



**Universitat  
Autònoma  
de Barcelona**

# **Color for Object Detection and Action Recognition**

A dissertation submitted by **Rao Muhammad  
Anwer** at Universitat Autònoma de Barcelona to  
fulfil the degree of **Doctor en Informàtica**.

Bellaterra, May 22, 2013

Director	<b>Dr. Antonio M. Lopez</b> Dep. Ciències de la Computació & Centre de Visió per Computador Universitat Autònoma de Barcelona
Co-director	<b>Dr. Joost van de Weijer</b> Dep. Ciències de la Computació & Centre de Visió per Computador Universitat Autònoma de Barcelona



This document was typeset by the author using L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub>.

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

Copyright © 2013 by Rao Muhammad Anwer. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.

ISBN:

Printed by

Dedicated to



# Abstract

Recognizing object categories in real world images is a challenging problem in computer vision. The deformable part based framework is currently the most successful approach for object detection. Generally, HOG are used for image representation with in the part-based framework. For action recognition, the bag-of-word framework has shown to provide promising results. With in the bag-of-words framework, local image patches are described by SIFT descriptor. Contrary to object detection and action recognition, combining color and shape has shown to provide the best performance for object and scene recognition.

In the first part of this thesis, we analyze the problem of person detection in still images. Standard person detection approaches rely on intensity based features for image representation while ignoring the color. Channel based descriptors is one of the most commonly used approaches in object recognition. This inspires us to evaluate incorporating color information using the channel based fusion approach for the task of person detection.

In the second part of the thesis, we investigate the problem of object detection in still images. Due to high dimensionality, channel based fusion increases the computational cost. Moreover, channel based fusion has been found to obtain inferior results for object category where one of the visual varies significantly. On the other hand, late fusion is known to provide improved results for a wide range of object categories. A consequence of late fusion strategy is the need of a pure color descriptor. Therefore, we propose to use Color attributes as an explicit color representation for object detection. Color attributes are compact and computationally efficient. Consequently color attributes are combined with traditional shape features providing excellent results for object detection task.

Finally, we focus on the problem of action detection and classification in still images. We investigate the potential of color for action classification and detection in still images. We also evaluate different fusion approaches for combining color and shape information for action recognition. Additionally, an analysis is performed to validate the contribution of color for action recognition. Our results clearly demonstrate that combining color and shape information significantly improve the performance of both action classification and detection in still images.

# Abstracto

Detectar objetos en imágenes es un problema central en el campo de la visión por computador. El marco de detección basado en modelos de partes deformable es actualmente el más eficaz. Generalmente, HOG es el descriptor de imágenes a partir del cual se construyen esos modelos. El reconocimiento de acciones humanas es otro de los tópicos de más interés actualmente en el campo de la visión por computador. En este caso, los modelos usados siguen la idea de conjuntos de palabras (visuales), en inglés *bag-of-words*, en este caso siendo SIFT uno de los descriptor de imágenes más usados para dar soporte a la formación de esos modelos. En este contexto hay una información muy relevante para el sistema visual humano que normalmente está infrautilizada tanto en la detección de objetos como en el reconocimiento de acciones, hablamos del color. Es decir, tanto HOG como SIFT suelen ser aplicados al canal de luminancia o algún tipo de proyección de los canales de color que también lo desechan. Globalmente esta tesis se centra en incorporar color como fuente de información adicional para mejorar tanto la detección objetos como el reconocimiento de acciones.

En primer lugar la tesis analiza el problema de la detección de personas en fotografías. En particular nos centramos en analizar la aportación del color a los métodos del estado del arte. A continuación damos el salto al problema de la detección de objetos en general, no solo personas. Además, en lugar de introducir el color en el nivel más bajo de la representación de la imagen, lo cual incrementa la dimensión de la representación provocando un mayor coste computacional y la necesidad de más ejemplos de aprendizaje, en esta tesis nos centramos en introducir el color en un nivel más alto de la representación. Esto no es trivial ya que el sistema en desarrollo tiene que aprender una serie de atributos de color que sean lo suficientemente discriminativos para cada tarea. En particular, en esta tesis combinamos esos atributos de color con los tradicionales atributos de forma y lo aplicamos de forma que mejoramos el estado del arte de la detección de objetos. Finalmente, nos centramos en llevar las ideas incorporadas para la tarea de detección a la tarea de reconocimiento de acciones. En este caso también demostramos cómo la incorporación del color, tal y como proponemos en esta tesis, permite mejorar el estado del arte.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Color for Object Detection and Action Recognition . . . . .	2
1.1.1	Color Features . . . . .	2
1.1.2	Combining Color and Shape for Object Detection . . . . .	3
1.1.3	Combining Color and Shape for Action Recognition . . . . .	4
1.2	Objectives and Approach . . . . .	5
1.3	Outline . . . . .	8
<b>2</b>	<b>Object Detection and Action Recognition in still images: A Review</b>	<b>9</b>
2.1	Object Detection in still images . . . . .	10
2.1.1	Holistic Approach . . . . .	11
2.1.2	Part-based Approach . . . . .	12
2.1.3	Object Detection Data Sets . . . . .	14
2.2	Action Recognition in still images . . . . .	16
2.2.1	Bag-of-words based action recognition . . . . .	16
2.2.2	Action Recognition Data Sets . . . . .	18
2.3	Color for Object Recognition . . . . .	20
2.3.1	Color Descriptors . . . . .	20
2.3.2	Combining color and shape for object recognition . . . . .	22
2.4	Conclusion . . . . .	24
<b>3</b>	<b>Color Contribution towards Person Detection</b>	<b>25</b>
3.1	Introduction . . . . .	25
3.2	Opponent colors space . . . . .	27
3.3	Coloring Person Detectors . . . . .	28
3.3.1	Holistic Person detector . . . . .	28
3.3.2	Part-based Person detector . . . . .	29
3.4	Experiments and Results . . . . .	30
3.4.1	INRIA Person dataset . . . . .	30
3.4.2	PASCAL VOC 2007 dataset . . . . .	33
3.5	Conclusions . . . . .	36
<b>4</b>	<b>Color Attributes for Object Detection</b>	<b>37</b>
4.1	Introduction . . . . .	37
4.2	Related work . . . . .	39
4.3	Color attributes for object detection . . . . .	40
4.3.1	Color descriptors . . . . .	40
4.3.2	Color descriptor evaluation . . . . .	41
4.4	Coloring object detection . . . . .	42
4.4.1	Coloring part-based object detection . . . . .	42
4.4.2	Coloring ESS object detection . . . . .	43
4.5	Cartoon character detection . . . . .	44
4.6	Experimental results . . . . .	45

4.6.1	Results on the PASCAL VOC datasets . . . . .	45
4.6.2	Results on the Cartoon dataset . . . . .	46
4.7	Conclusions . . . . .	47
<b>5</b>	<b>Coloring Action Recognition in Still Images</b>	<b>55</b>
5.1	Introduction . . . . .	55
5.2	Related Work . . . . .	57
5.3	Color Descriptors for Action Recognition . . . . .	58
5.4	Combining Color and Shape for Action Classification . . . . .	59
5.4.1	Standard Late Fusion . . . . .	61
5.4.2	Standard Early Fusion . . . . .	61
5.4.3	Channel-based Early Fusion . . . . .	61
5.4.4	Classifier-based Late Fusion . . . . .	61
5.4.5	Color Attention-based Late Fusion . . . . .	61
5.4.6	Portmanteau Vocabulary-based Fusion . . . . .	64
5.5	Combining Color and Shape for Action Detection . . . . .	65
5.6	Experiments . . . . .	66
5.6.1	Action Recognition Datasets . . . . .	66
5.6.2	Coloring Action Classification . . . . .	67
5.6.3	Coloring Action Detection . . . . .	72
5.6.4	Analysis of Action Recognition Results . . . . .	73
5.7	Discussion and Conclusion . . . . .	75
<b>6</b>	<b>Conclusions and Future Directions</b>	<b>79</b>
6.1	Conclusions . . . . .	79
6.2	Future Directions . . . . .	81
	<b>Bibliography</b>	<b>83</b>



# List of Tables

2.1	INRIA person dataset Statistics for training set and test set. . . . .	14
2.2	A brief summary on PASCAL VOC Datasets for training set and test set images. .	15
2.3	Comparison of Action data sets on still images in terms of number of actions, images and training/test splits. . . . .	18
3.1	Training and testing numbers per scenario: person windows (+); images without persons (-); initial background windows (-) after sampling 200 one per image without persons; number of images for testing as well as persons to be detected. . . . .	33
3.2	Average precision (AP) in % of the different trained and tested classifiers. Indoor/Countryside/Urban/Overall in the first column refer to the training step, while Indoor/Countryside/Urban in the second row refer to testing. Bold numbers indicate the higher APs comparing the counterpart RGB and OPP results. For the overall classifiers we not only include the overall APs, but also the APs corresponding to apply such classifiers only to specific scenarios during testing. . . . .	35
4.1	Comparison of feature dimensionality of different approaches. Our proposed CN-HOG feature increases dimensionality to only 42 dimensions. The early fusion extensions of HOG based on computing the HOG on multiple color channels result in dimensionality of 93 (notations are similar to [75]). The LBP-HOG approach combines the LBP and HOG using late fusion and increases overall dimensionality to 90. . . . .	43
4.2	Average precision results for the baseline HOG detector [25], color descriptors proposed in the literature [75] and our proposed CN-HOG approach on all 20 classes of the PASCAL VOC 2007 dataset. Note that our approach along among existing fusion methods outperforms shape alone on this dataset. Our approach provides a significant improvement of 2.5% mean AP over the standard HOG-based framework.	45
4.3	Comparison with state-of-the-art results on the PASCAL VOC 2007 dataset. Note that the approach of boosted LBP-HOG [100] combines HOG and LBP together using a boosting strategy for feature selection. However, our proposed combination of color names and HOGs (CN-HOG) is compact, computationally inexpensive and uses no feature selection. . . . .	47
4.4	Average precision results for the baseline HOG detector [25], our proposed CN-HOG approach and state-of-the-art results on all 20 classes of the PASCAL VOC 2009 dataset. Note that our approach provides an improvement of 1.4 mean AP over the standard HOG-based framework. . . . .	47
4.5	Comparison of different fusion approaches using the part-based approach on the Cartoon dataset. Our CN-HOG approach yields a significant gain of 14% in mean AP over the standard HOG-based approach. Compared to early fusion approaches, our approach results in a gain of 6.4%. . . . .	48
4.6	Comparison of different approaches within the ESS detection framework. Similar to our results using the part-based method, combining color attributes improves the overall performance by 4%. Our CN-SIFT method also yields superior performance compared to the well known color descriptors OpponentSIFT and C-SIFT. . . . .	48

4.7	Comparison of different detectors on Barney. . . . .	49
4.8	Comparison of different detectors on Daffy. . . . .	50
4.9	Comparison of different detectors on Roadrunner. . . . .	51
4.10	Comparison of different detectors on Shaggy. . . . .	52
4.11	Comparison of different detectors on Tweety. . . . .	53
4.12	Comparison of different detectors on Tom. . . . .	54
5.1	Performance evaluation of pure color descriptors on the three datasets. Performance is measured by mean AP over the action categories. Note that on all three datasets the color names descriptor yields the best performance. . . . .	68
5.2	SIFT and channel-based color descriptors on the three action datasets. RGB-SIFT yields the best results on the Willow dataset, while the best performance on the PASCAL VOC 2010 and Stanford-40 action datasets is achieved using RG-SIFT. . . . .	68
5.3	Comparison of our fusion combination approach with state-of-the-art results on the Willow dataset. On this dataset, our approach provides best results on 4 out of 7 action categories. Moreover, we achieve a gain of 2.5 mean AP over the best reported results. . . . .	71
5.4	Comparison with state-of-the-art results on the PASCAL VOC 2010 test set. Despite the simplicity, our approach which combines several color-shape fusion strategies still provides comparable results to best methods on this dataset. Note that, unlike our technique, state-of-the-art approaches typically use standard object detectors to model person-object interactions. Such approaches are complementary to our method and can be combined to further improve results. . . . .	71
5.5	Comparison of color fusion combination with state-of-the-art results on Stanford-40 dataset. Note that combining fusion approaches yields a significant gain of 6.2 in mean AP over the best reported results in the literature. . . . .	71
5.6	Comparison of different detection methods on the Stanford-40 dataset. The best performance is achieved by CN-HOG with a significant gain of 5.8 mean AP over standard HOG. . . . .	73
5.7	The best performing color descriptors on the three datasets used for action classification in this chapter. Note that for all three datasets the color name descriptor is the best choice. . . . .	76
5.8	Fusion approaches ranked by performance on the three datasets for action classification. Note that late fusion using color names consistently provides the best performance. For Stanford-40 dataset we excluded color attention due to its high dimensionality. . . . .	76

# List of Figures

1.1	Example images with a set of object categories from the PASCAL VOC dataset. Four different categories of pottedplant, person, aeroplane and sheep has shown in different image variations such as illuminations, pose, occlusion, scale and background clutter etc. . . . .	2
1.2	Example images from two different data sets under varying illumination, shadows and specularities. Top row: images from PASCAL VOC 2007 dataset of different categories. Bottom row: images from the Stanford-40 dataset of two different action categories (playing guitar, playing violin). . . . .	3
1.3	A color image with its Opponent color space channel transformed images. In case of OPPHOG, the HOG operation is performed on each opponent channel respectively and their concatenation is feeded to the classifier as compared to conventional approach. . . . .	4
2.1	Example images of different object categories from the PASCAL VOC 2007 and PASCAL VOC 2010 dataset. Left column: object detection (bounding box with dashed line). Middle column: action classification (bounding box with normal line). Right column: action detection (bounding box with dashed line). Dashed line means it is not available at testing time. . . . .	10
2.2	An illustration of a typical holistic approach for person detection. An image is represented by HOG features. A discriminative classifier is trained using linear SVM. Consequently an image is scanned at multiple locations and scales to verify the presence/absence of a person category instance. . . . .	11
2.3	Models learned on PASCAL VOC 2007 dataset. The root and part-filter, with the part filters placed at the centre of the allowable displacements. . . . .	13
2.4	Example of Positive (persons) and negative (background) windows from INRIA person dataset. Note that the dataset only contains upright person in static images. . . . .	14
2.5	Example images from the PASCAL VOC datasets for object detection task. Bounding boxes indicates instances of different classes in the images. A wide range of pose, scale, occlusion and imaging conditions exists in images. . . . .	15
2.6	Example images for different action categories from the PASCAL VOC 2010 dataset. All images contains person with large variations in pose, illumination and background clutter. . . . .	16
2.7	Bounding boxes showing a three level pyramid on the action recognition dataset. Separate histogram is constructed for each cell and at the end concatenated to form one final histogram for the bounding box. . . . .	18
2.8	Example images from Willow action dataset showing 7 action classes: interacting with computer, photographing, playing music, riding bike, riding horse, running and walking. . . . .	19
2.9	Example images from Stanford-40 dataset showing 40 diverse daily person actions such as brushing teeth, cleaning the floor, reading book, throwing a frisbee, etc. . . . .	19
2.10	Example images from detection and action recognition showing the large variations in illumination, shadows and specularities. . . . .	20

2.11	Early and late fusion schemes to combine color and shape information within bag-of-words framework. The $\alpha$ and $\beta$ parameters determine the relative weight of the two cues. . . . .	23
3.1	Annotation enrichment for PASCAL VOC 2007 dataset. First, second and third rows show images that we have annotated as <i>indoor</i> , <i>urban</i> and <i>countryside</i> , resp. .	26
3.2	As compared to Regular HOG, channel-based HOG is computed on each channel and final representation is the concatenation of all channel-based HOG. . . . .	29
3.3	As compared to Regular HOG, channel-based HOG is computed on each channel and final representation is the concatenation of all channel-based HOG. . . . .	30
3.4	Positive (persons) and negative (background) windows from INRIA dataset. . . . .	31
3.5	Per window (top) and per image (bottom) evaluation using logarithmic scales. Values at usual points of interest are included, <i>i.e.</i> , $10^{-4}$ FPPW and $10^0$ FPPI, resp. . . . .	32
3.6	Precision-recall (PR) curves obtained from the different experiments are shown: using RGB and OPP color spaces, for the indoor, countryside, urban and overall classifiers. The average precision (AP) of each PR curve is the number shown inside the respective parenthesis. The PRs of the specific classifiers are plotted together with the PRs of the overall classifiers applied only in the corresponding specific scenarios. . . . .	34
4.1	Augmenting existing intensity based part-based detector with color information. .	38
4.2	Find the Simpsons. On the left, the conventional part-based approach [25] fails to detect all four members of Simpsons. Only Bart and Lisa are correctly detected, while Homer is falsely detected as Lisa and Marge is not detected at all. On the right, our extension of the part-based detection framework with color attributes can correctly classify all four Simpsons. . . . .	39
4.3	KL-ratio for the PASCAL VOC 2007 and the Cartoon dataset. The graphs clearly show that the color attribute (CN) is superior to the HUE and OPP descriptors in terms of both compactness and discriminative power. . . . .	42
4.4	Visualization of learned part-based models with color attributes. Both the HOG and color attribute components of our trained models are shown. Each cell is represented by the color which is obtained by multiplying the SVM weights for the 11 CN bins with a color representative of the color names. Top row: the HOG and color attribute models for pottedplant and horse. Bottom row: Marge and Tweety models. In the case of horse, the brown color of the horse together with a person sitting on top of it is prominent. Similarly, the model is able to capture the blue hair of Marge and orange feet of Tweety. . . . .	43
4.5	Example images with annotations from the new Cartoon dataset. The dataset consists of images of 18 different cartoon characters. . . . .	44
4.6	Precision/recall curves of the various approaches on six different categories from the PASCAL VOC 2007 dataset. Other than the bus category, our approach provides significantly improved performance compared to others. . . . .	46
5.1	Example images for different action categories from the PASCAL VOC 2010 dataset. These images illustrate the complications related to color description due to the large variation in illumination, shadows and specularities. . . . .	57
5.2	We apply a three level pyramid on the bounding boxes of the action recognition datasets. Separate BOW histograms are constructed for each cell and are concatenated to form the final action descriptor. In this chapter we use a pyramid representation with three levels, yielding a total of 14 cells. . . . .	60

5.3	Pipelines for four different fusion methods. The fusion between color and shape is indicated by a 'plus' in case of concatenation of vectors or vocabulary histograms. In the case of classifier based fusion, the encircled multiplication and sum symbols refer to the two methods of classifier fusion investigated: summation and multiplication, respectively, of their outputs. The function $f(RGB)$ refers to a mapping of $RGB$ values to another color-space representation. The "vocabulary" modules refer to vocabulary assignment and have histograms as output. Methods which perform fusion before vocabulary assignment are called early fusion methods, otherwise they are late fusion approaches. . . . .	62
5.4	Example portmanteau clusters from the Willow and Stanford-40 datasets. Note that each portmanteau cluster constitutes a distinct pattern of shape and color. Moreover, several clusters are representative of humans and specific actions such as gardening. . . . .	64
5.5	Visualization of learned part-based models using CN-HOG on the Stanford-40 action dataset. Both the HOG and color names components of our trained models combined in a late fusion are shown. Each color cell is represented using the color obtained by multiplying the SVM weights for the 11 CN bins with a color representative of the color name. Top row: the HOG models for riding horse, playing guitar, riding bike and rowing boat. Bottom row: color models of the respective categories. In the case of horse riding, the brown color of the horse in the bottom with a person sitting on top of it is evident. . . . .	66
5.6	Example images from the three datasets used to evaluate color descriptors and fusion techniques. Top row: images from the Willow dataset. Middle row: images from PASCAL VOC 2010 action recognition dataset. Bottom row: example images from the Stanford-40 dataset. . . . .	67
5.7	Performance comparison of different approaches to fusing color and shape. The choice of color descriptor is crucial for portmanteau-based image presentations. On all three datasets, late fusion performs better than early fusion, and the best results are obtained using late fusion with color names. . . . .	69
5.8	. . . . .	70
5.9	Precision/recall curves of the various channel based approaches, HOG and CN-HOG on six different action categories from the Stanford-40 dataset. Other than the writing on a board action category, CN-HOG provides significantly improved performance over channel based methods. . . . .	74
5.10	Confusion matrix for late fusion of shape and color names on the Willow dataset. Superimposed are differences with the confusion matrix based on luminance alone for confusions where the absolute change is at least 3%. Late fusion reduces the confusion among different categories in general, but particularly so in the interacting with computer and playing music categories. . . . .	75
5.11	. . . . .	77
5.12	Analysis of detection errors for the Standford-40 dataset. The graph shows the decrease in errors which occur when going from standard HOG to CN-HOG (negative changes in the graph signify an increase in errors). Errors are split into errors caused by misaligned localization, confusion with other classes, and detections on the background. . . . .	78



# Chapter 1

## Introduction

One of the most outstanding accomplishments of human visual system is the ability to understand and analyse the visual world with apparent ease. Complex visual recognition tasks are performed by humans on a daily basis without a significant effort. On the other hand, one of the long standing goals of computer vision is to mimic the human visual system by performing automatic image understanding. In this thesis we focus on two important image understanding tasks, namely object detection and action recognition.

Object recognition problem poses some specific questions of varying difficulty about real world images. The most classical question about the image are "What is it?" and "Where is it?" The first one, "What is it?" termed as "image classification" in literature and is the problem of predicting whether an image contains a certain object category or not. The second one, "Where is it?", is referred as "object localization or detection" in computer vision. The task is to simultaneously classify and localize objects in images. Object detection is a challenging task due to large variations in viewpoint, illumination and scale, pose and background in real world images. Figure 1.1 shows images of different object categories from the PASCAL Visual Object Classes (VOC) dataset having large amount of variations within a class and among different object categories. For example the second row in Figure 1.1 shows example images from the person category with significant variation in background and lighting conditions. All these factors make the task of object detection in real world images extremely challenging.

Only recently action recognition in static images has gained a lot of attention. Here the task is to identify the action a human is performing from a single image. Given an image in which a person is indicated by a bounding box, the task is to associate it with an action from a set of action labels. Human action recognition can be further divided into action classification and detection. In action classification, bounding boxes of persons performing actions are provided both at train and test time. Whereas in action detection, bounding boxes information is only provided at training time. The task here is to simultaneously localize and classify the action. Recognizing human actions in still images is a difficult problem due to lack of temporal information and intra-class variability (pose, clothing and occlusions).

The two fields discussed above, object detection and action recognition, were until recently solely based on luminance information, and ignored color information. Generally, low level features describing shape are used in state-of-the-art object recognition frameworks. However, it is very unlikely for a single cue to have the same discriminative power for all object categories. Contrary to computational vision systems, in human perception, color plays a pivotal role in visual search mechanisms [30, 50, 88]. Studies related to human visual perception reveal that color performs well when used in combination with luminance [44, 71, 91]. To improve object detection and human



**Figure 1.1:** Example images with a set of object categories from the PASCAL VOC dataset. Four different categories of pottedplant, person, aeroplane and sheep has shown in different image variations such as illuminations, pose, occlusion, scale and background clutter etc.

action recognition, color should be combined with conventional intensity based frameworks. On object categories where luminance alone provides inferior results, a fusion of color and shape can be expected to provide improved performance. Therefore, in this thesis, we investigate the problem of incorporating color information for object detection and action recognition.

## 1.1 Color for Object Detection and Action Recognition

The state of the art approaches to object detection and action recognition [11, 25, 51] ignore color information. In Figure 1.2, we give several examples why color information is difficult. The images show large variations in color caused by changes in illumination across different object categories. Moreover, changes due to shadows, highlights and specularities, etc. makes the task of robust color description even more complicated. In conclusion, the variability of color information with acquisition camera, illuminant changes, object instances makes its usage a challenging problem.

Over the last 5 years color has been introduced for object recognition and has been proven to lead to improved results [31, 75, 76, 81]. Many color features have been proposed [1, 75, 76, 78] and several new approach to combine color with shape information [46, 48, 78]. We will look here in more detail into two of the main aspect of color information usage, namely, the choice of color features and the approach to combining color and shape information.

### 1.1.1 Color Features

The color features which are used can be roughly be divided in three groups. The first approach, called color space transformation, is based on transferring the RGB color space to another color space such as e.g. opponent color space. After transformation, the shape feature is computed separately on the different channels and combined. The rationale behind these methods is that the color space transformation decorrelates the color and luminance information before applying the shape descriptor [7, 75]. This method was used with success for object recognition.





**Figure 1.2:** Example images from two different data sets under varying illumination, shadows and specularities. Top row: images from PASCAL VOC 2007 dataset of different categories. Bottom row: images from the Stanford-40 dataset of two different action categories (playing guitar, playing violin).

The second approach focuses on photometric invariance. There exists scene-accidental variations such as illumination changes, varying shadows and changing camera position in color images which affect the performance of color representation. Photometric invariance theory which provides a road map to ensure invariance with respect to such events [75, 76]. But photometric invariance comes at the expense of discriminative power. So one has to be careful in choosing a color descriptor which have the property of photometric invariance without sacrificing discriminative power.

The third approach is based on assigning linguistic color labels to image pixels. English language has 11 basic color terms namely black, blue, brown, grey, green, orange, pink, purple, red, white and yellow as shown in linguistic studies [5]. In computer vision, color naming involves assigning a label to each pixel in an image. Weijer *et al.* [78] proposes to learn color names from Google images by computing a mapping between RGB values and color names. Color names are photometric invariant since several shades of color are mapped to the same color name. As well as color names also provides description for achromatic colors such as black, grey and white.

The suitability of these different approaches for color description has not yet been investigated for object detection and action recognition.

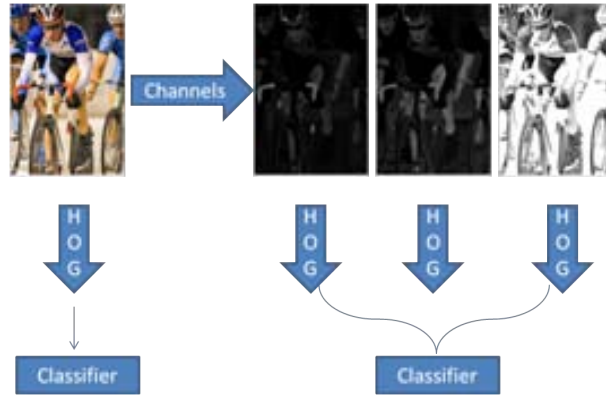
### 1.1.2 Combining Color and Shape for Object Detection

Deformable Part Model proposed by Felzenswalb *et al.* [25] has shown to provide state-of-the-art results for object detection. The part-based models focus on modelling an object as a collection of parts arranged in a deformable configuration using pictorial structure. The conventional part-based framework uses Histogram of Oriented Gradient(HOG) Dalal *et al.* [11] to represent an image. In such a framework color can be incorporated as channel based fusion or late fusion.

#### Channel based Fusion

In case of channel based fusion, color and shape features are combined at pixel level. Shape is computed on the color channels separately. The resulting feature vectors are concatenated into a single representation which is then an input to the classifier. This kind of approach has an implicit color description resulting into more discriminative power as compared to conventional approach. Channel based fusion scheme allows feature binding since both color and shape features are connected locally at the pixel level. However, channel based fusion scheme expands the feature dimensions exponentially, thus increasing the computation cost for the learning a model.

The channel based fusion approach has shown to provide excellent results for image classification



**Figure 1.3:** A color image with its Opponent color space channel transformed images. In case of OPPHOG, the HOG operation is performed on each opponent channel respectively and their concatenation is feeded to the classifier as compared to conventional approach.

task [75]. Several channel based color descriptor such as RGB-SIFT, Opponent-SIFT, RG-SIFT, C-SIFT (invariant of opponent) and HSV-SIFT are evaluated by Van de Sande *et al.* However, these color descriptors are yet to be evaluated for generic object detection task.

### Late Fusion

In contrast to early fusion, the two information sources are described separately from the beginning in late fusion. The separate histograms of color and shape are then combined into a single representation. The main advantage of such a scheme is its capability to learn the model independently from its elements of information. Furthermore, it is relatively easier to perform an adaptive cue weighting scheme for controlling the amounts of the contribution of each information. Late fusion holds the property of feature compactness since both color and shape have separate representations. This helps in recognizing man made classes such as car and chairs. The main disadvantage in late fusion is that the correlation among the source of information is lost as compared to channel based fusion [48].

In summary, combining color and shape has shown to provide excellent performance for object and scene recognition. The impact of color is yet to be evaluated for object detection. Combining color and shape information is expected to improve the performance of conventional intensity based part-based detection framework. However, the choice of color descriptor together with an optimal fusion strategy is crucial to obtain a performance gain.

### 1.1.3 Combining Color and Shape for Action Recognition

The bag-of-words approach is the most successful technique for action recognition in still images [12, 68]. The first stage called as, feature detection, involves detecting keypoint locations in an image. The second stage, feature extraction, involves describing the detected keypoint regions with local feature descriptors. Typically, the regions are described by extracting the shape information. In the vocabulary construction stage, the local shape features are then vector quantized into a fixed size visual vocabulary. A histogram over the visual words is then computed for each image which is then used to train a classifier for action recognition.

Contrary to action recognition, color has shown to provide excellent results when combined with shape for object recognition [1, 21, 46, 48, 75, 76]. Within the bag-of-words framework, a variety

of fusion approaches to combine color and shape information exist in literature. These fusion approaches differ in what stage the two visual cues are combined in the bag-of-words pipeline. The contribution of color for action recognition in still images is yet to be evaluated. Here we provide a brief introduction of different fusion approaches, popular in image classification, for combining color and shape cues.

### Early Fusion and Late Fusion

In case of early fusion, color and shape features are combined before the vocabulary construction stage. Separate color and shape features extracted in an image. The resulting features are concatenated into a single feature vector. This results in a joint color-shape visual vocabulary. Early fusion based visual vocabularies possess high discriminative power due to the fact that visual words are described by both color and shape cues. Early fusion provides impressive performance for categories where both visual cues are constant. For example many natural categories, such as flowers and animals, early fusion was found to provide improved results [48]. Whereas, late fusion involves combining visual cues at a later stage of the bag-of-words pipeline. In late fusion, separate visual vocabularies are constructed for color and shape cues. The two visual cues are then combined into a single representation at the histogram level. Late fusion is shown to provide impressive results for object categories where one of the visual cues varies significantly [48].

### Complex Fusion

An alternative approach to construct compact joint color-shape visual vocabularies is proposed by [46]. In this approach color and shape cues are processed separately and a product vocabulary is constructed. To counter the high dimensionality of product vocabulary, a discriminative compact visual vocabulary is learned by minimizing the loss in discriminative power caused by the clustering of visual words. Portmanteau fusion approach allows to weight the contribution of each visual cue efficiently.

