axis of maximum variation along the luminance direction, and two minor axes, attributed to chromatic changes. This means that changes in intensity are more probable than chromatic changes and therefore contain less information content. The color derivative distribution can be characterized by its second-order statistics, i.e. its covariance matrix $\mathbf{\Sigma}_{\mathbf{x}} = E[\mathbf{f}_{\mathbf{x}}\mathbf{f}_{\mathbf{x}}^T]$. When we apply a whitening transformation to the image derivatives according to, $\mathbf{g}_{\mathbf{x}} = \mathbf{\Sigma}_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{f}_{\mathbf{x}}$, this will result in a more homogeneous derivative distribution for $\mathbf{g}_{\mathbf{x}}$, in which the dominant variations in the intensity axes are suppressed, and the chromatic variations are enforced. As a result points with equal derivative strength, $\|\mathbf{g}_{\mathbf{x}}\|$, have similar information content.

Similar as in [93] we apply color boosting to compute a multi-scale contrast color attention map:

$$a_b(\mathbf{x}) = \sum_{\sigma \in S} \sum_{\mathbf{x}' \in N(\mathbf{x})} \left\| (\mathbf{\Sigma}_x^{\sigma})^{-\frac{1}{2}} \left(\mathbf{f}^{\sigma}(\mathbf{x}) - \mathbf{f}^{\sigma}(\mathbf{x}') \right) \right\| \tag{3.11}$$

where $\mathbf{f}^{\sigma}$ is the Gaussian smoothed image at scale σ, $N(\mathbf{x})$ is a 9x9 neighborhood window, moreover $S = \left[1, \sqrt{2}, 2, 2\sqrt{2}, ..., 32\right]$. We compute $\mathbf{\Sigma}_{\mathbf{x}}^{\sigma}$ from the derivatives at scale σ from a single image. The approach is an extension of the multi-contrast method by [52] to color. Examples of bottom-up attention maps are given in Fig. 3.5. These images demonstrate that the dominant colors are suppressed and the colorful, less frequent, edges are enhanced.

3.4.4 Multiple Cues

The proposed method can easily be extended to include multiple bottom-up and top-down attention cues. In this work we have also evaluated multiple top-down attention cues. For q top-down attention cues we compute

$$a(\mathbf{x}_{ij}, class) = a_t^1(\mathbf{x}_{ij}, class) \times ... \times a_t^q(\mathbf{x}_{ij}, class). \tag{3.12}$$

Note that the dimensionality of the image representation is independent of the number of attention cues. In the experiments, we shall provide results based on multiple color attention cues.

3.4.5 Relation to Interest Point Detectors

In bag-of-words two main approaches to feature detection can be distinguished [58]. Ignoring the image content *dense sampling* extracts features on a dense grid at multiple scales in the image. *Interest point* detectors adjust to the image by sampling more points from regions which are expected to be more informative. Examples of the most used interest point detectors are Harris-Laplace, Hessian and Laplace detectors. Here we show that interest point detectors can also be interpreted to be a shape-attention weighted version of a dense multi-scale feature detector.

Consider the following equation for attention based bag-of-words:

$$h\left(\mathrm{w}_n^s | I^i, class\right) \propto \sum_{j=1}^{M^i} a\left(\mathbf{x}_{ij\sigma}\right) \delta\left(\mathrm{w}_{\mathrm{ij}\sigma}^{\mathrm{s}}, \mathrm{w}_{\mathrm{n}}^{\mathrm{s}}\right), \tag{3.13}$$

where σ has been added to explicitly indicate that at every location multiple scales are taken into consideration. Interest point detectors can be considered as providing the function $a\left(\mathbf{x}_{ij\sigma}\right)$ which is one for feature locations and scales which were detected and zero otherwise. For example the Laplace detector computes the function $a\left(\mathbf{x}_{ij\sigma}\right)$ by finding the maxima in the Laplace scale-space representation of the image, and thereby providing a scale invariant blob detector. In these cases the shape-attention is bottom-up since the same detector is used invariably for all classes. The importance of interest point detectors versus dense sampling is much researched [56,58,64] and is not further investigated in this chapter.

Of interest here is the insight this gives us in the working of color attention. Although color attention does not have the hard assignment which is applied in traditional interest point detectors (selecting some features and ignoring others), the weights $a\left(\mathbf{x}_{ij}, class\right)$ could be understood as a color based 'soft' interest point detector, where some features have more weights than others. Furthermore, since the weights are class dependent, the resulting histograms can be interpreted as being formed by class-specific interest point detectors.

3.5 Experiments

In this section we first explain the experimental setup followed by an introduction to the data sets used in our experiments. The data sets have been selected to represent a varied importance of the two visual cues namely, color and shape. We then present the results of our proposed method on image classification. Finally, the results are compared to state-of-the-art methods fusing color and shape.

Figure 3.6: Examples from the four data sets. From top to bottom: Soccer, Flower, PASCAL VOC and Caltech-101 data sets.

3.5.1 Experimental Setup

To test our method, we have used a standard multiscale grid detector along with Harris-Laplace point detector [58] and a blob detector. We normalized all the patches to a standard size and descriptors are computed for all regions in the feature description step. A universal visual vocabulary representing all object categories in a data set is then computed by clustering the descriptor points using a standard K-means algorithm. In our approach the SIFT descriptor is used to create a shape vocabulary. A visual vocabulary of 400 is constructed for Soccer and Flower data sets. For Pascal VOC 2007 and 2009 data sets, a 4000 visual-word vocabulary is used. A visual vocabulary of 500 is employed for the Caltech-101 data set. To construct a color vocabulary, two different color descriptors, namely the color name (CN) descriptor [88, 89] and hue descriptor (HUE) [87]. Since color names has more discriminative power than hue we used a larger vocabulary for CN than for HUE for all datasets.

We abbreviate our results with notation convention $CA(\textit{descriptor cue}, \textit{attention cues})$ where CA stands for the integrated bottom-up and top-down components of color attention based bag-of-words and $TD(\textit{descriptor cue}, \textit{attention cue})$ where TD stands for Top-Down attention based bag-of-words representation. We shall provide results with one attention cue $CA(SIFT,\ HUE)$, $CA(SIFT,\ CN)$, and color attention with two attention cues $CA(SIFT, \{HUE, CN\})$ combined by using Eq. 3.12. The final image representation input to an SVM classifier is equal to the size of shape vocabulary times the number of object categories in the data set. In our experiments we use a standard non-linear SVM. A single γ and β parameter is learned for Soccer and Flower data set. For Caltech-101 parameters are learned globally for the

whole data set whereas for the PASCAL VOC data sets class-specific parameters are learned.

We compare our method with the standard methods used to combine color and shape features from literature: early fusion and late fusion. We perform early and late fusion with both CN and HUE descriptors. We also compare our approach with methods that combine color and shape at the classification stage by combining the multiple kernel responses. Recently, an extensive performance evaluation of color descriptors has been presented by [86]. We compare our results to the two descriptors reported to be superior. OpponentSIFT uses all the three channels $(O1, O2, O3)$ of the opponent color space. The $O1$ and $O2$ channels describe the color information in an image whereas $O3$ channel contains the intensity information in an image. The C-SIFT descriptor is derived from the opponent color space as $\frac{O1}{O3}$ and $\frac{O2}{O3}$, thereby making it invariant with respect to light intensity. Furthermore, it has also been mentioned by [86] that with no prior knowledge about object categories, OpponentSIFT descriptor was found to be the best choice.

3.5.2 Image Data Sets

We tested our method on five different and challenging data sets namely Soccer, Flower, PASCAL VOC 2007 and 2009 and Caltech-101 data sets. The data sets vary in the relative importance of the two cues, shape and color.

The Soccer data set [3] consists of 7 classes of different soccer teams [87]. Each class contains 40 images divided in 25 train and 15 test images per class. The Flower data set [4] consists of 17 classes of different variety of flower species and each class has 80 images. We use both the 40 training and 20 validation images per class (60) to train [61]. We also tested our approach on PASCAL VOC data sets [17, 18]. The PASCAL VOC 2007 data set [5] consists of 9963 images of 20 different classes with 5011 training images and 4952 test images. The PASCAL VOC 2009 data set [6] consists of 13704 images of 20 different classes with 7054 training images and 6650 test images. Finally, we tested our approach on Caltech-101 data set. The Caltech-101 data set [7] contains 9144 images of 102 different categories. The number of images per category varies from 31 to 800. Fig. 3.6 shows some images from the four data sets.

[3] The Soccer set at http://lear.inrialpes.fr/data

[4] The Flower set at http://www.robots.ox.ac.uk/vgg/

[5] The PASCAL VOC Challenge 2007 at http://www.pascal-network.org/challenges/VOC/voc2007/

[6] The PASCAL VOC Challenge 2009 at http://www.pascal-network.org/challenges/VOC/voc2009/

[7] The Caltech-101 data set at http://www.vision.caltech.edu/ImageDatasets/Caltech101/

3.5.3 Attention Cue Evaluation

We propose to combine color and shape by modulating shape features using color as an attention cue. The same framework can be used to modulate color features by exchanging the roles of color and shape. Table 3.1 provides results of our experiments where we investigate shape-shape attention, color-color attention, shape-color attention and color-shape attention. Experiments are performed on both Soccer and Flower data sets. The results in Table 3.1 suggest that color is the best choice as an attention cue, which coincides with the previous works done in visual attention literature [35, 104]. Therefore, in the following experiments color is used as an attention cue to modulate the shape features. [8]

$Attention-Cue$	$Descriptor-Cue$	*Soccer*	*Flower*
Shape	*Shape*	50	69
Color	*Color*	79	66
Shape	*Color*	78	69
Color	*Shape*	**87**	**87**

Table 3.1: Classification Score (percentage) on Soccer and Flower Set Data sets. The results are based on top-down color attention obtained by using different combinations of color and shape as attention and descriptor cues.

3.5.4 Soccer Data Set: color predominance

Image classification results are computed for the Soccer data set to test color and shape fusion under conditions where color is the predominant cue. In this data set the task is to recognize the soccer team present in the image. In this case, the color of the player's outfit is the most discriminative feature available.

The results on the Soccer data set are given in Table 3.2. The importance of color for this data set is demonstrated by the unsatisfactory results of shape alone where an accuracy of 50% is obtained. Color Names performed very well here due to their combination of photometric robustness and the ability to describe the achromatic regions. A further performance gain was obtained by combining hue and color name based color attention. In all cases combining features by color attention was found to outperform both early and late fusion. We also combine color and shape by taking the product of the two kernels obtaining a classification score of 91%. Note that also for both early and late fusion the relative weight of color and shape features is learned by cross-validation. The best results are obtained by combining the top-down and bottom-up attention demonstrating the fact that both types of attentions are important for obtaining best classification results.

Our method outperforms the best results reported in literature [88], where a

[8] In an additional experiment, we tried improving the results by using a color-shape descriptor cue and an attention cue. This was found to deteriorate the recognition performance.

score of 89% is reported, based on a combination of SIFT and CN in an early fusion manner. Further we compare to C-SIFT and Opp-SIFT [86] which provide an accuracy of 72% and 82% respectively. The below expected results for C-SIFT might be caused by the importance of the achromatic colors to recognize the team shirts (for example, Milan outfits are red-black and PSV outfits are red-white). This information is removed by the photometric invariance of C-SIFT. Our best results of 96% is obtained when color has greater influence over shape (γ=3) which is also analogous to the unsatisfactory results of shape alone. Moreover, top-down attention has more influence than bottom-up attention (β=0.6).

Method	(SIFT,HUE)	(SIFT,CN)	(SIFT,(CN,HUE))
$EarlyFusion$	84	88	90
$LateFusion$	81	86	88
TD	87	90	94
CA	**90**	**91**	**96**

Table 3.2: Classification scores (percentage) for various fusion approaches on Soccer Data set. The best results are obtained by CA outperforming the other fusion methods by 5%.

3.5.5 Flower Data Set: color and shape parity

Image classification results on the Flower data set show the performance of our method on a data set for which both shape and color information are essential. The task is to classify the images into 17 different categories of flower-species. The use of both color and shape are important as some flowers are clearly distinguishable by shape, e.g. daisies and some other by color, e.g. fritillaries.

Method	(SIFT,HUE)	(SIFT,CN)	(SIFT,(CN,HUE))
$EarlyFusion$	87	88	89
$LateFusion$	86	87	88
TD	90	90	91
CA	**93**	**94**	**95**

Table 3.3: Classification Scores (percentage) for various fusion approaches on Flower Data set. CA is shown to outperform existing fusion approaches by 6%.

The results on flower data set are given in Table 3.3. As expected on this data set early fusion provides better results compared to late fusion. [9]. Again combining color and shape by color attention obtains significantly better results than both early and late fusion. We also significantly outperform both C-SIFT and OpponentSIFT which provide classification scores of 82% and 85% respectively.

[9]We also performed an experiment for combining our color and shape features by using MKL. However, slightly better results of 86% were obtained by using a simple product of different kernel combinations which is similar to the results provided by [26].

