Fusing Color and Shape for Bag-of-Words based Object Recognition

Joost van de Weijer[1] and Fahad Shahbaz Khan[2]

[1] Computer Vision Center Barcelona, Edifici O, Campus UAB, 08193, Bellaterra, Spain,
joost@cvc.uab.es,
WWW home page: http://cat.uab.es/ joost/
[2] Computer Vision Laboratory, Linköping University, Sweden

Abstract. In this article we provide an analysis of existing methods for the incorporation of color in bag-of-words based image representations. We propose a list of desired properties on which bases fusing methods can be compared. We discuss existing methods and indicate shortcomings of the two well-known fusing methods, namely early and late fusion. Several recent works have addressed these shortcomings by exploiting top-down information in the bag-of-words pipeline: color attention which is motivated from human vision, and Portmanteau vocabularies which are based on information theoretic compression of product vocabularies. We point out several remaining challenges in cue fusion and provide directions for future research.

Keywords: object recognition, color features, bag-of-words, image classification

1 Introduction

Bag-of-words based object recognition has been among the most successful approaches to object recognition [1][2]. The method represents an image as an orderless collection of local regions, where in general local regions are discretized into a visual vocabulary. Images are represented as a histogram over the visual vocabulary. The method has been shown to obtain excellent results in image classification [1], object detection [3] and image retrieval [4].

The local regions in the images are generally represented by a shape descriptor, predominantly the SIFT descriptor[5]. Color was simultaneously introduced into bag-of-words in [6][7]. Van de Weijer and Schmid [6] proposed to extend the SIFT descriptor with photometric invariant color features. Bosch and Zisserman[7] applied the SIFT descriptor separately on the HSV channels, and concatenated the features of the channels into one single HSV-SIFT feature for each local feature. This idea was further developed and evaluated extensively by Van de Sande et al. [8]. These methods to fuse color and shape are called early fusion methods, because they combine the cues before the vocabulary construction.

Several methods explore the combination of multiple features at the classification stage, among which the well-known multi-kernel methods (MKL)[9]. A weighted linear combination of kernels is employed, where each feature is represented by multiple kernels. Gehler and Nowozin [10] showed that for image classification product of different kernels often provides comparable results to MKL. These methods are typically late fusion methods, because separate image representation for color and shape are constructed after which they are combined at the classifier stage. More recently, Fernando et al. [11] propose to compute a class specific vocabulary, where the visual words are selected from various vocabularies of different cues. The image representations which we discuss in this article can be used as input the MKL methods to further improve performance.

Much research has been dedicated to the investigation of what color feature is optimal to be combined with shape [12][6][8]. The performance gain obtained by color depends — not surprisingly — on the importance of color in the data set: changing from gains of up to 20% on e.g. sports and flower data sets to only a few percent on PASCAL VOC data set. The small gains obtained on the latter have triggered more research on how to optimally combine shape and color [13][14]. These works propose alternatives to the early and late fusion scheme.

In this paper, we analyze existing methods for combining shape and color. We start in Section 2 by listing a number of properties which are desirable for combination methods. Next, in Section 3 we discuss early and late fusion in more detail. A method motivated from human vision, called color attention[14], is analyzed in Section 4 and a special vocabulary construction method, known as Portmanteau vocabularies[13], is investigated in Section 5. We finalize with a discussion and future direction to further improve color and shape fusing.

2 Color in Bag-of-Words Image Representations

Traditionally, bag-of-word representations for object recognition are based on a single cue. In this case the features in the image are represented by a single visual vocabulary. Images are represented by the frequency histogram over the visual words. This representation is often improved with spatial pyramid to incorporate spatial information [15]. The image representations are subsequently provided to a classifier, predominantly an SVM, for image classification. Outside image classification, bag-of-words image representations have also been applied to object recognition [3] and to image retrieval [4].

For long, research focussed on finding the optimal color descriptor to combine with shape. Several evaluation articles exists, see e.g. [6][8]. In general features based on photometric invariance obtain good results [6][8]. Also, in several studies the color name descriptor[16][17], which is based on color terms which humans use to communicate, obtained excellent results [13]. Recently, a bio-inspired descriptor was shown to obtain excellent results [18]. An evaluation of the impact of color in the detection phase is available in [19]. In this article, we focuss on (after having picked a color feature) the optimal approach to fuse it with shape.