The color attention approach [48] follows a similar pipeline as late fusion. Separate histograms are constructed for both color and shape cues. In color attention, color is used as a top-down cue to modulate the shape features. The top-down cue means that the modulation is based on class-dependent. Regions in an image more likely to contain an object category instance have higher weighting in the shape histogram. The final representation is a concatenation of class-specific color attention histograms.

As mentioned above, all these fusion approaches have been investigated for object and scene recognition. However, their contribution for the task of action recognition in still images has yet to be investigated. An optimal fusion approach is expected to further improve the action recognition performance.

## 1.2 Objectives and Approach

As discussed above combining color and shape has shown to provide excellent results for object recognition. However, the contribution of color has yet to be investigated for the task of object detection and action recognition in still images. In this thesis we investigate the introduction of color in three application domains: person detection, general object detection and action recognition. The two main aspects - what color description to use and how to combine the information - will be analyzed in these three settings. This analysis has led us to the following objectives of this thesis research.

**Color Representation for Person Detection:** We first investigate the problem of person detection in still images. State-of-the-art person detection approaches rely on intensity based features for image representation. As mentioned earlier, channel based descriptors have shown to provide state-of-the-art results for object recognition. This prompts us to evaluate incorporating color information using the channel based fusion approach for the task of person detection.

Our proposed approach presented in chapter 3 is more closely related to that of Van de Sande *et al.* [75] as we investigate the contribution of color for detecting pedestrians, where K. van de Sande *et al.* show that applying a scale invariant feature transform (SIFT [58]) to OPP space is the best *a priori* option in the context of image category recognition. Note, that HOG is a SIFT inspired descriptor. Typically, in HOG computation, the gradient information is computed on top of each color channel of RGB color space R, G, B and then, at every pixel, only the maximum magnitude among the RGB channels is picked thus throwing away the color information. Based on this observation, in this chapter we evaluate the opponent colors (OPP) space as a biologically inspired alternative for pedestrian detection. In our proposal, HOG are applied to each color channel individually and, then, the final representation is the combination of all. By feeding OPP space in both baseline framework of Dalal *et al.* and part-based of Felzenswalb *et al.*, we will obtain better detection performance than by using RGB space.

Furthermore, we are also curious in measuring if person detection performance can be hampered by the type of scenario where it is performed. Three relevant types of scenarios such as indoor, countryside and urban scene has been chosen for evaluation. Consequently, a comparison between the OPP and RGB color space, when *pugged-in* for HOG-part-based person detection is performed in each scenarios. Based on this evaluation scheme, we found that the benefits of OPP with respect to RGB mainly come for indoor and countryside scenarios, those in which the human visual system was *designed* by evolution. Holistic approach with OPP color is tested on INRIA Person dataset where as part-based tried on the PASCAL VOC 2007 dataset (only person class). To evaluate different scenarios we also augmented the PASCAL VOC 2007 dataset annotations with urban, countryside and indoor scenes.

**Color Representation for Object Detection:** The task of generic object detection is very challenging. Standard object detection approaches rely on intensity information while ignoring the color. The channel based fusion scheme significantly increases the computational overhead due to high dimensionality. Furthermore, for object categories where one of the visual cues vary significantly, channel based fusion has been found to obtain inferior results. On the other hand, late fusion is shown to provide excellent results over wide range of object categories. A consequence of using late fusion strategy is the requirement to use a pure color descriptor. As discussed earlier, a successful should be compact, possesses a certain degree of photometric invariance while maintaining discriminative power. Therefore, a compact color representation should reduce the computational cost without deteriorating the detection performance over wide range of categories.

In extending object detection with color information a number of considerations should be taken into account. First, from image classification there have emerged two main approaches to combining color and shape information for building discriminative models, namely early and late fusion [48, 70]. In early fusion, shape and color are combined at pixels and processed together. For late fusion, shape and color of the image are described separately and the exact binding between the two is lost. If we consider the dimensionality of, respectively, the shape and color feature to be  $S$  and  $C$ , then early fusion features have dimensionality  $F * S$ , while late fusion features  $F + S$ . Early fusion features have higher discriminative power, due to their preservation of the binding property, however due to their high dimensionality require more training data to learn discriminative models. For image classification, early fusion was generally found to obtain the best results [75]. However, recent results show that when considering spatial pyramids, late fusion methods obtain better results [17]. This is due to the fact that once the spatial cells become smaller the uncertainty introduced by describing the shape and color separately is reduced. In the limit, where spatial cells represent a single pixel, both early and late fusion are the same. Due to the

importance of description of the smaller cells of spatial pyramids for object detection [11, 37], we expect late fusion to obtain better results than early fusion.

Second, the main challenge in color representation is its large variation caused by scene-accidental events such as changes in illumination color or varying shadows. Photometric invariance theory provides guidelines on how to obtain invariance with respect to such events, however photometric invariants come at the cost of a loss of discriminative power. A successful color descriptor for object detection must balance discriminative power and photometric invariance. A third consideration is the memory usage by the combine shape-color feature representation. In conclusion, given the vast learning time of object detection models, the compactness of the color descriptor should be taken into account.

Among several color representations, color names or attributes have been studied in many disciplines such as psychology, philosophy, anthropology, and linguistics [36]. Most of these studies are based on Sapir–Whorf hypothesis which states that our perception of stimuli is dependent on the allocation of names. The relationship between colors and their names is a classic case study of the Sapir–Whorf hypothesis. In chapter 4, we investigate the use of color names as a compact and powerful representation to combining with shape for object detection. These color attributes are combined with shape information to increasing overall detection performance while maintaining the extremely low computational costs. Experiments are performed on PASCAL VOC 2007 and 2009 datasets. We also introduce a new challenging data set to localize different cartoon characters in images. The results obtained from our experiments clearly suggest that late fusion of shape with an explicit color representation provides superior results compared to methods that fuse color and shape information in early fusion at the pixel level.

**Color Representation for Action Classification and Detection:** Action recognition in still images is one of the major field of research in computer vision. Action recognition involves both localization and /or classification of the action. Only recently, action recognition in static images has gained attention. In this formulation of the action classification problem, the task is to identify what *action* a human is performing from a single image. Another formulation, which is called as action detection, we have to do both localization as well as classify the action. The bag-of-words model is a successful approach for action recognition in still images. Within the bag-of-words framework the de facto choice for feature extraction is using SIFT descriptor. As evident from the recent success of color and shape fusion for image classification, color is also expected to improve the performance for action recognition. This motivates us to investigate the contribution of color for action recognition. A variety of color descriptors together with various fusion strategies exist in image classification literature. Therefore, a proper taxonomy is required to evaluate these variety of features and fusion approaches within the context of action recognition.

In chapter 5, we perform a comprehensive evaluation of state-of-the-art color descriptors for action classification in bag-of-words frameworks and for action detection in deformable part-based models. We also investigate six different fusion approaches namely early fusion, late fusion, channel-based fusion, classifier fusion, color attention and portmanteau vocabularies for combing shape and color features. We also investigate whether different color-shape fusion approaches contain complementary information and combining them appropriately can further improve action recognition performance. For action classification, the experiments were conducted on the three challenging datasets: Willow, PASCAL VOC 2010 and Standford-40. Action detection experiments were performed on Standford-40 dataset. Our experiments both on action classification and detection demonstrate that combining color and shape outperforms other pure shape based approaches.

### 1.3 Outline

This dissertation is organized in the following chapters. The second chapter focuses on state-of-the-art review at most popular object detection and action recognition frameworks. The third chapter is about coloring person detectors. We also re-evaluate their performance in different scene scenarios. Chapter 4 aims at solving the problem of generic object detection in still images. Chapter 5 focuses the problem of action recognition in still images. Finally, the thesis is concluded with chapter 6 which provides final conclusions and discussions about future work.

# Chapter 2

## Object Detection and Action Recognition in still images: A Review

---

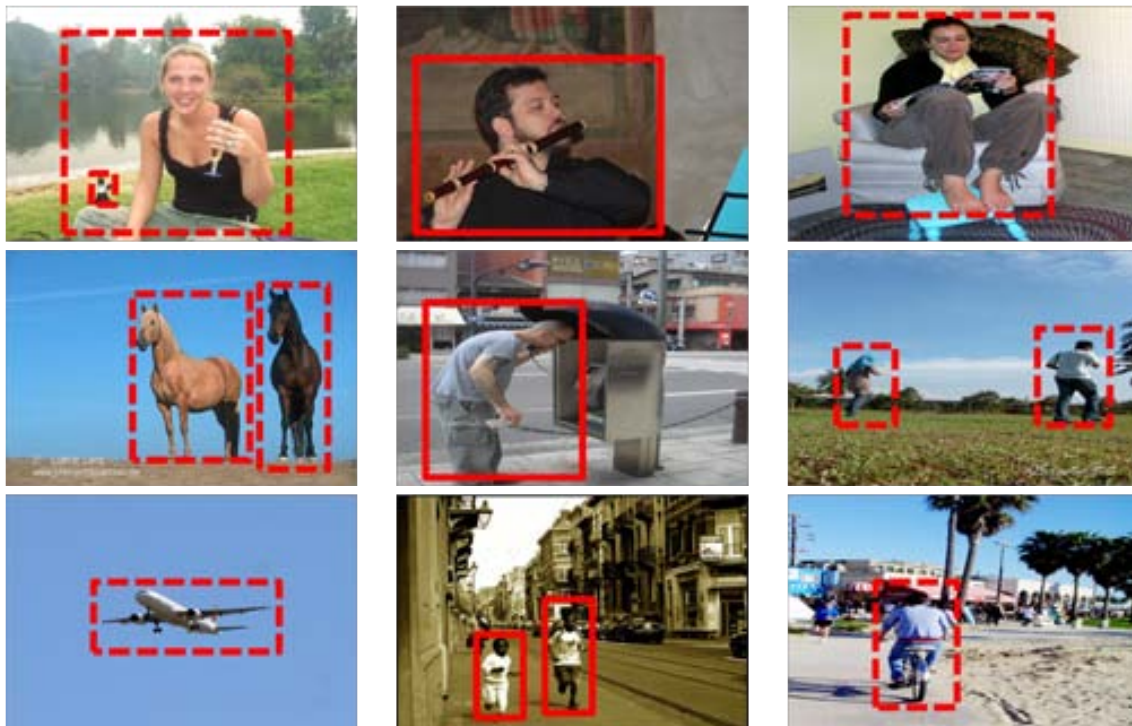
Regardless of many challenges, in recent years, object detection remains one of the most active research area in computer vision. Several approaches are proposed in literature includes holistic based, part-based, patch-based etc. The bag-of-words framework is most successful approach for action recognition in still images. Typically within the bag of words framework, local shape features are extracted for image representation. Contrary to action recognition, combining multiple cues specially color and shape have shown to provide excellent performance for image classification. In this chapter, we provide an overview on the commonly used approaches for object detection and action recognition in still images. We also provide an overview of variety of color descriptors commonly used for image classification.

---

Object detection is the task involving with determining the location of a concern object category in an image. Given an image, the task is to simultaneously classify and localize an object category instance. Significant amount of research have been done to develop the representation schemes and algorithms to identify generic objects in videos and images since many applications rely on object detection such as image retrieval, automotive safety and video surveillanc. However, object detection is a difficult task largely due to invariant with regard to changes in size, position and viewpoint of the object. Several other factors such as occlusion, intra-class variation and cluttered background also affect the overall performance of an object detector and thus make object detection even a more complicated problem.

The action recognition task addresses the problem of action classification and action detection in images or videos. Action classification involves predicting an action category given the bounding box of a person both at training and test time. The task is to associate an action label with each bounding box. In the case of action detection, the bounding box information is only available at training time. The action detection task is to simultaneously localize and classify the person associated action in an image. Action recognition is a difficult problem due to significant amount of pose, viewpoint and illumination variations.

In this chapter, we provide a detail overview on commonly used approaches for object detection and action recognition in still images. Figure 2.1 shows some example object categories in the PASCAL VOC 2007 and PASCAL VOC2010 datasets with ground truth localization. We start



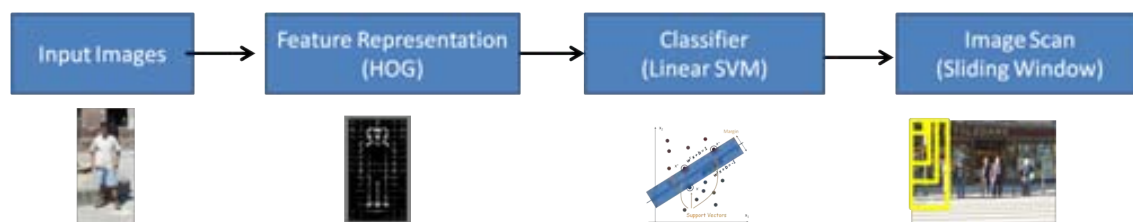
**Figure 2.1:** Example images of different object categories from the PASCAL VOC 2007 and PASCAL VOC 2010 dataset. Left column: object detection (bounding box with dashed line). Middle column: action classification (bounding box with normal line). Right column: action detection (bounding box with dashed line). Dashed line means it is not available at testing time.

with object detection approaches such as holistic, part-based in section 2.1. Whereas, section 2.2 provides an overview on the most commonly used approach namely the bag-of-words framework for action classification. Finally section 2.3 gives an insight on the variety of color descriptors commonly used in image classification literature.

## 2.1 Object Detection in still images

Despite many challenges, object detection problem has been investigated for a long time and many techniques have been proposed in the last decade or so to tackle the problem of object detection. These techniques can be grouped according to the kind of search method employed. Most often, a window-sliding approach is employed which treats the object detection problem as a binary classification task. The classifier is used to classify a window of an image, at all location and scales, as either it contains the specific object or not. At first, we describe the overall procedure of Dalal and Triggs [11] framework based on holistic approach. Baseline framework of Dalal and Triggs [11] relies on histogram of oriented gradients (HOG) as feature and linear support vector machines (linear SVM) as learning algorithm. After words, we explain the deformable part-based approach proposed by Felzenswalb *et al.* [25] which uses HOG as low level features along with latent support vector machines (LatSVM). Finally, an overview is provided on major object detection data sets used for object detection in still images.





**Figure 2.2:** An illustration of a typical holistic approach for person detection. An image is represented by HOG features. A discriminative classifier is trained using linear SVM. Consequently an image is scanned at multiple locations and scales to verify the presence/absence of a person category instance.

### 2.1.1 Holistic Approach

Holistic approach is the dominant paradigm in object detection and is conceptually based on scanning entire image at multiple position and scales by some fixed size of window [9, 11, 26, 59, 84]. An object detector is learnt from a training set by using some specific features and learning paradigm. Then, this trained classifier is used to examine the existence of an object in each window. Hence, the key factors involved in designing a holistic detector are 1) choice of good feature and 2) the selection and training of a proper classifier. A brief overview on feature extraction, learning classifier, scanning the image with non maxima suppression is shown in Figure 2.2 and described as follows:

#### Feature Extraction

Histogram of Oriented Gradient (HOG) [11] has shown superior performance to capture the edges or local shape information as compared to other low level features on multiple data sets. An object appearance is described in terms of local intensity gradients or edge directions in HOG feature. As a first step, the preprocessing of normalizing color and gamma values is performed on an image while computing HOG descriptor as mention in Dalal *et al.* [11]. As a next step, the gradient image is computed. The gradient image is divided into 8x8 grid of cells non overlapping to each other. Each cell then compile a 1D histogram of gradient orientations or edge orientations over pixels in that cell. The direction of the gradient together with the magnitude vote into a fixed number of predefined bins. In order to account for better invariance to illumination and contrast, the local group of cells must be locally normalized. A group of 2x2 cell, called blocks, are L2 normalized. Since the blocks overlap to each other, each cell contributes more than once in the final descriptor. Each normalized block descriptor is referred as Histogram of Oriented Gradient. Finally, all normalized block descriptor covering the detection window are combined into a final feature vector for use in the window classifier.

#### Learning a classifier

A linear SVM with HOG feature is used as a classifier which has been proved a competitive method. SVMs (Support Vector Machines) are a useful technique for data classification for the past decade. The basic SVM separate a set of input data by a hyperplane that maximises the margin between the object and non object class. For given a set of training examples, each of them belonging to one of two classes. The goal of SVM algorithm is to build a model based on the training data which predicts about the new examples one class or the other.

## Image scanning

The basic idea behind holistic approach is to scan entire image by each and every location with a fixed size window. In order to decide this exact position of an object within an image, we have to slide a window over each position but at multiple scales. To perform a multi-scale detection, the extended *pyramidal sliding window* strategy was proposed in Dalal [10]. In this process, the original image is scale by a factor  $s^i$  to obtain the image corresponding to the pyramid level  $i$ . For every pyramid level, we must slide the search window along the horizontal and vertical directions with a fixed stride pixels. Smaller the scale and stride pixel, the finer the sliding window search for better detection performance. However, this exhaustive search comes at the expense of higher processing time.

Since image scanning is performed at multiple scales, a single object can be detected several times at slightly different positions and scales. Multiple overlapped detections should be merged to get a unique detection per object by a clustering or *non-maximum-suppression* procedure.

### 2.1.2 Part-based Approach

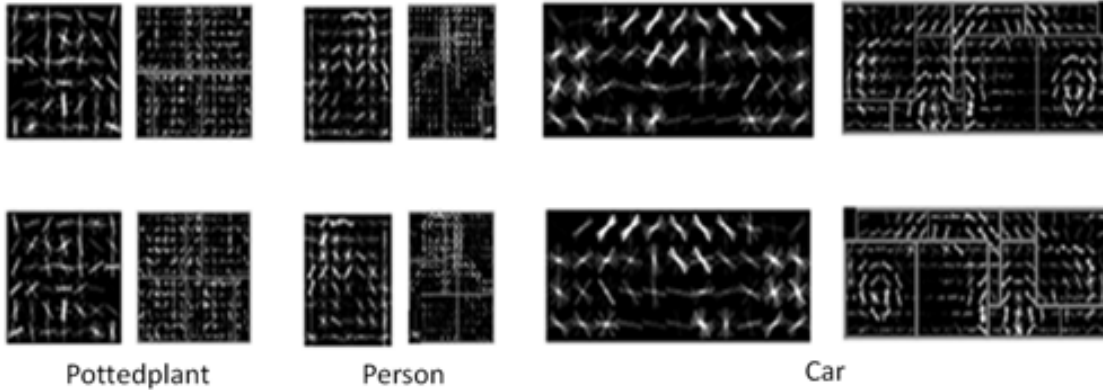
Since all objects are not in box shape, the idea to represent an object as collection of parts which have flexible spatial relationship was introduced by Fischler and Elschlager in 1973 [27]. The most successful approach for part-based is proposed by Felzenswalb *et al.* [25]. Deformable Part Model(DPM) of Felzenswalb *et al.* [25] uses pictorial structure framework as a building block for part based models. The DPM detectors and its invariant participated in PASCAL Visual Object Challenges between 2007 and 2012 have shown top performances as compared to other approaches.

The method consists of a collection of deformable part models for each class. A root model, which also serve as a reference point, is connected to a number of flexible parts in a star structure fashion. The root model in DPM is comparable to Dalal *et al.* [11] model but HOG feature used in DPM is changed. A summary of features and filters define for DPM is as follows

#### Features and Filters

To represent an object model, a new HOG feature is used with reduced dimensionality and high sensitivity to contrast direction which helps in improving accuracy and increasing the overall speed of the detector. Instead of 36 dimensional HOG features they used 13 dimensional feature, capturing 9 gradient orientation under single normalization as compared to four in Dalal *et al.* [11] and 4 additional features capturing texture gradient. For a final classifier, a 31 dimensional feature vector is used which consists of 9 contrast insensitive and 18 contrast sensitive orientations plus 4 texture gradients. A HOG feature pyramid is computed on an image pyramid going from coarser to finer resolution.

All the models, root and part, involve linear filters that is a rectangular template of weight vectors from HOG feature pyramid. The HOG feature pyramid coarser resolution, which captures the entire object, describe the root filter and finer resolution in HOG feature pyramid is used for defining deformable part filters which are more localized as compared to root. The score of model placement in HOG pyramid is a dot product between a set of weights and HOG feature within window are computed for both root and part filters.



**Figure 2.3:** Models learned on PASCAL VOC 2007 dataset. The root and part-filter, with the part filters placed at the centre of the allowable displacements.

### Model Learning

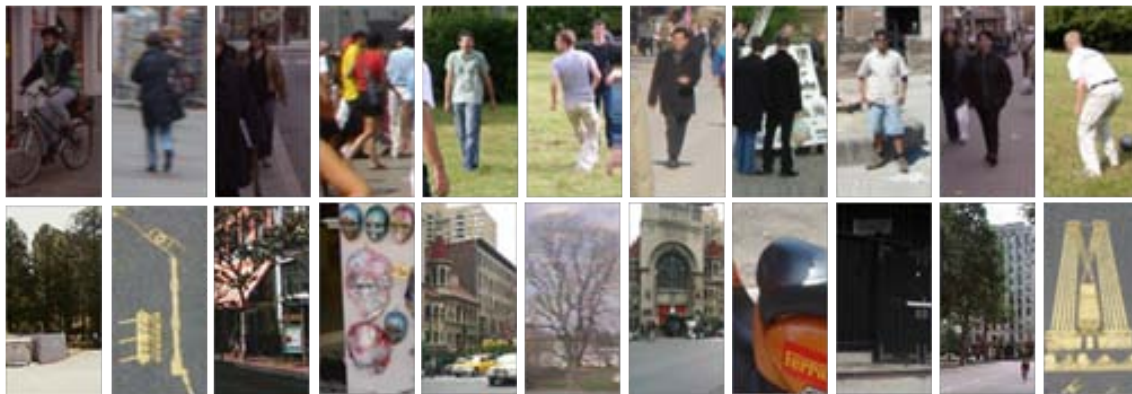
Modeling an object with parts, the training images only contain information about the bounding boxes of concern object. Since there is no information about the part locations of an object in training images, there must be a simple and effective strategy for learning effective parts. During a training procedure, DPM deal these parts location as a latent variables as well as the exact bounding box location of object.

To learn a model from weakly-labeled data, a generalized SVMs is introduced to handle latent variables such as part and object position which termed as latent variable SVM (LSVM). A LSVM is formulated from MI-SVM with introduction of extra latent variable. A triplet of  $\langle x_i, v_i, y_i \rangle$  is applied iteratively where  $v_i$  is selected latent variable for  $x_i$  under the model learned. The number of latent values for positive examples can be specified in LSVM, which makes training problem convex.

Within part-based model, each category is separated into multiple aspect ratios such as frontal, rear, side views etc. The poses and size of root filters are based on the statistics of bounding boxes in training images. The positive samples for initial root filters are scaled according to its size and aspect ratio and random sub-window negative examples are extracted from negative images. Multiple root filters with no latent variables are trained using an SVM. The trained root filters is used to find the best scoring place for the filters which significantly overlap with the bounding boxes in original positive images. The model is retrained with these new positive samples and original random negatives.

After retraining, multiple trained root models are merged into one mixture model with no parts. The combined model is retrain using latent SVM by alternating the latent mixture parameter for each positive example and estimating the latent objective function on the full dataset. In case of mixture model, component label and root location are only latent variables for each sample.

The parts for each component is initialized using a simple heuristic from the models trained above. The number of parts is fixed at 6 per component and are searched over by interpolating each root filter to twice the resolution. A rectangular area of size 6x6 is greedily selected from the root filter that has the most positive energy weights. Once a part is placed, the weights in that region of root filter are set to zero and the process is repeated until six parts are selected. The deformation cost for each part is also initialized. The model selected with parts for each component is trained with latent SVM. The model is updated ten times with positive and hard negative samples. The new positive samples are selected from positive images by applying the existing model at all position and scales with a 50% overlap of given bounding box. Whereas negative samples are selected from



**Figure 2.4:** Example of Positive (persons) and negative (background) windows from INRIA person dataset. Note that the dataset only contains upright person in static images.

negative images until a file size limit of 3GB memory is reached. To train such a complex model for a single category take over 3-5 hours on modern, multicore computer. Some of the final trained model is shown in Figure 2.3.

### 2.1.3 Object Detection Data Sets

A variety of dataset for object detection task exists in literatures [11, 19, 22, 33, 62, 93, 94]. Within these data sets, the INRIA person dataset and PASCAL Visual Object Classes (VOC) data sets are the defacto standard for person and object detection. Here, we provide an overview on INRIA person dataset and PASCAL VOC data sets.

#### INRIA Person dataset

The dataset is challenging due to large variations in pose, clothing, occlusion as well as complex background. The dataset contains upright person in static images as shown in Figure 2.4. The INRIA person dataset provides a 614 positive training images containing 1208 standing persons. Negative windows for training are extracted from the given 1218 background images. The testing set consists of 453 negative images and 288 positive images containing 566 persons in still images as shown in Table 2.1.

Train	Test
614 positive images	288 positive images
1218 negative images	453 negative images
1208 positive windows	566 positive windows

**Table 2.1:** INRIA person dataset Statistics for training set and test set.

**Evaluation Criteria:** The INRIA person dataset is widely used as benchmark to evaluate person detection. Two kind of evaluations are used namely *per window* and *per image*. In per window evaluation, person classifier is applied to the test truth positive windows and the test negative full images. Whereas, per images evaluation leads to better performance as compared to per window in real world applications [16]. In per image performance, a person detector is run as a whole on test positive and test negative images. Predicted detections are considered as correct if it overlap 50% with the ground truth.



**Figure 2.5:** Example images from the PASCAL VOC datasets for object detection task. Bounding boxes indicates instances of different classes in the images. A wide range of pose, scale, occlusion and imaging conditions exists in images.

### PASCAL VOC Data Sets

PASCAL Visual Object Classes (VOC) is a series of dataset introduced every years since 2005 till 2012. The aim of PASCAL VOC data sets was to provide a benchmark for object recognition, detection and segmentation with consistent annotation and standardized evaluation procedure. Since 2007 till 2012, the dataset contains 20 different classes namely aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, pottedplant, sheep, sofa, train and tv-monitor. The data sets provided under PASCAL VOC are challenging because large variations in object pose, illumination and appearance as shown in Figure 2.5. In our thesis, the PASCAL VOC data sets of 2007 and 2009 have been used for object detection in still images.

The PASCAL VOC 2007 dataset consists of 9963 images of 20 different classes with 5011 training images and 4952 test images. The PASCAL VOC 2007 dataset was the last year when annotation was released for test set. For scenario-based experiments, we further annotated the person class with three different scenarios namely indoor, urban and countryside both for training and testing sets. The PASCAL VOC 2009 dataset consists of the previous year's PASCAL VOC data sets images augmented with new images. It includes 13704 total images of 20 different classes with 7054 training images and 6650 test images. Since test set annotation is not available for this dataset, the results have to be submitted to the organizes directly. PASCAL VOC challenges data sets are given below in Table 2.2.

Dataset	Training	Testing	Total
PASCAL VOC 2007	5011	4952	9963
PASCAL VOC 2009	7054	6650	13704

**Table 2.2:** A brief summary on PASCAL VOC Datasets for training set and test set images.

**Evaluation Criteria:** The interpolated average precision (AP) is used to evaluate both classification and detection task. For detection task, the detection are judged true or false positive based on the area of overlap with ground truth bounding boxes. Detection would be marked correct if the predicted bounding box  $B_p$  with ground truth bounding box  $B_{gt}$  exceed 50% by the following



**Figure 2.6:** Example images for different action categories from the PASCAL VOC 2010 dataset. All images contains person with large variations in pose, illumination and background clutter.

formula:

$$a = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} \quad (2.1)$$

In summary, we analyze the most successful approaches namely holistic, part-based and benchmark data sets used for object detection in still images.

## 2.2 Action Recognition in still images

Action recognition involves the task of action classification and action detection in images. In action classification, the bounding box information is provided at the training and testing time. Whereas in action detection, the bounding box information is only provided at the training time. Given a test image, the task is to simultaneously classify and localize a person and the associated action. Figure 2.6 shows some examples of different actions performed in varying conditions. The deformable part-based models (DPM) object detector of Felzenszwalb has been used for action detection in still images as described above in section 2.1.2. For action classification, the bag-of-words framework has been the most applied approach providing promising results as explained below.

### 2.2.1 Bag-of-words based action recognition

The bag-of-words based models originally developed for text classification is an orderless representation that represent images as frequencies of words. The approach has shown to provide excellent results for object recognition and detection tasks [37, 51, 75, 81]. The framework pipeline includes following steps: feature detection, feature extraction, visual vocabulary and histogram construction. The resulting histogram are then used as an input to train a classifier to recognize object categories. An overview on each of these stages is as follows:

## Feature Detection

Feature detection is an initial low level image processing operation of detecting key points or regions in an image. Several feature detectors have been developed which can be divided into two broad groups of dense sampling and interest point sampling strategies. The dense sampling process an image at a single or multiple scale at fixed location forming a grid of rectangular windows. Whereas interest point focus on searching salient points in an image such as corners, blobs etc. Several interest points detectors have been proposed in literature like Harris-Laplace point detector, Laplacian, Laplacian-of-gaussian etc and color saliency boosting is the most commonly used color based interest point detectors.

## Feature Extraction

Once features have been detected, the next step involves describing the extracted local image patch around the feature. To perform a reliable object description, a descriptor should have the power to tackle intensity, rotation, scale and affine variations. In literature, many features such as shape, color, texture have been used to describe visual information.

Scale Invariant Feature Transform(SIFT) proposed by Lowe *et al.* [58] has proven successful within bag-of-words approaches for better discriminative shape information. SIFT descriptor are invariant to translation, rotation, scale and illumination. In practice SIFT descriptor are computed on grey level images thus ignoring color information in image. This method divides the normalized local patches into 4x4 grid of cells with an 8-bin orientation histogram for each cell resulting into 128 dimensional feature vector for each local patch.