On this data set our method surpassed the best results reported in literature [26, 63,66,107]. The results reported on this data set by [63] is 88.3% where shape, color and texture descriptors were combined along with the segmentation scheme proposed by [62]. On the other hand neither segmentation nor any bounding box knowledge have been used in our method. A more proximal comparison with our approach is that of [107] where a result of 89.02% was obtained by combining the spatial pyramids of SIFT with OpponentSIFT, C-SIFT, rgSIFT and RGBSIFT respectively using a bin-ratio dissimilarity kernel. [10]

In Fig. 3.7 the classification score as a function of γ and β is provided. Our best result of 95% is obtained with a significant color influence (γ=2). Moreover, for this data set bottom-up attention has the same influence as top-down attention (β=0.5). It can also be seen that bottom-up attention alone improves results from 69% to 76%.

3.5.6 PASCAL VOC Data Sets: shape predominance

We test our approach where the shape cue is predominant and color plays a subordinate role and report image classification results on the PASCAL VOC 2007 and 2009 data sets. The PASCAL VOC 2007 data set contains nearly 10,000 images of 20 different object categories. The 2009 PASCAL VOC data set contains 13704 images of 20 different categories. For these data sets the average precision is used as a performance metric in order to determine the accuracy of recognition results.

On this data set, shape alone provides a MAP of 53.7 on this data set. A MAP of 49.6 is obtained using C-SIFT. This drop in performance is caused by the categories having color-shape independency which effects early fusion based approaches. Table 3.4 shows the results of different color-shape fusion schemes. Among the existing approaches late fusion provides the best recognition performance of 56.0. Our proposed framework obtains significantly better results and doubles the gain obtained by color. Our best results of 58.0 is obtained by the combination of bottom-up and top-down attention. For categories such as plants and tvmonitor, color is more important than shape (γ=3) where as for categories like sheep, sofa and cars shape is more influential as compared to color (γ=1). For categories such as cow, dogs and bottle bottom-up attention plays an important role. However, for most categories top-down attention plays a larger role than bottom-up attention on this data set.

The results per object category are given in Fig. 3.8. It is worthy to observe that our approach performs substantially better over early and late fusion approaches on a variety of categories. Recall that early fusion approaches lack feature compactness and struggle with categories where one cue is constant and the other cue varies considerably. This behavior can be observed in object categories such as motorbike,

[10]The result reported by [31] is not the recognition score commonly used to evaluate the classification performance on Flower data set and therefore is not compared with our approach.

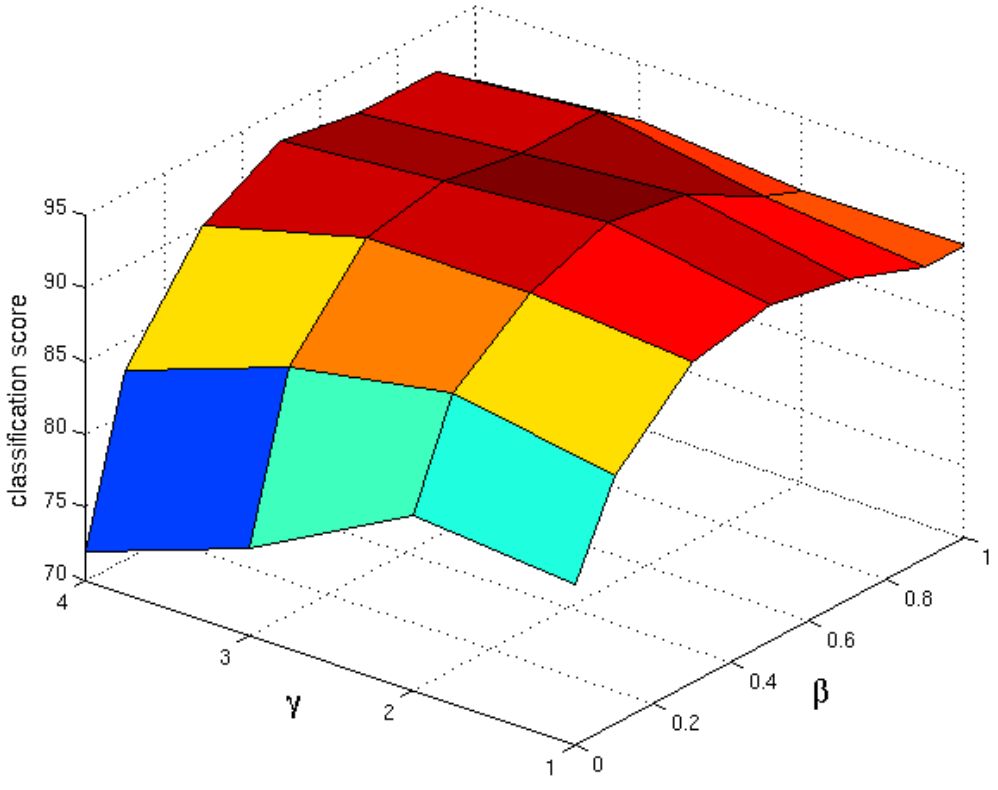


Figure 3.7: Recognition performance as a function of γ and β for the Flower data set. From a shape only representation (γ=0 and β=0) the score goes up from 69% to 95% by leveraging the influence of color versus shape and the two components of color attention.

bird etc. In such classes early fusion provides below-expected results. On the other hand, late fusion lacks feature binding as it struggles over categories characterized by both color and shape. This is apparant in categories such as cat, sheep, cow where early fusion provides better results over late fusion. Our approach, which combines the advantages of both early and late fusion, obtains good results on most type of categories in this data set.

To illustrate the strength of different image representations, Table 3.5 shows images of different object categories from the PASCAL VOC 2007 data set. For this data set the average precision is used as an evaluation criteria. To obtain an average precision for each object category, the ranked output is used to compute the precision/recall curve. Table 3.5 shows example images from bird, pottedplant, sofa and motorbike categories and their corresponding ranks obtained from different methods. Early fusion performs better than late fusion on the pottedplant image since color remains constant (color-shape dependency). For the motorbike image,

Method	(SIFT,HUE)	(SIFT,CN)	(SIFT,(CN,HUE))
EarlyFusion	54.6	54.8	55.7
LateFusion	55.3	55.6	56.0
TD	56.6	56.8	57.5
CA	**57.0**	**57.5**	**58.0**

Table 3.4: Mean Average Precision on PASCAL VOC 2007 Data Set. Note that our results significantly improve the performance over the conventional methods of combining color and shape namely, Early and Late feature fusion.

Ranking of Different Object Categories

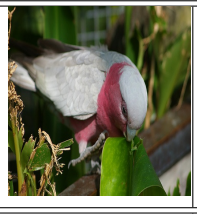
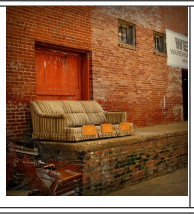
Method				
SIFT	1243	697	1325	155
Early Fusion	196	65	654	124
Late Fusion	183	164	64	30
Color Attention	10	13	36	87

Table 3.5: Images from bird, pottedplant, motorbike and sofa categories from the PASCAL VOC 2007 data set. The number indicates the rank for the corresponding object category. A lower number reflects higher confidence on the category label. The object category list contains 4952 elements in total. Color attention outperforms SIFT, early and late fusion on the bird, pottedplant and sofa category images. On motorbike category late fusion provides better ranking than color attention.

which possesses color-shape independency, late fusion performs best. Color attention outperforms other approaches on the first three example images.

The best entry in PASCAL 2007 VOC was by [56] where a mean average precision of 59.4 was reported by using SIFT, Hue-SIFT, spatial pyramid matching and a novel feature selection scheme. Without the novel feature selection scheme a mean average precision of 57.5 was reported. A similar experiment was performed by [86] where all the color descriptors (C-SIFT, rg-SIFT, OpponentSIFT and RGB-SIFT) were fused with SIFT and spatial pyramid matching to obtain a map of 60.5. Recently, [30] obtained a mean average precision of 63.5 by combining object classification and localization scores. A MAP of 64.0 is reported by [115] using shape alone with a superior coding scheme. This scheme yields a gain of 19.4% over standard vector-quantization used in our framework.

Table 3.6 shows the results obtained on 2009 PASCAL VOC data set. Our proposed approach outperforms SIFT over all the 20 categories.

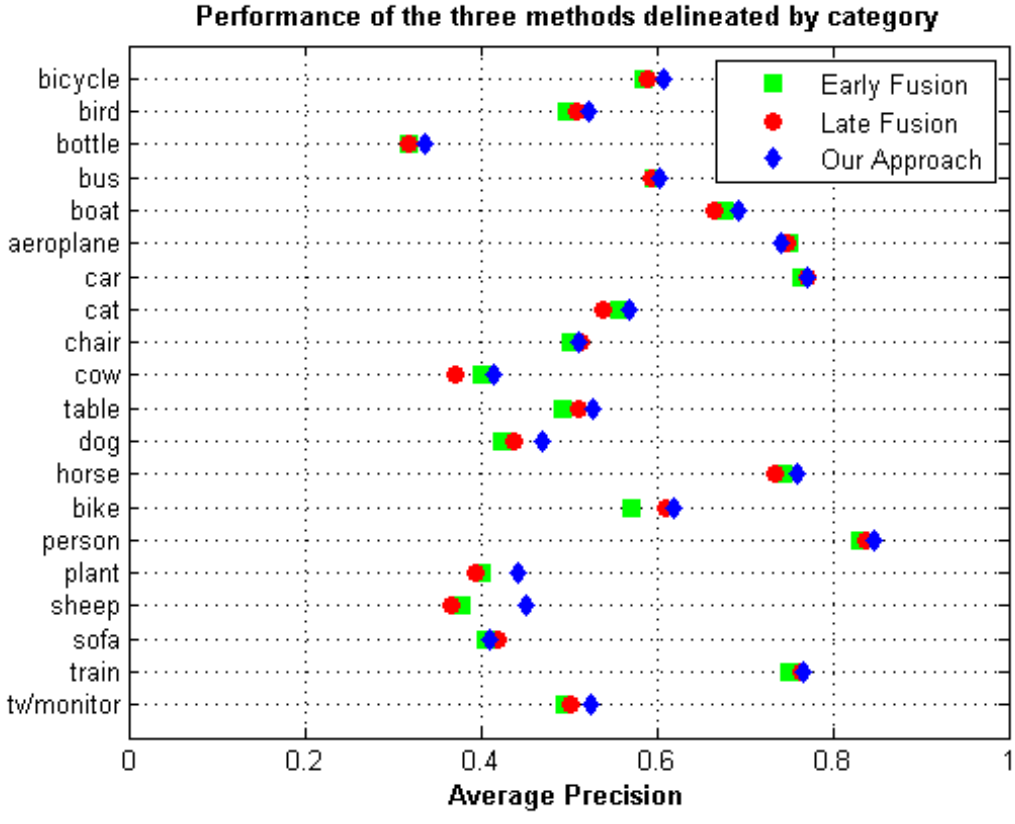


Figure 3.8: Results per category on PASCAL VOC 2007 data set: the results are split out per object category. Note that we outperform Early and Late Fusion in 16 out of 20 object categories.

For the PASCAL 2009 challenge submission, we further combine the color attention method with additional ColorSIFT [86], spatial pyramid matching and combining the classification scores with detection results [30]. We follow the classical bag-of-words pipeline where for each image different features are detected. A variety of feature extraction schemes such as GIST [65] are employed afterwards followed by vocabulary and histogram construction. Spatial information is captured using spatial pyramid histograms [47] by dividing the image into 2×2 (image quarters) and 1×3 (horizontal bars) subdivisions. We compressed the visual vocabularies using the agglomerative information bottleneck approach [22]. Finally, color attention is combined to provide as an input to the classifier. By using SIFT, we obtained a mean average precision (MAP) of 51.0 on the validation set. By adding color attention, we obtained a significant performance gain with a MAP score of 56.2. Finally we added additional descriptors to achieve a MAP of 59.4. Our final submission which also included the object localization results obtained best results on potted plants and tvmonitor category in the competition [11].

[11]PASCAL VOC 2009 at, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/results/

Method	Voc Size	Mean AP
$SIFT$	4000	52.1
$TD(SIFT, CN)$	$4000, 500$	55.1
$TD(SIFT, HUE)$	$4000, 300$	54.9
$TD(SIFT, \{CN, HUE\})$	$4000, \{500, 300\}$	56.1
$CA(SIFT, CN)$	$4000, 500$	55.6
$CA(SIFT, HUE)$	$4000, 300$	55.4
$CA(SIFT, \{CN, HUE\})$	$4000, \{500, 300\}$	**56.4**

Table 3.6: Mean Average Precision on PASCAL VOC 2009 dataset. Note that our results significantly improve the performance over the conventional SIFT descriptor.