Incorporating multiple cues (in our case fusing shape and color) into the bag-of-words representation can be done in many ways, and we will discuss several of them. Before doing so, we enumerate the properties which are expected to be of importance for a successful fusing method:

- Cue compactness: cues are represented by separate vocabularies. This prevents difficulties which arise from learning vocabularies in combined feature spaces. In addition, when categories are only constant over one of the cues (for example cars are constant over shape but vary over color), then cue compactness ensures that the representation is not spread-out.
- Cue binding: cues are combined at the pixel, meaning that if both cues at the same location are positively correlated with a certain class they will reinforce each other.
- Cue weighting: the relative weight of color and shape in the final image representation can be tuned. This is often achieved by means of cross validation.
- Category scalability: the representation scales for large-scale classification problems, which typically contain hundreds of classes. Desirable is that the representation size is independent of the number of classes.
- Multi-cue scalability: the representation allows for multiple cues. Next to color one could for example also consider texture, optical flow, etc.

We will discuss the presence and absence of these properties for several combination methods in the following sections.

3 Early and Late Fusion

In this section, we review the two most popular methods to combine color and shape in the bag-of-words framework. They are called early and late fusion. The nomenclature early or late is dependent on whether the fusion is done before or after the vocabulary construction. We will discuss advantages and disadvantages of both methods.

We start by introducing some mathematical notations. In bag-of-words a number of local features f_{ij}, j=1...M^i are detected in each image I^i, i=1,2,...,N , where M^i is the total number of features in image i. The local features are represented in visual vocabularies which describe various image cues such as shape and color. We assume that visual vocabularies for the cues are available, $\mathrm{W^k} = \{\mathrm{w^k_1}, ..., \mathrm{w^k_{V^k}}\}$, with the visual words $\mathrm{w^k_n}$, n=1,2,...,V^k and $k \in \{s, c, sc\}$ for the two separate cues shape and color and for the combined visual vocabulary of color and shape.

In the case of late fusion, the features f_{ij} are quantized into a pair of visual words $(\mathrm{w^s_{ij}}, \mathrm{w^c_{ij}})$. Separate frequency histograms for shape and color ($k \in \{s, c\}$) are constructed according to:

$$n\left(\mathrm{w^k_n}|I^i\right) \propto \sum_{j=1}^{M^i} \delta\left(\mathrm{w^k_{ij}}, \mathrm{w^k_n}\right) \tag{1}$$

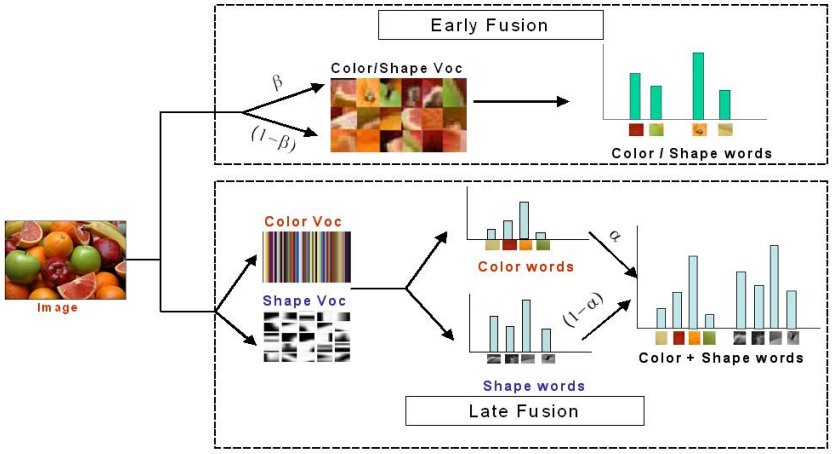

Fig. 1. Early and late fusion schemes to combine color and shape information. The α and β parameters determine the relative weight of the two cues.

with

$$\delta(x, y) = \begin{cases} 0 & \text{for } x \neq y \\ 1 & \text{for } x = y \end{cases} \tag{2}$$

The final representation of the image is then a concatenation of the shape and the color frequency histogram. Often a parameter balancing the relative weight of color and shape is introduced when concatenating the two histograms. This parameter is typically optimized with cross-validation.