## Visual Vocabulary and Histogram Construction

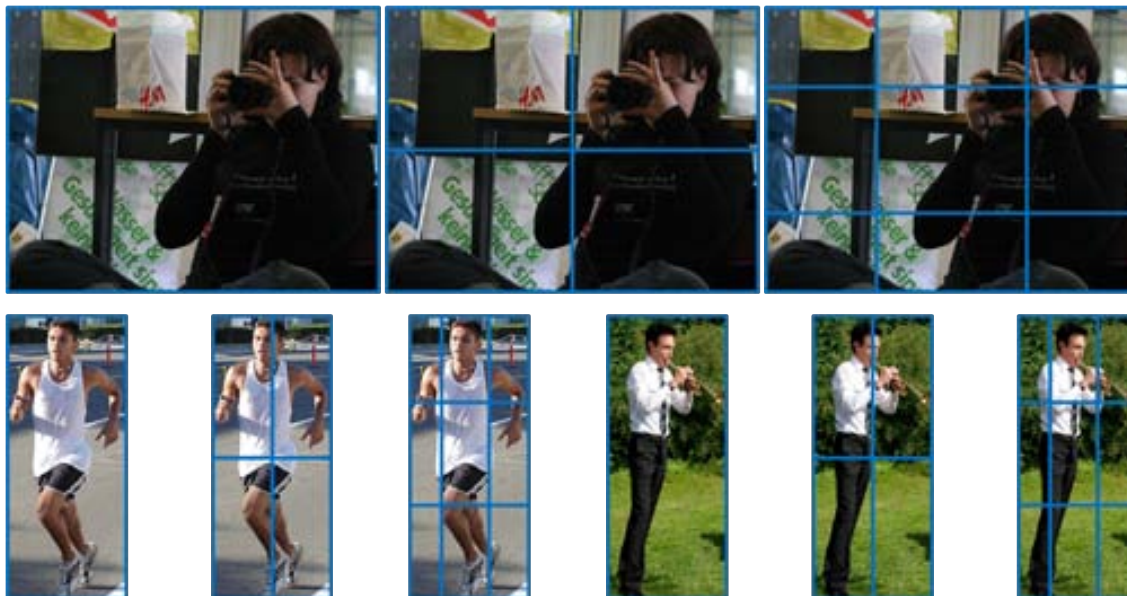
The next step within the bag-of-words framework is conversion of visual feature vectors of several local patches into visual vocabulary. A visual vocabulary is constructed on training set by performing k-means clustering over all the feature vectors. K-means clustering initialize k cluster centers by randomly selecting descriptor points. Each descriptor point is assign to the nearest center and cluster centers are then updated according to mean of all points in the cluster. The process is repeated for a fixed amount of iterations. Visual vocabulary size is equal to the number of clusters in k-means algorithm.

Finally, the image can be represented by the histogram of visual words from computed visual vocabulary. The histogram is constructed by counting the number of instances of visual words in an image which are then used as input for training a classifier.

As a final step, the constructed histogram for training set are provided to a machine learning algorithm for classification with class labels. In most cases, Support Vector Machines are used in bag-of-words framework for classification whereas we apply a nonlinear SVM with the  $\chi^2$  kernel [99]. The trained model is then employed on a given test image to predict he category label of the image.

## Spatial Pyramid Representation

The conventional bow approach represents an image as an orderless histogram of visual words. The representation ignores the spatial information crucial for image classification. To incorporates spatial information with in the bow Lazebnik *et al.* [54] proposed to construct spatial pyramid based image representation. The spatial pyramid techniques works by repeatedly dividing an image into sub-windows. Consequently a histogram is computed for each sub-window. The final representation is the concatenation of spatial pyramid histogram of each level. This simple yet



**Figure 2.7:** Bounding boxes showing a three level pyramid on the action recognition dataset. Separate histogram is constructed for each cell and at the end concatenated to form one final histogram for the bounding box.

powerful representation has shown to provide state of the results and is the key ingredients in every image classification framework [8, 17, 75, 95, 101].

Figure 2.7 shows a spatial pyramid representation for action recognition task. Typically, the bounding box information of a person is provided at both training and test time. Local features are computed within the bounding box consequently spatial pyramid representation is used to obtain the final representation.

## 2.2.2 Action Recognition Data Sets

This section briefly introduces the three challenging publicly available and commonly used standard action recognition data sets. We make use of Willow, PASCAL VOC 2010 and Stanford-40 action recognition data sets. The bounding boxes for the person performing an action are provided with the data sets except test set labels for PASCAL VOC 2010 dataset is not available. We perform action classification on all three data sets whereas action detection task is done on only Stanford-40 dataset. A comparison of action data sets is shown in Table 2.3.

Dataset	No of Actions	No of Images	Training	Testing
Willow	7	911	490	421
PASCAL VOC 2010	9	908	454	454
Stanford-40	40	9532	4000	5532

**Table 2.3:** Comparison of Action data sets on still images in terms of number of actions, images and training/test splits.

**Willow Action Dataset:** The Willow action dataset composed of 7 different classes for action classification in still images namely interacting with computer, photographing, playing music, riding bike, riding horse, running and walking as shown in Figure 2.8. The dataset has a total of 911 images. Each action class contains at least 109 persons, split into 70 examples per class for





**Figure 2.8:** Example images from Willow action dataset showing 7 action classes: interacting with computer, photographing, playing music, riding bike, riding horse, running and walking.



**Figure 2.9:** Example images from Stanford-40 dataset showing 40 diverse daily person actions such as brushing teeth, cleaning the floor, reading book, throwing a frisbee, etc.

training and the rest left is for testing. The mean average precision (mAP) is used for performance evaluation for action classification.

**PASCAL VOC 2010:** The PASCAL VOC 2010 dataset holds 908 images equally divided in to training and test sets. The dataset is challenging with 9 different action categories: phoning, playing instrument, reading, riding bike, riding horse, running, taking photo, using computer and walking. Figure 2.6 shows examples from PASCAL VOC 2010 dataset. The average precision (AP) is used as principal quantitative measure for action recognition in still images. Test set ground truth is not available for this dataset, the results have to be submitted to the organizes directly.

**Stanford-40:** This is one of the most challenging dataset used for action classification having large variations in person pose, appearance, and background clutter. There are 9532 images in total with 180 – 300 images for each action class. Stanford-40 action dataset has a total of 4000 training images and 5532 testing images in 40 action classes such as jumping, repairing a car, cooking, applauding, brushing teeth, cutting vegetables, throwing a frisbee, etc. The large number of action categories make this dataset particularly challenging for both action classification and action detection task in still images. Examples of the images are shown in Figure 2.9.

In summary, we analyze the most successful approaches for action classification namely bag-of-words approach and data sets used for action recognition task in still images.



**Figure 2.10:** Example images from detection and action recognition showing the large variations in illumination, shadows and specularities.

## 2.3 Color for Object Recognition

Conventionally intensity based descriptors have been employed to perform object detection and action recognition while ignoring color information. Surprisingly color information have shown to provide improved performance when combined with intensity based descriptors in image classification. Color descriptors has received relatively low attention in object detection and action recognition. This suggest the need of investigating the theoretical implications of color in object detection and action recognition in still images.

Even though there is a general agreement that color information should be exploited for object recognition there are several factors which complicate its usage. Color measurements vary with scene accidental variations, such as shadows, shading specularities and illuminant changes. Figure 2.10 shows some examples images from object and action recognition datasets. The image clearly illustrate the complications related to illumination, shadows and specularities for each category. In addition, color measurements vary with different acquisition system and compression algorithms used. In this section we review existing color descriptors and several methods to introduce color into an object recognition system.

### 2.3.1 Color Descriptors

Variety of color descriptors exists in image classification literature. This can be further divided into different categories namely Physics based color descriptors, Linguistic color descriptors and

Channel based color descriptors. Here below, we discuss each of these descriptors as follows

### Physics-based color descriptors

In this section, we concentrate on color descriptors derived from physical models. As mentioned in [76], local color descriptor should have ability to handle photometric changes such as shadows, shading, specularities and object reflectance changes. These events are mostly modeled by the dichromatic reflection model introduced by Shafer [66]. The model composed of two additive components of specular reflectance and a body reflectance parts. The specular reflectance describes the light portion that is bounced back immediately from the surface of the object causing specularities whereas body reflectance describes the light which is reflected after interacting with material body. So the model in the mathematical equation can be written as

$$f = m_b * C_b + m_s * C_b \quad (2.2)$$

The parameter  $f$  is the  $(R, G, B)$  factor, and  $m_b, m_s$  are the geometrical terms denoting geometry changes in scene (angle of incidence light, objects orientation, viewing, etc.). The dichromatic reflection model can be used to systematically derive the photometric invariance of color features.

The opponent color can be defined as

$$\begin{pmatrix} O1 \\ O2 \\ O3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (2.3)$$

Opponent color is invariant with respect to specularities, blur, diffuse illumination as shown in [35, 76] but still variant for lighting geometry variations. Lighting geometry and specularities invariance can be obtained by hue,

$$hue = \arctan \left( \frac{O1_x}{O2_x} \right), \quad (2.4)$$

where  $O1_x$  and  $O2_x$  are two chromatic opponent channels. The hue descriptor [76] on an image patch is constructed from the corresponding RGB values of each pixel according to:

$$hue = \arctan \left( \frac{\sqrt{3}(R - G)}{R + G - 2B} \right). \quad (2.5)$$

The histogram is weighted by the corresponding saturation pixel to counter instabilities in hue. In hue descriptor, error propagation analysis is performed to the hue transformation which shows that hue is inversely proportional to saturation. The hue descriptor is invariant for specularities and shadow effects while assuming white illumination. Hue descriptor has 36 dimensions.

### Linguistic color descriptors

There is a wealth of evidence showing that color lexicons vary in size in different languages as shown by Berlin and Kay [5]. All languages have 2 to 11 basic color terms except Russian and Hungarian. English has 11 basic color terms namely black, blue, brown, grey, green, orange, pink, purple, red, white and yellow.

Color names (CN) also known as color term, is a word that point towards a specific color. In computer vision, a color name descriptor is designed to mimic the usage of color terms in human language by Weijer *et al.* [78]. Color name descriptor is constructed by mapping learned from Google images to transform RGB to a probability over the color names. Thus allows to

represent patches as histogram over the eleven color names. Color name display a certain amount of photometric invariance since values with similar hue and saturation are mapped to the same color name. As well as it provide description of achromatic colors such as black, grey and white which are not photometrically invariant, but which improve the overall discrimination power of the descriptor. They have the additional advantage of being a very compact representation at only 11 dimensions.

### Channel based color descriptors

Channel based color descriptors are an other way of computing shape descriptors on different color spaces. In image classification, Bosch *et al.* [6] computed SIFT descriptors over all three channels in HSV color space. Weijer *et al.* [76] proposed to concatenate SIFT descriptor either with weighted hue or opponent angle histograms. Whereas van de sande *et al.* [75] performed a relative study on color descriptors. In this work, several channel based descriptors such as RGB-SIFT, Opponent-SIFT, RG-SIFT, C-SIFT(invariant of opponent) and HSV-SIFT are computed on respective color space channels. Experimental evaluation has shown that an Opponent-SIFT provide superior performance for object recognition task. Opponent-SIFT is computed on the three channels of opponent color spaces as described in Eq.2.3.

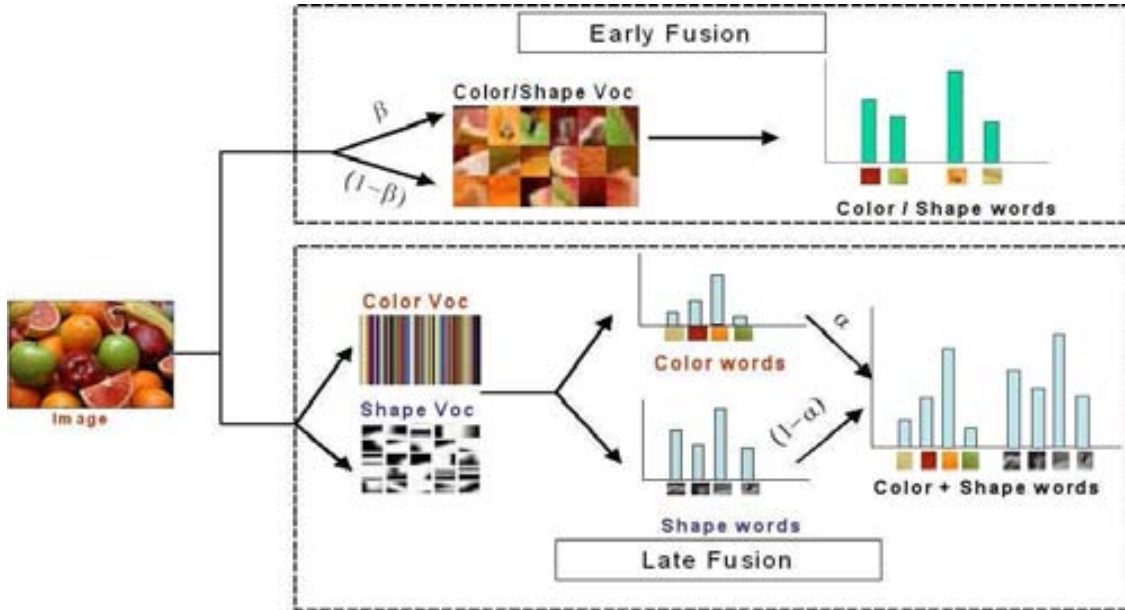
### 2.3.2 Combining color and shape for object recognition

Generally bag-of-words framework is the most commonly used approach for object recognition. Initially intensity based features such as SIFT have been used for image representations. However, in recent years, a lot of work has been done on incorporating color information with in the bag-of-words framework. The two main questions are: what color representation to use and how to fuse it with shape information. Combining color and shape descriptors have shown to provide excellent results in object recognition. A variety of color descriptors and fusion techniques have been suggested in literature [6, 7, 17, 46, 48, 75, 76]. The gain which is obtained by color in image classification varying from 3-4% on PASCAL dataset to 5-10% on colorful data sets such as flower and birds [7, 75, 76]. This motivates looking into power of color representation for object detection and action recognition task in still images. Here, we discuss variety of fusion approaches popular in object recognition.

#### Early and Late Fusion

Early and late feature fusion are two of the main approaches to combine color and shape information. The early fusion approach combines color and shape at the feature level. Image is represented by pure color and shape descriptor. The two representations are concatenated before the vocabulary construction stage. This results in a joint color-shape vocabulary having high discriminative power. Early fusion possesses the property of feature binding as both color and shape are combined locally at the feature level. Early fusion is shown to be more suitable for natural object categories where color and shape are constant [48].

Late fusion combines color and shape at the image level. Separate color and shape histograms are constructed for an image. The resulting histograms are concatenated into a single representation before the classification stage. Late fusion possesses the property of feature compactness as sept visual vocabulary are constructed for color and shape. However, it lacks the property of feature binding since the connection between color and shape feature is lost. Late fusion is shown to provide improved results for man made categories where one of the two visual cues varies significantly [48]. Figure 2.11 show a comparison between late and early fusion within the bag-of-words frameworks.



**Figure 2.11:** Early and late fusion schemes to combine color and shape information within bag-of-words framework. The  $\alpha$  and  $\beta$  parameters determine the relative weight of the two cues.

An important aspect when combining multiple cues is to leverage the contribution of individual cues in the final representation. This is crucial since in many object categories the contribution of individual cues vary significantly. A common way to allow cue weighting is to introduce a weighting parameter between the individual cues. Generally the weighting parameter is learned through cross validation performed over the validation set. In early fusion, this involves introducing the weight vector results in constructing multiple vocabularies and histogram. This is laborious since the whole pipeline has to be repeated several times. In case of late fusion, cue weighting is less cumbersome since the weighting parameter is introduced after the histogram construction.

### Complex fusion Schemes

Recently several methods have been introduced to introduce feature binding in late fusion [46–48]. The color attention approach separately process the color and shape cues. Color attention maps are used to modulate shape features. Both top-down and bottom-up color information is used to weight shape histogram. The top-down color attention maps are obtained using the class information over the training set whereas the bottom-up attention maps are based on natural image statistics [79,80]. More features are sampled with in regions of an image that is more likely to contain an object. The color attention approach is inspired from biologically theories of visual attention [73,90] .

Another way to introduce feature binding in late fusion is to construct a product vocabulary over color and shape features. However, these product vocabularies suffer with the problem of high dimensionality. The portmanteau approach counter this problem by introducing a information theoretic clustering approach to compress these unwieldy product vocabularies. This results in a compact joint color-shape visual vocabulary. Since, color and shape are process lately the portmanteau approach is also convenient for cue weighting.

## **2.4 Conclusion**

In this chapter, we have provided an overview on most popular object detection pipelines such as holistic, part-based approach. The bag-of-words framework is explained for action recognition. The stages within each approach have been discussed in brief. Finally, we provide a brief overview on color descriptors. In the following chapters, we will investigate several aspect of introducing color information in object detection and action recognition.

# Chapter 3

## Color Contribution towards Person Detectors<sup>1</sup>

---

Person detection is a key component in fields such as advanced driver assistance and video surveillance. However, even detecting non-occluded standing person remains a challenge of intensive research. Finding good features to build person models for further detection is probably one of the most important issues to face. Currently, shape, texture and motion features have deserved extensive attention in the literature. However, color-based features, which are important in other domains (*e.g.*, image categorization), have received much less attention. In fact, the use of RGB color space has become a kind of choice *by default*. The focus has been put in developing first and second order features on top of RGB space (*e.g.*, HOG and co-occurrence matrices, resp.). In this chapter we evaluate the opponent colors (OPP) space as a biologically inspired alternative for person detection. We investigate augmenting HOG feature with color information both in the framework of Dalal *et al.* and part-based detector of Felzenszwalb *et al.* Our experiments demonstrate that OPP outperforms RGB space. This is a relevant result since, up to the best of our knowledge, OPP space has not been previously used for person detection. This suggests that in the future it could be worth to compute co-occurrence matrices, self-similarity features, etc., also on top of OPP space, *i.e.*, as we have done with HOG in this chapter. We also investigate possible differences among types of scenarios: indoor, urban and countryside. Interestingly, our experiments suggest that the benefits of OPP with respect to RGB mainly come for indoor and countryside scenarios, those in which the human visual system was *designed* by evolution.

---

### 3.1 Introduction

Camera-based person detection is of great interest for applications in the fields of content management, video-surveillance [43, 83, 87] and driver assistance [18, 29, 32]. Person detection is difficult because the great variety of backgrounds (scenarios, illumination) in which persons are present, as well as their intra-class variability (pose, clothe, occlusion). Even detecting non-occluded persons that are standing, is still a hot topic of research. In order to improve person detection results we can focus on *classification*, *i.e.*, on building a classifier that given an image window decides if it contains a person or not. Nowadays, most successful classification processes for person detection follow the learning-from-examples paradigm [18, 32]. For instance, Dalal *et al.* [11] proposed

---

<sup>1</sup>Part of this Chapter appeared in CAIP(2011) and IbPRIA(2011).



**Figure 3.1:** Annotation enrichment for PASCAL VOC 2007 dataset. First, second and third rows show images that we have annotated as *indoor*, *urban* and *countryside*, resp.

a holistic classifier that relies on histograms of oriented gradients (HOG) as features and linear support vector machines (linear SVM) as learning algorithm, which still remains as a competitive baseline method for comparison with new person classifiers [16, 18]. Whereas discriminative part-based approaches [24, 25], that heavily rely on dynamic part detection, constitute the state of the art for detecting persons. The part-based person detectors generally use the histograms of oriented gradients (HOG) introduced in [11] by Dalal *et al.* as low-level features for building person models.

Finding good features for developing a person classifier is a major key for its success. Focusing on person appearance, different sets of features try to exploit (combinations of) cues such of shape and texture [87]. However, although color information deserves special attention in domains such as segmentation and category recognition [42, 75], it has not been explored in deep for person detection. In fact, the baseline classifier of Dalal *et al.* [11] uses standard RGB. In particular, gradient information is computed individually for each color channel and then, at each pixel, only the gradient information corresponding to the maximum magnitude among the RGB channels is used for computing the HOG. Dalal *et al.* reported that similar results were obtained using LAB space. This approach has been the common way of using color for person detection since then [24, 55, 63, 87, 89, 92] and, as a matter of fact, it has been considered as pretty similar to the use of the image intensity in cases where color information was not available [18].

Human beings do not rely on long (L), middle (M) and short (S) wavelength channels (RGB-like) separately for color perception. In order to increase subsistence, evolution provided the human retina with ganglion cells that combine L, M and S channels to work in *opponent-colors-space* mode for enhancing the visual detection of events of interest as well as compressing the color information of L, M and S *acquisition cells* [38, 41, 49]. Such compressed color information is sent through the optical nerve to the brain for later decompression and interpretation. Accordingly, in this chapter we evaluate the opponent colors (OPP) space as a biologically inspired alternative for person detection. On the other hand, in the context of image categorization [75] it has been demonstrated the usefulness of the so-called *opponent color space* (OPP) when working with the



so-called SIFT descriptor [58]. Since HOG are SIFT-inspired, we think it is worth to test the use of opponent colors for person detection, *i.e.*, replacing the RGB color space by the OPP one in the baseline framework of Dalal *et al.* and the part-based person detection method described in [25]. Moreover, we are interested in assessing if person detection performance can be affected by the type of scenario where it is performed. In other words, we want to perform a scenario-based comparison between the OPP and RGB color spaces, when *pugged-in* for HOG-part-based person detection.

As scenarios we have chosen three relevant types: indoor, countryside and urban. In order to conduct our scenarios experiments we use the class *person* included in the popular PASCAL visual object classes (VOC) challenge [22]. We have enriched the annotation with the *indoor*, *countryside* and *urban* labels, both for training and testing data as shown by some example images in Figure 3.1. As we will see, our experiments suggest that the benefits of OPP with respect to RGB mainly come for indoor and countryside scenarios, those in which the human visual system was *designed* by evolution.

For evaluation color in the baseline framework of Dalal *et al.* [11], we have used the so-called INRIA person dataset. This dataset contains color images and still is widely used for benchmarking. To support our claim we not only present so-called *per window* evaluation on INRIA person dataset, but also *per image* evaluation as highly recommended in [16].

We argue that altogether is a relevant result since, up to the best of our knowledge, OPP space was not previously used for person detection. Thus, with the aim of enriching feature space for person classifiers, our work suggests that in the future it could be worth to compute co-occurrence matrices, self-similarity features, etc., on top of OPP space, *i.e.*, as we have done here with HOG.

The rest of the chapter is organized as follows. In section 3.2 we define the OPP color space. In section 3.3 we summarize the details of the person detectors developed for our experiments. In section 3.4 we draw the experiments and discuss the corresponding results. Finally, section 3.5 draws the conclusions and future work.

## 3.2 Opponent colors space

In the late 19th century, E. Hering noted that the four hues red, green, yellow and blue are fundamental in the sense that they cannot be described as mixtures of other hues. Then, he stated that there were three types of photo receptors: white-black, yellow-blue and red-green [38]. Nowadays we know that there are not such *image acquisition cells* in human vision. However, Hering was right in postulating the *computation of opponent colors* (*i.e.*, red *vs* green and yellow *vs* blue) in human color vision.

Contemporary science of human vision states that color photo receptors at the retina (*i.e.*, cones) are sensitive to long (L-cone), middle (M-cone) and short (S-cone) wavelengths. A single cone is color blind since its activation depends on both the wavelengths and intensity of the stimulus. A comparison of the signals from different classes of photo receptors is therefore the most basic computational requirement of a color vision system. The existence of cone-opponent retinal ganglion cells that perform such comparisons is well established for human vision.

In particular, opponent process theory postulates that yellow-blue and red-green information is represented by two parallel channels in the visual system that combine cone signals differently. It is now accepted that at an early stage in the red-green opponent pathway, signals from L and M cones are opposed, and in the yellow-blue pathway signals from S cones oppose a combined signal from L and M cones [49]. In addition, there is a third luminance or achromatic mechanisms in which retinal ganglion cells receive L- and M- cone input. Thus, L, M and S belong to a first layer of the retina whereas luminance and opponent colors belong to a second layer of it, forming

the basis of chromatic input to the primary visual cortex. Note also that this mechanism is not random since human color vision evolved for increasing the probability of subsistence [41].

Seeing the RGB space used for codifying color in digital images as the LMS color space of the first layer of human retina, we can also compute an opponent colors (OPP) space as follows [75]:

$$\begin{aligned} \text{red-green} &: O_1 = (R - G)/\sqrt{2} , \\ \text{yellow-blue} &: O_2 = ((R + G) - 2B)/\sqrt{6} , \\ \text{luminance} &: O_3 = (R + G + B)/\sqrt{3} , \end{aligned} \quad (3.1)$$

for R, G and B running on values in  $[0, 1]$ .

### 3.3 Coloring Person Detectors

In this section we show how two detection methods can be augmented with opponent color space for person detection. We start by coloring the baseline framework of Dalal *et al.* [11], then we show how color can enhance the performance of a part-based detection framework [25].

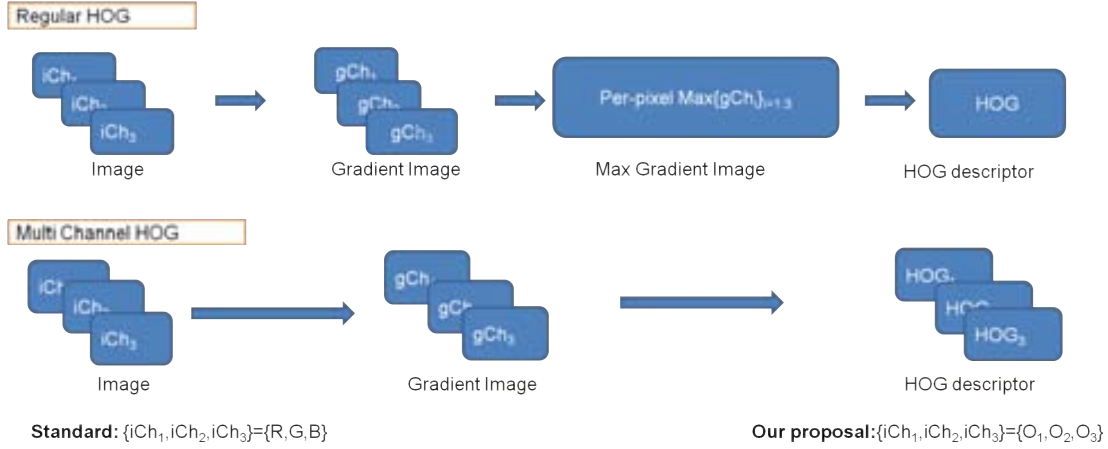
#### 3.3.1 Holistic Person detector

A *person detector* is composed of a *person classifier* learnt from a training set by using specific *features* and a *learning machine*. With this classifier we *scan a given image* looking for persons. Since multiple detections can be produced by a single person, we also need a mechanism to *select the best detection*. The procedures we use for feature extraction, machine learning, scanning the images, as well as selecting the best detection from a cluster of them, are briefly reviewed in this section.

**Person classifier:** We follow the settings suggested by Dalal *et al.* for computing HOG features and learning the person classifier using a linear SVM. Such approach remains competitive [16, 18] and, in fact, is the core from which many new proposals are developed [24, 87]. However, Dalal *et al.* as well as in many following works [24, 55, 63, 87, 89, 92], compute HOG on top of RGB space. More specifically, gradient information is computed individually for each color channel and then, at each pixel, only the gradient information corresponding to the maximum magnitude among the RGB channels is used for computing the HOG. We argue that the *max* operation basically is throwing away the color information, *i.e.*, only some sort of luminance contrast is captured by HOG. Accordingly, we propose to replace the features considered by Dalal *et al.* so that color information is also captured as shown in Figure 3.2 .

Our proposal is twofold. First, we remove the *max* operation, *i.e.*, HOG are applied to each color channel separately and, then, the corresponding feature vectors are concatenated to form a single feature vector. Such three-channels HOG are then the input that the linear SVM will use to learn the person classifier. Second, we propose the use of OPP space instead of RGB one. We will see that both ideas are essential to improve person classification performance.

**Image scanning:** In order to perform multi-scale person detection we use the extended *pyramidal sliding window* strategy as proposed in Dalal's PhD [10]. The original image is scaled by a factor  $s^i$  to obtain the image corresponding to the pyramid level  $i$ . Then, given a pyramid level, we must shift the search window along the horizontal and vertical directions with a given stride  $\Delta = (\delta_x, \delta_y)$  pixels. The smaller the  $s$  and  $\Delta$  parameters, the finer the sliding window search. Using a finer



**Figure 3.2:** As compared to Regular HOG, channel-based HOG is computed on each channel and final representation is the concatenation of all channel-based HOG.

search we can expect better detection performance. However, this is to the expense of a higher processing time. Dalal set  $s = 1.2$  and  $\Delta = (8, 8)$ . In our work we found  $s = 1.05$  and  $\Delta = (4, 4)$  pixels a better tradeoff between processing time and detection performance.

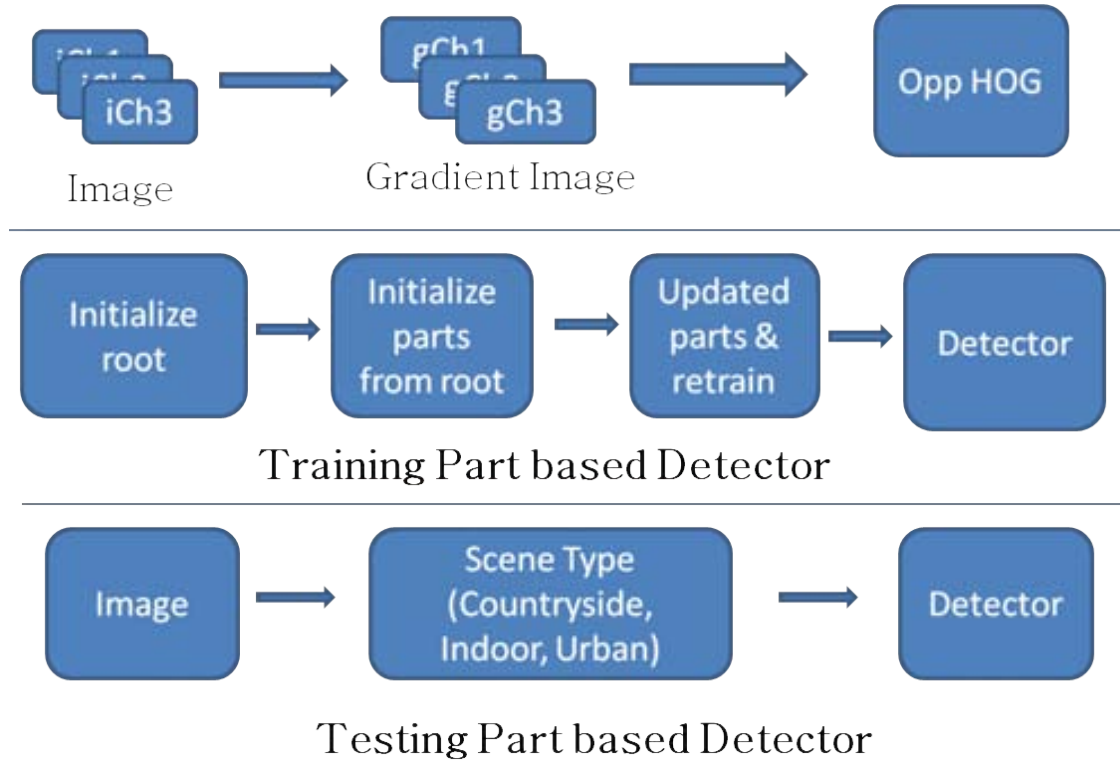
**Select the best detection:** In multi-scale person detection a single person can be detected several times at slightly different positions and scales. Since a unique detection per person is desired, multiple overlapped detections should be grouped by a clustering or *non-maximum-suppression* procedure. In this case, we don't follow the Dalal's proposal in [10]. Instead, we rely on the iterative confidence- and overlapping clustering approach of Laptev [53], which is a simpler and faster technique than Dalal's proposal and yields similar results.