3.5.7 Caltech-101 Data Set: color and shape co-interference

Finally, our approach is tested in a scenario where combining color with shape has shown to consistently deteriorate the results in literature [5,26,91,94]. Several factors hamper the performance of color features in this data set: low image quality, number of grayscale images (5%), many graphics-based images in different object categories (i.e. garfield, pigeon, panda etc.) and several object categories (i.e. scissors, Buddha etc.) containing the object placed on a variable color background.

The Caltech-101 data set contains 9000 images divided into 102 categories. We followed the standard protocol [5, 26, 47] for our experiments by using 30 images per category for the training and upto 50 images per category for testing. Multi-way image classification is obtained by empolying a one-vs-all SVM classifier. A binary classifier is learned to distinguish each class from the rest of the categories. For each test image, the category label of the classifier is assigned that provides the maximum response. We provide results over all 102 categories and the final recognition performance is measured as the mean recognition rate per category.

Method	Voc Size	Score
$SIFT$	500	73.3
$EarlyFusion(SIFT, CN)$	1000	70.6
$LateFusion(SIFT, CN)$	$500 + 500$	74.9
$OpponentSIFT$	1000	66.3
$C - SIFT$	1000	59.7
$TD(SIFT, CN)$	$500, 500$	74.7
$CA(SIFT, CN)$	$500, 500$	**76.2**

Table 3.7: Recognition results on Caltech-101 Set. Note that conventional early fusion based approaches to combine color and shape provide inferior results compared to the results obtained using shape alone.

Table 3.7 shows the results obtained using spatial pyramid representations upto level 2. Among the existing approaches, only late fusion provides a gain over shape alone. For all early fusion approaches inferior results are obtained compared to

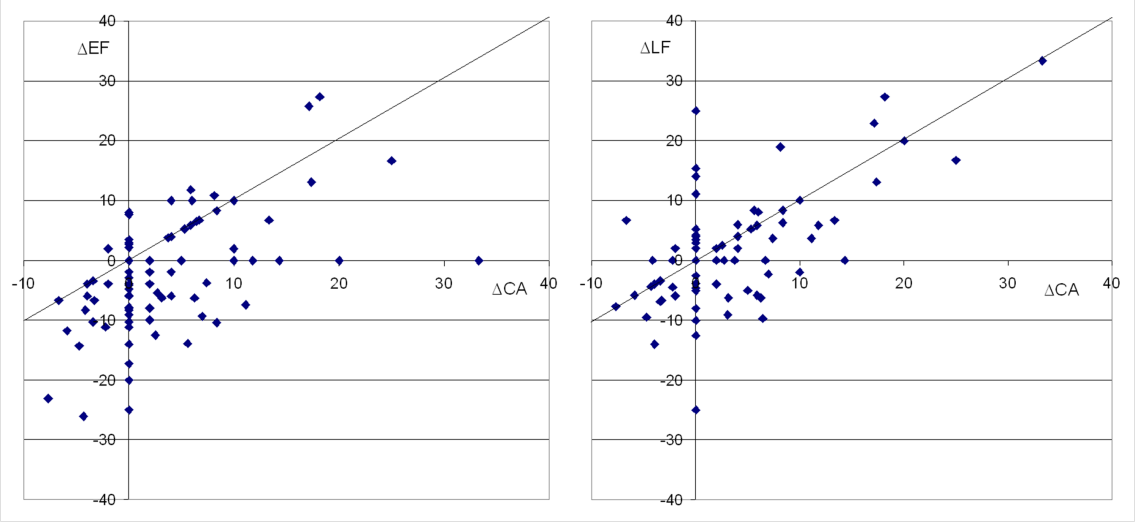


Figure 3.9: Left figure: comparison of gain over shape obtained by early fusion (ΔEF) to gain obtained by color attention (ΔCA). Every dot represents one of the Caltech-101 categories. All points above the origin show an advantage of early fusion over shape. All points on the right of origin depict a gain of color attention over shape. For all points below the diagonal color attention outperforms early fusion. Similar results for late fusion are shown in the figure on the right.

shape alone. Our approach that combines the strength of both early and late fusion improves the recognition performance on this data set. Introducing color information is beneficial for some categories such as flamingo-head, pizza, lobster, dolphin etc. whereas recognition performance of categories such as hedgehog, gramophone, pigeon, emu etc. are hampered by combining color and shape.

In Fig. 3.9, a performance comparison of early and late fusion versus color attention is given. For all the categories below the diagonal, color attention outperforms early and late fusion. As illustrated in Fig. 3.9 for most of the object categories in this data set, the best results are obtained using color attention.

The best results reported on this data set is 82.1% by [26] using variants of multiple kernel learning to combine 49 different kernel matrices of 8 different types of features such as SIFT, ColorSIFT, HOG, LBP, V1S+ etc. Our proposed approach can be further employed together with previously used features to further boost the results. In Table 3.8 we compare to other approaches which combine color and shape cues. Note that we do not learn class-specific weights of the spatial pyramid levels which has been shown to improve the results significantly [5, 26, 94] mainly due to the fact that objects are always in the center of the image. Results show that early fusion combination of color and shape deteriorates results significantly upto 12%. Our approach improves the overall performance on this data set compared to shape alone.

Method	Shape	Color-Shape	Score
[1]	71.6	68.2	−3.4
[91]	52.8	40.8	−12.0
[94]	73.0	63.0	−10.0
[26]	66.4	55.0	−11.4
OurApproach	73.3	76.2	**+2.9**

Table 3.8: Comparison in performance of shape and color-shape approaches reported in literature with our proposed approach. Note that our method improves the overall recognition performance over shape alone on Caltech-101 data set.

3.6 Conclusions

In this chapter we have performed an analysis on two existing approaches (early and late fusion) that combine color and shape features. Experimental results clearly demonstrate that both these approaches are sub-optimal for a subset of object categories. This analysis leads us to define two desired properties for feature combination: *feature binding* and *feature compactness*, which in a standard bag-of-words approach are mutually exclusive.

We present a new image representation which combines color and shape within the bag-of-words framework. Our method processes color and shape separately and then combines it by using both bottom-up and top-down attention. The bottom-up component of color attention is obtained by applying a color saliency method whereas the top-down component is obtained by using learned category-specific color information. The bottom-up and top-down attention maps are then used to modulate the weights of local shape features. Consequently, a class-specific image histogram is constructed for each category.

Experiments are conducted on standard object recognition data sets. On the two data sets, Soccer and Flower, where color plays a pivotal role, our method obtains state-of-the-art results increasing classification rate over 5% compared to early and late fusion. On the PASCAL VOC data sets, we show that existing methods based on early fusion underperform for classes with shape-color independency, including many man-made classes. Results based on color attention show that also for these classes color does contribute to overall recognition performance. Performance comparison of our approach to existing fusion approaches has been shown in Table 3.9.

The dimensionality of color attention histogram is equivalent to the number of object categories times the size of the shape vocabulary. Therefore as a future research direction, we aim to look at dimensionality reduction techniques such as PCA and PLS to reduce the dimensionality of color attention histograms. Another interesting future research line includes looking into other visual features that can be used as an attention cue. Recently, [48, 49] have applied our model to incorporate motion features as an attention cue and demonstrated its effectiveness for event recognition. We believe that top-down guidance can also improve the performance

Data set	SIFT	Early Fusion	Late Fusion	CA
Soccer	50	90	88	96
Flower	69	89	88	95
PASCAL VOC	53.7	55.7	56.0	58.0
Caltech-101	73.3	70.6	74.9	76.2

Table 3.9: Comparison of our approach with existing fusion approaches on various data sets. Note that our approach outperforms early and late fusion on all data sets.

in several other applications such as object detection and action recognition.

Chapter 4

Discriminative Compact Pyramids for Object and Scene Recognition[1]

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 chapter, we present a novel framework for obtaining compact pyramid representation. Firstly, 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 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 datasets.

4.1 Introduction

Bag-of-words based image representation is one of the most successful approaches for object and scene recognition [1, 7, 12, 14, 21, 46, 59, 72, 86, 113]. 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 [53]. 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

[1] Accepted for publication by Pattern Recognition Journal [15].

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 [47] 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, [16, 19, 42, 105, 106] benefit form 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. [109]). If these large pyramid representation 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 [47] 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 [22, 45, 101]. 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 [13, 78]. Slonim and Tishby [78] 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, authors in [22] 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. 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 [1, 7, 10, 26, 86, 87] 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, we investigate which fusion approach is more appropriate within the spatial pyramids framework.

In summary, the objective of this chapter is twofold. Firstly, 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 Clustering (DITC) technique [13] 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. Secondly, 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 chapter is organized as follows: next section describes the datasets 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 chapter and describes the most important lines of future research.

4.2 Datasets and Implementation Details

In this section we provide details about the datasets which will be used, 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. 4.1 shows some example images from the five data sets.

4.2.1 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 [50] contains 8 sports event 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

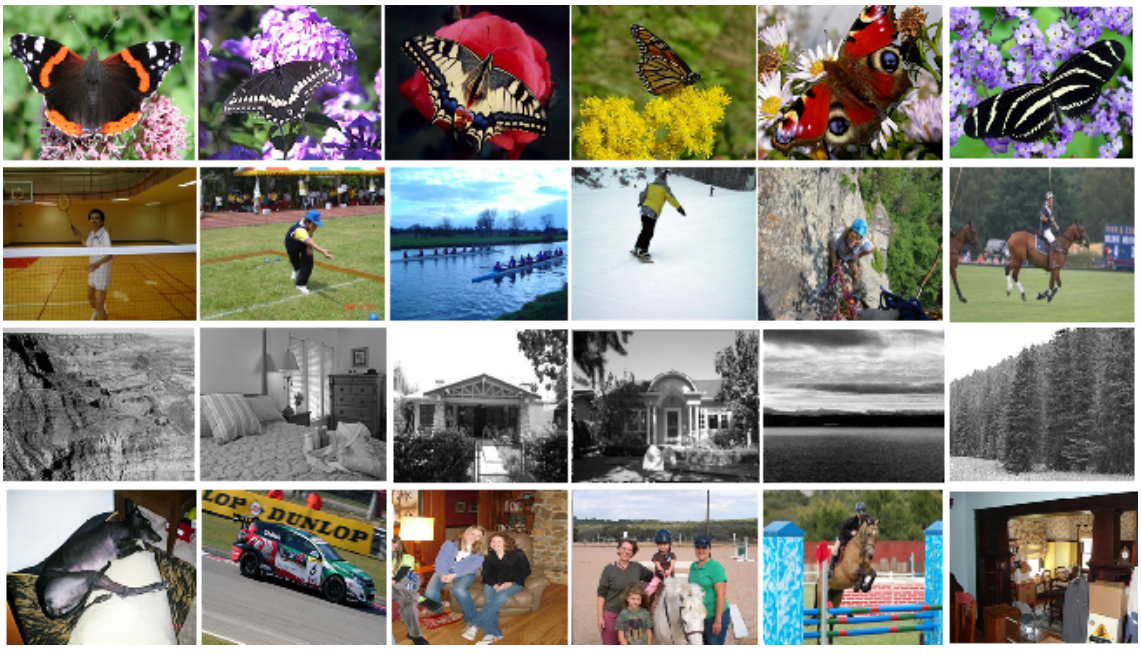


Figure 4.1: Example images from the data sets. From top to down: Butterflies, Sports Events, 15 class Scenes and PASCAL VOC data sets.

selected images are used for training and 60 are chosen for testing.

The 15 class Scenes recognition data set [47] is composed of fifteen scene categories. Each category has 200 to 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 [44] and Pascal VOC 2007 and 2009 data sets [16]. 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 13704 images of 20 different object categories with 7054 training images and 6650 test images.

4.2.2 Implementation Details

We shortly discuss the implementation details we use for the bag-of-words based image classification. We apply a standard multiple-scale grid detector along with interest point detectors (Harris-Laplace and blob detector). In the feature extraction step, we use SIFT descriptor [53] for shape features, Color Names [89] descriptor for color features and the SelfSimilarity descriptor [76] 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 [55].

4.2.3 Image Representation using Spatial Pyramids

Spatial pyramid scheme proposed by [47] have 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 [2].

4.3 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 [13,78]. Only AIB has been applied for compact image representation [22], 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 [50] and 15 class Scenes [47] data sets to demonstrate that our proposed compact pyramid representation maintain 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 [42, 43], or large scale image retrieval [70]). 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 work we consider that the final representation size is an input parameter to the compression algorithm.