In the case of early fusion, the features of color and shape are concatenated before assignment to the vocabulary, after which the image representation is built with:

$$n\left(\mathrm{w}_n^{\mathrm{sc}} | I^i\right) \propto \sum_{j=1}^{M^i} \delta\left(\mathrm{w}_{\mathrm{ij}}^{\mathrm{sc}}, \mathrm{w}_{\mathrm{n}}^{\mathrm{sc}}\right) \tag{3}$$

The final representation of the image is a single multi-cue histogram $n\left(\mathrm{w}_n^{\mathrm{sc}} | I^i\right)$. Here, also a weighting parameter between shape and color could be considered. However, because this parameter changes the vocabulary construction, it is often considered unfeasible due to the high computational cost.

Product vocabularies, which are a special case of early fusion vocabularies, are a good way to understand the differences between early and late fusion. A product vocabulary is constructed by combining every word in the shape vocabulary with every word in the color vocabulary. An example is provided in Figure 2. Now consider early and late fusion for this simple case. The late fusion representation would consist out of a histogram over the shapes (circle, star, and triangle), and a separate histogram over the colors (red, yellow, and blue). The early fusion representation would be a histogram over the nine words which are

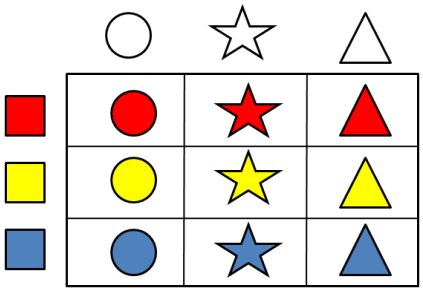

Fig. 2. The product vocabulary combines every shape word with every color word. Product vocabularies help understand the differences between early and late fusion. Furthermore, they are at the bases of Portmanteau vocabularies. See text for more information.

formed by combining all shapes with all colors. Consider now the case that we want to find all images which contain yellow stars (e.g. in children drawings). This is difficult for late fusion since we have one histogram telling us of the presence of a star in the image, and another tells us of the presence of yellow, but we are not sure whether both events happened at the same location in the image. From early fusion, which has a single word for yellow stars, it is easy to infer its presence. If we instead, we want to find all images containing balloons (represented by colored circles), late fusion would provide a good representation, since all balloons are assigned to the circle shape. In this case early fusion would complicate the task of the classifier, since balloons are now represented by multiple words (red circles, blue circles, etc.). In general, classes which have color-shape dependency (like the yellow star) are better represented by early fusion. Instead, classes which have color and shape independency, like most man-made classes (and our balloon example), are better represented by late fusion.

In Table 1 an overview of the properties of early and late fusion is provided. The joined vocabulary which is used in early fusion results in the absence of cue compactness, however it ensures cue binding. In theory feature weighting is possible, but since it is computationally costly, in practice we do not attribute this property to early fusion. Late fusion, does have cue compactness, but lacks cue binding. However, it does allow for feature binding. Both, methods scale relatively well with the number of categories. Late fusion does further scale with the number of cues, which is not the case for early fusion. In the case the of early fusion, the problems which are already becoming evident when constructing a vocabulary for two cues, are only expected to augment for multiple cues.

4 Color Attention

The color attention approach [14] to combining shape and color is motivated from human vision research, where it is widely believed that the basic features

of visual objects such as color and shape are loosely bundled into objects before the arrival of attention [20]. The two well known theories providing the evidence that attention is involved to bind the basic features into a recognizable object are Feature Integration Theory [21] and Guided Search [22]. It is further asserted from these two models that the basic features are initially represented separately before they are integrated at a later stage in the presence of attention. Among several properties of visual stimuli, only few are used to control the deployment of visual attention [23]. Color is one such attribute which is undoubtedly used to guide visual attention [23].