### 3.3.2 Part-based Person detector

The part-based paradigm, introduced by Fischler and Elshlager dates back to 1973 [27]. It provides an elegant way of representing an object category and is particularly efficient for object localization. This model has been built and extended in many direction according to different problems in the computer vision field. Here, we will briefly overview the main principles of part-based methods.

In part-based models, the focus remains on modeling an object as having a number of parts arranged in a deformable configuration. Each part captures the appearance of the object at local level and there is some flexibility in object-parts placement to account for global deformations. The best configuration of such a model is framed on an image as an energy minimization problem which measures the score for each part and deformation score for each pair of connected parts. Part-based models can be separated into many categories depending upon the connection structure to represent the parts: constellation model, star-shaped, tree-shaped, bag of features, etc. Recently, [24, 25] has adopted the star-structured part-based model, which has shown to provide excellent results on person detection [22]. The appearance of an object is represented by histograms of oriented gradients (HOG) features in a 31-dimensional feature vector. HOG of part filters are captured at twice the resolution of the root (full-body) filter to model appearance at multiple scales. Here we follow the implementation associated to [25], whose code has been kindly put publicly available by the authors.

In this implementation, and many others derived from it, HOG features are computed on top of RGB color space. Or more precisely, on top of the *max-gradient* operation on RGB color space



**Figure 3.3:** As compared to Regular HOG, channel-based HOG is computed on each channel and final representation is the concatenation of all channel-based HOG.

(*i.e.*,  $\max\{\nabla R, \nabla G, \nabla B\}$ ). This way of computing HOG derives from the original work by Dalal *et al.* [11] where HOG features were defined in the context of a holistic person detector. Since HOG are SIFT-inspired, we think it is worth to test the use of opponent colors for person detection, *i.e.*, replacing the RGB color space by the OPP one in the part-based person detection method in [25]. The computed detectors are then tested in each scenarios, the overall procedure is explained in Figure 3.3.

## 3.4 Experiments and Results

Here we first present our results on the INRIA Person Dataset using the baseline framework of Dalal *et al.* [11] and in section 3.4.2 results on PASCAL VOC 2007 dataset (only on person class) using part-based detector of Felzenszwalb *et al.* [25].

### 3.4.1 INRIA Person dataset

We rely on the widely used INRIA person dataset of color images for our experiments. This dataset shows a wide range of person variations in pose, clothing, occlusions as well as complex backgrounds. Moreover, the dataset is divided in separated sets of null intersection for training and testing.

The training set contains 2,416 *positive* samples consisting in image windows (original and vertical mirror), each one containing a person framed by certain amount of background. Positives



**Figure 3.4:** Positive (persons) and negative (background) windows from INRIA dataset.

are of the same size (*canonical detection window*), although many of them come from an isotropic down scaling. We term this set of windows as  $\mathcal{W}_+^{\text{train}}$ . For collecting *negative* samples, *i.e.*, image windows that do not contain persons, there are available 1,218 person-free images. We term this set of images as  $\mathcal{I}_-^{\text{train}}$ . The testing set consists of: (1)  $\mathcal{I}_-^{\text{test}}$ : 453 person-free images; (2)  $\mathcal{I}_+^{\text{test}}$ : 288 images containing labeled persons (ground truth); (3)  $\mathcal{W}_+^{\text{test}}$ : 1,126 positives analogous to the ones in  $\mathcal{W}_+^{\text{train}}$  after cropping and mirroring the ground truth of  $\mathcal{I}_+^{\text{test}}$ .

### 3.4.1.1 Training

We use the standard training procedure for the INRIA dataset [10,11]. First, we collect random negative windows from the images in  $\mathcal{I}_-^{\text{train}}$  (10 windows per image to have 12,180 negatives) and down scale them to the size of the canonical detection window; let's call this set of windows  $\mathcal{W}_-^{\text{train}}$ . Then, given the sets  $\mathcal{W}_+^{\text{train}}$  and  $\mathcal{W}_-^{\text{train}}$ , we compute the HOG of such labelled windows on top of the desired color space, and learn the person classifier using the linear SVM. Finally, we run the corresponding person detector on  $\mathcal{I}_-^{\text{train}}$  in order to follow the recommended *bootstrapping* technique, *i.e.*, to append the set  $\mathcal{W}_-^{\text{train}}$  with *hard negative windows* and re-train the person classifier. We apply two bootstrapping iterations. Figure 3.4 shows positive and negative training samples.

### 3.4.1.2 Evaluation

In our experiments we use two widely extended methods of evaluation: *per window* and *per image*. In per window evaluation we assess the results of the person classifier when applied to the  $\mathcal{W}_+^{\text{test}}$  and the images in  $\mathcal{I}_-^{\text{test}}$ . Let  $P^\#$  be the cardinality of  $\mathcal{W}_+^{\text{test}}$ , and let's term as  $P^{\text{TP}}$  the number of elements in  $\mathcal{W}_+^{\text{test}}$  classified as *persons* (*i.e.*, total of so-called true positives). Let  $N^\#$  be the total number of windows processed by applying the pyramidal sliding window technique to the images in  $\mathcal{I}_-^{\text{test}}$  (for each image more than one million of windows are usually processed), and let's term as  $N^{\text{FP}}$  the number of such windows classified as *persons* (*i.e.*, total of so-called false positives). Then, we define the per window detection rate as  $\text{DR}^{\text{PW}} = P^{\text{TP}}/P^\#$ ,  $\text{DR}^{\text{PW}} \in [0, 1]$ . Corresponding miss rate is defined as  $\text{MR}^{\text{PW}} = 1 - \text{DR}^{\text{PW}}$ . Analogously, we define the false positives per window as  $\text{FPP}^{\text{PW}} = N^{\text{FP}}/N^\#$ ,  $\text{FPP}^{\text{PW}} \in [0, 1]$ . We remark that for any given image window, the person classifier returns a real value that we threshold with a fixed value  $t$  in order to classify the window as of type *person* or *non-person*. Thus,  $\text{DR}^{\text{PW}}$  and  $\text{FPP}^{\text{PW}}$  are functions of  $t$ . This allows to plot evaluation curves  $\text{E}^{\text{PW}}(t) = (\text{FPP}^{\text{PW}}(t), \text{MR}^{\text{PW}}(t))$  (so-called ROCs) that show the tradeoff between the miss rate and the false positives per window for each  $t$ .

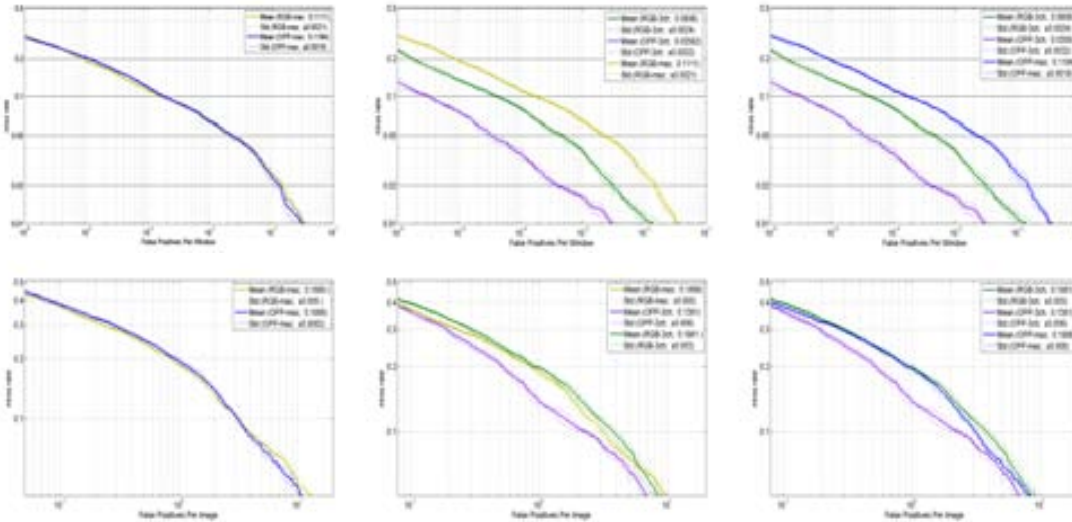
However, some researchers show that it may be more realistic to follow per image evaluation [16]. In this case, not only the person classifier is evaluated but the whole person detector. In particular, the sets  $\mathcal{I}_+^{\text{test}}$  and  $\mathcal{I}_-^{\text{test}}$  are seen as a single set of images,  $\mathcal{I}^{\text{test}}$ , where the person detector is run. Then, the set of detections is compared with the ground truth for counting how many of such detections are true positives ( $T^{\text{TP}}$ ) and how many are false positives ( $T^{\text{FP}}$ ). If  $I^\#$  is the cardinality of  $\mathcal{I}^{\text{test}}$  and  $H^\#$  the number of labeled persons in  $\mathcal{I}_+^{\text{test}}$ , then we can define the per image detection rate as  $\text{DR}^{\text{Pi}} = T^{\text{TP}}/H^\#$  ( $\text{DR}^{\text{Pi}} \in [0, 1]$ ); per image miss rate  $\text{MR}^{\text{Pi}} = 1 - \text{DR}^{\text{Pi}}$  and

the false positives per image as  $\text{FP}^{\text{Pi}} = \text{T}^{\text{FP}}/\text{I}^{\#}$ . In order to determine if a detection overlaps sufficiently with a labeled person of  $\mathcal{I}_+^{\text{test}}$  we follow the so-called PASCAL criteria [16] (also for bootstrapping during training). Now, analogously to  $\text{E}^{\text{PW}}(t)$  we can define the evaluation curve  $\text{E}^{\text{Pi}}(t) = (\text{FP}^{\text{Pi}}(t), \text{MR}^{\text{Pi}}(t))$ ;  $\text{FP}^{\text{Pi}}(t)$  can be greater than one.

### 3.4.1.3 Devised experiment

We train four types of classifiers: *RGB-max*; *RGB-3ch*; *OPP-max* and *OPP-3ch*. The *OPP vs RGB* refers to the used color space. The *3ch* stands for computing HOG for each color channel separately and then concatenate the three feature vectors into a single one. The *max* stands for computing HOG by taking into account, at each pixel, only the gradient of highest magnitude among the color channels, *i.e.*, the usual approach introduced by Dalal *et al.*

Since collecting negatives during training involves a random selection, obtained classifiers can vary from train to train. Therefore, for each type of classifier we repeat the training and further evaluation five times. This gives five curves per classifier (20 curves), thus, we condense the results for each classifier in the respective mean  $\pm$  standard deviation curves for both per window ( $\text{E}^{\text{PW}}(t)$ ) and per image ( $\text{E}^{\text{Pi}}(t)$ ) evaluation. Figure 3.5 summarizes the obtained results.



**Figure 3.5:** Per window (top) and per image (bottom) evaluation using logarithmic scales. Values at usual points of interest are included, *i.e.*,  $10^{-4}$  FPPW and  $10^0$  FPPI, resp.

### 3.4.1.4 Discussion

We point out two main observations: (1) *OPP-3ch* clearly outperforms *RGB-3ch/max*; (2) the *max* operation throws away the color information. Let us argue these observations. Per image and per window evaluation show that *OPP-3ch* outperforms *RGB-3ch/max*, especially at the usual points of interest, *i.e.*,  $\text{FP}^{\text{PW}} = 10^{-4}$  and  $\text{FP}^{\text{Pi}} = 10^0$ . At  $\text{FP}^{\text{PW}} = 10^{-4}$  *OPP-3ch* has an average miss rate of 3.6%, while for *RGB-3ch* is 8.1% and for *RGB-max* 11.1%. At  $\text{FP}^{\text{Pi}} = 10^0$  *OPP-3ch* has an average miss rate of 13.9%, while for *RGB-3ch* is 19.8% and for *RGB-max* 18.9%. Moreover, the *max* operation removes the difference between RGB and OPP spaces. Besides, per image evaluation shows that both the *3ch* and the *max* configurations are similar for the RGB case, but quite different for OPP, where *3ch* clearly wins. For instance, the average miss rate of *OPP-3ch* at  $\text{FP}^{\text{Pi}} = 10^0$  is 13.9% while for *OPP-max* it is 19.6%.

	Training			Testing	
	Windows (+)	Images (-)	Initial Windows (-)	Images	Windows (+)
Indoor	(45.5%) 4268	(36.0%) 516	(36.0%) 103200	(41.0%) 2031	(49.1%) 2252
Countryside	(18.8%) 1762	(29.0%) 414	(29.0%) 82800	(29.5%) 1463	(22.2%) 1004
Urban	(35.7%) 3350	(35.0%) 501	(35.0%) 100200	(29.5%) 1458	(28.1%) 1272
Overall	9380	1431	286200	4952	4528

**Table 3.1:** Training and testing numbers per scenario: person windows (+); images without persons (-); initial background windows (-) after sampling 200 one per image without persons; number of images for testing as well as persons to be detected.

### 3.4.2 PASCAL VOC 2007 dataset

The PASCAL VOC 2007 dataset consists of 9963 images of 20 different object classes with 5011 training images and 4952 test images. We will use the *person* class of the PASCAL VOC detection challenge of 2007. The reason for using the data from 2007 is that it was the last time that test set ground-truth were provided. We need such annotations to enrich them with the different scenarios we have mentioned in Figure 3.1. After doing such enrichment for training and testing data, we obtain the numbers of training windows and testing images per scenario summarized in Table 3.1. In order to evaluate the obtained results, we follow the PASCAL VOC 2007 protocol, which is based on *precision-recall* (PR) curves and the associated *average precision* (AP). The average precision is proportional to the area under a precision-recall curve. Please, refer to [22] for more details about such protocol.

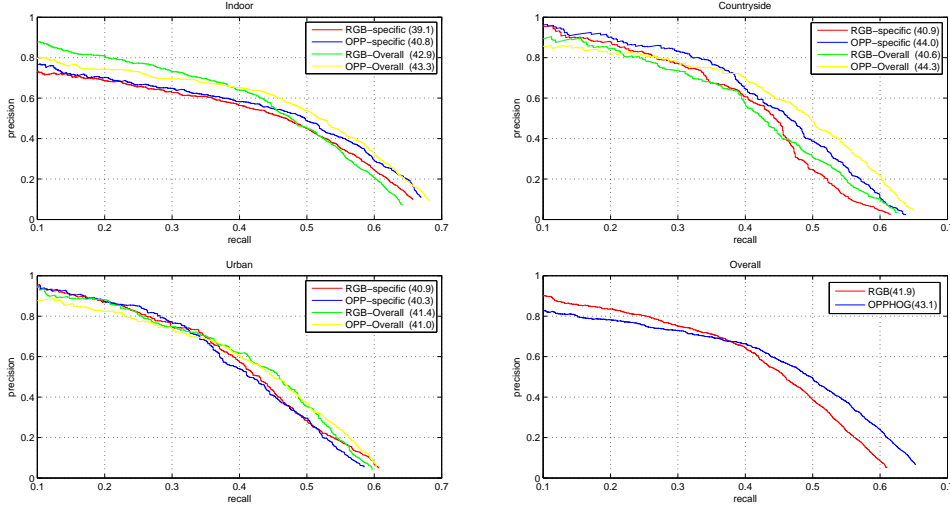
#### 3.4.2.1 Devised Experiments

In this section, we want to address the following specific questions in the context of part-based person detection: (**Q1.**) *if our detector must work in specific scenarios, is it better to use OPP or RGB?*. This is useful to know it for specific systems that must work in specific locations (e.g., intruder detection, person detection, etc.) rather than as general computer vision systems. (**Q2.**) *if we don't know a priori the scenario in which our detector must work, is it better to use OPP or RGB?*. This question is more related to general systems that must work in a broad spectrum of environments (e.g., automatically detecting people for focusing before a camera shot).

In order to answer **Q1** we will run experiments where person classifiers, based on RGB and OPP, are trained and tested in specific scenarios. We have selected three different and relevant scenarios: indoor, countryside and urban as shown in Figure 3.1. In particular we will run the part-based method summarized in section 3.3.2, with the only difference of the input color space used before computing the HOG descriptors: we run equivalent experiments for RGB and OPP.

In order to answer **Q2** we run experiments analogous to the scenario-based ones, but without actually distinguishing the scenario. In other words, we perform the type of experiments that PASCAL VOC challenge participants do, for the cases of RGB and OPP color spaces. Additionally, we not only present the overall result on the full testing dataset but also the results of applying the overall classifiers (*i.e.*, the ones trained without taking into account the scenarios) to each considered scenario separately.

It is worth to mention that Felzenszwalb *et al.* method computes the HOG over the *max-gradient* as we have seen in section 3.3.2, however, we compute separate HOG features for each OPP channel. Thus, our features are of a dimension three times higher than the usually used



**Figure 3.6:** Precision-recall (PR) curves obtained from the different experiments are shown: using RGB and OPP color spaces, for the indoor, countryside, urban and overall classifiers. The average precision (AP) of each PR curve is the number shown inside the respective parenthesis. The PRs of the specific classifiers are plotted together with the PRs of the overall classifiers applied only in the corresponding specific scenarios.

for HOG computation. Nevertheless, for a fair comparison we did similar experiments using the separate R, G and B channels in an analogous use to the three OPP channels. The results were basically analogous to the use of *max-gradient* for RGB, thus, the conclusions of this paper do not change. Accordingly, here on we will only focus on the usual procedure found in the literature, *i.e.*, computing the *max-gradient* for RGB. Note that while RGB channels are highly correlated ones, OPP ones are not.

For the training of any classifier we apply the bootstrapping method to collect hard negatives. We follow the scheme provided by the publicly available software of Felzenszwalb *et al.*, which collects all possible hard negatives until filling 3GB of working memory. In practice, this means to perform about ten bootstrapping.

In summary, the experiments to be done are:

- *Indoor*, *countryside* and *urban* classifiers: they are learnt from indoor images and applied to indoor images. The same for countryside and urban ones.
- *Overall* classifier: it is learnt from all the images but tested in different ways: on all the test images; only in the test images classified as *indoor*; only *countryside*; only *urban*.

These experiments must be run for OPP and RGB color spaces. Thus, we get a total of 14 PR curves and corresponding APs. Figure 3.6 shows all the PR curves in a meaningful way and in Table 3.2 presents the corresponding APs. Additionally, we also applied each scenario-specifically-trained classifier to the other scenarios (not trained). We do not plot the corresponding PR curves for the sake of simplicity but we include the respective APs in Table 3.2.

### 3.4.2.2 Discussion

Results summarized in Figure 3.6 and Table 3.2 allow to answer the questions **Q1** and **Q2** stated in Devised experiments.



	RGB			OPP		
	Indoor	Countryside	Urban	Indoor	Countryside	Urban
Indoor	39.1	21.8	21.2	<b>40.8</b>	23.4	22.8
Countryside	22.0	40.9	31.1	24.9	<b>44.0</b>	33.4
Urban	29.9	34.9	<b>40.9</b>	33.3	39.8	40.3
Overall	41.9			<b>43.1</b>		
	42.9	40.6	<b>41.4</b>	<b>43.3</b>	<b>44.3</b>	41.0

**Table 3.2:** Average precision (AP) in % of the different trained and tested classifiers. Indoor/Countryside/Urban/Overall in the first column refer to the training step, while Indoor/Countryside/Urban in the second row refer to testing. Bold numbers indicate the higher APs comparing the counterpart RGB and OPP results. For the overall classifiers we not only include the overall APs, but also the APs corresponding to apply such classifiers only to specific scenarios during testing.

Table 3.2 shows that AP in indoor scenarios is 1.7 points higher for OPP than for RGB when using only such type of scenarios for training. In the case of countryside the difference is even higher, 3.1 points. However, in urban scenarios RGB performs 0.6 points better.

A closer look to the PR curves in Figure 3.6 for indoor, urban and countryside scenarios gives more detailed insight. In the case of indoor scenarios we appreciate that for the specifically trained and tested classifiers the difference between OPP and RGB is higher for higher recall. This fact is not captured by the AP computation method used in PASCAL VOC 2007 detection challenge. Note, that detection systems are usually interested in having higher recall. In countryside scenarios we observe an analogous situation, but with higher differences. In the case of urban scenarios we see that the specifically trained classifiers are pretty similar along the whole PR plot.

From these observations we conclude that the answer to question **Q1** is: *for indoor and countryside scenarios OPP color space performs better than RGB, while for urban scenarios it seems that there is not a clear preference for mid-to-high recalls.* The major benefit of OPP is for countryside scenarios. Interestingly, OPP color space is the result of human evolution inside primitive indoor and countryside environments, not urban ones, where humans were targets of interest among others. Primitive indoor scenarios are of different background than modern ones. However, countryside colors remain constant. Of course, we don't argue here that our experiments are supporting psychological/evolutionary claims about the human vision system, we only want to point out here what in our modest opinion is an interesting fact.

Regarding question **Q2**, Table 3.2 shows that when jointly using all human windows and backgrounds for training, the AP is 1.2 points higher for OPP than for RGB. Again, by a closer look to PR curves Figure 3.6 for the overall case, we observe that the major benefit of OPP comes for recalls over 40%, *e.g.*, for a recall of the 50% we obtain about ten points more of precision with OPP. We can also assess the performance of these overall classifiers focused on our specific scenarios. We observe then that for the indoor ones, for recalls below the 40% RGB is giving higher precision, however, over such recall the situation changes. The AP is 0.4 points higher for OPP than for RGB. The case of countryside scenarios is analogous but here the OPP starts to offer better precision before, approximately for recalls higher than the 22%. The AP is 3.7 points higher for OPP. In urban scenarios precision is higher with RGB than with OPP for recalls lower than approximately the 30%, however, over such recall OPP and RGB behave pretty similar. The AP is 0.4 points higher for RGB.

From these observations we conclude that the answer to question **Q2** is: *combining data coming from different scenarios during training helps to potentially obtain benefits from OPP over RGB, however, the final benefits will only be obtained if the classifier is used in indoor and countryside scenarios.* Note that the best scenario for OPP, *i.e.*, countryside according to our experiments, is the

less represented in the training of overall classifiers Table 3.1. During testing, countryside and urban scenarios are, basically, equally represented, but indoor scenarios gain in testing presence Table 3.1, which probably is the reason for OPP offering an overall improvement over RGB (countryside cases help AP for OPP while urban cases help RGB).

In summary, using OPP for human detection is worth out of urban scenarios, specially for countryside. Examining Table 3.2 one could be also tempted to conclude that overall detectors outperform the specifically trained ones, however, we think that this can be only an effect of the number of examples and counter-examples during training. What is clear (and expected), however, is that classifiers trained only in one type of scenario perform poorly in the other types of scenarios.

## 3.5 Conclusions

In this chapter we have investigated the effect of using the opponent color space, which is based on the human vision system, for person detection. We have taken as baseline HOG+LinearSVM person detector proposed by Dalal *et al.*, we obtain better results than by following practice of using RGB color space. This conclusion is based on per-window and per-image evaluation over the widely used INRIA person dataset. Furthermore, we have seen that by feeding such a color space in part-based method proposed by Felzenszwalb *et al.*, we obtain better performance. Then, by following the protocols of the PASCAL VOC challenge of 2007, applied to the *person* class, we have collected experimental results that state that opponent color space is a better choice for computing HOG in indoor and, specially, countryside environments. In urban scenarios, there is no clear benefit. Interestingly, indoor and countryside scenarios, those in which the human visual system was *designed* by evolution. The combination of opponent color scape and Felzenszwalb *et al.* method as well as the scenario-based study are new up to the best of our knowledge.

# Chapter 4

## Color Attributes for Object Detection<sup>1</sup>

---

State-of-the-art object detectors typically use shape information as a low level feature representation to capture the local structure of an object. This paper shows that early fusion of shape and color, as is popular in image classification, leads to a significant drop in performance for object detection. Moreover, such approaches also yields sub-optimal results for object categories with varying importance of color and shape.

In this chapter we propose the use of color attributes as an explicit color representation for object detection. Color attributes are compact, computationally efficient, and when combined with traditional shape features provide state-of-the-art results for object detection. Our method is tested on the PASCAL VOC 2007 and 2009 datasets and results clearly show that our method improves over state-of-the-art techniques despite its simplicity. We also introduce a new dataset consisting of cartoon character images in which color plays a pivotal role. On this dataset, our approach yields a significant gain of 14% in mean AP over conventional state-of-the-art methods.

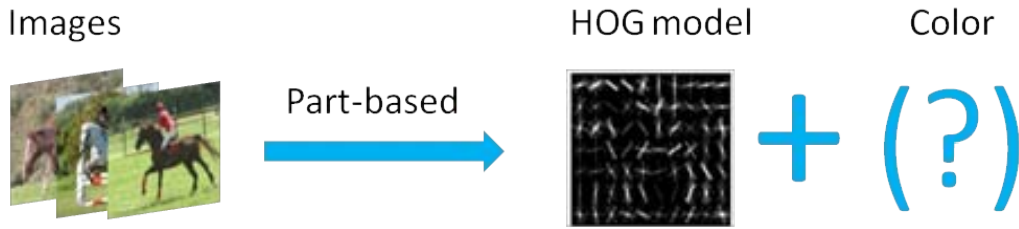
---

### 4.1 Introduction

Object detection is one of the most challenging problems in computer vision. It is difficult due to the significant amount of variation between images belonging to the same object category. Other factors, such as changes in viewpoint and scale, illumination, partial occlusions and multiple instances further complicate the problem of object detection [11, 22, 25, 74, 81, 100]. Most state-of-the-art approaches to object detection rely on intensity-based features that ignore color information in the image [11, 25, 51, 100]. Figure 4.1 shows the lack of color information in part-based approach. This exclusion of color information is usually due to large variations in color caused by changes in illumination, compression, shadows and highlights, etc. These variations make the task of robust color description especially difficult. On the other hand, and in contrast to object detection, color has been shown to yield excellent results in combination with shape features for image classification [46, 48, 75]. The few approaches which do apply color for object detection focus on a single class such as pedestrians [3, 85, 92]. However, the problem of generic object detection is more challenging and the contribution of color to object detection on standard benchmark datasets such as the PASCAL VOC [22] is yet to be investigated. In this chapter, we investigate extending color information in two existing methods for object detection, specifically the part-

---

<sup>1</sup>Appeared in Twenty-Fifth IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2012).



**Figure 4.1:** Augmenting existing intensity based part-based detector with color information.

based detection framework [25] and the Efficient Subwindow Search approach [51]. The failure of existing approaches motivates us to distinguish three main criteria which should be taken into account when choosing an approach to integrating color into object detection.

**Feature Combination:** There exist two main approaches to combining shape and color information: early and late fusion [48, 70, 75]. Early fusion combines shape and color at the pixel level, which are then processed together throughout the rest of the learning and classification pipelines [48, 70]. In late fusion, shape and color are described separately from the beginning and the exact binding between the two features is lost. Early fusion, in general, results in more discriminative features than late fusion since it preserves the spatial binding between color and shape features. Due to its high discriminative power, early fusion has traditionally been a very successful tool for image classification [75]. Recent results, however, have shown that when incorporating spatial pyramids, late fusion methods often obtain better results [17]. This is due to the fact that once spatial cells become smaller the uncertainty introduced by describing shape and color separately is reduced. In the limit, where spatial cells represent a single pixel, early and late fusion are equivalent. The importance of smaller cells of spatial pyramids for object detection has been amply demonstrated [11, 37], and therefore our intuition is that late fusion of color and shape will yield better object detection performance than early fusion.

**Photometric invariance:** One of the main challenges in color representation is the large variation in features caused by scene-accidental effects such as illumination changes and varying shadows. Photometric invariance theory provides guidelines on how to ensure invariance with respect to such events [75, 76], however photometric invariance comes at the cost of discriminative power. The choice of the color descriptor used should take into consideration both its photometric invariance as well as its discriminative power.

**Compactness:** Existing luminance-based object detection methods use complex representations. For example the part-based method of Felzenswalb [25] models an object as a collection of parts, where each part is represented by a number of histograms of gradient orientations over a number of cells. Each cell is represented by a 31-dimensional vector. Training such a complex model, for just a single class, can require over 3GB of memory and take over 15 hours on a modern, multi-core computer. When extending these cells with color information it is therefore imperative to use a color descriptor as compact as possible both because of memory usage and because of total learning time.

This chapter investigates the incorporation of color for object detection based on the above mentioned criteria. We demonstrate the advantages of combining color with shape on the two most popularly used detection frameworks, namely part-based detection with deformable part models [25] and Efficient Subwindow Search (ESS) for object localization [51]. In contrast to conventional fusion approaches that compute shape features on the color channels independently, we propose the use of color attributes as an explicit color representation. The resulting image representations are compact and computationally efficient while providing excellent detection performance



**Figure 4.2:** Find the Simpsons. On the left, the conventional part-based approach [25] fails to detect all four members of Simpsons. Only Bart and Lisa are correctly detected, while Homer is falsely detected as Lisa and Marge is not detected at all. On the right, our extension of the part-based detection framework with color attributes can correctly classify all four Simpsons.

on challenging datasets. Figure 4.2 provides some examples of how our extension correctly detects challenging object classes where state-of-the-art techniques using shape information alone fail.

## 4.2 Related work

Most successful approaches to object detection are based on the learning-from-examples paradigm and rely on shape or texture information for image representation [11, 25, 100]. Conventionally, a sliding window approach is used which exhaustively scans an image at multiple locations and scales. An SVM is then trained using positive and negative examples from each object category. Given a test image, a classifier then selects the candidate windows most likely to contain an object instance. Among various features, histograms of oriented gradients (HOG) proposed by Dalal and Triggs [11] are the most commonly used features for object detection.

Recently, discriminative, part-based approaches [25, 100] have been shown to provide excellent performance on the PASCAL VOC datasets [20].

Felzenszwalb et al. [25] propose a star-structured part-based detector where HOGs are used for image representation and latent support vector machines for classification. A boosted HOG-LBP detector is proposed by [100], where LBP descriptors are combined with HOGs to incorporate texture information. A boosting technique is employed for feature selection and their approach yields improved performance for objects. In this paper, we incorporate color information within the part-based framework of Felzenszwalb et al. [25]. Contrary to the approach presented by [100], our approach requires no feature selection to identify relevant features for part representation.

In contrast to part-based detection methods, the bag-of-words model has also been used for

object detection [37, 51, 74, 81]. These methods are based on the bag-of-words framework where features are quantized into a visual vocabulary. Vedaldi et al. [81] use a multiple kernel learning framework with powerful visual features for object detection. Harzallah et al. [37] use a two-stage cascade classifier for efficient detection. Their approach also combines object localization and image classification scores. The sliding window approach together with the bag-of-words framework is computationally expensive. Alexe et al. [2] propose an objectness measure to select a few candidate regions likely to contain an object instance in an image. Van de Sande et al. [74] propose the use of hierarchical segmentation as a selective search strategy for object detection. Alternatively, Lampert et al. [51] propose an Efficient Subwindow Search strategy (ESS) to counter the problem of exhaustively scanning sliding windows. In this paper, we also investigate the contribution of color when used in combination with shape features in the ESS framework.