[2] The winners of Pascal VOC 2007 [56] 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 $\times 8$

4.3.1 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 subregions in the spatial pyramid. For a level two pyramid constructed from a 1000 word vocabulary, this will lead to a final representation of $(1 + 4 + 16) \times 1000 = 21000$ words. We will consider clustering these 21000 words into a smaller set where each cluster represents words with 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,

$$I(C,W) = \sum_i \sum_t p(c_i, w_t) log \frac{p(c_i, w_t)}{p(c_i)p(w_t)}, \tag{4.1}$$

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 mutual information. Thus, we aim at finding word clusters that minimize the decrease in the mutual information:

$$I(C,W) - I(C,W^C). \tag{4.2}$$

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. (4.2) can be rewritten as

$$\sum_i \sum_t \pi_t p(c_i|w_t) log \frac{p(c_i|w_t)}{p(c_i)} - \sum_i \sum_j \sum_{w_t \in W_j} \pi_t p(c_i|w_t) log \frac{p(c_i|W_j)}{p(c_i)} \tag{4.3}$$

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

In the seminal work [13], Dhillon et al. prove that this is equal to

$$I(C,W) - I(C,W^C) = \sum_j \sum_{w_t \in W_j} \pi_t KL((p(C|w_t)), (p(C|W_j))) \tag{4.4}$$

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

$$KL(p_1, p_2) = \sum_{x \in X} p_1(x) log \frac{p_1(x)}{p_2(x)}. \tag{4.5}$$

Eq. (4.4) 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 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. (4.4).

AIB Compression [78]: 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. (4.1). The decrease in the mutual information is monotonically reduced after each merge. 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. (4.4).

DITC Compression [13]: Other than AIB which iteratively reduces the number of words until then 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 decline in mutual information as given by Eq. (4.4), 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. (4.4), DITC iteratively performs the following steps:

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

$$p(C|W_j) = \sum_{w_t \in W_j} \frac{\pi_t}{\pi(W_j)} p(C|w_t), \tag{4.6}$$

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

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

$$j^*(w_t) = argmin_j KL(p(C|w_t), p(C|Wj)) \tag{4.7}$$

   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 [13].

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

Method	Level	Size	Sports Events	15 Scenes
$Pyramid$	2	21000	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 4.1: Classification Score (percentage) on both the Sports Events and 15 class Scenes Data sets. The results demonstrates that by applying the AIB compression [22] a considerable loss in performance occurred for compact vocabularies.

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. (4.7) 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. (4.7), 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 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. (4.7) to be computed for every pair, $P(C|w_t)$ and $p(C|W_j)$ at a cost of $O(mkc\tau)$, where generally $k << m$. 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 20000 words on a data set with 15 classes, AIB (using [22]) took 14460 seconds while DITC converged in 234 seconds using a standard PC.

4.3.2 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 [45] propose a method for discriminative vocabulary construction which uses ideas of the theory of DITC [13]. However, the word clusters where 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 [22].

Table 4.1 shows numerical results obtained by applying AIB on both the Sports

Method	Level	Size	Sports Events	15 Scenes
$Pyramid$	2	21000	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

Table 4.2: Classification Score (percentage) on both the Sports Events and 15 class Scenes data sets. The results demonstrates that DITC successfully compresses the vocabularies while preserving their discriminative power.

Events and 15 Scenes data sets for different sizes. We started by using vocabulary of size 1000 for constructing a three level pyramid of 21000 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 result in a sub-optimal word clusters because it greedily merges just two word clusters at every step and it does not directly optimize the global objective function of Eq. (4.4).

Table 4.2 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 21000 word vocabulary.

Classification accuracies of both compression approaches are shown Figure 4.2 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 achieve 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. Figure 4.3 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 technqiues. It is worthy to mention that DITC also provides better performance compared to both PCA and PLS with a very small compact pyramid representations (50 bins).

The performance difference between DITC and AIB becomes especially appar-

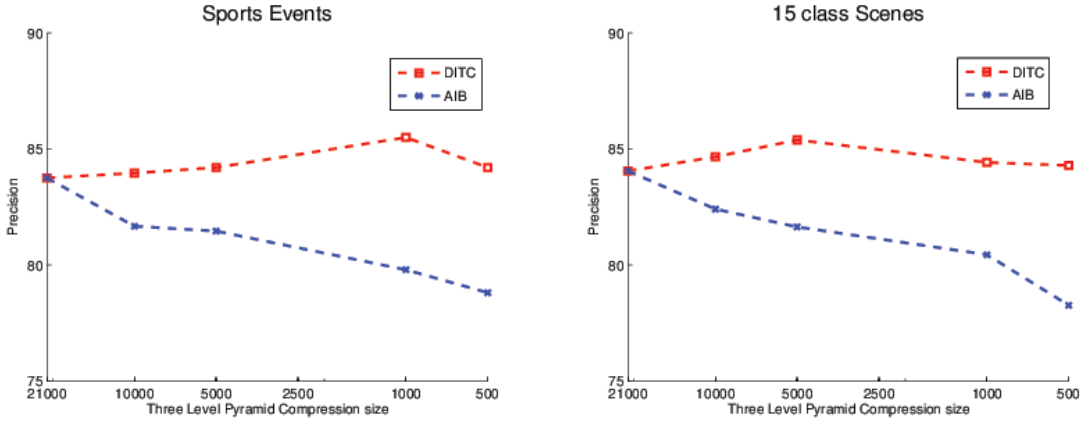


Figure 4.2: Sports Events data set (left) and 15 class Scenes data set (right) classification accuracy for compressing the whole pyramid representation leading to a more compact pyramid representation using the two compression approaches considered namely: DITC vs. AIB.

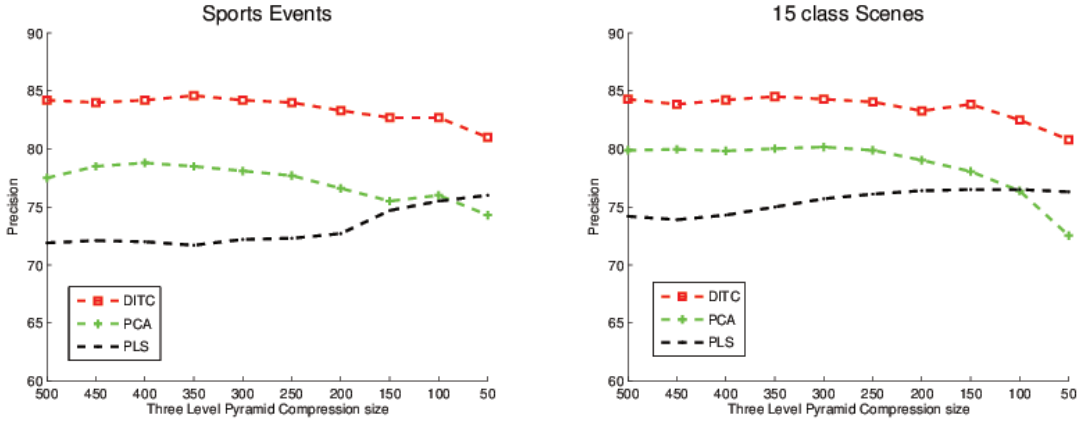


Figure 4.3: Sports Events data set (left) and 15 class Scenes data set (right) classification accuracy for compressing the whole pyramid to a compact representation using approaches namely: DITC, PLS and PCA. Note that DITC based compression also provides superior performance for very compact pyramid representations.

	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

Table 4.3: 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 25500 compressed to size 200. The second row shows the compression results using the AIB [22] and the third row shows the results using DITC [13].

ent for high compression. An initial pyramid representation of the PASCAL dataset of 25500 words is compressed to 200 clusters. Table 4.3 shows a 14% higher Mean Average-Precision for having compact pyramid representations based on DITC compared to those obtained using AIB on object recognition.

4.3.3 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 4.3.1 and denoted as *PyrComp* hereafter).

Table 4.4 shows the results obtained using both of the considered proposed designs on 15 class Scenes and the Sports Events datasets. To compare the classification scores obtained from the 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.

Method	Level	Size	Sports Events	15 Scenes
Pyramid	2	21000	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**

Table 4.4: Classification score on the Sports Events and 15 class Scenes datasets 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).

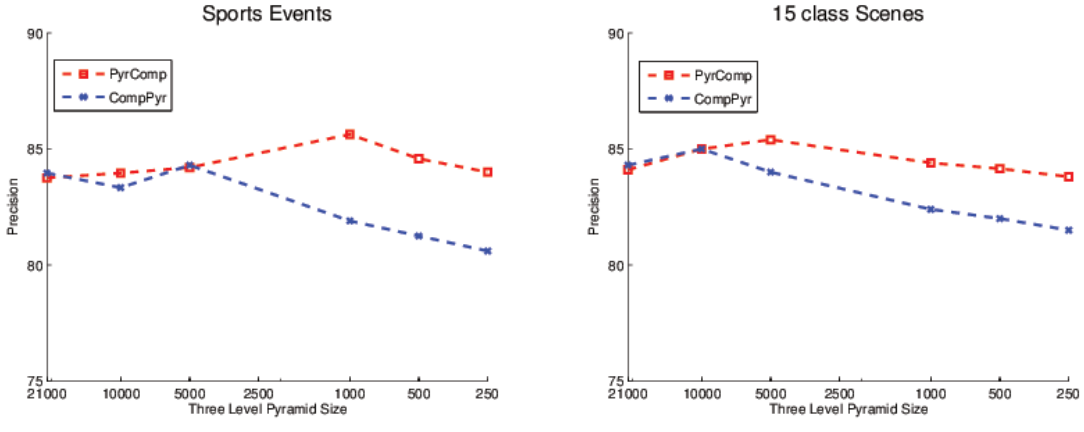


Figure 4.4: Classification comparison between *PyrComp* and *CompPyr* strategies for (left) 15 class Scenes and (right) Sports Events datasets.

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 slightly improves performance. Additionally Figure 4.4 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 [45,54,108]. Our experiment show that compressing the vocabulary within the spatial pyramid, significantly improves 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.

4.4 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 [19]. 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.

4.4.1 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. Example of such categories are man made objects such as car, buses and chairs etc.

Typically within the bag-of-words framework a number of local features f^c_{mn}, m=1...M^n are extracted from training images I_n. Where $n = 1, 2, ..., N$, 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 :

$$f^{1\&2}_{mn} = \big(\beta\ f^1_{mn}, (1-\beta) f^2_{mn}\big) \tag{4.8}$$

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 obtained by concatenating the separate histograms:

$$h^{(V_1,V_2)}(I) = \left[\beta\ h^{V_1}(I), (1-\beta)\ h^{V_2}(I)\right] \tag{4.9}$$

Note that we have introduced a weight parameter β for both early and late fusion which allows 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. [7] computes the SIFT descriptor on the H,S,V channels and then concatenates the final descriptor into a single representation. Van de Weijer and Schmid [87] compare photometrically invariant representations in combination with SIFT for object recognition. Recently, Van de Sande et al. [86] 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, [2, 6, 25, 74, 91] combine kernel combinations of different visual features. A weighted linear combination of kernels is employed, where each feature is represented by multiple kernels. Beside 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 [26]. 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. [39]. 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.) [7, 86, 87]. We refer to these spatial pyramids that are based on early fusion scheme as *early fusion spatial pyramids* and

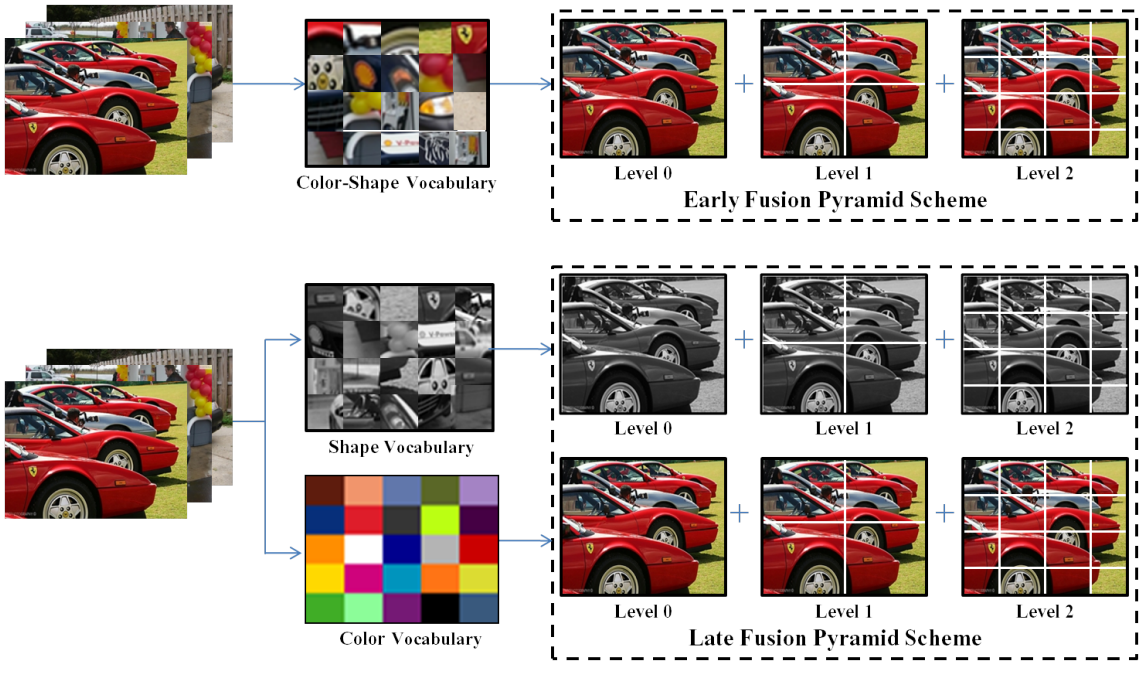


Figure 4.5: Early and Late fusion pyramid schemes. In the early fusion pyramid scheme a combined color-shape vocabulary is constructed as a result of which a single pyramid representation is obtained. To construct a late fusion pyramid, a separate vocabulary is constructed for color and shape and spatial pyramids are obtained for each cue. We show that late fusion is the recommended approach for combining multiple features.

the spatial pyramids that are based on late fusion as a *late fusion spatial pyramids*. Figure 4.5 highlights the two spatial pyramid matching approaches.