The idea of attention can be introduced into the bag-of-words framework with:

$$n\left(\mathrm{w}_n^s|I^i, class\right) \propto \sum_{j=1}^{M^i} p\left(class|\mathrm{w}_{\mathrm{ij}}^{\mathrm{c}}\right) \delta\left(\mathrm{w}_{\mathrm{ij}}^{\mathrm{s}}, \mathrm{w}_{\mathrm{n}}^{\mathrm{s}}\right), \tag{4}$$

where $p\left(class|\mathrm{w}_{\mathrm{ij}}^{\mathrm{c}}\right)$ is the probability of the class given the color word of the j^{th} local feature of the i^{th} image and is dependent on both the location $\mathbf{x}_{ij}$ and the *class*. In practice $p\left(class|\mathrm{w}_{\mathrm{ij}}^{\mathrm{c}}\right)$ is measured from the labeled training set. In color attention-based bag-of-words the functionality of shape and color have been separated. The color cue is used as the attention cue, and modulates the shape feature (which is called the descriptor cue). The weights $p\left(class|\mathrm{w}_{\mathrm{ij}}^{\mathrm{c}}\right)$ can be interpreted as attention maps, which for every pixel in the image give the probability of the class given the color at that location. An overview of the method is provided in Figure 3. The main difference to standard bag-of-words is that shape-features have more importance in regions with high attention. Note that all histograms are based on the same set of detected features and only the weighting varies for each *class*. As a consequence a different distribution over the same shape words is obtained for each *class* as shown in Fig. 3.

If we look at what properties color attention has, we see that it possesses both cue compactness and cue binding, since it is based on separate vocabularies for shape and color and they are combined directly at the location. As consequence, shape and color features at the same location, which provide evidence of the same class, reinforce each other. Color attention also has the possibility of cue weighting (detailed in [14]). It also scales well with multiple cues, since increasing the number of cues does not influence the final representation size. The main disadvantage of color attention is that the image representation increases linearly with the number of classes. The properties are summarized in Table 1.

5 Portmanteau Vocabularies

Portmanteau vocabularies are a special way to construct visual vocabularies [13]. They are based on the basic insight that product vocabularies (see also Fig. 2) combine the properties cue compactness and cue binding.

The vocabularies for shape and color are given by W^s and W^c with respective vocabulary sizes V^s and V^c. Then the product vocabulary is given by

$$= \{\{\mathrm{w}_{\mathrm{i}}^{\mathrm{s}}, \mathrm{w}_{\mathrm{j}}^{\mathrm{c}}\} \mid 1 \leq \mathrm{i} \leq \mathrm{V}^{\mathrm{s}}, 1 \leq \mathrm{j} \leq \mathrm{V}^{\mathrm{c}}\}, \tag{5}$$

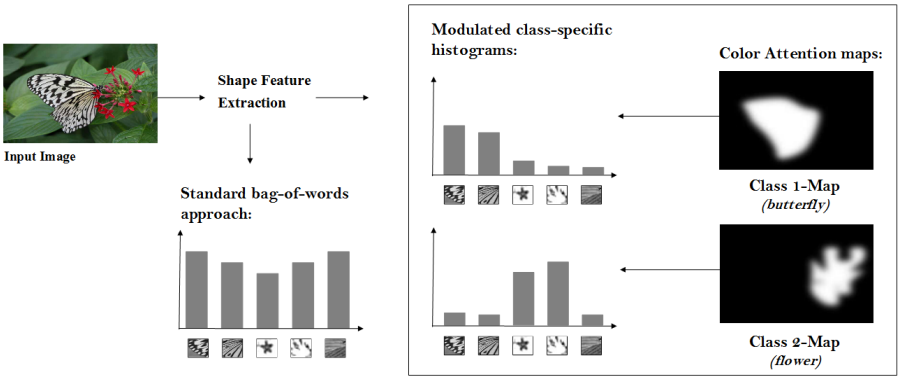

Fig. 3. Top-down control of visual attention based on color. In standard bag-of-words the image representation, here as distribution over visual shape words, is constructed in a bottom-up fashion. Here a top-down class-specific color attention is applied to modulate the impact of the shape-words in the image on the histogram construction. Consequently, a separate histogram is constructed for all categories, where the visual words relevant to each category (in this case flowers and butterflies) are accentuated. Figure taken from [14]
.