### 4.3 Color attributes for object detection

In this section we describe the color descriptors we will use to augment the shape-based feature descriptors used for object detection. Based on the analysis in the introduction section, our approach will apply a late fusion of shape and color.

#### 4.3.1 Color descriptors

A consequence of our choice of late fusion is that we require a pure color descriptor. In addition, we would like this color descriptor to be discriminative, to possess photometric invariance to some degree, and to be compact. Several color descriptors have been proposed in literature. We consider three of them here.

**Robust hue descriptor (HUE) [76]:** image patches are represented by a histogram over hue computed from the corresponding RGB values of each pixel according to:

$$hue = \arctan \left( \frac{\sqrt{3}(R - G)}{R + G - 2B} \right). \quad (4.1)$$

To counter instabilities in hue, its impact in the histogram is weighted by the saturation of the corresponding pixel. The hue descriptor is invariant with respect to lighting geometry and specularities when assuming white illumination.

**Opponent derivative descriptor (OPP) [76]:** image patches are represented by a histogram over the opponent angle:

$$ang_{\mathbf{x}}^O = \arctan \left( \frac{O1_{\mathbf{x}}}{O2_{\mathbf{x}}} \right) \quad (4.2)$$

where  $O1_{\mathbf{x}}$  and  $O2_{\mathbf{x}}$  are the spatial derivatives in the chromatic opponent channels. The opponent angle is weighted by the chromatic derivative strength  $\sqrt{O1_{\mathbf{x}}^2 + O2_{\mathbf{x}}^2}$ . The opponent angle is invariant with respect to specularities and diffuse lighting.

**Color names (CN) [78]:** color names, or color attributes, are linguistic color labels which humans assign to colors in the world. Based on several criteria with respect to usage and uniqueness, Berlin and Kay [5] in a linguistic study concluded that the English language contains eleven basic color terms: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow. In computer vision, color attributes involve the assignment of linguistic color labels to pixels in an image. This requires a mapping between RGB values and color attributes [4]. In this paper, we use the mapping learned from Google images in [78] as a color descriptor. Color names display a certain amount of photometric invariance because several shades of a color are mapped to the same color name. They also provide an added advantage of allowing the description of achromatic colors such as black,

grey and white which are impossible to distinguish from a photometric invariance perspective. Color names have been found to be a successful color feature for image classification [48]. They have the additional advantage of being a very compact representation.

The color name descriptor is defined as a vector containing the probability of a color name given an image region  $R$ :

$$CN = \{p(cn_1|R), p(cn_2|R), \dots, p(cn_{11}|R)\} \quad (4.3)$$

with

$$p(cn_i|R) = \frac{1}{P} \sum_{x \in R} p(cn_i|\mathbf{f}(x)), \quad (4.4)$$

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  $p(cn_i|\mathbf{f})$  is the probability of a color name given a pixel value. The probabilities  $p(cn_i|\mathbf{f})$  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, the PLSA approach is used [78].

To summarize, color names possess some degree of photometric invariance. However, they also can encode achromatic colors such as black, grey and white, leading to higher discriminative power.

### 4.3.2 Color descriptor evaluation

To select one of the color descriptors described above we performed the following experiment. The histograms of a  $2 \times 2$  spatial pyramid for all the bounding boxes of each object category are extracted. To compare discriminative power, for each histogram we compute KL-ratio between the Kullback-Leibler (KL) divergence of each histogram with members of the other classes and the KL-divergence with members of its own class:

$$\text{KL-ratio} = \frac{\sum_{k \in C^m} \min_{j \notin C^m} KL(p_j, p_k)}{\sum_{k \in C^m} \min_{i \in C^m, i \neq k} KL(p_i, p_k)}, \quad (4.5)$$

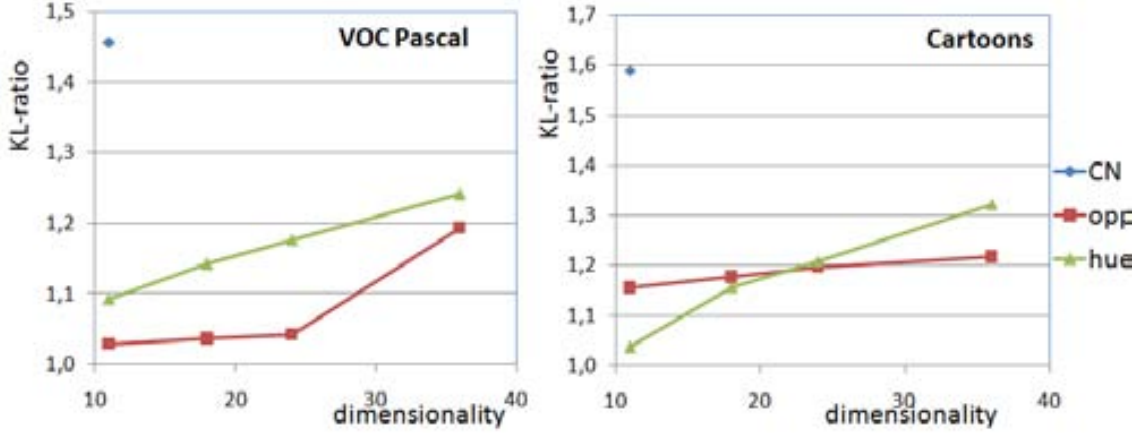
where

$$KL(p_i, p_j) = \sum_{x=1}^N p_i(x) \log \frac{p_i(x)}{p_j(x)}, \quad (4.6)$$

and  $p_i$  is the histogram of bounding box  $i$  over the  $N$  visual words  $x$ . Indices  $i \in C^m$  represent bounding boxes which belong to class  $m$ , while  $j \notin C^m$  are random samples of the bounding boxes which are not the same class as  $m$ . We choose the number of negative samples  $j$  to be three times the size of the positive samples  $i \in C^m$ . A higher KL-ratio reflects a more discriminative descriptor, since the average intra-class KL-divergence is lower than the inter-class KL-divergence.

In Figure 4.3, we report the average KL-ratio over the classes for each color features, HUE, OPP and CN, on the two data sets we use in our experimental evaluation: PASCAL VOC 2007 and a new data set Cartoons. For both HUE and OPP we vary the number of bins in the histogram from 36 as used in [76] to eleven bins which is the size of the CN descriptor. Lowering the dimensionality of the OPP and HUE descriptors leads as expected to lower KL-ratios. As can be seen, the CN descriptor obtains higher KL-ratio even compared to the 36 dimensional representation of the HUE and OPP descriptors.

Based on this experiment we select the CN descriptor as the color feature to use for object detection. It is a pure color feature, therefore allowing us to use it for late fusion with shape features, and based on the KL-ratio it was demonstrated to be more discriminative and compact than the HUE and OPP color descriptors.



**Figure 4.3:** KL-ratio for the PASCAL VOC 2007 and the Cartoon dataset. The graphs clearly show that the color attribute (CN) is superior to the HUE and OPP descriptors in terms of both compactness and discriminative power.

## 4.4 Coloring object detection

In this section we show how two detection methods can be augmented with color attributes for object detection. We start by coloring a part-based detection framework [25], then we show how color can enhance the performance of ESS-based object localization [51].

### 4.4.1 Coloring part-based object detection

In part-based object detection each object is modeled as a deformable collection of parts with a root model at its core [25]. The root filter can be seen as analogous to the HOG-based representation of Dalal and Triggs [11]. Learning in the part-based framework is performed by using a latent SVM formulation. The detection score for a window is obtained by concatenating the root filter, the part filters and the deformation cost of the configuration of all parts. Both the root and the parts are represented by a dense grid of 8x8 non-overlapping cells. A one-dimensional histogram of HOG features is computed over all the pixels in a cell, capturing the local intensity changes.

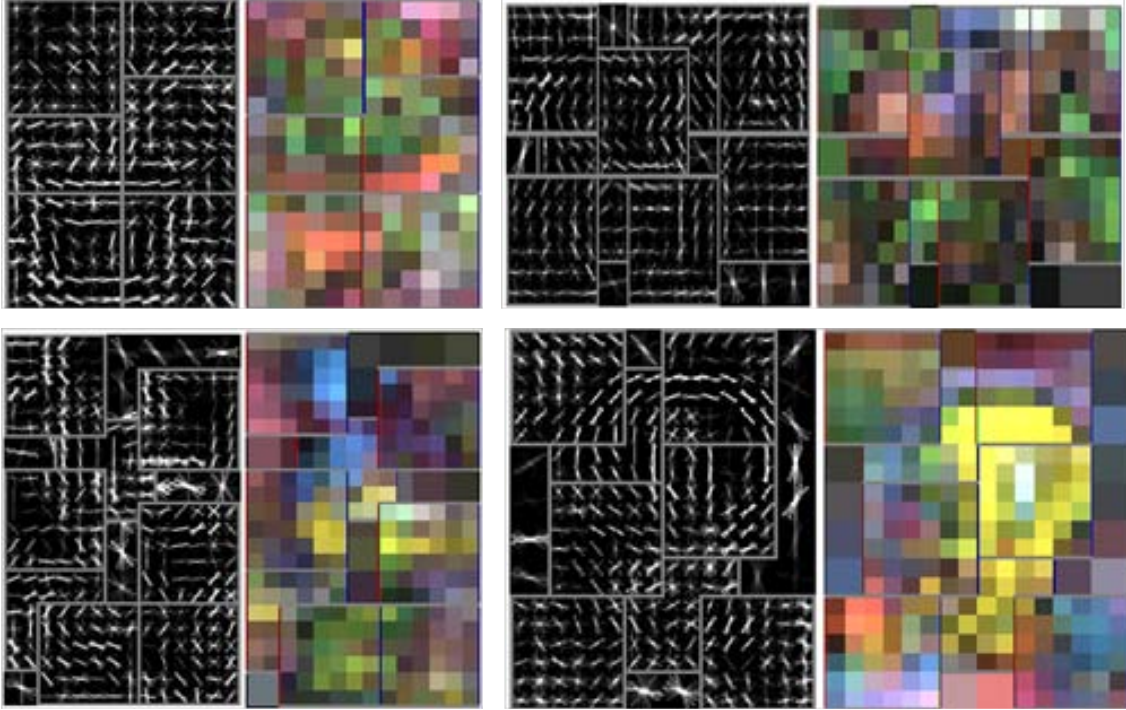
Conventionally, HOGs are computed densely to represent an image. An image is divided into 8x8 non-overlapping pixel regions known as cells. We follow a similar procedure to compute color attributes for each cell, resulting in a histogram representation. We extend the 31-dimensional HOG vector with the eleven-dimensional color attributes vector. For cell  $C_i$ , the representation is obtained by concatenation:

$$C_i = [HOG_i, CN_i], \quad (4.7)$$

and this concatenated representation thus has dimensionality 42. This is still significantly more compact than an early fusion approach where the HOG would be computed on multiple color channels. Such an approach slows the whole detection pipeline significantly by increasing both time complexity and memory usage. Table 4.1 shows a comparison of feature dimensions of different extensions of the part-based method.

Throughout the learning of the deformable part-based model both appearance and color are used. Therefore, the introduction of color leads to models which can significantly differ from the models learned on only luminance-based appearance. Examples of four models are provided in Figure 5.5.





**Figure 4.4:** Visualization of learned part-based models with color attributes. Both the HOG and color attribute components of our trained models are shown. Each cell is represented by the color which is obtained by multiplying the SVM weights for the 11 CN bins with a color representative of the color names. Top row: the HOG and color attribute models for pottedplant and horse. Bottom row: Marge and Tweety models. In the case of horse, the brown color of the horse together with a person sitting on top of it is prominent. Similarly, the model is able to capture the blue hair of Marge and orange feet of Tweety.

#### 4.4.2 Coloring ESS object detection

The Efficient Subwindow Search (ESS) object localization framework [51] offers an efficient alternative to the computationally expensive bag-of-words approach to sliding window object detection. ESS relies on a branch and bound strategy in order to globally optimize a quality criterion across all sub-windows in an image. ESS is based on a bag-of-words representation of the image. Typically, a number of local features are extracted from each image, and these local features are then quantized into a visual vocabulary from which histograms are generated.

A shape-based visual vocabulary of SIFT features is usually used for detection using the ESS framework [51]. Color can be incorporated using early or late fusion for image representation. Both extensions are straightforward. In early fusion a single combined color-shape vocabulary is created and extracted patches are represented by a color-shape visual word. In late fusion, a separate

Feature	HOG	OPPHOG	RGBHOG	C-HOG	LBP-HOG [100]	CN-HOG
Dimension	31	93	93	93	90	42

**Table 4.1:** Comparison of feature dimensionality of different approaches. Our proposed CN-HOG feature increases dimensionality to only 42 dimensions. The early fusion extensions of HOG based on computing the HOG on multiple color channels result in dimensionality of 93 (notations are similar to [75]). The LBP-HOG approach combines the LBP and HOG using late fusion and increases overall dimensionality to 90.



**Figure 4.5:** Example images with annotations from the new Cartoon dataset. The dataset consists of images of 18 different cartoon characters.

shape and color vocabulary are learned, and patches are represented by two indexes, one for the shape vocabulary and one for the color vocabulary. Though we use late fusion to incorporate CN features into ESS, we will also compare with ESS results based on early fusion.

## 4.5 Cartoon character detection

The PASCAL VOC dataset for object detection is predominantly shape-oriented and color plays a subordinate role [48]. To evaluate the potential contribution of color to object detection, we present a new, publicly available dataset of cartoon character images.<sup>2</sup> The dataset consist of 586 images of 18 popular cartoon characters collected from Google. The 18 cartoon characters in the dataset are: The Simpsons (Bart, Homer, Marge, Lisa), the Flintstones (Fred and Barney), Tom, Jerry, Sylvester, Tweety, Bugs, Daffy, Scooby, Shaggy, Roadrunner, Coyote, Donald Duck and Micky Mouse. The dataset contains a variable number of images for each character, ranging from 28 (Marge) to 85 (Tom). Each class is equally divided into training and testing sets where the number of images per category vary. The dataset is challenging as the images come from sources of different types, such as graphics, wallpapers, sketches, etc. Figure 4.5 shows some example images from the dataset. Note the variable quality, appearance and scale of the various cartoon characters. To evaluate detection performance, we follow the PASCAL VOC evaluation criteria [22].

<sup>2</sup>The dataset is available at <http://www.cat.uab.cat/Research/object-detection>

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean AP
HOG [25]	28.9	59.5	10.0	15.2	<b>25.5</b>	<b>49.6</b>	57.9	19.3	<b>22.4</b>	<b>25.2</b>	23.3	11.1	56.8	48.7	41.9	12.2	17.8	<b>33.6</b>	45.1	41.6	32.3
OPPHOG	29.2	54.2	10.7	14.5	17.9	45.8	53.5	21.7	19.3	22.8	21.7	12.3	57.4	46.0	41.2	15.6	19.2	25.0	42.2	41.2	30.6
RGBHOG	33.9	56.5	6.8	13.7	22.9	46.2	56.6	14.9	20.4	22.8	19.3	11.7	57.1	46.7	40.6	13.3	19.2	31.6	47.5	43.4	31.3
C-HOG	29.1	54.7	9.8	14.3	17.9	44.8	55.2	16.0	19.5	25.1	19.6	11.8	58.5	46.6	27.1	15.2	19.0	26.9	44.0	46.6	30.1
CN-HOG ( <i>This paper</i> )	<b>34.5</b>	<b>61.1</b>	<b>11.5</b>	<b>19.0</b>	22.2	46.5	<b>58.9</b>	<b>24.7</b>	21.7	25.1	<b>27.1</b>	<b>13.0</b>	<b>59.7</b>	<b>51.6</b>	<b>44.0</b>	<b>19.2</b>	<b>24.4</b>	33.1	<b>48.4</b>	<b>49.7</b>	<b>34.8</b>

**Table 4.2:** Average precision results for the baseline HOG detector [25], color descriptors proposed in the literature [75] and our proposed CN-HOG approach on all 20 classes of the PASCAL VOC 2007 dataset. Note that our approach along among existing fusion methods outperforms shape alone on this dataset. Our approach provides a significant improvement of 2.5% mean AP over the standard HOG-based framework.

## 4.6 Experimental results

Here we first present our results on the PASCAL VOC datasets and in section 4.6.2 results on the Cartoon dataset.

### 4.6.1 Results on the PASCAL VOC datasets

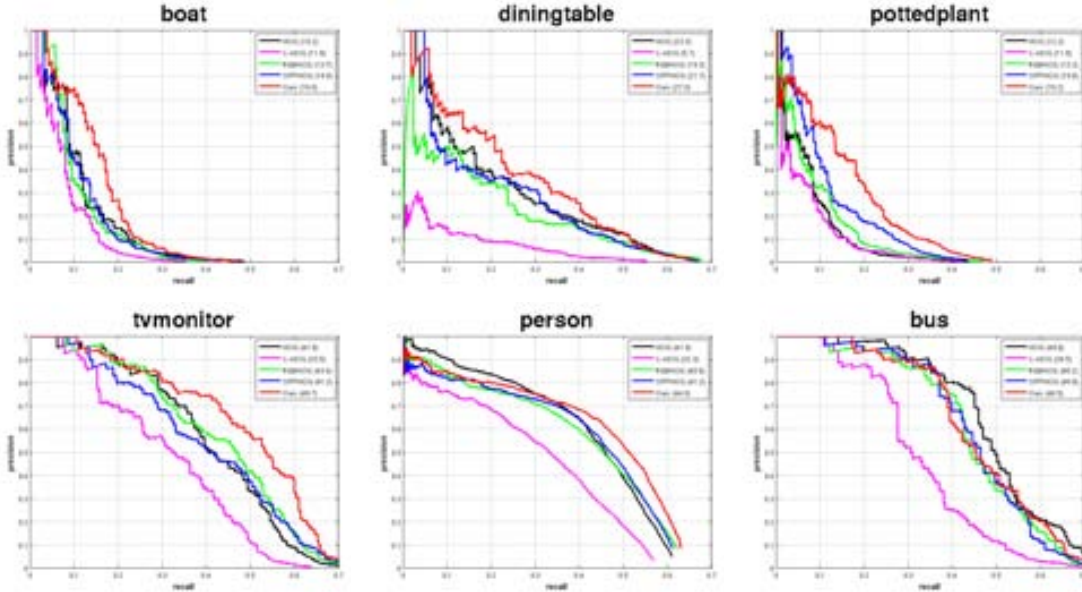
The PASCAL VOC 2007 dataset consists of 9963 images of 20 different object classes with 5011 training images and 4952 test images. The 2009 PASCAL VOC dataset contains 13704 images of 20 different categories. We first provide results of our approach based on the part-based approach [25] discussed in Section 4.4.1. Afterwards, we show the results obtained using ESS for object detection.

**Coloring part-based object detection:** The conventional part-based framework [25] is based on HOGs for feature description. We start by comparing our approach with [25] where HOGs with no color information are used as a feature descriptor. We first perform an experiment to compare our proposed approach with existing color descriptors. Recently, a comprehensive evaluation of color descriptors has been presented by Van de Sande et al. [75]. In this evaluation, opponentSIFT and C-SIFT were shown to yield superior performance on image classification.

We also perform experiments using the standard RGB color space. In case of early fusion, HOG features are computed on the color channels and the resulting feature vectors are concatenated in a single representation. To the best of our knowledge, the performance of these color descriptors have not been evaluated before on the task of object detection within a part-based framework. Table 4.2 shows the results on all 20 categories of the PASCAL VOC 2007 dataset. None of the three color-based methods, namely opponentHOG, RGBHOG and C-HOG, improve the performance over the standard HOG features. Our proposed approach, which has the additional advantage of being compact and computationally efficient, results in a significant improvement of 2.5% on the mean average precision over the baseline HOG. Figure 4.6 shows the precision/recall curves for these color descriptors as well as the baseline HOG on six different object categories from the PASCAL VOC 2007 dataset.

Finally, Table 4.4 shows results obtained on the PASCAL VOC 2009 dataset. Our proposed approach obtains a gain of 1.4% in mean AP over standard HOG based framework. Our method provides superior results on 15 out of 20 object categories compared to standard HOG based framework. Moreover, independently weighting the contribution of color and shape is expected to improve on categories where adding color provides inferior performance.

**Coloring ESS-based object detection:** The ESS-based approach has been shown to provide good localization results on the cat and dog categories of the PASCAL VOC 2007 dataset. We only report the results on cat and dog categories since ESS results in similar or better results compared to part-based methods on these two classes. To evaluate the performance of our proposed approach, we construct a 4000 visual word shape vocabulary based on SIFT features. A visual vocabulary of 500 color-words is constructed using the CN descriptor described above.



**Figure 4.6:** Precision/recall curves of the various approaches on six different categories from the PASCAL VOC 2007 dataset. Other than the bus category, our approach provides significantly improved performance compared to others.

On the cat category, our proposed approach provides an AP of 22.3% compared to 20.7% obtained using shape alone. Similar results are obtained on the dog category where shape alone and our approach provide score of 13.8 and 15.8 respectively.

**Comparison with state-of-the-art results:** Table 4.3 shows a comparison of our approach with the state-of-the-art results reported in literature. Firstly, our proposed color attribute-based approach improves the baseline part-based approach on 15 out of the 20 object categories. The results reported by [81] are obtained by using the bag-of-words framework with multiple features combined using a multiple kernel learning framework. The boosted HOG-LBP approach [100] combines HOG and LBP features while employing boosting as a feature selection mechanism. It is further reported by [100] that without this feature selection strategy, the naive feature combination provides inferior results. In contrast to these approaches, no feature selection strategy is used in our approach, though a selection strategy can be easily incorporated which is expected to further improve results. The approach proposed by [74] provides a mean AP of 33.9% using multiple color spaces, specifically RGB, opponent, normalized rgb and hue for segmentation. Moreover, a dense representation based on SIFT, opponentSIFT and RGBSIFT is used within the bag-of-words framework. Our approach provides the best mean AP reported on this dataset in the detection literature<sup>3</sup> [15, 20, 25, 74, 100, 102]. Finally, on the PASCAL VOC 2009 dataset our approach provides best results on 7 object categories.

## 4.6.2 Results on the Cartoon dataset

Here we report the results obtained on our new Cartoon dataset in which color plays an important role. We first show results using the part-based framework and then follow with a comparison of several approaches using the ESS-based object detection framework.

**Coloring part-based object detection:** Table 4.5 shows the results obtained using the part-

<sup>3</sup>We do not compare our results with methods combining image classification and detection. Such approaches can be seen as complementary to our approach.

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean AP
HOG [25]	28.9	59.5	10.0	15.2	25.5	49.6	57.9	19.3	22.4	25.2	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6	32.3
Best 2007 [20]	26.2	40.9	9.8	9.4	21.4	39.3	43.2	24.0	12.8	14.0	9.8	16.2	33.5	37.5	22.1	12.0	17.5	14.7	33.4	28.9	23.3
UCI [15]	28.8	56.2	3.2	14.2	<b>29.4</b>	38.7	48.7	12.4	16.0	17.7	24.0	11.7	45.0	39.4	35.5	15.2	16.1	20.1	34.2	35.4	27.1
LEO [102]	29.4	55.8	9.4	14.3	28.6	44.0	51.3	21.3	20.0	19.3	25.2	12.5	50.4	38.4	36.6	15.1	19.7	25.1	36.8	39.3	29.6
Oxford-MKL [81]	<b>37.6</b>	47.8	<b>15.3</b>	15.3	21.9	<b>50.7</b>	50.6	<b>30.0</b>	17.3	<b>33.0</b>	22.5	<b>21.5</b>	51.2	45.5	23.3	12.4	23.9	28.5	45.3	48.5	32.1
LBP-HOG [100]	36.7	59.8	11.8	17.5	26.3	49.8	58.2	24.0	<b>22.9</b>	27.0	24.3	15.2	58.2	49.2	<b>44.6</b>	13.5	21.4	<b>34.9</b>	47.5	42.3	34.3
CN-HOG ( <i>This paper</i> )	34.5	<b>61.1</b>	11.5	<b>19.0</b>	22.2	46.5	<b>58.9</b>	24.7	21.7	25.1	<b>27.1</b>	13.0	<b>59.7</b>	<b>51.6</b>	44.0	<b>19.2</b>	<b>24.4</b>	33.1	<b>48.4</b>	<b>49.7</b>	<b>34.8</b>

**Table 4.3:** Comparison with state-of-the-art results on the PASCAL VOC 2007 dataset. Note that the approach of boosted LBP-HOG [100] combines HOG and LBP together using a boosting strategy for feature selection. However, our proposed combination of color names and HOGs (CN-HOG) is compact, computationally inexpensive and uses no feature selection.

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean AP
HOG [25]	38.0	<b>47.2</b>	11.3	12.5	28.1	40.5	38.4	23.4	<b>17.3</b>	20.6	15.7	13.8	41.0	43.2	41.7	12.8	24.4	16.6	42.4	31.9	28.0
Oxford-MKL [81]	<b>47.8</b>	39.8	<b>17.4</b>	<b>15.8</b>	21.9	42.9	27.7	<b>30.5</b>	14.6	20.6	<b>22.3</b>	<b>17.0</b>	34.6	43.7	21.6	10.2	<b>25.1</b>	16.6	<b>46.3</b>	<b>37.6</b>	27.7
UOCTTI	39.5	46.8	13.5	15.0	<b>28.5</b>	<b>43.8</b>	37.2	20.7	14.9	22.8	8.7	14.4	38.0	42.0	41.5	12.6	24.2	15.8	43.9	33.5	27.9
HOG-LBP	11.4	27.5	6.0	11.1	27.0	38.8	33.7	25.2	15.0	14.4	16.9	15.1	36.3	40.9	37.0	13.2	22.8	9.6	3.5	32.1	21.9
CN-HOG ( <i>This paper</i> )	36.3	46.6	12.5	15.2	27.8	43.5	<b>39.0</b>	26.3	16.8	<b>23.2</b>	18.8	15.0	<b>41.4</b>	<b>46.7</b>	<b>43.3</b>	<b>14.7</b>	23.0	18.3	43.6	35.5	<b>29.4</b>

**Table 4.4:** Average precision results for the baseline HOG detector [25], our proposed CN-HOG approach and state-of-the-art results on all 20 classes of the PASCAL VOC 2009 dataset. Note that our approach provides an improvement of 1.4 mean AP over the standard HOG-based framework.

based framework. The conventional part-based approach using HOG features provides a mean AP of 27.6%. The early fusion based approaches yield similar results<sup>4</sup>. Our approach, however, results in a significant gain of 14% in mean AP compared to standard HOG. Moreover, our approach gives the best performance on 9 out of the 18 cartoon categories compared to HOG, opponentHOG, C-HOG and RGBHOG. On categories such as Daffy and Tom, our approach results in a gain of 25.9% and 9.2%, respectively, compared to the second-best approach. This significant gain can be credited to the fact that color names have the additional capability of encoding achromatic colors such as black, grey and white.

**Coloring ESS-based object detection:** Here we compare our approach with shape alone and the two best color descriptors reported in the literature, namely opponentSIFT and C-SIFT [75]. We perform an experiment to compare the performance of different fusion approaches. We use visual vocabularies of 500 shape words and 100 color names. For opponentSIFT and C-SIFT, a visual vocabulary of 500 words is constructed. A comparison of different approaches using the ESS framework is shown in table 4.6. On this dataset approaches based on color-shape fusion improve performance over standard ESS using SIFT alone.

Table 4.6 shows the comparative performance of our approach. The results obtained using the ESS-based framework is inferior to that obtained using the part-based method. Both opponentSIFT and C-SIFT yield improved results compared to shape alone. Our approach using the ESS framework gives the best performance on the Sylvester category. Interestingly, on the Cartoon dataset ESS again achieves the best results on cats and dogs. Example detection results from different cartoon categories are shown below.

## 4.7 Conclusions

This chapter investigates the problem of incorporating color for object detection. Most state-of-the-art object detectors rely on shape while ignoring color. Recent approaches to augmenting intensity-based detectors with color often provide inferior results for object categories with varying importance of color and shape. We propose the use of color attributes as an explicit color representation for object detection. Color attributes are compact, computationally efficient, and possess some degree of photometric invariance while maintaining discriminative power. We show that our approach can significantly improve detection performance on the challenging PASCAL

<sup>4</sup>We also performed an experiment with a 36-dimensional hue-saturation color descriptor concatenated with a HOG. This yielded a MAP of 34.2%, significantly lower than the 41.7% of our compact representation.

	bart	homer	marge	lisa	fred	barney	tom	jerry	syvester	tweety	bugs	daffy	scooby	shaggy	roadrunner	coyote	donaldduck	mickymouse	mean AP
HOG	72.5	47.7	22.7	82.3	54.5	48.8	6.0	20.5	28.3	18.1	<b>25.2</b>	5.5	14.6	12.0	3.5	2.4	11.1	20.7	27.6
OPPHOG	<b>74.2</b>	37.6	43.2	86.2	61.6	51.2	13.4	30.7	15.8	<b>54.4</b>	25.0	8.8	16.2	21.4	<b>32.5</b>	10.9	17.1	16.0	34.2
RGBHOG	71.1	<b>60.6</b>	33.0	84.7	62.5	47.6	13.0	34.3	<b>36.1</b>	41.9	21.7	6.50	11.5	16.6	29.2	<b>11.1</b>	13.1	40.2	35.3
C-HOG	65.5	46.7	26.4	84.1	62.4	<b>60.4</b>	23.6	38.7	32.8	33.2	20.6	9.7	<b>23.4</b>	25.1	29.3	4.6	21.3	25.9	35.2
CN-HOG ( <i>This paper</i> )	72.3	40.4	<b>43.4</b>	<b>89.8</b>	<b>72.8</b>	55.1	<b>32.8</b>	<b>52.3</b>	32.9	51.4	22.2	<b>35.6</b>	19.8	<b>25.2</b>	21.9	10.0	<b>27.9</b>	<b>45.3</b>	<b>41.7</b>

**Table 4.5:** Comparison of different fusion approaches using the part-based approach on the Cartoon dataset. Our CN-HOG approach yields a significant gain of 14% in mean AP over the standard HOG-based approach. Compared to early fusion approaches, our approach results in a gain of 6.4%.

	bart	homer	marge	lisa	fred	barney	tom	jerry	syvester	tweety	bugs	daffy	scooby	shaggy	roadrunner	coyote	donaldduck	mickymouse	mean AP
Shape	19.0	7.9	8.8	23.6	4.2	1.6	17.5	0.01	24.4	0.2	10	9.8	5.0	2.1	5.8	<b>7.4</b>	5.4	5.5	8.8
CSIFT	16.3	4.4	<b>17.2</b>	17.7	5.7	3.8	20.0	0.8	24.9	1.9	<b>16.2</b>	10.8	<b>7.4</b>	8.1	<b>15.3</b>	5.6	1.9	7.0	10.3
OPPSIFT	12.8	5.7	14.8	<b>28.6</b>	6.3	0.8	24.4	0.6	18.5	<b>5.1</b>	6.5	3.7	3.6	<b>12.1</b>	1.2	4.3	5.2	<b>12.5</b>	9.3
CN-SIFT ( <i>This paper</i> )	<b>28.7</b>	<b>13.6</b>	9.9	19.2	<b>18.2</b>	<b>5.1</b>	<b>24.5</b>	<b>1.9</b>	<b>37.4</b>	0.1	9.9	<b>13.3</b>	7.1	11.7	14.0	3.5	<b>9.8</b>	4.0	<b>12.9</b>

**Table 4.6:** Comparison of different approaches within the ESS detection framework. Similar to our results using the part-based method, combining color attributes improves the overall performance by 4%. Our CN-SIFT method also yields superior performance compared to the well known color descriptors OpponentSIFT and C-SIFT.