4.4.2 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 recognition. 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 [89] 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 fusion we use the two standard implementations as given by Eqs. (4.8) and (4.9). 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 [86]. Table 4.5 shows the results obtained on Sports Events data set. For this data set, shape is an important cue and color plays

Method	Level	Size	Score
Shape	0	800	80.6
Color	0	300	53.9
Opp − SIFT	0	1100	**82.9**
EarlyFusion	0	1100	80.6
LateFusion	0	1100	81.8
Opp − SIFT	1	5500	82.3
EarlyFusion	1	5500	80.8
LateFusion	1	5500	**82.7**
Opp − SIFT	2	23100	80.8
Earlyfusion	2	23100	82.7
Latefusion	2	23100	**84.4**

Table 4.5: Classification Score (percentage) on Sports Events Data set.

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 degrade (the performance of OpponentSIFT at 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 [26]. However, results were found to be inferior compared to the late fusion spatial pyramid scheme.

Table 4.6 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 fusion 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.

4.5 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 is 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

Method	Level	Size	Score
Shape	0	1000	79.4
Color	0	300	53.3
$Opp-SIFT$	0	1500	78.7
$EarlyFusion$	0	1500	79.6
$LateFusion$	0	1300	**81.9**
$Opp-SIFT$	1	7500	79.6
$EarlyFusion$	1	7500	81.7
$LateFusion$	1	6500	**84.4**
$Opp-SIFT$	2	31500	81.0
$Earlyfusion$	2	31500	83.3
$Latefusion$	2	27300	**87.9**

Table 4.6: Classification Score (percentage) on Butterflies Data set.

Data Sets	Best Score		PS		PS_C		$PS_C + PC_C + PSS_C$	
	Size	**Score**	Size	**Score**	Size	**Score**	Size	**Score**
Sports	6K	84.2 [105]	21K	83.8	1K	85.6	2K	**87.1**
15 Scenes	21K	84.3 [8]	21K	84.1	1K	84.4	2K	**86.7**
Butterflies	2K	90.6 [44]	21K	89.5	1K	89.0	2K	**91.4**
Pascal 2007	160K	**60.5** [86]	84K	57.4	15K	57.2	25K	59.5
Pascal 2009	4194K	**64.6** [109]	84K	55.7	15K	55.2	25K	57.6

Table 4.7: Classification Score (percentage) on Sports Events, 15 class Scenes, Butterflies, Pascal VOC 2007 and 2009 Data sets.

feature separately and then combine them in a late fusion manner.

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 literature. Table 4.7 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 21000 gives 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 [8, 98, 105–107]. The final accuracy is obtained with our compact histogram of dimensionality 2000 reduced from the original pyramid histograms of dimensionality 42000.

For the **15 category Scenes data set**, we followed the standard protocol of splitting the data set in to training and testing 5 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 21000 gives 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 dimensionality of 42000. We obtained a classification accuracy of 86.7 which is to the best of our knowledge the best performance on this data set [8, 98, 105–107].

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 [44].

The results on the **Pascal VOC 2007** show 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 [86], 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 [30], where authors include additional information of object bounding boxes from object detection to improve image classification.

For the **Pascal VOC 2009**, 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.

4.6 Conclusions

A major drawback of spatial pyramids is the high dimensionality of their image representation. 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 recognition, 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 [30, 42]. The application of bag-of-words based detection has been particularly advanced due to the efficient sub-window search (ESS) algorithm proposed by Lampert et al. [42]. 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 [54, 108], which has been shown excellent results in recent works in image restoration and face recognition [34, 111]. 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.

Chapter 5

Portmanteau Vocabularies for Multi-Cue Image Representation[1]

We describe a novel technique for feature combination in the bag-of-words model of image classification. Our approach builds discriminative compound words from primitive cues learned independently from training images. Our main observation is that modeling joint-cue distributions independently is more statistically robust for typical classification problems than attempting to empirically estimate the dependent, joint-cue distribution directly. We use Information theoretic vocabulary compression to find discriminative combinations of cues and the resulting vocabulary of *portmanteau*[2] words is compact, has the cue binding property, and supports individual weighting of cues in the final image representation. State-of-the-art results on both the Oxford Flower-102 and Caltech-UCSD Bird-200 datasets demonstrate the effectiveness of our technique compared to other, significantly more complex approaches to multi-cue image representation.

5.1 Introduction

Image categorization is the task of classifying an image as containing an objects from a predefined list of categories. One of the most successful approaches to this problem is the bag-of-words (BOW) [7, 47]. In the bag-of-words model an image

[1] Appeared in Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS 2011) [38].

[2] A *portmanteau* is a combination of two or more words to form a neologism that communicates a concept better than any individual word (e.g. Ski resort + Konference = *Skonference*). We use the term to describe our vocabularies to emphasize the connotation with combining color and shape words into new, more meaningful representations.

is first represented by a collection of local image features detected either sparsely or in a regular, dense grid. Each local feature is then represented by one or more cues, each describing one aspect of a small region around the corresponding feature. Typical local cues include color, shape, and texture. These cues are then quantized into visual words and the final image representation is a histogram over these visual vocabularies. In the final stage of the BOW approach the histogram representations are sent to a classifier.

The success of BOW is highly dependent on the quality of the visual vocabulary. In this chapter we investigate visual vocabularies which are used to represent images whose local features are described by both shape and color. To extend BOW to multiple cues, two properties are especially important: cue binding and cue weighting. A visual vocabulary is said to have the *binding property* when two independent cues appearing at the same location in an image remain coupled in the final image representation. For example, if every local patch in an image is independently described by a shape word and a color word, in the final image representation using compound words the binding property ensures that shape and color words coming from the same feature location are coupled in the final representation. The term *binding* is borrowed from the neuroscience field where it is used to describe the way in which humans select and integrate the separate cues of objects in the correct combinations in order to accurately recognize them [82]. The property of *cue weighting* implies that it is possible to adapt the relevance of each cue depending on the dataset. The importance of cue weighting can be seen from the success of Multiple Kernel Learning (MKL) techniques where weights for each cue are automatically learned [3, 9, 63, 74, 91, 92].

Traditionally, two approaches exist to handle multiple *cues* in BOW. When each cue has its own visual vocabulary the result is known as a *late fusion* image representation in which an image is represented as one histogram over shape-words and another histogram over color-words. Such a representation does not have the cue binding property, meaning that it is impossible to know exactly which color-shape events co-occurred at local features. Late fusion does, however, allow cue weighting. Another approach, called *early fusion*, constructs a single visual vocabulary of joint color-shape words. Representations over early fusion vocabularies have the cue binding property, meaning that the spatial co-occurrence of shape and color events is preserved. However, cue weighting in early fusion vocabularies is very cumbersome since must be performed before vocabulary construction making cross-validation very expensive. Recently, Khan et al. [39] proposed a method which combines cue binding and weighting. However, their final image representation size is equal to number of vocabulary words times the number of classes, and is therefore not feasible for the large data sets considered in this chapter.

A straightforward, if combinatorially inconvenient, approach to ensuring the binding property is to create a new vocabulary that contains one word for each combination of original shape and color feature. Considering that each of the original shape and color vocabularies may contain thousands of words, the resulting joint

vocabulary may contain millions. Such large vocabularies are impractical as estimating joint color-shape statistics is often infeasible due to the difficulty of sampling from limited training data. Furthermore, with so many parameters the resulting classifiers are prone to overfitting. Because of this and other problems, this type of joint feature representation has not been further pursued as a way of ensuring that image representations have the binding property.

In recent years a number of vocabulary compression techniques have appeared that derive small, discriminative vocabularies from very large ones [13, 22, 78]. Most of these techniques are based on information theoretic clustering algorithms that attempt to combine words that are equivalently discriminative for the set of object categories being considered. Because these techniques are guided by the discriminative power of clusters of visual words, estimates of class-conditional visual word probabilities are essential. These recent developments in vocabulary compression allow us to reconsider the direct, Cartesian product approach to building compound vocabularies.

These vocabulary compression techniques have been demonstrated on single-cue vocabularies with a few tens of thousands of words. Starting from even moderately sized shape and color vocabularies results in a compound shape-color vocabulary an order of magnitude larger. In such cases, robust estimates of the underlying class-conditional joint-cue distributions may be difficult to obtain. We show that for typical datasets a strong independence assumption about the joint color-shape distribution leads to more robust estimates of the class-conditional distributions needed for vocabulary compression. In addition, our estimation technique allows flexible cue-specific weighting that cannot be easily performed with other cue combination techniques that maintain the binding property.

5.2 Portmanteau vocabularies

In this section we propose a new multi-cue vocabulary construction method that results in compact vocabularies which possess both the cue binding and the cue weighting properties described above. Our approach is to build *portmanteau vocabularies* of discriminative, compound shape and color words chosen from independently learned color and shape lexicons. The term portmanteau is used in natural language for words which are a blend of two other words and which combine their meaning. We use the term *portmanteau* to describe these compound terms to emphasize the fact that, similarly to the use of neologistic portmanteaux in natural language to capture complex and compound concepts, we create groups of color and shape words to describe semantic concepts inadequately described by shape or color alone.

A simple way to ensure the binding property is by considering a product vocabulary that contains a new word for every combination of shape and color terms.

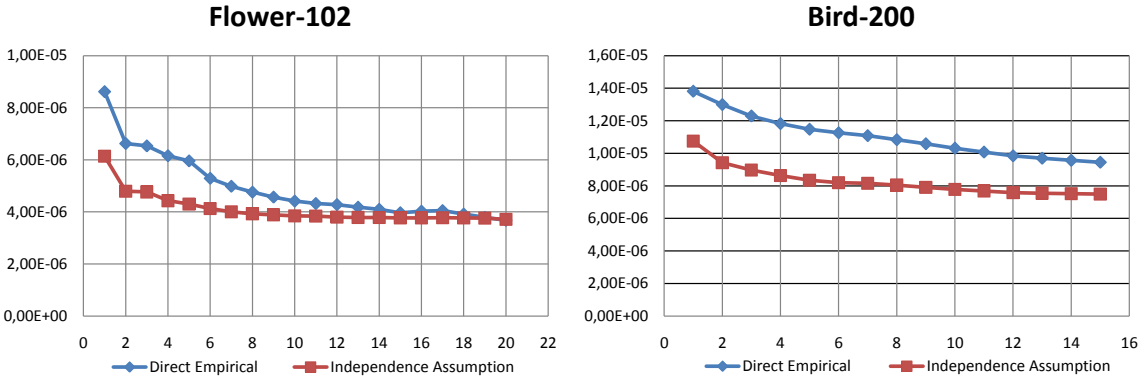


Figure 5.1: Comparison of two estimates of the joint cue distribution $p(S, C|R)$ on two large datasets. The graphs plot the Jenson-Shannon divergence between each estimate and the true joint distribution as a functions of the number of training images used to estimate them. The true joint distribution is estimated empirically over all images in each dataset. Estimation using the independence assumption of equation (5.2) yields similar or better estimates than their empirical counterparts.

Assume that $S = \{s_1, s_2, ..., s_M\}$ and $C = \{c_1, c_2, ..., c_N\}$ represent the visual shape and color vocabularies, respectively. Then the product vocabulary is given by

$$W = \{w_1, w_2, ..., w_T\} = \{\{s_i, c_j\} \mid 1 \leq i \leq M, 1 \leq j \leq N\},$$

where $T = M \times N$. We will also use the the notation s_m to identify a member from the set S.

A disadvantage of vocabularies of compound terms constructed by considering the Cartesian product of all primitive shape and color words is that the total number of visual words is equal to the number of color words times the number of shape words, which typically results in hundreds of thousands of elements in the final vocabulary. This is impractical for two reasons. First, the high dimensionality of the representation hampers the use of complex classifiers such as SVMs. Second, insufficient training data often renders robust estimation of parameters very difficult and the resulting classifiers tend to overfit the training set. Because of these drawbacks, compound product vocabularies have, to the best of our knowledge, not been pursued in literature. In the next two subsections we discuss our approach to overcoming these two drawbacks.

5.2.1 Compact Portmanteau Vocabularies

In recent years, several algorithms for feature clustering have been proposed which compress large vocabularies into small ones [13,22,78]. To reduce the high-dimensionality of the product vocabulary, we apply Divisive Information-Theoretic feature Clustering (DITC) algorithm [13], which was shown to outperform AIB [78]. Furthermore,

DITC has also been successfully employed to construct compact pyramid representations [15].