where $T = V^s \times V^c$. A disadvantage of product vocabularies is that they are very large. A typical SIFT vocabulary of 1000 combined with a color vocabulary of 100 would yield a product vocabulary of 100.000. Apart from being impractical from a memory point of view, there is also a danger of overfitting due to insufficient training data. A solution to these problems can be found by considering vocabulary compression techniques which have been presented in recent years (see e.g. [24]). Several methods, based on information theory, provide means to fuse the visual words of the product vocabulary into a compact image representation. For Portmanteau vocabularies the DITC algorithm [24] is applied. The method fuses words based on the $p\left(class|\mathrm{w}_{\mathrm{ij}}^{\mathrm{sc}}\right)$ which is obtained from the training data. Words are joined in such a way as to minimize the drop in discriminative power. In Fig. 4 examples of local regions attributed to the same Portmanteau word are shown.

As said, image representation constructed from Portmanteau vocabularies, possess cue compactness and cue binding. They also allow for cue weighting (see [13] for details). They scale relatively well with the number of categories: the number of words used in the final representation is a user input and compact image representations have been used for problems with up to two hundred classes [13]. However, extending them to multiple cues is currently infeasible. The product vocabulary explodes even further, making the statistics for $p\left(class|\mathrm{w}_{\mathrm{ij}}\right)$ which are at the base of the method unreliable. The properties are summarized in Table 1.

Fig. 4. Example of Portmanteau vocabulary. Each of the large boxes contains 100 image patches sampled from one Portmanteau word on the Oxford Flower-102 dataset.

6 Challenges and Future Research

In this section, we compare the fusion methods (summarized in Table 1), and discuss several possible directions for future research into cue fusing.

We focussed on the fusion of color and shape, but much of the discourse would be equally valid for the incorporation of other cues such as texture, optical flow, self-similarity, etc. Also for these cues cue-binding could be important, and a late fusion of the cues would lead to suboptimal results. However, late fusion remains the most common approach to join multiple feature representations, as for example in [10]. A notable exception is Li et al. [25], who have applied the color attention method to fuse motion cues with SIFT for event detection. Late fusion is ideal for cues which are not spatial such as text annotation or audio, in which case the cue binding property is irrelevant.

Most of the first approaches to fuse color and shape were based on early fusion [6][7][8], and suffered especially from the lack of cue compactness. For classes, which have color-shape independency (like cars, busses, etc) these methods often performed worse then bag-of-words based on only luminance SIFT. In principle, the finding of the best representation per class could be left to an MKL algorithm, which would automatically learn to lower the weight of early fusion representations for classes with shape-color independency, while balancing the weight of late fusion, Portmanteau and color attention-based representations.

The two methods color attention and Portmanteau vocabularies, combine the advantages of early and late fusion, namely feature binding and feature compactness. They were explicitly designed for this, and do so by using top-down information in the form of $p\,(class|\mathrm{w})$. Extending these representations with spatial information in the form of spatial pyramids [15] needs further investigation. One

Method	Compactness	Binding	Weighting	Cue Scal.	Category Scal.
Early Fusion	No	Yes	No	No	Yes
Late Fusion	Yes	No	Yes	Yes	Yes
Color Attention	Yes	Yes	Yes	Yes	No
Portmanteau	Yes	Yes	Yes	No	Yes

Table 1. Overview of properties for several methods to combine multiple cues into the bag-of-word framework. See text for discussion of table.

way would be to estimate $p\,(class|\mathrm{w}, cell)$ but because the statistics for pyramid-cells is less abundant than for images this could have negative influence on these methods. The main disadvantage for color attention, which is that its representation size scales linearly with the number of classes, makes the method unrealistic for large scale data sets. However, information theoretic methods [24] could be applied to reduce $p\,(class|\mathrm{w})$ matrix. Also, one against all representations could be considered for color attention.