VOC datasets where existing color-based fusion approaches have shown to provide below-expected results. Finally, we introduce a new dataset of cartoon characters where color plays an important role.

Barney Detector

HOG	OpponentHOG	RGBHOG	CHOG	CNHOG ( <i>This paper</i> )
				
				
				
				
				

Table 4.7: Comparison of different detectors on Barney.

## Daffy Detector



HOG	OpponentHOG	RGBHOG	CHOG	CNHOG( <i>This paper</i> )
				
				
				
				
				

Table 4.8: Comparison of different detectors on Daffy.



Roadrunner Detector


























HOG	OpponentHOG	RGBHOG	CHOG	CNHOG ( <i>This paper</i> )
				
				
				
				
				

Table 4.9: Comparison of different detectors on Roadrunner.

## Shaggy Detector


























HOG	OpponentHOG	RGBHOG	CHOG	CNHOG ( <i>This paper</i> )
				
				
				
				
				

Table 4.10: Comparison of different detectors on Shaggy.

Tweety Detector

HOG	OpponentHOG	RGBHOG	CHOG	CNHOG ( <i>This paper</i> )
				
				
				
				
				

Table 4.11: Comparison of different detectors on Tweety.

Tom Detector


























HOG	OpponentHOG	RGBHOG	CHOG	CNHOG( <i>This paper</i> )
				
				
				
				
				

Table 4.12: Comparison of different detectors on Tom.

# Chapter 5

## Coloring Action Recognition in Still Images<sup>1</sup>

---

In this chapter, we investigate the problem of human action recognition in static images. By action recognition we intend a class of problems which includes both action classification and action detection (i.e. simultaneous localization and classification). Bag-of-words image representations yield promising results for action classification, and deformable part models perform very well object detection. The representations for action recognition typically use only shape cues and ignore color information. Inspired by the recent success of color in image classification and object detection, we investigate the potential of color for action classification and detection in static images. We perform a comprehensive evaluation of color descriptors and fusion approaches for action recognition. Experiments were conducted on the three datasets most used for benchmarking action recognition in still images: Willow, PASCAL VOC 2010 and Stanford-40. Our experiments demonstrate that incorporating color information considerably improves recognition performance, and that a descriptor based on color names outperforms pure color descriptors. Our experiments demonstrate that late fusion of color and shape information outperforms other approaches on action recognition. Finally, we show that the different color-shape fusion approaches result in complementary information and combining them yields state-of-the-art performance for action classification.

---

### 5.1 Introduction

Action category recognition in still images is a major emerging problem in computer vision.<sup>2</sup> The general problem of action recognition encompasses both the localization and classification of actions in images or video. Only recently has action recognition in static images, where the objective is to identify the action a human is performing from a single image, gained attention from the computer vision research community. In one formulation of action recognition in still images, which we refer to as *action classification*, bounding boxes of humans performing actions are provided both at training and test time. The underlying premise of action classification is that person detectors are reliable enough to correctly localize persons. Another formulation, which we refer to as *action detection*, aims to simultaneously localize and classify the action [14]. Recognizing human actions in

---

<sup>1</sup>Accepted to IEEE JOURNAL OF COMPUTER VISION (IJCV 2013).

<sup>2</sup>As evidenced by the First Workshop on Action Recognition and Pose Estimation in Still Images held in conjunction with ECCV 2012:<http://vision.stanford.edu/apsi2012/index.html>

static images is a difficult problem due to the significant amount of pose, viewpoint and illumination variation. In this work, we investigate the potential of color features for enhancing both action classification and action detection in still images.

In general, the bag-of-words framework is the most applied framework for action classification [12, 67, 68]. State-of-the-art approaches to action classification typically make use of intensity-based features to represent local patches. Color-based features have in most cases been excluded up to now due to large variations in color caused by changes in illumination, shadows and highlights. Such variations complicate the problem of robust color description as can be seen in Figure 5.1. Color has, however, led to significantly improved results on other recognition tasks, such as image classification and object detection [1, 31, 46, 48, 75, 76]. Here we investigate both color features and fusion methods to optimally incorporate color into the human action classification pipeline.

Approaches to action detection, on the other hand, must both localize and classify actions in images or video. A number of techniques have been proposed recently for this problem, and until very recently the emphasis has been on action detection in video. [28] proposed an approach based on a sequence of atomic action units to detect actions in videos. [72] introduced a structural learning approach to action detection in unconstrained videos. A multiple-instance learning framework was proposed by [40] for learning action detectors based on imprecise action locations. [98] propose a naive Bayes mutual information maximization framework for matching patterns in videos. Recently, [14] investigated the problem of action detection in still images. In this chapter, we also investigate the problem of action detection in still images.

Deformable part-based models [25] have demonstrated excellent results for object detection. The conventional part-based framework uses HOG features [11] for image representation. Several works recently have aimed at combining multiple features for object detection [45, 81, 100]. [100] proposed a combination of HOG and LBP features for object detection, and [45] evaluated a variety of color descriptors for object detection. Inspired by the success of color-enhanced object detection, we believe color can also help to improve part-based models for action detection. Therefore, in this chapter we also perform an evaluation of color descriptors for the problem of action detection in still images.

The contribution of this work is twofold. First, we provide a comprehensive evaluation of local color descriptors for human action classification and human action detection in still images. Second, we evaluate different fusion approaches: early fusion, late fusion, channel-based fusion, classifier fusion, color attention [48] and portmanteau vocabularies [46] for combining color and shape features in action recognition. Based on extensive experiments on three action recognition datasets, our results suggest that careful selection of the color descriptor, together with an optimal fusion strategy, yield state-of-the-art results for both action classification and action detection. We conclude with a set of recommendations on the suitability of color descriptors and fusion approaches for action recognition in still images.

Additionally, in this chapter we perform an analysis of the contribution of color for action recognition. We find that color information from objects accompanying actions (such as horses or guitars) can considerably improve classification. In addition, an analysis of action detection errors shows that color information increases the number of localization errors, but that increase is more than compensated by a drop in errors due to confusion with other classes and false detections on the background.

The rest of this chapter is organized as follows. In the next section we discuss work related to the problem of action recognition. In Section 5.3 we give an overview of state-of-the-art color descriptors. We describe a number of approaches to fusing shape and color for action classification in Section 5.4, and in Section 5.5 we show how to incorporate color into a part-based detection framework for action detection. In Section 5.6 we present extensive experimental results on three challenging action recognition datasets, with a comparative evaluation with respect to the state-of-the-art. We finish in Section 5.7 with concluding remarks and general recommendations for



**Figure 5.1:** Example images for different action categories from the PASCAL VOC 2010 dataset. These images illustrate the complications related to color description due to the large variation in illumination, shadows and specularities.

selecting color descriptors and fusion approaches for human action recognition problems.

## 5.2 Related Work

Action recognition in static images has gained a lot of attention recently [12, 60, 65, 68, 96, 97]. Recognizing human actions in static images is difficult due to the lack of temporal information and to large variations in human appearance and pose. Most successful approaches to action recognition adopt the bag-of-words (BOW) approach popular in object recognition [12, 68]. The bag-of-words approach involves detecting keypoint regions which are then described with local feature descriptors. Typically, SIFT descriptors are used to describe image features in intensity images, and these local features are then quantized against a learned visual vocabulary. A histogram over these visual words is then constructed to obtain the final image representation, and finally these histograms are used to train classifiers for recognition.

Other than the BOW approach, several methods have recently been proposed which focus on finding human-object interactions to improve action recognition. [65] propose a human-centric approach that works by first localizing a human and then finding an object and its relationship to it. A poselet activation vector was proposed by [60] that captures the pose in multi-scale manner. The approach captures the 3D pose of a human and the corresponding action from the static images. A discriminatively trained model representing human-object interactions was used by [13]. Their model is constructed using spatial co-occurrences of objects and individual body parts. They further propose a discriminative learning procedure to solve the problem of the large number of possible interaction pairs. [96] propose to use attributes and parts by learning a set of sparse attribute and part bases for action recognition. The approach we propose in this chapter is complementary to the aforementioned techniques and can be used in combination with any of them to improve action recognition.

The use of color for object recognition has been extensively studied [1, 21, 46, 48, 75, 76]. A variety of color descriptors and approaches to combining color and shape cues for object recognition have been proposed in the literature [6, 75–77, 82]. [6] propose to compute SIFT descriptors directly on HSV channels of color images. A set of robust and photometrically invariant color descriptors was proposed by [76]. [64] propose an approach to matching a region between an image and a query image that is based on the integral P-channel representation obtained by computing image features on the pixels. Real-time view-based pose recognition and interpolation based on P-channels was proposed by [23]. P-channel based image representations combine the advantages of histograms

and local linear models. A low dimensional color descriptor based on color names was proposed by [78]. See [75] for a comprehensive study and evaluation of a large number of color descriptors.

The discriminative, deformable part-based framework [25, 100] yields excellent state-of-the-art results for object detection. A star-structured deformable part method was proposed by [25] in which latent support vector machines are employed for classification. The part-based method uses HOG features for image representation and yields excellent performance on the PASCAL VOC datasets [22], especially on the person category. Recently [45] proposed augmenting the standard part-based approach with color information, which results in significant improvement in performance. In this chapter we investigate the contribution of color within a part-based framework for action detection.

As mentioned above, color in object and scene recognition has received significant attention in recent years. However, color has yet to be evaluated in the context of action recognition. This chapter extends our earlier work [45] to action detection and we investigate the potential of combining color and shape for both action classification and action detection. Beyond the work in [45] we here perform an extensive comparison of fusion methods for color and shape. In addition we analyze the contribution of color for action recognition (both classification and detection) in detail. Based on an extensive experimental evaluation, we categorize the different approaches and provide recommendations on the choice of color descriptor and fusion approach for a variety of action recognition problems.

### 5.3 Color Descriptors for Action Recognition

In this section, we introduce the pure color descriptors used in our evaluation. We use the term *pure* to emphasize the fact that these descriptors do not code any shape information about the local patch.

**RGB descriptor (RGB):** As the most simple baseline we use the RGB descriptor, which is just the concatenation of the average R, G and B values of the local patch.

**C descriptor (C):** The C descriptor is defined as  $C = \begin{pmatrix} O1 & O2 & O3 \end{pmatrix}^T$ , where  $O1$ ,  $O2$  and  $O3$  are derived from the opponent color space as [56]:

$$\begin{pmatrix} O1 \\ O2 \\ O3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \quad (5.1)$$

The first two dimensions of C, which are invariant with respect to shadow and shading, are combined with the luminance channel. The final descriptor for a patch is three dimensional and is computed by averaging the C values over the patch. This descriptor was originally proposed by [34].

**Hue-saturation descriptor (HS):** The HS descriptor is computed by first applying a polar coordinate transform to the chromatic channels of the opponent color space (see Eq. 5.1) to obtain the hue and saturation channels:

$$H = \arctan\left(\frac{O1}{O2}\right), \quad (5.2)$$

$$S = \sqrt{O1^2 + O2^2}, \quad (5.3)$$

and then constructing a hue-saturation histogram over the values in the patch. This descriptor is invariant to luminance variations and has 36 dimensions (nine bins for hue times four for saturation).

**Robust hue descriptor (HUE):** The robust hue descriptor was proposed by [76]. To counter instabilities in hue, its impact in the histogram is weighted by the saturation of the corresponding



pixel. The descriptor is derived from an error analysis of the hue representation which shows that the saturation is proportional to the certainty of the hue measurement. As a consequence, the update of the robust hue histogram for achromatic colors (with near zero saturation), where the hue is ill defined, is close to zero. The hue descriptor is invariant with respect to lighting geometry and specularities when assuming white illumination. The final descriptor also has 36 dimensions.

**Opponent derivative descriptor (OPP):** In contrast to the other color descriptors, which are based on the (transformed) RGB values of the image, this descriptor is based on image derivatives. It is based on the opponent angle, which is defined as:

$$ang_{\mathbf{x}}^O = \arctan\left(\frac{O1_{\mathbf{x}}}{O2_{\mathbf{x}}}\right), \quad (5.4)$$

where  $O1_{\mathbf{x}}$  and  $O2_{\mathbf{x}}$  are the derivatives of the chromatic opponent channels. The opponent angle becomes unstable when the derivative in the chromatic plane  $O1_w = \sqrt{O1_{\mathbf{x}}^2 + O2_{\mathbf{x}}^2}$  goes to zero. To counter this, the histogram of  $ang_{\mathbf{x}}^O$  is constructed, using the corresponding  $O1_w$  value to update bins when constructing the histogram. The opponent angle is invariant with respect to specularities, diffuse lighting and blur [76]. The final descriptor has 36 dimensions.

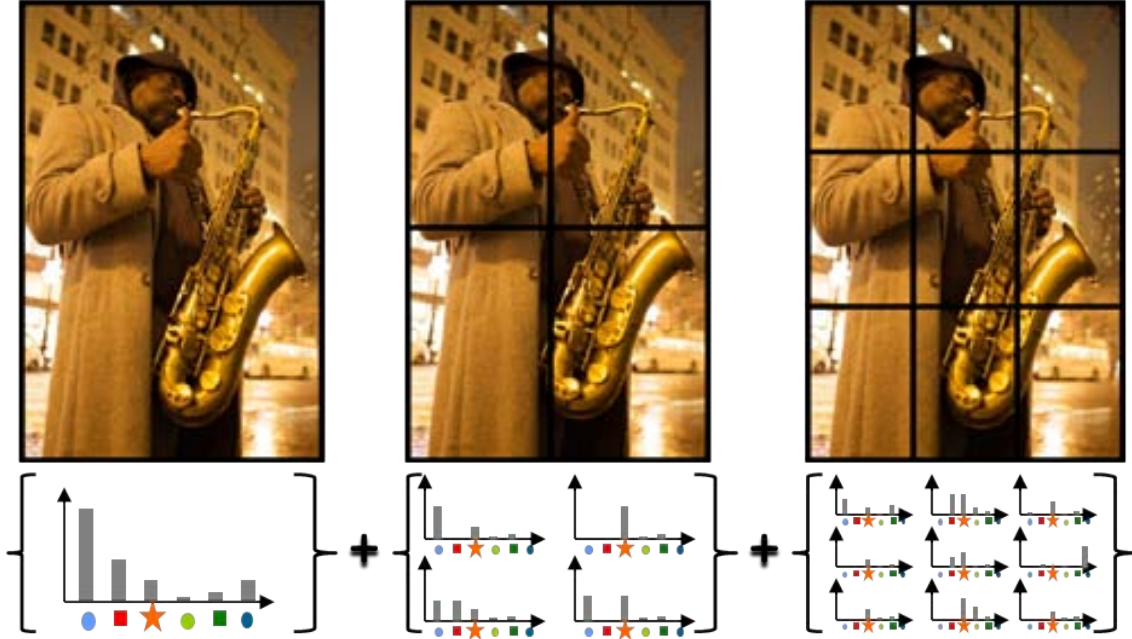
**Color names (CN):** The above descriptors are designed to have specific photometric invariance properties. Instead, the color names descriptor is designed to mimic the usage of color terms in human language [78]. Color names are terms used by humans to communicate color, such as “green”, “black”, and “crimson”. A linguistic study identified that the English language has eleven basic color terms: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow [5].

The color name descriptor is based on the eleven basic color terms. We use the mapping learned from Google images to transform RGB to a probability over the color names [78]. This allows us to represent patches as histogram over the eleven color names. If we look at the shape of color names in the RGB cube we see that in general they form a wedge, like a slice of cake, on the chromatic plane (formed by  $O1$  and  $O2$ ) and that they are elongated along the intensity ( $O3$ ) axis [4]. This means that they have a certain amount of photometric invariance since values with similar hue and saturation are mapped to the same color name. However, there are also achromatic color names (‘black’, ‘grey’ and ‘white’), which are not photometrically invariant, but which improve the discriminative power of the descriptor. Possibly, because of this mixture of photometric invariance and discriminative power, color names were successful in both image classification [46] and object detection [45]. Finally, they have the additional advantage of being a very compact representation at only 11 dimensions.

## 5.4 Combining Color and Shape for Action Classification

As mentioned earlier, for action classification the bounding box information of humans performing actions are available both at training and test time. Given a test image, the task is to predict an action category label for each human bounding box. For action classification we concentrate the popular bag-of-words framework which has shown promising results on action classification in still images [12, 67, 68]. Here we discuss, within the context of action classification, different fusion approaches proposed in literature for combining color and shape cues within the bag-of-words framework. Throughout this chapter we use the SIFT descriptor for describing the shape of local image patches [58].

Figure 5.2 shows the bag-of-words action representation which is considered in this chapter. The bounding boxes of people in action are provided with each dataset and are used as input to the action classification algorithm at both training and test time. Throughout the chapter we will ignore background information and only describe the information within the bounding box of the



**Figure 5.2:** We apply a three level pyramid on the bounding boxes of the action recognition datasets. Separate BOW histograms are constructed for each cell and are concatenated to form the final action descriptor. In this chapter we use a pyramid representation with three levels, yielding a total of 14 cells.

person in action.<sup>3</sup> For all image representations, we incorporate spatial information via a spatial pyramid [54]. A histogram over a visual vocabulary is constructed for each of the cells of the pyramid, which has been found to yield excellent action classification results [12].

In the BOW representation for action classification color and shape can be fused at different stages. We categorize fusion techniques as early or late fusion methods based on whether fusion is performed before or after the vocabulary assignment stage.<sup>4</sup> Pipelines for several fusion methods are illustrated in Figure 5.3.

Before discussing the various fusion methods we introduce some mathematical notation. The final representation of an action region is obtained by concatenating the  $C$  cells of the pyramid representation into a single histogram  $H = [h_1, \dots, h_C]$ , where  $h_i$  corresponds to the histogram of visual words in spatial pyramid cell  $i$ . Visual vocabularies are denoted by  $W^k = \{w_1^k, \dots, w_{V^k}^k\}$ , where  $w_i^k$  represents the  $i$ -th visual word from visual vocabulary  $k$ , and  $V^k$  is the total number of visual words in vocabulary  $k$ . The superscript  $k \in \{s, c, sc\}$  indicates the visual vocabulary used:  $s$  for shape,  $c$  for color and  $sc$  for a combined shape color vocabulary.<sup>5</sup> The features in the image can be assigned to these vocabularies and we use  $x_j^k$  to denote the assignment of the feature indexed by  $j$  to vocabulary  $W^k$ . We use  $j \in c_i$  to indicate all indexes of the features which are part of cell  $i$ . Then, the histogram of cell  $i$  for cue  $k$  is given by

$$h_i^k(w_n^k) \propto \sum_{j \in c_i} \delta(x_j^k, w_n^k), \quad (5.5)$$

<sup>3</sup>Only in experiment 5.6.2.3 do we add additional information from the background.

<sup>4</sup>Note that the terminology of early and late fusion varies. In some communities early fusion refers to combination before the classifier and late fusion to combination after the classifier [52].

<sup>5</sup>The combined vocabulary  $sc$  is constructed by concatenating the shape and color features before constructing the vocabulary in the combined feature-space.

where  $\delta$  is the Dirac delta function:

$$\delta(x, y) = \begin{cases} 0 & \text{for } x \neq y \\ 1 & \text{for } x = y \end{cases} \quad (5.6)$$

### 5.4.1 Standard Late Fusion

Late feature fusion involves combining the color and shape after vocabulary assignment. The two visual cues are represented by a histogram over their corresponding visual vocabularies. The two histograms are concatenated into a single representation before training and classification. Thus, the final histogram of cell  $i$  is  $h_i = [h_i^s, h_i^c]$ . Late fusion was found to be beneficial for man made categories where color and shape features are more likely to be independent [48].

### 5.4.2 Standard Early Fusion

Early fusion involves combining color and shape at an early stage of the BOW pipeline. The histogram of cell  $i$  is given by  $h_i = h_i^{sc}$ . Early fusion is based on a joint color-shape visual vocabulary. Visual vocabularies based on early fusion possess high discriminative power due to the fact that visual words are described by both color and shape cues. Early fusion was found to be a good representation for natural classes, such as flowers and animals, where color and shape cues are dependent [48].

### 5.4.3 Channel-based Early Fusion

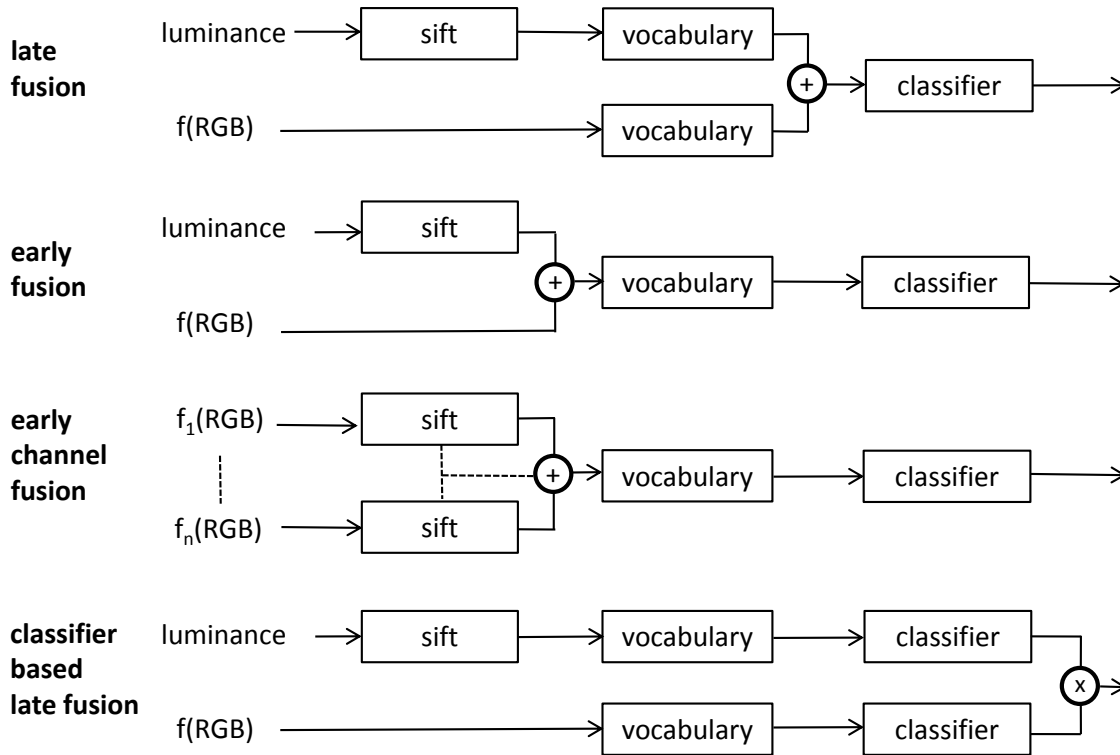
Channel-based fusion for color and shape was first proposed by [1] and later extensively investigated and tested by [75]. First, a color space transform is performed, after which the SIFT descriptor is computed on each channel. The resulting SIFT descriptors are concatenated for all channels before vocabulary assignment. The histograms of each cell are similar to standard early fusion and are given by  $h_i = h_i^{sc}$ . However, while in standard early fusion the SIFT feature is combined with a pure color descriptor, in channel-based early fusion SIFT descriptors are computed on different color representations of the image, after which the various SIFT descriptors are concatenated. We follow [75] and evaluate five different channel-based descriptors: RGB-SIFT, RG-SIFT, OPP-SIFT, C-SIFT and HSV-SIFT.

### 5.4.4 Classifier-based Late Fusion

Another form of late fusion commonly used in image classification is combination of multiple cues at the kernel level. In these approaches, separate classifiers are trained for each visual cue and the results are combined to obtain the final classification score. In our case, with separate color and shape cues, the inputs to the classifiers are individual histograms  $H^s = [h_1^s, \dots, h_C^s]$  and  $H^c = [h_1^c, \dots, h_C^c]$ . In the work of [31] it was shown that addition and product of different kernels yield excellent classification performance comparable to more complicated Multiple Kernel Learning (MKL) methods. In this work we also evaluate these two kernel combination approaches.

### 5.4.5 Color Attention-based Late Fusion

The next two fusion methods which we discuss both aim to introduce the feature binding property into late fusion methods. Feature binding is the property that color and shape are fused at the



**Figure 5.3:** Pipelines for four different fusion methods. The fusion between color and shape is indicated by a 'plus' in case of concatenation of vectors or vocabulary histograms. In the case of classifier based fusion, the encircled multiplication and sum symbols refer to the two methods of classifier fusion investigated: summation and multiplication, respectively, of their outputs. The function  $f(\text{RGB})$  refers to a mapping of  $\text{RGB}$  values to another color-space representation. The "vocabulary" modules refer to vocabulary assignment and have histograms as output. Methods which perform fusion before vocabulary assignment are called early fusion methods, otherwise they are late fusion approaches.

feature level and remain coupled throughout the BOW pipeline. In late fusion this property is lost because, after the separate histograms over the shape and color words are constructed, it is impossible to infer what color word was associated with what shape word in the original images. For example, we know that there are circles and squares in the images and red and blue features, but after discarding the location of each feature, we can no longer say if there are red circles or blue squares present. Early fusion possesses the feature binding property because a combined shape-color vocabulary is used, possibly with a separate words for red circles and blue squares. However, combined shape-color vocabularies yield inferior results for classes were one cue varies considerably, as is often the case with man-made objects. Both color attention and portmanteau vocabularies aim to introduce feature binding into the late fusion pipeline.

The color attention [48] method follows the same pipeline as late fusion (see the first pipeline in Figure 5.3), however the concatenation operator is replaced by a color attention algorithm. In color attention the color cue is used to modulate the shape histogram. This modulation is class dependent, and the final representation of cell  $i$  is given by concatenating class-specific histograms:

$$h_i = [h_i^{cl_1}, \dots, h_i^{cl_m}], \quad (5.7)$$

where  $m$  is the number of classes. For each class the histogram is computed as:

$$h_i^{cl_t}(w_n^s) \propto \sum_{j \in c_i} p(cl_t | w_j^c) \delta(x_j^k, w_n^s), \quad (5.8)$$

where the only difference with respect to computing a shape histogram according to Eq. 5.5 is the modulation with  $p(cl_t | w_j^c)$ . This is the probability of the class  $cl_t$  given the color word  $w_j^c$ . As a consequence, shape words are distributed over the class-specific histograms according to  $p(cl_t | w_j^c)$ . For example, in a two class problem of oranges and apples, a shape word which coinciding with an orange feature will end-up primarily in the orange histogram. The advantage of this representation is that it has the property of feature binding since color and shape are combined at the feature level, while a drawback is that it scales with the number of classes. For more details on color attention-based representations, see [48].

The probability  $p(cl_t | w_j^c)$  can be computed in several ways. We consider three scenarios. In the first we compute a different probability for each of the cells indicated by  $i$ :

$$p_i(cl_t | w_n^c) = \frac{\sum_{s \in cl_t} \sum_{j \in c_i^s} \delta(x_j^c, w_n^c)}{\sum_s \sum_{j \in c_i^s} \delta(x_j^c, w_n^c)}, \quad (5.9)$$

where the occurrence of color feature  $w_n^c$  in cell  $i$  for class  $cl_t$  is divided by the occurrence of the same feature in cell  $i$  of all classes. The second scenario uses the same  $p(cl_t | w_j^c)$  for all cells of the object:

$$p(cl_t | w_n^c) = \frac{\sum_{s \in cl_t} \sum_i \sum_{j \in c_i^s} \delta(x_j^c, w_n^c)}{\sum_s \sum_i \sum_{j \in c_i^s} \delta(x_j^c, w_n^c)}, \quad (5.10)$$

which removes the dependence on  $i$ . The cell-dependent probability can learn a richer color model, for example that the gold of the trumpet-playing action is more common in the top part of the image. The second representation is less noisy since it is based on the combined statistics of all cells. The third scenario which we evaluated uses the average of the two probabilities. We found this to obtain the best results and use it in all experiments on color attention-based fusion of shape and color.



**Figure 5.4:** Example portmanteau clusters from the Willow and Stanford-40 datasets. Note that each portmanteau cluster constitutes a distinct pattern of shape and color. Moreover, several clusters are representative of humans and specific actions such as gardening.

#### 5.4.6 Portmanteau Vocabulary-based Fusion

A second approach to introducing feature binding into the late fusion representation is through portmanteau vocabularies [46]. Portmanteau vocabularies are based on the observation that a simple way to obtain feature binding is by considering a product vocabulary of shape and color:

$$\begin{aligned} W &= \{w_1, w_2, \dots, w_T\} \\ &= \{\{w_q^s, w_r^c\} \mid 1 \leq q \leq V^s, 1 \leq r \leq V^c\}. \end{aligned} \quad (5.11)$$

The main drawback of this is that this leads to very large vocabularies of size  $T = V^s \times V^c$ . In [46] this is countered by discriminatively learning a compact vocabulary starting from the product vocabulary. The compact vocabulary is chosen to minimize the loss in discriminative power caused by the clustering of words. The clustering is based on  $p(cl_t | w_n)$ . Similarly as for color attention, we tested three scenarios: different discriminative vocabularies for each cell based on statistics only from the cell, one discriminative vocabulary for all cells, and discriminative vocabularies for each cell based on an average of cell statistics and whole bounding box statistics. Again, we found the last strategy to perform best and we use it in our experiments. Figure 5.4 shows example portmanteau clusters from the Willow and Stanford-40 datasets. The clusters show homogeneity among color and shape cues. Moreover, they also encode high level information. For instance the first cluster in the bottom row of Figure 5.4, containing many patches with hands and plants, clearly encodes information about the gardening class.

## 5.5 Combining Color and Shape for Action Detection

Action detection is the problem of simultaneously localizing and classifying an action. In this task, the bounding box information is only available at the training time. To investigate the influence of color for action detection, we incorporate color into the popular part-based object detection method of [25]. Instead of learning one model for each object class, we use the method to learn a model for each action class.