The DITC algorithm is designed to find a fixed number of clusters which minimize the loss in mutual information between clusters and the class labels of training samples. In our algorithm, loss in mutual information is measured between original product vocabulary and the resulting clusters. The algorithm joins words which have similar discriminative power over the set of classes in the image categorization problem. This is measured by the probability distributions $p\left(R|w_t\right)$, where $R = \{r_1, r_2, ..r_L\}$ is the set of L classes.

More precisely, the drop in mutual information I between the vocabulary W and the class labels R when going from the original set of vocabulary words W to the clustered representation $W^R = \{W_1, W_2, ..., W_J\}$ (where every W_j represents a cluster of words from W) is equal to

$$I\left(R;W\right) - I\left(R;W^R\right) = \sum_{j=1}^{J} \sum_{w_t \in W_j} p\left(w_t\right) KL\left(p\left(R|w_t\right) || \; p\left(R|W_j\right)\right), \tag{5.1}$$

where KL is the Kullback-Leibler divergence between two distributions. Equation (5.1) states that the drop in mutual information is equal to the prior-weighted KL-divergence between a word and its assigned cluster. The DITC algorithm minimizes this objective function by alternating computation of the cluster distributions and assignment of compound visual words to their closest cluster. For more details on the DITC algorithm we refer to Dhillon et al. [13]. Here we apply the DITC algorithm to reduce the high-dimensionality of the compound vocabularies. We call the compact vocabulary which is the output of the DITC algorithm the *portmanteau vocabulary* and its words accordingly *portmanteau words*. The final image representation $p(W^R)$ is a distribution over the portmanteau words.

5.2.2 Joint distribution estimation

In solving the problem of high-dimensionality of the compound vocabularies we seemingly further complicated the estimation problem. As DITC is based on estimates of the class-conditional distributions $p(S, C|R) = p(W|R)$ over product vocabularies, we have increased the number of parameters to be estimated to $M \times N \times L$. This can easily reach millions of parameters for standard image datasets. To solve this problem we propose to estimate the class conditional distributions by assuming independence of color and shape, given the class:

$$p(s_m, c_n|R) \propto p(s_m|R)p(c_n|R). \tag{5.2}$$

Note that we do not assume independence of the cues themselves, but rather the less restrictive independence of the cues given the class. Instead of directly estimating the empirical joint distribution $p(S, C|R)$, we reduce the number of parameters

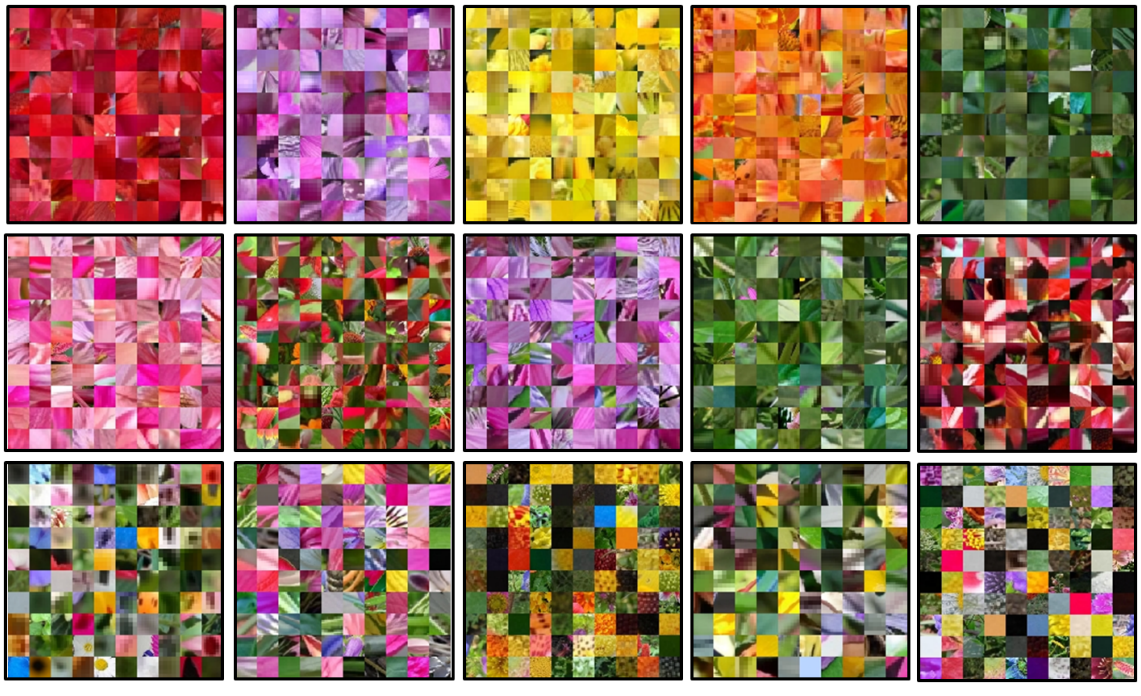


Figure 5.2: The effect of α on DITC clusters. Each of the large boxes contains 100 image patches sampled from one Portmanteau word on the Oxford Flower-102 dataset. Top row: five clusters for $\alpha = 0.1$. Note how these clusters are relatively homogeneous in color, while shape varies considerably within each. Middle row: five clusters sampled for $\alpha = 0.5$. The clusters show consistency over both color and shape. Bottom row: five clusters sampled for $\alpha = 0.9$. Notice how in this case shape is instead homogeneous within each cluster.

to estimate to $(M + N) \times L$, which in the vocabulary configurations discussed in this chapter represents a reduction in complexity of two orders of magnitude. As an additional advantage, we will show in section 5.2.3 that estimating the joint distribution $p(S, C|R)$ allows us to introduce cue weighting.

To verify the quality of the empirical estimates of equation (5.2) we perform the following experiment. In figure 5.1 we plot the Jensen-Shannon (JS) divergence between the empirical joint distribution obtained from the test images and the two estimates: direct estimation of the empirical joint distribution $p(S, C|R)$ on the training set, and an approximate estimate made by assuming independence as in equation (5.2). Results are provided as a function of the number of training images for two large datasets. A low JS-divergence means a better estimate of the true joint-cue distribution. The plotted lines show the curves for a color cue vocabulary of 100 words and a shape cue vocabulary of 5,000 words, resulting in a product vocabulary of 500,000 words. On both datasets we see that the independence assumption actually leads to a better or equally good estimate of the joint distribution. Increasing the number of training samples, or starting with smaller color and shape vocabularies and hence reducing the number of parameters to esti-

mate, will improve direct empirical estimates of $p(S, C)$. However, figure 5.1 shows that for typical vocabulary settings on large datasets the independence assumption results in equivalently good or better estimates of the joint distribution.

5.2.3 Cue weighting

Constructing the compact portmanteau vocabularies based on the independence assumption significantly reduces the number of parameters to estimate. Furthermore, as we will see in this section, it allows us to control the relative contribution of color and shape cues in the final representation.

We introduce a weighting parameter $\alpha \in [0, 1]$ in the estimate of $p(C, S)$:

$$p^{\alpha}(s_m, c_n|R) \propto p(s_m|R)^{\alpha} p(c_n|R)^{1-\alpha} \tag{5.3}$$

where an α close to zero results in a larger influence of the color words, and a α close to one leads to a vocabulary which focuses predominantly on shape.

To illustrate the influence of α on the vocabulary construction, we show samples from portmanteau words obtained on the Oxford Flower-102 dataset (see figure 5.4) in figure 5.2. The DITC algorithm is applied to reduce the product vocabulary of 500,000 compound words to 100 portmanteau words. For settings of $\alpha \in \{0.1, 0.5, 0.9\}$ we show five of the hundred words. Each word is represented by one hundred randomly sampled patches from the dataset which have been assigned to the word. The effect of changing the α can be clearly seen. For low α the Portmanteau words exhibit homogeneity of color but lack within-cluster shape consistency. On the other hand for high α the words show strong shape homogeneity such as low and high frequency lines and blobs, while color is more uniformly distributed. For a setting of $\alpha = 0.5$ the clustering is more consistent in both color and shape.

Additionally, another parameter β is introduced:

$$p^{\alpha,\beta}(s_m, c_n|R) \propto \left(p(s_m|R)^{\alpha} p(c_n|R)^{1-\alpha}\right)^{\beta} \tag{5.4}$$

To illustrate the influence of β consider the following experiment on synthetic data. We generate a set of 100 words which have random discriminative power $p(R|w_t)$ over $L = 10$ classes. In figure 5.3 we show the $p(R|w_t)$ for a subset of 20 words in grey, and $p(R|W_j) \propto \sum_{w_t \in W_j} p(w_t)p(R|w_t)$ for the ten portmanteau words in color. We observe that increasing the β parameter directs DITC to find clusters which are each highly discriminative for a single class, rather than being discriminative over all classes. We found that higher β values often lead to image representations which improve classification results.

These weighting parameters are learned through cross validation on the training set. In practice we found α to change with the data set according to the importance

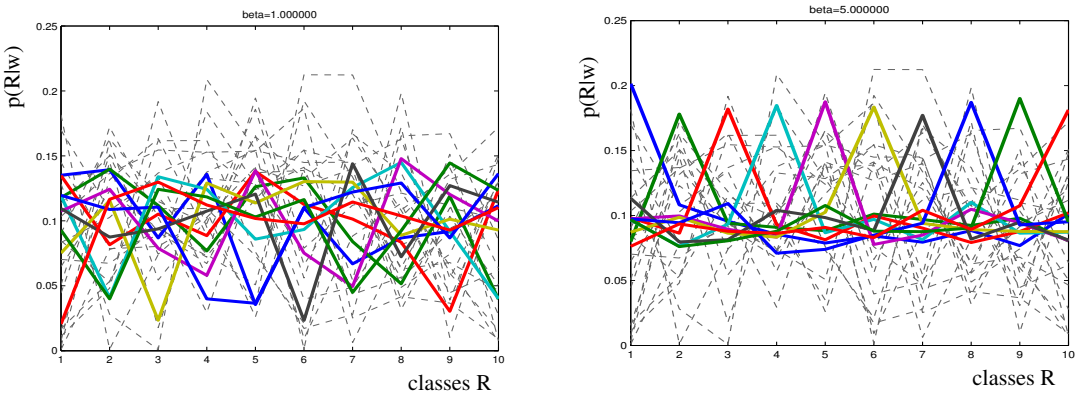


Figure 5.3: The effect of β on DITC clusters. For 20 words $p(R|w_t)$ is plotted in dotted grey lines. DITC is used to obtain ten portmanteau means $p(R|W_j)$ are plotted in different colors. On the left is shown the final clustering for $\beta = 1.0$. Note that none of the portmanteau means are especially discriminative for one particular class. On the right, however, for $\beta = 5.0$ each portmanteau concentrates on discriminating one class.

of color and shape. The β parameter was found to to be constant at a value 5 for the two datasets evaluated in this chapter. Both parameters were found to significantly improve results on both datasets.

5.2.4 Image representation with portmanteau vocabularies

We summarize our approach to constructing portmanteau vocabularies for image representation. We emphasize the fact that our approach is fundamentally about deriving compact multi-cue image representations and, as such, can be used as a drop-in replacement in any bag-of-words pipeline.

Image representation by portmanteau vocabulary built from color and shape cues follows these steps:

1. Independent color and shape vocabularies are constructed by standard K-means clustering over color and shape descriptors extracted from training images.

2. Empirical class-conditional word distributions $p(S|R)$ and $p(C|R)$ are computed from the training set, the joint cue distribution $P(S, C|R)$ is estimated assuming conditional independence as in equation (5.4).

3. The portmanteau vocabulary is computed with the DITC algorithm. The output of the DITC is a list of indexes which, for each member of the compound vocabulary maps to one of the J portmanteau words.

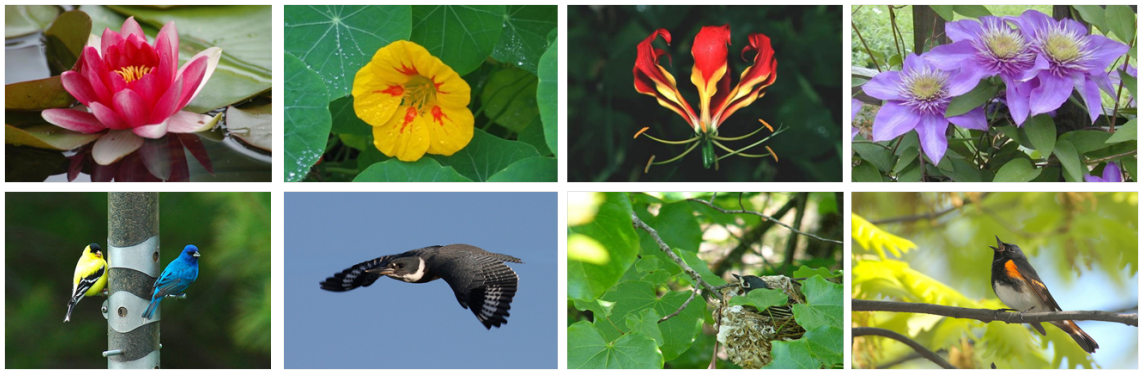


Figure 5.4: Example images from the two datasets used in our experiments Top: images from four categories of the Flower-102 dataset. Bottom: four example images from the Bird-200 dataset.