Late fusion is a very simple method to implement and only suffers from the lack of cue-binding. However, Elfiky et al. [26] have shown that within spatial pyramids the lack of cue-binding becomes less important. In a spatial pyramid representation the image is represented by histograms over local cells in the image. In the extreme case where each cell would only have a single feature late fusion would possess the cue-binding property. This can also be seen for object detection based on bag-of-words. For these methods, a several level pyramid is used and accordingly late fusion was found to obtain excellent results [27]. As a consequence, it seems that after localization of the objects which requires some form of cue binding, the actual classification of the objects could be done in a late fusion fashion together with a spatial pyramid representation.

Acknowledgements

This work is funded by the Project MEC TIN2009-14173, and the Ramon y Cajal Program of Spanish Ministry of Science.

References

1. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE conference on Computer Vision and Patter Recognition. Volume 2. (June 2003) 264–271
2. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. on Pattern Analysis and Machine Intelligence **27**(10) (2005) 1615–1630
3. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 606–613

4. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2. CVPR '06, IEEE Computer Society (2006) 2161–2168
5. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60**(2) (2004) 91–110
6. van de Weijer, J., Schmid, C.: Coloring local feature extraction. In: Proc. of the European Conference on Computer Vision. Volume 2., Graz, Austria (2006) 334–348
7. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via plsa. Computer Vision–ECCV 2006 (2006) 517–530
8. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. PAMI **32**(9) (2010) 1582–1596
9. Bach, F.: Exploring large feature spaces with hierarchical multiple kernel learning. In: NIPS. (2008)
10. Gehler, P.V., Nowozin, S.: On feature combination for multiclass object classification. In: In Proc. International Conference on Computer Vision. (2009)
11. Fernando, B., Fromont, E., Muselet, D., Sebban, M.: Discriminative feature fusion for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3434–3441
12. Burghouts, G., Geusebroek, J.: Performance evaluation of local colour invariants. Computer Vision and Image Understanding **113**(1) (2009) 48–62
13. Khan, F., Van de Weijer, J., Bagdanov, A., Vanrell, M.: Portmanteau vocabularies for multi-cue image representation. In: Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS 2011). (2011)
14. Khan, F.S., van de Weijer, J., Vanrell, M.: Modulating shape features by color attention for object recognition. International Journal of Computer Vision (IJCV) **98**(1) (2012) 49–64
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE conference on Computer Vision and Patter Recognition. (2006) 2169–2178
16. van de Weijer, J., Schmid, C.: Applying color names to image description. In: IEEE International Conference on Image Processing (ICIP), San Antonio, USA (2007)
17. van de Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing **18**(7) (july 2009) 1512–1524
18. Zhang, J., Barhomi, Y., Serre, T.: A new biologically inspired color image descriptor. European Conference on Computer Vision (2012)
19. Rojas-Vigo, D., Khan, F.S., van de Weijer, J., Gevers, T.: The impact of color on bag-of-words based object recognition. In: Int. Conference on Pattern Recognition (ICPR). (2010)
20. Treisman, A.: The binding problem. Current Opinion in Neurobiology **6** (1996) 171–178
21. Treisman, A., Gelade, G.: A feature integration theory of attention. Cogn. Psych **12** (1980) 97–136
22. Wolfe, J.M.: Visual Search. (1998) in Attention, edited by H. Pashler, Psychology Press Ltd.
23. Wolfe, J.M., Horowitz, T.: What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience **5** (2004) 1–7

24. Dhillon, I., Mallela, S., Kumar, R.: A divisive information-theoretic feature clustering algorithm for text classification. Journal of Machine Learning Research (JMLR) **3** (2003) 1265–1287
25. Li, L., Yuan, C., Hu, W., Li, B.: Top-down cues for event recognition. Computer Vision–ACCV 2010 (2011) 691–702
26. Elfiky, N., Khan, F.S., van de Weijer, J., Gonzalez, J.: Discriminative compact pyramids for object and scene recognition. Pattern Recognition (PR) **45**(4) (April 2012) 1627–1636
27. Khan, F., Anwer, R., van de Weijer, J., Bagdanov, A., Vanrell, M., Lopez, A.: Color attributes for object detection. In: IEEE conference on Computer Vision and Patter Recognition. (2012)