In part-based object detection such as that of [25], each object is modeled as a deformable collection of parts with a root model at its core. The root filter can be seen similar to the standard HOG-based representation of [11]. To learn a classifier in the part-based framework a latent SVM formulation is employed. The root filter, the part filters and the deformation cost of the configuration of all parts are concatenated to obtain a detection score for a window. To represent the root and the parts, a dense grid of 8x8 non-overlapping cells is used. For each cell, a one-dimensional histogram of HOG features is computed over all the pixels.

Conventionally, HOGs are computed densely to represent an image capturing local intensity changes. We evaluate two methods of incorporating color into object detection<sup>6</sup>. The first method, which we call channel fusion, computes HOGs separately on the three channels and concatenates the result:

$$D_i = [HOG_i^R, HOG_i^G, HOG_i^B], \quad (5.12)$$

where  $D_i$  is the representation of HOG cell  $i$ . We evaluate the channel fusion approach for RGB, RG, OPP, HSV and C color spaces. The original HOG representation of 31-dimensions is thus extended to a representation of 93-dimensions for all color spaces except for RG-HOG which has 62 dimensions.

The second combination method we consider, which we call late fusion, concatenates the HOG cell representation and a color representation:

$$D_i = [HOG_i, C_i], \quad (5.13)$$

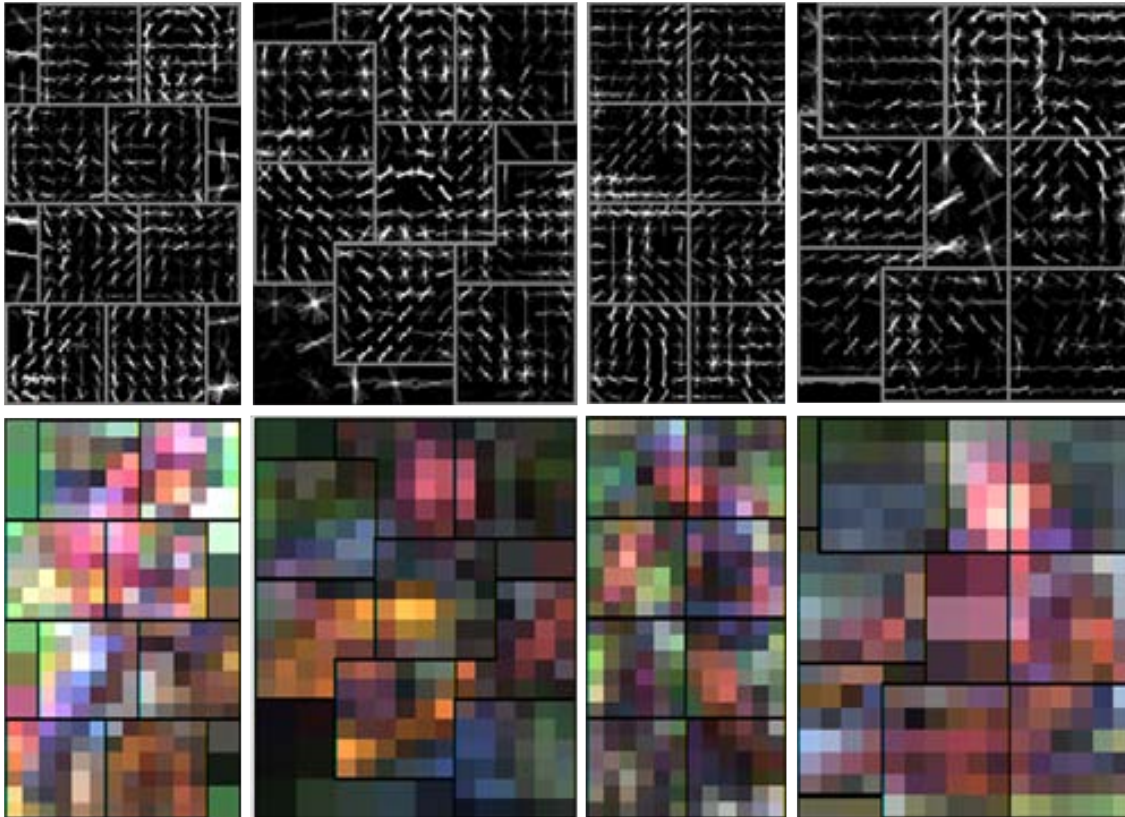
where  $C_i$  is a color descriptor. This concatenated representation thus has dimensionality 31 plus the dimension of the color feature. We evaluate this fusion method for all descriptors described in Section 5.3. All pure color descriptors except color names have 36 dimensions. For the *RGB* and *C* descriptor we learned a visual vocabulary of 36 words, and appended a histogram over these words to the HOG representation.

Part-based detection using luminance features is already a computationally demanding task. Training the part-based model for just a single class can require over 3GB of memory and take over 5 hours on a modern, multi-core computer. When extending HOG descriptors with color information it is therefore imperative to use a color descriptor as compact as possible both because of memory usage and because of total training time. In Table 5.6 we compare the dimensionality of the feature dimensions of different extensions of the part-based method.

Also note that throughout the learning of the part-based model both shape and color are employed. Therefore, augmenting the part-based framework with color information yields significantly different models than those obtained using shape alone. Examples of four models from the Stanford-40 dataset are given in Figure 5.5. One can see that the color model picks up the skin color as well as the color of accompanying objects or context such as horse, guitar and water.

---

<sup>6</sup>Due to the absence of a vocabulary stage, several of the fusion methods explained in Section 5.4 cannot be applied to part-based object detection.



**Figure 5.5:** Visualization of learned part-based models using CN-HOG on the Stanford-40 action dataset. Both the HOG and color names components of our trained models combined in a late fusion are shown. Each color cell is represented using the color obtained by multiplying the SVM weights for the 11 CN bins with a color representative of the color name. Top row: the HOG models for riding horse, playing guitar, riding bike and rowing boat. Bottom row: color models of the respective categories. In the case of horse riding, the brown color of the horse in the bottom with a person sitting on top of it is evident.

## 5.6 Experiments

In this section we introduce the datasets used in the experiments and present our results on color descriptors and fusion techniques for action classification and detection.

### 5.6.1 Action Recognition Datasets

For our experimental evaluation, we use three standard action recognition datasets: Willow, PASCAL VOC 2010 and Stanford-40. The Willow dataset is a dataset consisting of 7 different action categories: interacting with computer, photographing, playing music, riding bike, riding horse, running and walking.<sup>7</sup>

The PASCAL VOC 2010 dataset consists of 9 different action categories: phoning, playing instrument, reading, riding bike, riding horse, running, taking photo, using computer and walking.<sup>8</sup>

Finally, we also present results on Stanford-40, which is one of the most challenging action

<sup>7</sup>The Willow dataset is available at:<http://www.di.ens.fr/willow/research/stillactions/>

<sup>8</sup>PASCAL 2010 is available at:<http://www.pascal-network.org/challenges/VOC/voc2010/>





**Figure 5.6:** Example images from the three datasets used to evaluate color descriptors and fusion techniques. Top row: images from the Willow dataset. Middle row: images from PASCAL VOC 2010 action recognition dataset. Bottom row: example images from the Stanford-40 dataset.

recognition datasets currently available.<sup>9</sup> Stanford-40 consists of 9532 images of 40 different action categories such as jumping, repairing a car, cooking, applauding, brushing teeth, cutting vegetables, throwing a frisbee, etc. The large number of action categories make this dataset particularly challenging. Figure 5.6 shows some of example images from the three datasets.

### 5.6.2 Coloring Action Classification

Here we present our experimental evaluation for the problem of action classification. As mentioned earlier, action classification involves predicting the action category given the bounding box of a person both at training and testing time. We present results using pure color descriptors, a variety of fusion techniques, and a combination of different fusion techniques.

We follow the standard bag-of-words pipeline for all experiments on action classification. Dense sampling at multiple scales is used to extract descriptors from image regions. For shape representation we use the SIFT descriptor, now the de facto standard for shape description in BOW models. For color descriptor evaluation, we use the six pure color descriptors discussed in Section 5.3 above. For shape we construct a visual vocabulary of 1000 words, and for color we use a visual vocabulary of 500 words. In case of early fusion, portmanteau and channel-based representations, we use a larger visual vocabulary of 1500 visual words. For early fusion, the histogram representations are normalized before concatenation. The RGB and C descriptors are normalized to be in the range  $[0, 1]$ , whereas in the case of channel-based fusion the normalization is applied per channel. In all cases, the final image representation is based on a spatial pyramid of three levels ( $1 \times 1$ ,  $2 \times 2$ , and  $3 \times 3$ ), yielding a total of 14 cells [54]. For classification, we use a nonlinear SVM with a  $\chi^2$  kernel [99]. For classifier fusion we use the addition of different kernel responses since in all our experiments it was shown to provide superior results compared to multiplication of kernels. In our experiment we do not use a weighting parameter to tune the trade-off between color and shape. Since the test set for the PASCAL VOC 2010 dataset is withheld by the organizers, for pure color descriptors and fusion strategies experiments are performed on the validation set. However, the final results using different fusion methods are obtained by performing experiments on the PASCAL VOC 2010 test set.

Method	Dimensions	Vocabulary size	Willow	PASCAL VOC 2010	Stanford-40
RGB	3	500	40.0	31.2	15.9
HUE	36	500	38.3	30.8	13.7
Opp-Angle	36	500	32.7	25.9	10.7
HS	36	500	32.9	28.8	10.9
C	3	500	40.0	32.3	15.6
CN	11	500	<b>44.7</b>	<b>34.4</b>	<b>17.6</b>

**Table 5.1:** Performance evaluation of pure color descriptors on the three datasets. Performance is measured by mean AP over the action categories. Note that on all three datasets the color names descriptor yields the best performance.

Method	Dimensions	Vocabulary size	Willow	PASCAL VOC 2010	Stanford-40
SIFT	128	1000	64.9	54.1	38.6
RGB-SIFT	384	1500	<b>65.6</b>	53.7	39.4
RG-SIFT	256	1500	65.0	<b>54.6</b>	<b>39.6</b>
Opp-SIFT	384	1500	63.0	49.8	35.3
HSV-SIFT	384	1500	59.2	50.6	37.0
C-SIFT	384	1500	62.6	52.7	37.6

**Table 5.2:** SIFT and channel-based color descriptors on the three action datasets. RGB-SIFT yields the best results on the Willow dataset, while the best performance on the PASCAL VOC 2010 and Stanford-40 action datasets is achieved using RG-SIFT.

### 5.6.2.1 Pure Color Descriptors for Action Classification

We compare six different color descriptors using the same experimental settings. Table 5.1 shows the results on all three datasets. On the Willow dataset, a significant gain of 4.7 in mean AP is obtained using the color names descriptor compared to the second best color descriptor. On the PASCAL VOC 2010 dataset, the best results are achieved again by using the color name descriptor yielding a mean AP of 34.4. Finally, on the Stanford-40 dataset, both the RGB and C descriptors yield similar results of 15.9 and 15.6, respectively. Similar to the previous two datasets, and despite the great diversity in action categories in Stanford-40 and the compactness of the descriptor, the best performance of 17.6 is achieved again by using the color name descriptor.

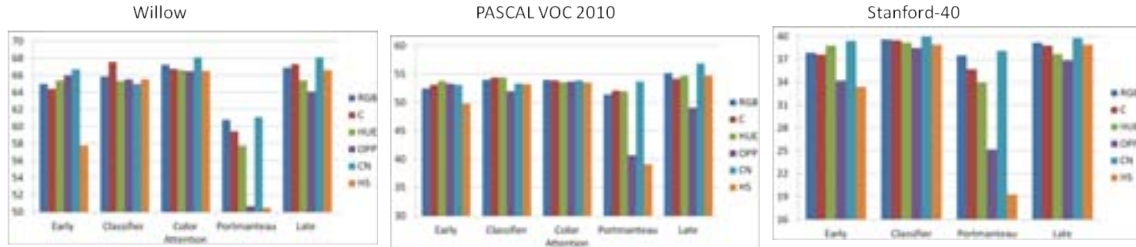
In summary, the color names descriptor significantly outperform other pure color descriptors on all three datasets. As previously mentioned, color names possess a certain degree of photometric invariance with the additional ability to encode achromatic colors, which leads to higher discriminative power than other descriptors. This further strengthens the argument that a balance in photometric invariance and discriminative power is essential when incorporating color descriptors in recognition pipelines.

### 5.6.2.2 Fusion Techniques for Action Classification

Here we present results obtained on the three datasets using different approaches fusing color and shape cues. We present first the results using channel-based representations.

For all channel based color descriptors we construct a visual vocabulary of 1500 words and build a spatial pyramid for the final image representation. Table 5.2 shows the results of using different channel-based fusion approaches on the three datasets. On Willow, shape alone yields a mean AP of 64.9. The best results are achieved using RGB-SIFT which provides an improvement of 0.7 in mean AP over shape alone. On the PASCAL VOC 2010, shape alone yields a mean AP of 54.1.

<sup>9</sup>The Stanford-40 dataset is available at <http://vision.stanford.edu/Datasets/40actions.html>



**Figure 5.7:** Performance comparison of different approaches to fusing color and shape. The choice of color descriptor is crucial for portmanteau-based image presentations. On all three datasets, late fusion performs better than early fusion, and the best results are obtained using late fusion with color names.

The best performance of 54.6 is obtained using RG-SIFT on this dataset. On the more challenging Stanford-40 dataset, shape alone provides a mean AP of 38.6, Opp-SIFT and C-SIFT 37.0 and 37.6, respectively. Finally like the PASCAL VOC 2010 dataset, the best results are obtained using RG-SIFT.

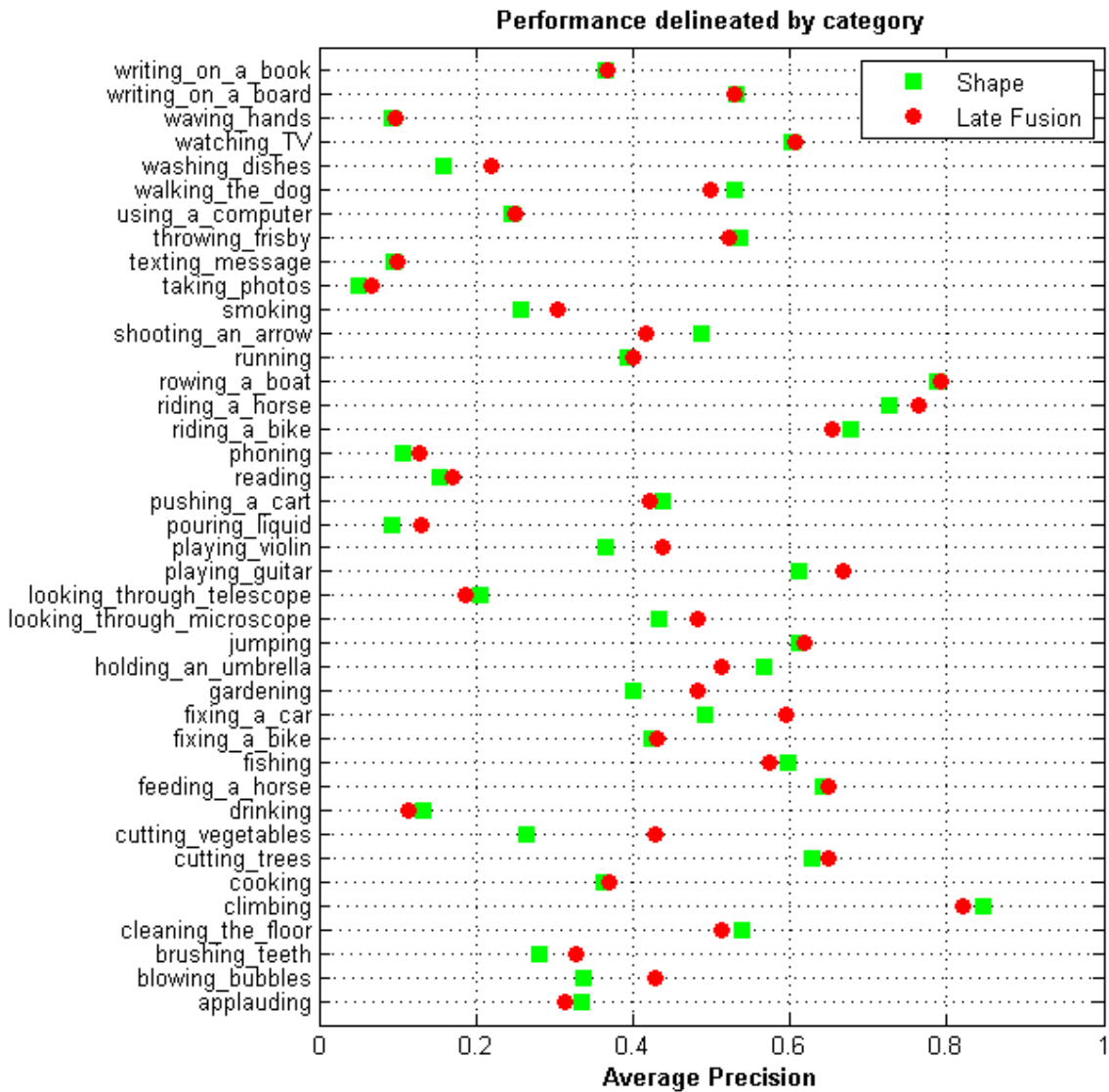
In conclusion, our experimental results suggest that, unlike image classification, both Opp-SIFT and C-SIFT provides inferior performance. RG-SIFT and RGB-SIFT provide the best performance on the action recognition datasets. Channel based fusion approaches fail to provide a significant improvement over shape alone on both Willow and PASCAL VOC datasets.

Figure 5.7 presents the results of different fusion strategies on the three datasets. On the Willow dataset, a combination of shape with the color name descriptor provides the best performance. Among all the different fusion strategies, the best results are obtained using color attention and late fusion. It is worthwhile mentioning that in both cases the best results are obtained with the color name descriptor. For portmanteau-based image representation the choice of color descriptor is extremely crucial with the best performance provided by color names.

On the PASCAL VOC 2010 dataset, in both early and classifier fusion settings, the best performance is achieved using the HUE descriptor and shape. For color attention, the choice of color descriptor is not crucial since all of them provide similar performance. The choice of color descriptor is most crucial for portmanteau-based image representations where color names provide significantly improved results. On this dataset again late fusion yields the best performance. Moreover the best result of 56.9 is obtained using late fusion of color names and shape. Picking the right fusion strategy (late fusion) together with the best color descriptor (color names) provides a significant performance gain of 2.8 over shape alone.

On the Stanford-40 dataset, combining color with shape at a later stage provides best results as shown in Figure 5.7. We do not compare the color attention approach on this dataset due to its very high dimensionality. In all cases combining shape with color names provides the best performance. Among all fusion approaches, both late fusion and classifier-based methods yield the best performance. Figure 5.8 gives a per-category performance comparison between late fusion with color names and shape alone. Note that for most of the action categories a combination of color and shape improves the results compared to shape alone. A significant improvement is obtained on categories such as cutting vegetables, fixing a car, gardening, looking through a microscope and playing a violin. This shows that despite the large variation in classes, color is still able to improve performance. However, the right choice of color descriptor together with the correct fusion strategy is crucial to obtain the optimal performance gain.

Our experimental evaluation of different fusion approaches shows that color, when combined with shape, improves performance for action classification in still images. On the Willow dataset, the best fusion approach yields a mean AP of 68.1 compared to 64.9 obtained using shape alone.



**Figure 5.8:** Per-category comparison between late fusion and shape alone. Here late fusion refers to fusion of color names and shape. On many action categories combining color and shape improves performance over shape alone.

	int. computer	photographing	playingmusic	ridingbike	ridinghorse	running	walking	mean AP
[12]	58.2	35.4	73.2	82.4	69.6	44.5	54.2	59.6
[13]	56.6	37.5	72.0	90.4	75.0	59.7	57.6	64.1
[68]	59.7	42.6	74.6	87.8	84.2	56.1	56.5	65.9
[69]	<b>64.5</b>	40.9	75.0	<b>91.0</b>	<b>87.6</b>	55.0	59.2	67.6
Our approach	61.9	<b>48.2</b>	<b>76.5</b>	90.3	84.3	<b>64.7</b>	<b>64.6</b>	<b>70.1</b>

**Table 5.3:** Comparison of our fusion combination approach with state-of-the-art results on the Willow dataset. On this dataset, our approach provides best results on 4 out of 7 action categories. Moreover, we achieve a gain of 2.5 mean AP over the best reported results.

	phoning	playingmusic	reading	ridingbike	ridinghorse	running	takingphoto	usingcomputer	walking	mean AP
[60]	49.6	43.2	27.7	<b>83.7</b>	89.4	85.6	31.0	59.1	67.9	59.7
[67]	45.5	54.5	31.7	75.2	88.1	76.9	32.9	64.1	62.0	59.0
[13]	48.6	53.1	28.6	80.1	<b>90.7</b>	85.8	33.5	56.1	69.6	60.7
[96]	42.8	60.8	41.5	80.2	90.6	87.8	<b>41.4</b>	<b>66.1</b>	<b>74.4</b>	<b>65.1</b>
[65]	<b>55.0</b>	<b>81.0</b>	<b>69.0</b>	71.0	90.0	59.0	36.0	50.0	44.0	62.0
Our approach	52.1	52.0	34.1	81.5	90.3	<b>88.1</b>	37.3	59.9	66.5	62.4

**Table 5.4:** Comparison with state-of-the-art results on the PASCAL VOC 2010 test set. Despite the simplicity, our approach which combines several color-shape fusion strategies still provides comparable results to best methods on this dataset. Note that, unlike our technique, state-of-the-art approaches typically use standard object detectors to model person-object interactions. Such approaches are complementary to our method and can be combined to further improve results.

On the PASCAL VOC 2010 validation set, the best fusion strategy yields a mean AP of 56.9 compared to 54.1 obtained using shape alone. A mean AP of 40.0 is obtained using color-shape fusion, compared to 38.6 obtained using shape alone on the Stanford-40 dataset.

In summary, the best performance is achieved when the color names descriptor is used as the color representation, and late fusion consistently yields superior performance gains on all three datasets compared to other fusion approaches. It was shown by [48] that late fusion yields superior performance for object categories where one of the visual cues changes significantly. This is true with most of the action categories such as riding bike, riding horse, cutting vegetables where color changes significantly. The success of late fusion over early fusion at the spatial pyramid level has also been observed by [17]. This superiority is due to the fact that, as we move to finer and finer levels of spatial pyramid representation, late and early fusion become equivalent. In other words, the loss of the binding property in late fusion is less of a disadvantage when using a pyramid representation where the uncertainty of the spatial origin of the feature is limited by the cell size. As a consequence, the demonstrated advantages of color attention and portmanteau vocabularies for image classification are not seen for action recognition.

### 5.6.2.3 Combining Fusion Techniques for Action Classification

Method	Object Bank	LLC	Sparse Bases	EPM	Ours
mAP	32.5	35.2	45.7	42.2	<b>51.9</b>

**Table 5.5:** Comparison of color fusion combination with state-of-the-art results on Stanford-40 dataset. Note that combining fusion approaches yields a significant gain of 6.2 in mean AP over the best reported results in the literature.

In this section we analyze the potential of combining fusion approaches and determine if these strategies are complementary in nature. We combine portmanteau, color attention, early, late and channel-based fusion approaches. Except for channel-based fusion, we use the color names descriptor in all fusion approaches. All the color-shape fusion approaches are trained separately and the final probabilities are summed to form the final decision. As mentioned earlier, we do not performed any feature weighting. However, such color-shape weighting parameters can easily

be introduced for combining different color-shape fusion methods in a multiple kernel learning framework.

Table 5.3 shows results of combining different fusion methods and a comparison to state-of-the-art results on the Willow dataset. The final combination achieves a mean AP of 70.3, which is the best result reported on this dataset [12, 13, 65, 69]. A mean AP of 64.1 is reported by [13] with an approach that models complex interactions between persons and objects. The interactions are modeled using external data to train body part detectors. [68] report a mean AP of 65.9 using a technique determining spatial saliency and an improved version of spatial pyramids. Color-shape fusion approaches, despite their simplicity, improve the state-of-the-art by 2.5 mean AP on this dataset.

A comparison of our fusion combination approach to the state-of-the-art on the PASCAL VOC 2010 dataset is shown in Table 5.4. Most state-of-the-art approaches rely on detection techniques to find human-object relationships. [60] report a mean AP of 59.7 using a poselet detector that captures the pose in multi-scale manner. A mean AP of 62.0 is reported by [65] using a human-centric approach to localize humans and find object-human relationships. The best result of 65.1 is reported by [96] using a technique that learns a sparse basis of attributes and parts. Combining multiple fused color-shape representations using a classical bag-of-words framework without detection information provides comparable results to these more complex methods. It is worth mentioning that the color-based models are complementary to detection-based techniques and the two approaches can be combined to further improve action recognition performance.

Table 5.5 shows a comparison with state-of-the-art performance reported on the Stanford-40 dataset [57, 86, 96]. In order to improve the overall performance on this large dataset we increase the vocabulary size for shape to 4000. Recently [69] report a mean AP of 42.2 based on learning a discriminative collection of part templates. The previous best result of 45.7 was obtained using attributes and parts, where attributes represents human actions and parts are model objects and poselets. This technique is complementary to our color fusion combination and could be used in combination with it. Surprisingly, despite the simplicity of our approach which combines multiple fused color-shape models, the final performance significantly surpasses the state-of-the-art results on this large dataset. A significant gain of 6.2 in mean AP is achieved over the best results reported in the literature [96].

### 5.6.3 Coloring Action Detection

Here we evaluate the performance of color descriptors for action detection. In action detection only training images are labeled with a person. Given a test image, the task is to simultaneously localize and classify the actions being performed by humans in it. All the experiments are performed on the Stanford-40 action dataset. To the best of our knowledge, this is the first time the problem of action detection in still images has been investigated on such a large scale dataset. As in action classification, performance is evaluated in terms of average precision which is the standard way of evaluating classification and detection approaches on the PASCAL VOC datasets.

The deformable part-based approach yields state-of-the-art results for generic object and person detection [22]. Here we investigate this approach for the task of action detection. We augment the conventional part-based framework with color information using channel and late feature fusion<sup>10</sup>. In channel based fusion, HOGs are computed independently on different color spaces. Similar to action classification, we evaluate five different color spaces: RGB, RG, Opponent, C and HSV. Note that channel based fusion results in a high dimensional image representation thereby slowing the whole detection framework. In the case of late fusion, a pure color descriptor is concatenated with a HOG for image representation. We evaluate late fusion approach for all the pure color

<sup>10</sup>We also performed experiments replacing HOG with pure color descriptors but significantly inferior results were obtained.

Method	Dimension	mean AP	Method	Dimension	mean AP
HOG	31	21.7	HS+HOG	67	22.3
OPP-HOG	93	25.7	HUE+HOG	67	25.1
RGB-HOG	93	22.1	OPP+HOG	67	23.8
HSV-HOG	93	24.3	C+HOG	67	23.6
RG-HOG	62	21.7	RGB+HOG	67	23.1
C-HOG	93	24.5	CN+HOG	42	<b>27.5</b>

**Table 5.6:** Comparison of different detection methods on the Stanford-40 dataset. The best performance is achieved by CN-HOG with a significant gain of 5.8 mean AP over standard HOG.

descriptors described in Section 5.3.

In Table 5.6 the results obtained on the Stanford-40 action dataset are presented. The conventional HOG-based deformable part model yields a mean AP of 21.7. A significant performance gain is obtained using most of the color based detectors. Among channel based fusion approaches, the best results are obtained using the OPP-HOG descriptor with a mean AP of 25.7. Most of the late fusion methods also improve the results over luminance alone. The best performance is achieved using the CN-HOG method with a significant performance gain of 5.8 mean AP over standard HOG. It is worthy to mention that color names, while having only 11 dimensions, also provided the best results for the action classification as shown earlier. For 38 out of 40 action categories introducing color information improves the performance. For most action categories the introduction of color information improves performance by a significant margin. For example, on the riding horse category color improves the performance from 49 to 74 AP. The CN-HOG model learns the brown color of the horse (see Figure 5.5) which gives it an advantage over luminance-based detection.

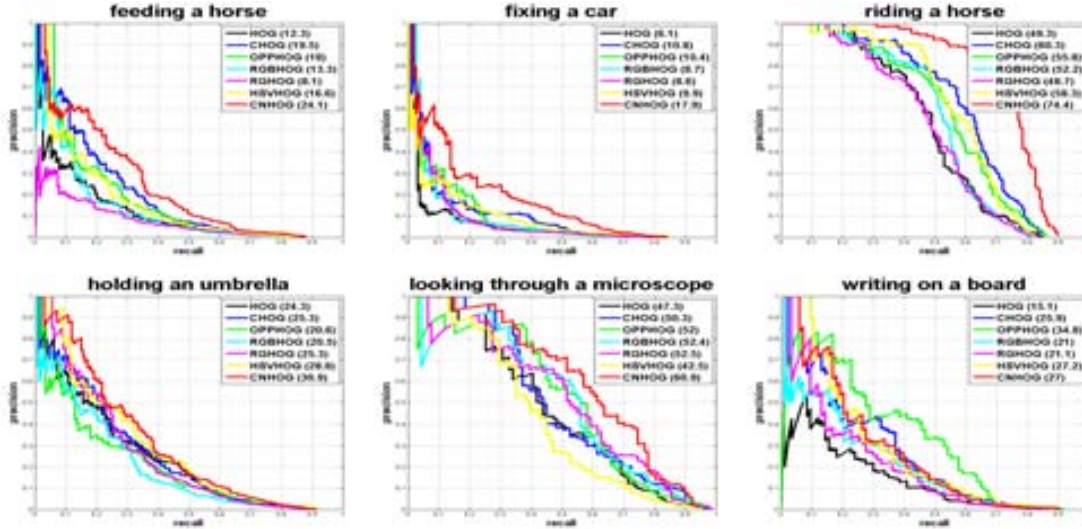
Figure 5.9 shows precision/recall curves on six different action categories from the Stanford-40 dataset. Introducing color information improves performance compared to shape alone on all six categories. Other than the writing on a board class, CN-HOG provides the best performance. In summary, the results clearly suggest that incorporating color information within the part-based framework significantly improves the overall action detection performance. As with action classification, late fusion using color names yields the best performance. This demonstrates that color names, apart from being very compact, are superior to other color descriptors for both action classification and detection.

#### 5.6.4 Analysis of Action Recognition Results

In this section we analyze how color improves action recognition results, with the aim of better understanding what extra information is provided by color. To do so we compare results obtained using the color name descriptor with late fusion, which was found to be superior for both classification and detection, to standard luminance-based recognition.