4. Using the index list output by DITC, the original image features are revisited and the index corresponding the compound shape-color word at each feature is used to represent each image as a histogram over the portmanteau vocabulary.

5.3 Experimental results

We follow the standard bag-of-words approach. We use a combination of interest-point detectors along with a dense multi-scale grid detector. The SIFT descriptor [53] is used to construct a shape vocabulary. For color we use the color name descriptor, which is computed by converting sRGB values to color names according to [89] after which each patch is represented as a histogram over the eleven color names. The shape and color vocabularies are constructed using the standard K-means algorithm. In all our experiments we use a shape vocabulary of 5000 words and a color vocabulary of 100 words. Applying Laplace weighting was not found to influence the results and therefore not used in the experiments. The classifier is a non-linear, multi-way, one-versus-all SVM using the χ^2 kernel [113]. Each test image is assigned the label of the classifier giving the highest response and the final classification score is the mean recognition rate per category.

We performed several experiments to validate our approach to building multi-cue vocabularies by comparing with other methods which are based on exactly the same initial SIFT and CN descriptors:

- **Shape and Color only**: a single vocabulary of 5000 SIFT words and one of 100 CN words.
- **Early fusion**: SIFT and CN are concatenated into single descriptor. The relative weight of shape and color is optimized by cross-validation. Note that

cross-validation on cue weighting parameters for early fusion must be done over the *entire* BOW pipeline, from vocabulary construction to classification. Vocabulary size is 5000.

- **Direct empirical**: DITC based on the empirical distribution of $p(S, C|R)$ over a total of 500.000 compound words estimated on the training set.
- **Independence assumption**: where $p(S, C|R) = p(S|R)p(C|R)$ is assumed. We also show separate results with and without using α and β.

In all cases the color-shape visual vocabularies are compressed to 500 visual words and spatial pyramids are constructed for the final image representation as in [47]. All of the above approaches were evaluated on two standard and challenging datasets: Oxford Flower-102 and Caltech-UCSD Bird-200. The train-test splits are fixed for both datasets and are provided on their respective websites.[3] and the [4]

5.3.1 Results on the Flower-102 and Bird-200 datasets

The Oxford Flower-102 dataset contains 8189 images of 102 different flower species. It is a challenging dataset due to significant scale and illumination changes (see figure 5.4). The results are presented in table 5.1(a). We see that shape alone yields results superior to color. Early fusion is reasonably good at 70.5%. This is however obtained through laborious cross validation to obtain the optimal balance between CN and SIFT cues. Since our cue weighting is done after the initial vocabulary and histogram construction, cross-validation is significantly faster than for early fusion.

The bottom three rows of table 5.1(a) give the results of our approach to image representation with portmanteau vocabularies in a variety of configurations. The direct empirical estimation of the joint shape-color distribution provides slightly better results than estimation based on the independence assumption. However, weighting the two visual cues using the α parameter described in equation (5.3) in the independent estimation of $p(s, c|\text{class})$ improves the results significantly. In particular, the gain of almost 7% obtained by adding β is remarkable. The best recognition performance were obtained for $\alpha = 0.8$ and $\beta = 5$.

The Caltech-UCSD Bird-200 dataset contains 6033 images from 200 different bird species. This dataset contains many bird species that closely resemble each other in terms of color and shape cues, making the recognition task extremely difficult. Table 5.1(a) contains test results for our approach on Bird-200 as well. Interestingly, on this dataset color outperforms shape alone and early fusion yields only a small improvement over color. Results based on portmanteau vocabularies outperform early fusion, and estimation based on the independence assumption provide better results than direct empirical estimation. These results are further improved by the

[3]The Flower-102 dataset at http://www.robots.ox.ac.uk/vgg/research/flowers/
[4]Birds-200 set at http://www.vision.caltech.edu/visipedia/CUB-200.html

Method	Flower	Bird
Shape only	60.7	12.9
Color only	48.5	16.8
Early Fusion	70.5	17.0
Direct empirical	64.6	18.9
Independent	63.5	19.8
Independent + α	66.4	21.6
Independent + α + β	**73.3**	**22.4**

(a)

Method	Bird	Flower
OpponentSIFT	14.0	69.2
C-SIFT	13.9	65.9
MKL [63]	–	72.8
MKL [9]	19.0	–
Random Forest [112]	19.2	–
Saliency [37]	–	71.0
Our Approach	**22.4**	**73.3**

(b)

Table 5.1: Comparative evaluation of our approach. (a) Classification score on Flower-102 and Bird-200 datasets for individual features, early fusion and several configurations of our approach. (b) Comparison of our approach to the state-of-the-art on the Bird-200 and Flower-102 datasets.

introduction of cue weighting with a final score of 22.4% obtained with $\alpha = 0.7$ and $\beta = 5$ outperforming all others.

5.3.2 Comparison with the state-of-the-art

Recently, an extensive performance evaluation of color descriptors was presented by van de Sande et al. [86]. In this evaluation the OpponentSIFT and C-SIFT were reported to provide superior performance on image categorization problems. We construct a visual vocabulary of 5000 visual words for both OpponentSIFT and C-SIFT and apply the DITC algorithm to compress it to 500 visual words. As shown in table 5.1(b), Our approach provides significantly better results compared to both OpponentSIFT and C-SIFT, possibly due to the fact neither supports cue weighting.

In recent years, combining multiple cues using Multiple Kernel Learning (MKL) techniques has received a lot of attention. These approaches combine multiple cues and multiple kernels and apply per-class cue weighting. Table 5.1(b) includes two recent MKL techniques that report state-of-the-art performance. The technique described in [9] is based on geometric blur, grayscale SIFT, color SIFT and full image color histograms, while the approach in [63] also employs HSV, SIFT int, SIFT bd, and HOG descriptors in the MKL framework of [91]. Despite the simplicity of our approach, which is based on only two cues and a single kernel, it outperforms these complex multi-cue learning techniques. Also note that both MKL approaches are based on learning class-specific weighting for multiple cues. This is especially cumbersome when there exist several hundred object categories in a dataset (e.g. the Bird-200 dataset contains 200 bird categories). In contrast to these approaches, we learn a global, class-independent cue weighting parameters to balance color and shape cues.

On the Flower-102 dataset, our final classification score of 73.3% is comparable to

the state-of-the-art recognition performance [31,37,63][5] obtained on this dataset. It should be noted that Nilsback and Zisserman [63] obtain a classification performance of 72.8% using segmented images and a combination of four different visual cues in a multiple kernel learning framework. Our performance, however, is obtained on unsegmented images using only color and shape cues. On the Bird-200 dataset, our approach significantly outperforms state-of-the-art methods [9, 98, 112].

5.4 Conclusions

In this chapter we propose a new method to construct multi-cue, visual *portmanteau* vocabularies that combine color and shape cues. When constructing a multi-cue vocabulary two properties are especially desirable: cue binding and cue weighting. Starting from multi-cue product vocabularies we compress this representation to form discriminative compound terms, or portmanteaux, used in the final image representation. Experiments demonstrate that assuming independence of visual cues given the categories provides a robust estimation of joint-cue distributions compared to direct empirical estimation. Assuming independence also has the advantage of both reducing the complexity of the representation by two orders of magnitude and allowing flexible cue weighting. Our final image representation is compact, maintains the cue binding property, admits cue weighting and yields state-of-the-art performance on the image categorization problem.

We tested our approach on two datasets, each with more than one hundred object categories. Results demonstrate the superiority of our approach over existing ones combining color and shape cues. We obtain a gain of 2.8% and 5.4% over the early fusion approach. Our approach also outperforms methods based on multiple cues and MKL with per-class parameter learning. This leaves open the possibility of using our approach to multi-cue image representation within an MKL framework.

[5] From correspondence with the authors of [31] we learned that the results reported in their paper are erroneous and they do not obtain results better than [63].

Chapter 6

Conclusions and Future Directions

In this thesis, we aim at improving the bag-of-words approach by proposing efficient image representations to combine multiple cues especially color and shape for object and scene recognition. In this chapter we summarize the approaches proposed in this thesis to improve the bag-of-words based object and scene recognition. The chapter ends with future research directions.

6.1 Conclusions

In this thesis, we have investigated methods to combine color and shape features within the bag-of-words framework for object recognition. We performed a theoretical analysis of existing approaches to combine color and shape. Early fusion has the feature binding property and helps for categories which possess constancy over color and shape cues. Late fusion has the property of feature compactness. This is especially desirable for object categories where one of the two visual cues varies significantly. To counter the problems of early and late fusion, we propose a novel approach to combine color and shape cues in chapter 3.

In the second part of the thesis, we focused on the problem of constructing discriminative and compact spatial pyramid representations for object and scene recognition. Spatial pyramid scheme encodes the spatial information missing in the orderless bag-of-words based representation. The technique works by dividing an image into increasingly finer sub-regions as a result of which a multi-resolution histogram is constructed. Although spatial pyramids provides excellent performance, the resulting histogram is very high dimensional thereby increasing the classification time significantly. Furthermore, it is still the unclear how early and late fusion of color and shape works at the spatial pyramid levels. Therefore in chapter 4, we propose an approach to construct compact and discriminative spatial pyramids which preserves the recognition performance while reducing the dimensionality of spatial pyramids significantly.

Finally, in the last part of the thesis, we have investigated the problem of combining color and shape cues for data sets with several hundred object categories. We focused on constructing a compact and discriminative color-shape visual vocabulary. Early fusion based visual vocabularies are the most common way of constructing a joint color-shape visual vocabulary. However, weighting the importance of the two visual cues which is highly desirable is extremely cumbersome in early fusion vocabularies. We propose to construct compound visual words from primitive visual cues using information theoretic vocabulary compression technique in chapter 5.

The methods proposed and the results obtained in this thesis are summarized in the paragraphs below:

Chapter 3: Modulating Shape Features by Color Attention for Object recognition. To counter the problem of early and late feature fusion, we proposed a novel image representation to combine color and shape cues. Our approach separately processes the shape and color cues and combines them by modulating the shape features by category-specific color attention. Color is used to compute bottom-up and top-down attention maps. Subsequently, the color attention maps are deployed to modulate the weights of the shape features. Shape features are given more weight in regions of an image that are more likely to contain an object instance. We have compared our approach with existing methods that combine color and shape cues. The results obtained clearly demonstrate that our proposed approach significantly outperforms existing methods for combining color and shape.

Chapter 4: Discriminative Compact Pyramids for Object and Scene Recognition. Spatial pyramid scheme has been successfully applied to incorporate spatial information within the bag-of-words framework. However, a major drawback is that it leads to high dimensional image representations. By reducing the dimensionality of spatial pyramids can further allow to incorporate multiple cues such as color and shape for improved recognition. To counter the high dimensionality problem of spatial pyramids, we have presented a novel framework for obtaining compact pyramid representation. Firstly, we have investigated the usage of the divisive information theoretic feature clustering (DITC) algorithm in creating a compact pyramid representation. In many scenarios this method is shown to reduce the size of a high dimensional pyramid representation up to an order of magnitude with little or no loss in accuracy. Moreover, we have also investigated the optimal combination of multiple features in the context of our compact pyramid representation. The experiments have showed that our method can obtain state-of-the-art results on several challenging data sets.

Chapter 5: Portmanteau Vocabularies for Multi-Cue Image Representation. Although early fusion based visual vocabularies possess the feature binding property, yet the best results are obtained by weighting the different visual cues. This parametric weighting is normally done using cross-validation thereby increasing the time complexity of the problem. Late fusion, on the other hand, allows efficient feature weighting but lacks feature binding. To this end we have described

a novel technique for feature combination within the bag-of-words model for image classification. Our approach is based on constructing discriminative compound words from visual cues learned independently from training images. We have used Information theoretic vocabulary compression to find discriminative combinations of visual cues and the resulting visual vocabulary is compact, has the cue binding property, and supports individual weighting of cues in the final image representation. State-of-the-art results on standard object recognition data sets demonstrate the effectiveness of our technique compared to other, significantly more complex approaches to multi-cue image representation.

6.2 Future Directions

Combining color and shape cues using the methods proposed in this thesis has shown excellent performance for object recognition task. In future, We aim at applying the image representations proposed in chapter 3 and 5 for object detection and action recognition task. The problem of combining multiple cues in these applications is still open to debate. Most of the existing approaches [97, 114] combines multiple cues such as shape and texture in a late fusion manner for object detection. Late fusion based approaches suffer when both visual cues are constant. Since both color attention and portmanteau allows feature binding, it would be interesting to investigate them for object detection in future.

The approaches presented in chapter 3 and 5 of this paper are shown to combine color and shape visual cues successfully. However, other visual cues such as texture, optical flow etc. can also be combined using the mentioned approaches. In recent works by [48,49] color attention is used to incorporate motion features as an attention cue for event recognition. Therefore, we aim at extending it for other visual cues for applications such as object detection and action recognition.

The compact pyramid representation introduced in chapter 4 allows efficient combination of multiple visual cues. For complex problems, such as large scale image retrieval and object detection, reducing the computational complexity and memory usage is of paramount importance. In such applications, compact pyramid representations of multiple visual cues will allow to achieve higher recognition performance without increasing the complexity. Therefore, we aim at applying the proposed approach for large scale image classification data sets such as ImageNet.

Bibliography

[1] A.Bosch, A.Zisserman, and X.Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[2] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proc. International Conference on Machine Learning*, 2004.

[3] Francis Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *NIPS*, 2008.

[4] Brent Berlin and Paul Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA, 1969.

[5] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[6] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, 2007.

[7] A. Bosch, A. Zisserman, and X.Munoz. Scene classification via plsa. In *Proc. European Conf. on Computer Vision*, 2006.

[8] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. Learning mid-level features for recognition. In *Proc. Computer Vision and Pattern Recognition*, 2010.

[9] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[10] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local colour invariants. *CVIU*, 113:48–62, 2009.

[11] Hongping Cai, Fei Yan, and Krystian Mikolajczyk. Learning weights for codebook in image classification and retrieval. In *CVPR*, 2010.

[12] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[13] Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *J. of Machine Learning Research*, 3:1265–1287, 2003.

[14] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Proc. IEEE Int. Conf. on Computer Vision*, 2003.

[15] Noha Elfiky, Fahad Shahbaz Khan, Joost van de Weijer, and Jordi Gonzalez. Discriminative compact pyramids for object and scene recognition. *Pattern Recognition Journal*, 2011.

[16] M. Everingham, L. Van Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The PASCAL Visual Object Classes challenge 2007 results.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J.Winn, and A. Zisserman. The pascal visual object classes challenge 2007 results., 2007.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J.Winn, and A. Zisserman. The pascal visual object classes challenge 2009 results., 2009.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes challenge 2008 (voc2008) results.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2008 (voc2008) results. [online]. available: http://www.pascal-network.org/challenges/voc/voc2008/, 2008.

[21] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. Computer Vision and Pattern Recognition*, 2005.

[22] B. Fulkerson, A. Vedaldi, and S. Soatto. Localizing objects with smart dictionaries. In *Proc. European Conf. on Computer Vision*, 2008.

[23] Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *PAMI*, 31(6):989–1005, 2009.

[24] K.R. Gegenfurtner and J. Rieger. Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology.*, 10:805–808, 2000.

[25] P. V. Gehler and S. Nowozin. Let the kernel figure it out:principled learning of pre-processing for kernel classifiers. In *Proc. Computer Vision and Pattern Recognition*, 2009.

[26] Peter Vincent Gehler and Sebastian Nowozin. On feature combination for multiclass object classification. In *Proc. IEEE Int. Conf. on Computer Vision*, 2009.

[27] T. Gevers and A. W. M. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, 1999.

[28] T. Gevers and Harro M. G. Stokman. Robust histogram construction from color invariants for object recognition. *PAMI*, 26:113–117, 2004.

[29] K. Grauman, , and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. *Proc. IEEE Int. Conf. on Computer Vision*, pages 1458–1465, 2005.

[30] Hedi Harzallah, Frederic Jurie, and Cordelia Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009.

[31] Satoshi Ito and Susumu Kubota. Object classification using hetrogeneous co-occurrence features. In *ECCV*, 2010.

[32] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, Nov 1998.

[33] Arnold W. M. Smeulders Jan-Mark Geusebroek Jan van Gemert, Cor J. Veenman. Visual word ambiguity. *PAMI*, 32(7):1271–1283, 2010.

[34] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proc. International Conference on Machine Learning*, 2010.

[35] Timothi Jost, Nabil Ouerhani, Roman von Wartburg, Reni Miri, and Heinz Higli. Assessing the contribution of color in visual attention. *CVIU*, 100(1–2):107–123, 2005.

[36] Frederic Jurie and Bill Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.

[37] Christopher Kanan and Garrison Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. 2010. In Proc. CVPR.

[38] Fahad Shahbaz Khan, Joost van de Weijer, Andrew Bagdanov, and Maria Vanrell. Portmanteau vocabularies for multi-cue image representation. In *Twenty-Fifth Annual Conference on Neural Information Processing Systems*, 2011.

[39] Fahad Shahbaz Khan, Joost van de Weijer, and Maria Vanrell. Top-down color attention for object recognition. In *Proc. IEEE Int. Conf. on Computer Vision*, 2009.

[40] Fahad Shahbaz Khan, Joost van de Weijer, and Maria Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision*, 2011.

[41] A. Kristjansson. Independent and additive repetition priming of motion direction and color in visual search. *Psychological Research*, (73):158–166, 2009.

[42] C.H. Lampert, M.B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *Proc. Computer Vision and Pattern Recognition*, 2008.

[43] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2129–2142, 2009.

[44] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization and segmentation. *Image and Vision Computing*, 27(5):523–534, 209.

[45] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1294–1309, 2009.

[46] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

[47] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. Computer Vision and Pattern Recognition*, 2006.

[48] Li Li, Weiming Hu, Bing Li, Chunfeng Yuan, Pengfei Zhu, and Wanqing Li. Event recognition based on top-down motion attention. 2010. In Proc. ICPR.

[49] Li Li, Chunfeng Yuan, Weiming Hu, and Bing Li. Top-down cues for event recognition. In *ACCV*, 2010.

[50] Li-Jia Li and Li Fei-Fei. What,where and who? classifying events by scene and object recognition. In *Proc. IEEE Int. Conf. on Computer Vision*, 2007.

[51] D.T Lindsey, A.M Brown, E. Reijnen, A.N Rich, Y.I Kuzmova, and J.M Wolfe. Color channels, not color categories, guide visual search for desaturated color targets.

[52] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. In *CVPR*, 2007.

[53] D. G. Lowe. Distinctive image features from scale-invariant points. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.

[54] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proc. International Conference on Machine Learning*, 2009.

[55] Subhransu Maji, Alexander C. Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *Proc. Computer Vision and Pattern Recognition*, 2008.

[56] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representation for visual object class recognition. In *Visual recognition Challenge Workshop, in conjuncture with ICCV*, 2007.

[57] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *PAMI*, 28(5):802–817, May 2006.

[58] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, , and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65(1–2):43–72, 2005.

[59] Krystian Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[60] Chetan Nandakumar and Jitendra Malik. Understanding rapid category detection via multiply degraded images. *Journal of Vision*, 9(19):1–8, 2009.

[61] M-E Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006.

[62] M-E Nilsback and A. Zisserman. Delving into the whorl of flower segmentation. In *BMVC*, 2007.

[63] M-E Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[64] Eric Nowak, Frederic Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006.

[65] Aude Oliva and Antonio B. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[66] Francesco Orabona, Jie Luo, and Barbara Caputo. Online-batch strongly convex multi kernel learning. In *CVPR*, 2010.

[67] Marco Pedersoli, Jordi Gonzalez, Andrew Bagdanov, and Juan Jose Villanueva. Recursive coarse-to-fine localization for fast object detection. In *ECCV*, 2010.

[68] Marius V. Peelen, Li Fei-Fei, and Sabine Kastner. Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, pages 94–97, 2009.

[69] Florent Perronnin. Universal and adapted vocabularies for generic visual categorization. *PAMI*, 30(7):1243–1256, 2008.

[70] Florent Perronnin, Jorge Sánchez, and Yan Liu. Large-scale image categorization with explicit data embedding. In *Proc. Computer Vision and Pattern Recognition*, 2010.

[71] Robert J. Peters and Laurent Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *CVPR*, 2007.

[72] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modelling scenes with local descriptors and latent aspects. In *Proc. IEEE Int. Conf. on Computer Vision*, 2005.

[73] Pedro Quelhas and Jean-Marc Odobez. Natural scene image modeling using color and texture visterms. In *CIVR*, 2006.

[74] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning. In *Proc. International Conference on Machine Learning*, 2007.

[75] Steven A Shafer. Using color to separate reflection components. *Color Research and Application*, 10(4):210–218, 1985.

[76] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Proc. Computer Vision and Pattern Recognition*, 2007.

[77] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[78] N. Slonim and N. Tishby. Agglomerative information bottleneck. In *Advances in Neural Information Processing Systems*, 1999.

[79] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM MM*, 2005.

[80] J. Stottinger, A. Hanbury, Th. Gevers, and N. Sebe. Lonely but attractive: Sparse color salient points for object retrieval and categorization. In *CVPR Workshops*, 2009.

[81] A. Treisman. The binding problem. *Current Opinion in Neurobiology*, 6:171–178, 1996.

[82] Anne Treisman. Feature Binding, Attention and Object Perception. *Philosophical Transactions: Biological Sciences*, 353(1373):1295–1306, 1998.

[83] J.K. Tsotsos, W.Y. Wai S.M. Culhan and, Y.H. Lai, N. Davis, and F. Nuflo. Modeling visual-attention via selective tuning. *Artif. Intell..*, 78:507–545, 1995.

[84] Tinne Tuytelaars and Cordelia Schmid. Vector quantizing feature space with a regular lattice. In *ICCV*, 2007.

[85] K. van de Sande, Th. Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. In *CVPR*, 2008.

[86] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[87] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *Proc. European Conf. on Computer Vision*, 2006.

[88] J. van de Weijer and C. Schmid. Applying color names to image description. In *ICIP*, 2007.

[89] J. van de Weijer, C. Schmid, Jakob J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1524, 2009.

[90] Joost van de Weijer, Theo Gevers, and Andrew D. Bagdanov. Boosting color saliency in image feature detection. *PAMI*, 28(1):150–156, 2006.

[91] M. Varma and D. Ray. Learning the discriminative powerinvariance trade-off. In *Proc. IEEE Int. Conf. on Computer Vision*, 2007.

[92] Manik Varma and Bodla Rakesh Babu. More generality in efficient multiple kernel learning. In *ICML*, 2009.

[93] Eduard Vazquez, Theo Gevers, Marcel Lucassen, Joost van de Weijer, and Ramon Baldrich. Saliency of color image derivatives: A comparison between computational models and human perception. *Journal of the Optical Society of America A (JOSA)*, 27(3):1–20, 2010.

[94] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

[95] J. Vogel and B.Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007.

[96] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, 2006.

[97] Xioayu Wang, Tony X. Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*, 2009.

[98] Zhengxiang Wang, Yiqun Hu, and Liang-Tien Chia. Image-to-class distance metric learning for image classification. In *Proc. European Conf. on Computer Vision*, 2010.

[99] Dietrich Wettschereck, David W. Aha, and Takao Mohri. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11:273–314, 1997.

[100] Felix A. Wichmann, Lindsay T. Sharpe, and Karl R. Gegenfurtner. The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology:Learning, Memory, and Cognition.*, 28:509–520, 2002.

[101] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proc. IEEE Int. Conf. on Computer Vision*, 2005.

[102] J. M. Wolfe. *Visual Search.* 1998. in Attention, edited by H. Pashler, Psychology Press Ltd.

[103] J. M. Wolfe. *The Deployment of Visual Attention:Two Surprises.* Search and Target Acquisition, edited by NATO-RTO, NATO-RTO., 2000.

[104] J. M. Wolfe and T.S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5:1–7, 2004.

[105] J. Wu and J.M. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Proc. IEEE Int. Conf. on Computer Vision*, 2009.

[106] Jianxin Wu. A fast dual method for hik svm learning. In *Proc. European Conf. on Computer Vision*, 2010.

[107] Nianhua Xie, Haibin Ling, Weiming Hu, and Xiaoqin Zhang. Use bin-ratio information for category and scene classification. In *Proc. Computer Vision and Pattern Recognition*, 2010.

[108] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. Computer Vision and Pattern Recognition*, 2009.

[109] Jianchao Yang, Kai Yu, and Thomas Huang. Efficient highly over-complete sparse coding using a mixture model. In *Proc. European Conf. on Computer Vision*, 2010.

[110] Liu Yang, Rong Jin, Rahul Sukthankar, and Frederic Jurie. Unifying discriminative visual codebook generation with classifier training for object category recognition. In *CVPR*, 2008.

[111] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *Proc. Computer Vision and Pattern Recognition*, 2011.

[112] Bangpeng Yao, Aditya Khosla, and Li Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.

[113] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object catergories: An in-depth study. a comprehensive study. *Int. Journal of Computer Vision*, 73(2):213–218, 2007.

[114] Junge Zhang, Kaiqi Huang, Yinan Yu, and Tieniu Tan. Boosted local structured hog-lbp for object localization. In *CVPR*, 2010.

[115] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.