First we look in more detail at image classification. Figure 5.10 shows the confusion matrix obtained on the Willow dataset using late fusion and the color name descriptor<sup>11</sup>. Overall, color reduces the confusion among categories. The most notable reduction in confusion is between interacting with computer and playing music. Adding color improves performance on most action categories except for riding bike. The remaining confusions are logical such as between running and walking.

<sup>11</sup>The confusion matrix is constructed by assigning each image to the class for which it gets the highest classification score.



**Figure 5.9:** Precision/recall curves of the various channel based approaches, HOG and CN-HOG on six different action categories from the Stanford-40 dataset. Other than the writing on a board action category, CN-HOG provides significantly improved performance over channel based methods.

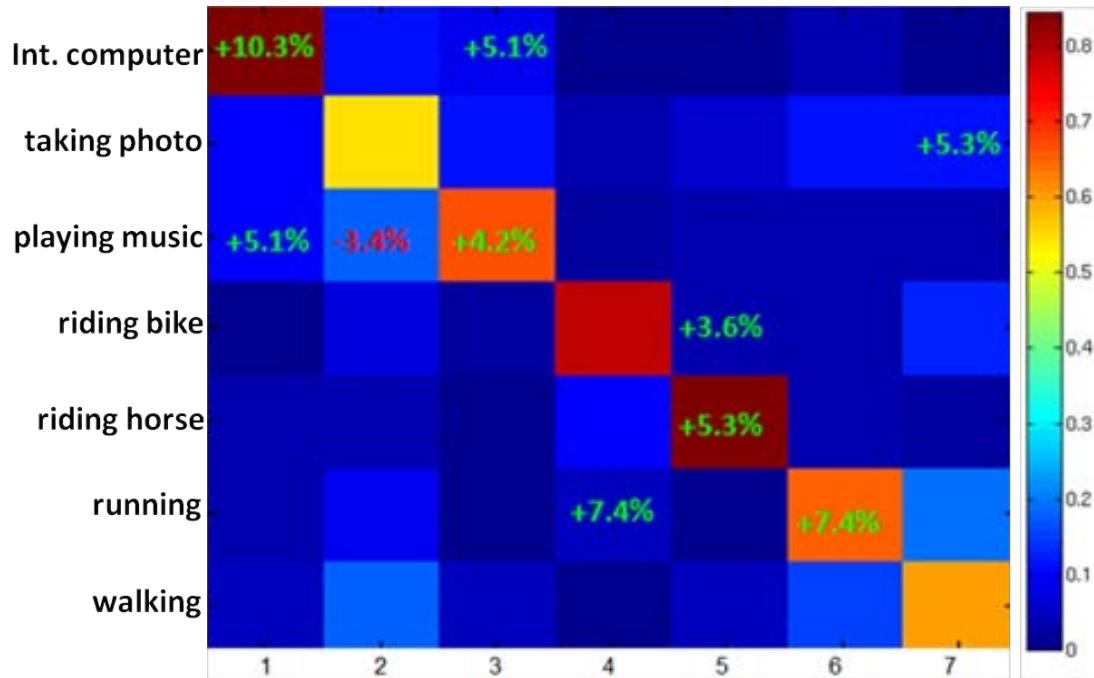
To illustrate further the contribution of color information for action classification we generated heat maps of classifier responses. Heat maps help to identify regions in an image which are discriminative for a particular category. The maps are constructed by projecting the weights of a linear SVM classifier learned for a specific category to the dense grid of feature locations in an image. Figure 5.11 shows heat maps using shape features (second row) and color-shape features (third row) for riding horse, playing guitar and using computer categories. In both “playing music” and “riding horse”, the heatmap of combined shape and color shows that the classifier puts more weight on the discriminating object (the instrument or horse) which defines the action. We also include an example where color deteriorates results: the shape only heatmap for the image of the “using computer” class puts relatively more weight on the keyboard which is important for distinguishing this class.

To better understand the performance improvement obtained by adding color to action detection, we follow the procedure described by [39] for the diagnosis of errors in generic object detectors. This analysis divides the errors made by object detectors into a number of categories, allowing us to analyze which errors are reduced by the introduction of color. We divide the false positive errors which are made in the “top-ranked” detections<sup>12</sup> into three categories. The first category contains errors caused by localization. These occur when the label is correct but the bounding box is misaligned ( $0.1 \leq \text{overlap} \leq 0.5$ ). The second category of errors we consider is due to confusion with other classes. These happen when the bounding box has at least an overlap of 0.1 with an instance of another class in the dataset. The third category is confusion with the background, which we consider to be all false positives which are not in one of the other categories. Typically these occur on textured background areas in the scenes.

Figure 5.12 shows the results of this analysis on the Stanford-40 dataset. There are several observations which can be made from this graph. For most classes the number of errors is reduced by adding color. In the graph this can be observed by noting that the sum of contributions from the three categories of error is positive. Localization errors, however, increase for 24 classes, stay the same for 6, and improve for 10. On the whole dataset adding color resulted in 76 more false detection due to localization errors. However, these additional localization errors are more than

<sup>12</sup>Top ranked detections are the top  $N_j$  detections of a class, where  $N_j$  is equal to the number of positive examples for that class.





**Figure 5.10:** Confusion matrix for late fusion of shape and color names on the Willow dataset. Superimposed are differences with the confusion matrix based on luminance alone for confusions where the absolute change is at least 3%. Late fusion reduces the confusion among different categories in general, but particularly so in the interacting with computer and playing music categories.

compensated by the drop in false positives due to confusion with other classes (304) and the drop in detections on the background (80).

In conclusion, adding color improves action detection mainly because of the drop in errors due to confusions with other classes. However, at the same time adding color increases error due to localization errors. We believe this is caused by the fact that the HOG description is edge based, whereas the color name description is based on RGB values. As a result the “color template” is less localized, meaning that small changes will not lead to drastic changes in the detection score. It is interesting to note that in the human visual system the spatial resolution of color is significantly lower than for luminance [61], implying that for precise localization the human vision system relies on luminance.

## 5.7 Discussion and Conclusion

In this chapter we have performed an extensive evaluation of the contribution of color to action classification and action detection in still images. Inspired by the recent success of color in object and scene recognition, we evaluated a variety of color descriptors and different fusion approaches for action recognition. Experiments on action recognition datasets clearly suggest that color improves performance for action classification and detection. However, as shown in this chapter, a naive combination of color with shape can negatively affect action recognition performance. Therefore, careful selection of color descriptor together with an optimal fusion strategy is crucial to obtaining gains in performance.

Table 5.7 ranks the various color descriptors with respect to their performance for the action classification task. The RGB and C descriptors provide similar performance, while the color name

<b>Willow</b>	<b>PASCAL 2010</b>	<b>Stanford-40</b>
1. CN	1. CN	1. CN
2,3. RGB/C	2. C	2. RGB
—	3. RGB	3. C
4. HUE	4. HUE	4. HUE
5. HS	5. HS	5. HS
6. OPP	6. OPP	6. OPP

**Table 5.7:** The best performing color descriptors on the three datasets used for action classification in this chapter. Note that for all three datasets the color name descriptor is the best choice.

descriptor significantly outperforms all other pure color descriptors and consistently yields the best results on all three datasets.

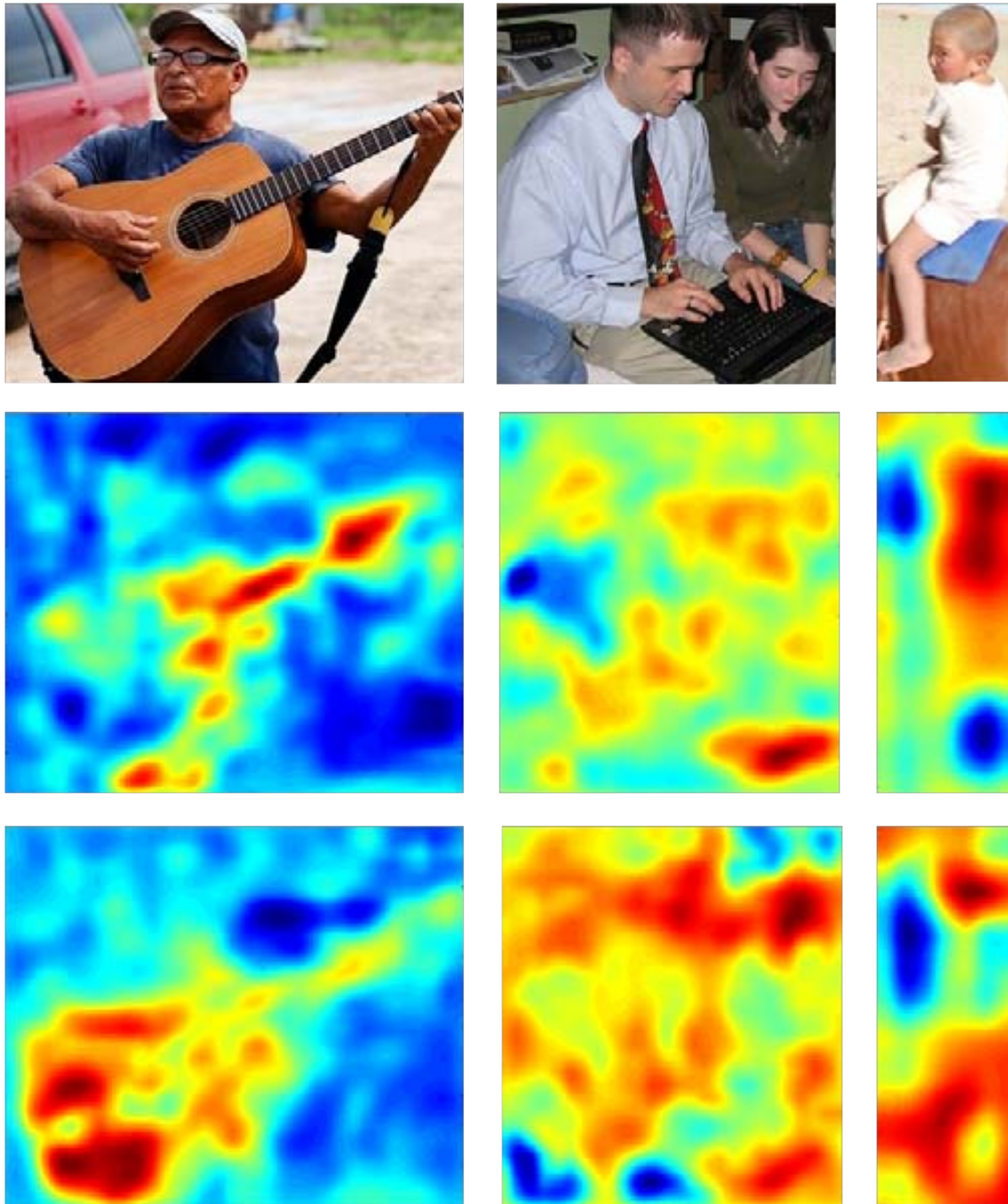
<b>Willow</b>	<b>PASCAL 2010</b>	<b>Stanford-40</b>
1,2. LF/CA	1. LF	1,2. LF/CLF
—	2. CLF	—
3. CLF	3. ColorSIFT	3. EF
4. EF	4. CA	4. Port
5. ColorSIFT	5. EF	5. ColorSIFT
6. Port	6. Port	

**Table 5.8:** Fusion approaches ranked by performance on the three datasets for action classification. Note that late fusion using color names consistently provides the best performance. For Stanford-40 dataset we excluded color attention due to its high dimensionality.

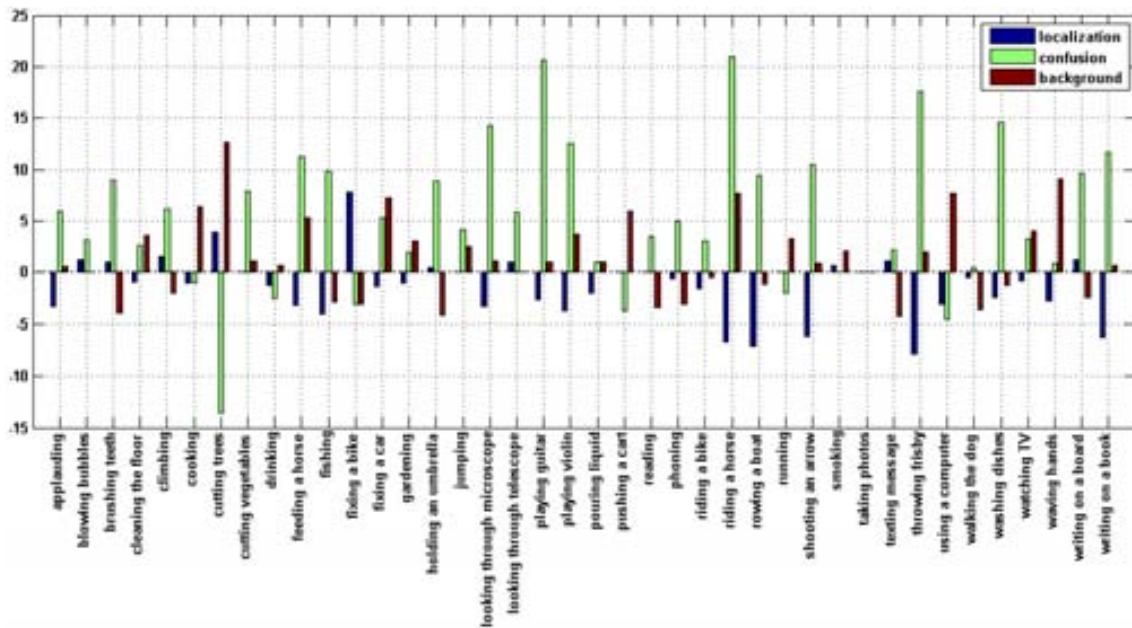
In Table 5.8 we order the different fusion approaches evaluated for action classification in this chapter. We exclude color attention on the Stanford-40 dataset due to its high dimensionality. On all the three datasets, late fusion of color and shape yields better results than early feature fusion.

We have shown that the different fused color representations are complementary in nature and that a naive combination of these different fusions of color and shape further improve performance for action classification. Note that nowhere in this chapter do we use any weighting strategy to leverage the contribution of color and shape. However such a weighting could be learned in an MKL framework using the image representations discussed in this work.

Finally, we also investigated the contribution of color for action detection. In the action detection task, bounding boxes of actors are only available at training time. The problem involves simultaneously classifying and localizing the person performing a specific action. We investigated the incorporation of color in a deformable part-based framework for action detection. A variety of color descriptors were evaluated on the Stanford-40 action dataset and our results clearly suggest that color yields a significant improvement in action detection performance. As with action classification, the color names descriptor results in the best performance for action detection. This further strengthens our conclusion that color names, with its balance of photometric invariance and discriminative power, is the best choice for action recognition.



**Figure 5.11:** Heat maps of classifier responses for playing guitar, using computer and riding horse categories. The top row contains the original images. The second row shows heat maps using shape alone. The third row contains heat maps using fused color-shape features. Despite their being no location information coded into the classifiers, adding color information helps to localize the horse and guitar in the images.



**Figure 5.12:** Analysis of detection errors for the Stanford-40 dataset. The graph shows the decrease in errors which occur when going from standard HOG to CN-HOG (negative changes in the graph signify an increase in errors). Errors are split into errors caused by misaligned localization, confusion with other classes, and detections on the background.

# Chapter 6

## Conclusions and Future Directions

In this thesis work, we focus on improving object detection and action recognition performance in still images by incorporating the color information. This chapter described the main conclusion of the thesis by summarizing the approaches proposed and demonstrating the usefulness of these for object detection and action recognition in still images.

### 6.1 Conclusions

Traditionally, the approaches used in object detection and action recognition are based on luminance information only. The successful deformable part-based models approach for object detection uses HOG for image representation. While the bag-of-words framework widely used for action recognition task describe local image patches by SIFT descriptor. Therefore, we performed a detail analysis of various existing approaches to combine color and shape information for these tasks. The main contribution of this thesis is to enrich existing luminance based object detection and action recognition approaches with color information.

In the first part of the thesis, we focus on the problem of person detection in still images. The approaches used to detect person in images heavily rely on intensity based features for image representation while ignoring the color information. The color information combined with shape has shown to improve performance in other domains such as object and scene recognition. Channel based description is one of the most commonly used approaches for object recognition. We evaluated channel based fusion approaches to combine color and shape information for person detection task in still images. Furthermore, we also investigate the contribution of color in different scenarios such as indoor, urban and countryside. Our experiments reveals that the real benefit of color information comes from indoor and countryside scenarios.

In the second part of the thesis, we analyze the problem of object detection in still images. Channel based fusion approaches previously used for person detection increases the computational cost of the learned model due to increment in feature dimensionality. In addition, channel based fusion decreases results for object categories where one of the visual cue varies significantly. Whereas, late fusion is known to provide better results for a broad range categories in image classification. A consequence of late fusion approach is the use of a pure color descriptor. Therefore, a color attributes is proposed to use as an explicit color representation for object detection. The main benefit of color attribute is that they are compact as well as computationally efficient. Consequently color attributes when combined with shape information provides superior results for object detection task.

Finally, in the last part of the thesis, we investigate the problem of action detection and classification in still images. In action classification task, we have to identify the action being performed by a person in an image. The bounding box of person performing actions are provided both at train and test time for classification. In action detection, we have to simultaneously localize and classify the person action in still images. The bounding box information is only available for training time. Traditionally, the luminance based approaches are used for action detection and classification in still images. We focus on the problem of incorporating color information for action recognition task in still images. In addition, we also analyze different color-shape fusion approaches for action recognition. The experiments performed clearly demonstrate that combining color and shape information improve the performance in action recognition in still images.

The results obtained in this thesis are summarized in the paragraph below:

**Chapter 3: Color Contribution towards Person Detection.** In this chapter, we investigate the use of opponent color space for person detection task. Opponent color space is based on the human vision system. The framework of Dalal *et al.* and part-based detector of Felzenswalb *et al.* are used as baseline for color evaluation. Both frameworks relies on histograms of oriented gradients (HOG) as feature representation. HOG is computed on three opponent channel and concatenated into a final feature vector for the classifier. We obtain better results on per-window and per-image evaluation over the widely used INRIA person dataset in HOG+LinearSVM person detector of Dalal *et al.*. The part-based method augmented with color information also improve the performance on the *person* class of the PASCAL VOC 2007 dataset. Furthermore, to enhance the scene understanding for person detection, we re-annotated the PASCAL VOC 2007 dataset person class according to three scenarios: indoor, urban and countryside. The detector performance improved in indoor and countryside scenarios whereas there is no clear improvement in urban scene.

**Chapter 4: Color Attributes for Object Detection.** In this chapter augmenting color information within object detection frameworks is proposed. Most state-of-the-art object detectors are based on shape or texture information while ignoring the color. Recent approaches to incorporate intensity based detectors with color often provides inferior results for object categories with varying importance of color and shape. Due to the complexity of the object detection task, a careful consideration is required when choosing an approach to augment color into object detection. This chapter introduce the usage of color attributes as an explicit color representation for object detection. Color attributes are compact, computationally efficient, possess some degree of photometric invariance and are capable of encoding achromatic colors. Experimental results shows a significant improvement on the challenging PASCAL VOC datasets where existing color-based fusion approaches have shown to provide below-expected results. Moreover, we introduce a new dataset of cartoon characters where color plays an important contribution. On this dataset, our approach provides a gain of 14% over state-of-the-art intensity based detector.

**Chapter 5: Coloring Action Recognition in Still Images.** This chapter addresses first time an extensive evaluation of the contribution of color for action recognition in still images. The color evaluation for action recognition is inspired by the recent success of color information in other domains such as object and scene recognition. Several color descriptors and different fusion approaches is investigated for action recognition in still images. Experiments conducted on three challenging action recognition datasets clearly show that color improves the performance for action classification and action detection in still images. Combining color and shape in a naive manner can hamper the recognition performance. This further leads to results inferior to shape alone in some cases. Therefore, a careful selection of color feature and the right fusion strategy is required for obtaining significant improvement in performance.

## 6.2 Future Directions

Combining color and shape information using the methods proposed in this thesis has shown to provide excellent results for object detection and action recognition in still images. Existing approaches integrate texture and shape information for object detection task [87, 100]. As a future work, incorporating texture with color and shape in object detection and action recognition in static images is still an open debate.

An interesting future direction will be to investigate how to combine fused color-shape representations with approaches based on object detection and pose estimation. Many approaches to action recognition rely on modeling human-object interactions and we expect that the integration of fused color-shape representations with such approaches will further improve the recognition performance.

Another possible future work is to investigate the potential of color features for object detection and action recognition in videos. Motion features are generally used for video description. Following the success of color for object detection and action recognition in still images, it is worth investigating how color performs for video description.





# 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] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *CVPR*, 2010.
- [3] Rao Muhammad Anwer, David Vazquez, and Antonio M. Lopez. Color contribution to part-based person detection in different types of scenarios. In *CAIP*, 2011.
- [4] R. Benavente, M. Vanrell, and R. Baldrich. Parametric fuzzy sets for automatic color naming. *JOSA*, 25(10):2582–2593, 2008.
- [5] Brent Berlin and Paul Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA, 1969.
- [6] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pls. In *Proc. European Conf. on Computer Vision*, 2006.
- [7] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local colour invariants. *CVIU*, 113:48–62, 2009.
- [8] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.
- [9] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007.
- [10] N. Dalal. *Finding people in images and videos*. PhD Thesis, Institut National Polytechnique de Grenoble / INRIA Rhône-Alpes, 2006.
- [11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Proc. Computer Vision and Pattern Recognition*, 2005.
- [12] Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.
- [13] Vincent Delaitre, Josef Sivic, and Ivan Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011.
- [14] Chaitanya Desai and Deva Ramanan. Detecting actions, poses, and objects with relational phraselets. 2012. In *Proc. ECCV*.
- [15] Chaitanya Desai, Deva Ramanan, and Charless Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.
- [16] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: a benchmark. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

- [17] Noha Elfiky, Fahad Shahbaz Khan, Joost van de Weijer, and Jordi Gonzalez. Discriminative compact pyramids for object and scene recognition. *Pattern Recognition Journal*, 2011.
- [18] M. Enzweiler and D. Gavrilu. Monocular pedestrian detection: survey and experiments, 2009.
- [19] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *Proc. IEEE Int. Conf. on Computer Vision*, 2007.
- [20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes challenge 2007 results.
- [21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2009 results., 2009.
- [22] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. Journal on Computer Vision* , 88(2):303–338, 2010.
- [23] Michael Felsberg and Johan Hedborg. Real-time view-based pose recognition and interpolation for tracking initialization. *J. Real-Time Image Processing*, 2(3):103–115, 2007.
- [24] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.
- [25] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.
- [26] Vittorio Ferrari, Loic Fevrier, Frédéric Jurie, and Cordelia Schmid. Groups of adjacent contour segments for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 36–51, 2008.
- [27] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.
- [28] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Action sequence models for efficient action detection. 2011. In *Proc. CVPR*.
- [29] T. Gandhi and M.M. Trivedi. Pedestrian protection systems: issues, survey, and challenges. *IEEE Trans. on Intelligent Transportation Systems* , 8(3):413–430, 2007.
- [30] K.R. Gegenfurtner and J. Rieger. Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology.*, 10:805–808, 2000.
- [31] Peter Vincent Gehler and Sebastian Nowozin. On feature combination for multiclass object classification. In *Proc. IEEE Int. Conf. on Computer Vision*, 2009.
- [32] D. Gerónimo, A.M. López, A.D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence* , 32(7):1239–1258, 2010.
- [33] David Geronimo, Angel Sappa, Antonio Lopez, and Daniel Ponsa. Adaptive image sampling and windows classification for on-board pedestrian detection.
- [34] Jan-Mark Geusebroek, Rein van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *PAMI*, 23(12):1338–1350, 2001.
- [35] T. Gevers and A. W. M. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, 1999.

- [36] C. L. Hardin and Luisa Maffi. *Color Categories in Thought and Language*. Cambridge University Press, 1997.
- [37] Hedi Harzallah, Frederic Jurie, and Cordelia Schmid. Combining efficient object localization and image classification. In *Proc. IEEE Int. Conf. on Computer Vision*, 2009.
- [38] E. Hering. *Outlines of a theory of the light sense (translated by L.M. Hurvich and D. Jameson)*. Harvard University Press, 1964.
- [39] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV*, 2012.
- [40] Yuxiao Hu, Liangliang Cao, Fengjun Lv, Shuicheng Yan, Yihong Gong, and Thomas S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. 2009. In *Proc. ICCV*.
- [41] J.D. Mollon. "tho' she kneel'd in that place where they grew ..." the uses and origins of primate colour vision. *Journal of Experimental Biology*, 146(1):21–38, 1989.
- [42] J.M. Álvarez, T. Gevers, and A.M. López. Learning photometric invariance for object detection. *Int. Journal on Computer Vision*, 90(1):45–61, 2008.
- [43] M. Jones and D. Snow. Pedestrian detection using boosted features over many frames. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.
- [44] Timothy Jost, Nabil Ouerhani, Roman von Wartburg, Reni Miri, and Heinz Higl. Assessing the contribution of color in visual attention. *CVIU*, 100(1–2):107–123, 2005.
- [45] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Andrew D. Bagdanov, Maria Vanrell, and Antonio M. Lopez. Color attributes for object detection. In *CVPR*, 2012.
- [46] 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.
- [47] 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.
- [48] Fahad Shahbaz Khan, Joost van de Weijer, and Maria Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision*, 2012.
- [49] J. Krauskopf, D.R. Williams, and D.W. Heeley. Cardinal directions of color space. *Vision Research*, 22(9):1123–1132, 1982.
- [50] A. Kristjansson. Independent and additive repetition priming of motion direction and color in visual search. *Psychological Research*, (73):158–166, 2009.
- [51] C.H. Lampert, M.B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008.
- [52] Zhen-Zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G. Hauptmann. Double fusion for multimedia event detection. In *MMM*, 2012.
- [53] I. Laptev. Improving object detection with boosted histograms. *Image and Vision Computing*, 27(5):535–544, 2009.
- [54] 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.

- [55] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3D scene analysis from a moving vehicle. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, 2007.
- [56] Reiner Lenz, Thanh Hai Bui, and Javier Hernandez-Andres. Group theoretical structure of spectral spaces. *Journal of Mathematical Imaging and Vision*, 23(3):297–313, 2005.
- [57] Li-Jia Li, Hao Su, Eric P. Xing, and Fei-Fei Li. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.
- [58] D. G. Lowe. Distinctive image features from scale-invariant points. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.
- [59] 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.
- [60] Subhransu Maji, Lubomir D. Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.
- [61] K. T. Mullen. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of Physiology*, (359):381–400, 1985.
- [62] S. Munder and D. Gavrila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.
- [63] L. Oliveira, U. Nunes, and P. Peixoto. On exploration of classifier ensemble synergism in pedestrian detection. *IEEE Trans. on Intelligent Transportation Systems*, 11(1):16–27, 2010.
- [64] Alain Pagani, Didier Stricker, and Michael Felsberg. Integral p-channels for fast and robust region matching. In *ICIP*, 2009.
- [65] Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari. Weakly supervised learning of interactions between humans and objects. *PAMI*, 34(3):601–614, 2012.
- [66] Steven A Shafer. Using color to separate reflection components. *Color Research and Application*, 10(4):210–218, 1985.
- [67] Nataliya Shapovalova, Wenjuan Gong, Marco Pedersoli, Francesc Xavier Roca, and Jordi Gonzalez. On importance of interactions and context in human action recognition. In *IbPRIA*, 2011.
- [68] Gaurav Sharma, Frederic Jurie, and Cordelia Schmid. Discriminative spatial saliency for image classification. In *CVPR*, 2012.
- [69] Gaurav Sharma, Frederic Jurie, and Cordelia Schmid. Expanded parts model for human attribute and action recognition in still images. In *CVPR*, 2013.
- [70] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM MM*, 2005.
- [71] Ian Spence, Patrick Wong, Maria Rusan, and Naghme Rastegar. How color enhances visual memory for natural scenes. *Psychological Science.*, 17:1–6, 2006.
- [72] Du Tran and Junsong Yuan. Max-margin structured output regression for spatio-temporal action localization. 2012. In *Proc. NIPS*.
- [73] A. Treisman and G. Gelade. A feature integration theory of attention. *Cogn. Psych.*, 12:97–136, 1980.

- [74] K. van de Sande, Jasper R. R. Uijlings, Th. Gevers, and Arnold W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.
- [75] 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.
- [76] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *Proc. European Conf. on Computer Vision*, 2006.
- [77] J. van de Weijer and C. Schmid. Applying color names to image description. In *ICIP*, 2007.
- [78] 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.
- [79] Joost van de Weijer, Theo Gevers, and Andrew D. Bagdanov. Boosting color saliency in image feature detection. *PAMI*, 28(1):150–156, 2006.
- [80] 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.
- [81] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.
- [82] David Augusto Rojas Vigo, Fahad Shahbaz Khan, and Joost van de Weijer and Theo Gevers. The impact of color on bag-of-words based object recognition. In *ICPR*, 2010.
- [83] P. Viola, M.J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *Int. Journal on Computer Vision*, 63(2):153–161, 2005.
- [84] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features, 2001. *CVPR*.
- [85] Stefan Walk, Nikodem Majer, Konrad Schindler, and Bernt Schiele. New features and insights for pedestrian detection. In *CVPR*, 2010.
- [86] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.
- [87] Xioayu Wang, Tony X. Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*, 2009.
- [88] 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.
- [89] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.
- [90] J. M. Wolfe. *Visual Search*. 1998. in *Attention*, edited by H. Pashler, Psychology Press Ltd.
- [91] 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.
- [92] W.R. Schwartz, A. Kembhavi, D. Harwood, and L.S. Davis. Human detection using partial least squares analysis. In *Int. Conf. on Computer Vision*, Kyoto, Japan, 2009.
- [93] Bo Wu and Ramakant Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proc. IEEE Int. Conf. on Computer Vision*, 2005.

- [94] Bo Wu and Ramakant Nevatia. Cluster boosted tree classifier for multi-view, multi-pose object detection. In *Proc. IEEE Int. Conf. on Computer Vision*, 2007.
- [95] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.
- [96] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas J. Guibas, and Fei-Fei Li. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.
- [97] Bangpeng Yao and Fei-Fei Li. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *PAMI*, 34(9):1691–1703, 2012.
- [98] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative video pattern search for efficient action detection. *PAMI*, 33(9):1728–1743, 2011.
- [99] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: An in-depth study. a comprehensive study. *Int. Journal of Computer Vision*, 73(2):213–218, 2007.
- [100] Junge Zhang, Kaiqi Huang, Yinan Yu, and Tieniu Tan. Boosted local structured hog-lbp for object localization. In *CVPR*, 2010.
- [101] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.
- [102] Long Zhu, Yuanhao Chen, Alan L. Yuille, and William T. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010.