Coloring Action Recognition in Still Images

Fahad Shahbaz Khan, Muhammad Anwer Rao, Joost van de Weijer, Andrew Bagdanov,
Antonio Lopez and Michael Felsberg

Linköping University Post Print

N.B.: When citing this work, cite the original article.

The original publication is available at www.springerlink.com:

Fahad Shahbaz Khan, Muhammad Anwer Rao, Joost van de Weijer, Andrew Bagdanov, Antonio Lopez and Michael Felsberg, Coloring Action Recognition in Still Images, 2013, International Journal of Computer Vision, (105), 3, 205-221.
http://dx.doi.org/10.1007/s11263-013-0633-0

Copyright: Springer Verlag (Germany)
http://www.springerlink.com/?MUD=MP

Postprint available at: Linköping University Electronic Press
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-97463

Noname manuscript No.
(will be inserted by the editor)

Coloring Action Recognition in Still Images

Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Andrew D. Bagdanov, Antonio M. Lopez, Michael Felsberg

Received:

Abstract In this article 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.

Rao Muhammad Anwer, Joost van de Weijer, Antonio M. Lopez:
[1]Computer Vision Centre Barcelona, Universitat Autonoma de Barcelona, Spain
Fahad Shahbaz Khan, Michael Felsberg:
[2]Computer Vision Laboratory, Linköping University, Sweden
Andrew D. Bagdanov:
[3]Media Integration and Communication Center, University of Florence, Italy

Keywords Color features, image representation, action recognition.

1 Introduction

Action category recognition in still images is a major emerging problem in computer vision.[1] 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 (Desai and Ramanan, 2012). Recognizing human actions in 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 (Sharma et al, 2012; Delaitre et al, 2010; Shapovalova et al, 2011). State-of-the-art approaches to action classification typically make use of intensity-based features to represent

[1] 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`

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 1. Color has, however, led to significantly improved results on other recognition tasks, such as image classification and object detection (van de Weijer and Schmid, 2006; van de Sande et al, 2010; Bosch et al, 2008; Gehler and Nowozin, 2009; Khan et al, 2012b, 2011). 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. Gaidon et al (2011) proposed an approach based on a sequence of atomic action units to detect actions in videos. Tran and Yuan (2012) introduced a structural learning approach to action detection in unconstrained videos. A multiple-instance learning framework was proposed by Hu et al (2009) for learning action detectors based on imprecise action locations. Yuan et al (2011) propose a naive Bayes mutual information maximization framework for matching patterns in videos. Recently, Desai and Ramanan (2012) investigated the problem of action detection in still images. In this paper, we also investigate the problem of action detection in still images.

Deformable part-based models (Felzenszwalb et al, 2010) have demonstrated excellent results for object detection. The conventional part-based framework uses HOG features (Dalal and Triggs, 2005) for image representation. Several works recently have aimed at combining multiple features for object detection (Khan et al, 2012a; Zhang et al, 2010; Vedaldi et al, 2009). Zhang et al (2010) proposed a combination of HOG and LBP features for object detection, and Khan et al (2012a) 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 paper 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 (Khan et al, 2012b) and portmanteau vocabularies (Khan et al, 2011) 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.

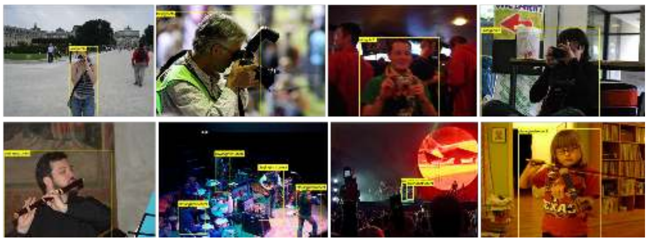

Fig. 1 Example images for different action categories from the PASCAL VOC 2010 dataset. These images illustrate the complications related to color description due to the large variation in illumination, shadows and specularities.

Additionally, in this paper 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 paper is organized as follows. In the next section we discuss work related to the problem of action recognition. In Section 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 4, and in Section 5 we show how to incorporate color into a part-based detection framework for action detection. In Section 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 7 with concluding remarks and general recommendations for selecting color descriptors and fusion approaches for human action recognition problems.

2 Related Work

Action recognition in static images has gained a lot of attention recently (Sharma et al, 2012; Prest et al, 2012; Delaitre et al, 2010; Yao and Li, 2012; Yao et al, 2011; Maji et al, 2011). 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 (Sharma et al, 2012; Delaitre et al, 2010). 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. Prest et al (2012) 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 Maji et al (2011) 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 Delaitre et al (2011). 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. Yao et al (2011) 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 paper 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 (van de Weijer and Schmid, 2006; van de Sande et al, 2010; Bosch et al, 2008; Everingham et al, 2009; Khan et al, 2012b, 2011). A variety of color descriptors and approaches to combining color and shape cues for object recognition have been proposed in the literature (van de Weijer and Schmid, 2007, 2006; van de Sande et al, 2010; Bosch et al, 2006; Vigo et al, 2010). Bosch et al (2006) propose to compute SIFT descriptors directly on HSV channels of color images. A set of robust and photometrically invariant color descriptors was proposed by van de Weijer and Schmid (2006). Pagani et al (2009) 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 Felsberg and Hedborg (2007). 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 van de Weijer et al (2009). See (van de Sande et al, 2010) for a comprehensive study and evaluation of a large number of color descriptors.

The discriminative, deformable part-based framework (Zhang et al, 2010; Felzenszwalb et al, 2010) yields excellent state-of-the-art results for object detection. A star-structured deformable part method was proposed by Felzenszwalb et al (2010) 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 (Everingham et al, 2010), especially on the person category. Recently (Khan et al, 2012a) proposed augmenting the standard part-based approach with color information, which results in significant improvement in performance. In this paper 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 paper extends our earlier work Khan et al (2012a) on action detection. In this paper we focus on action recognition and we investigate the potential of combining color and shape for both action classification and action detection. Beyond the work in Khan et al (2012a) 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.

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 $C = \left(\frac{O1}{O3}\ \frac{O2}{O3}\ O3\right)^{\mathrm{T}}$, where $O1$, $O2$ and $O3$ are derived from the opponent color space as (Lenz et al, 2005):

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

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 Geusebroek et al (2001).

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. 1) to obtain the hue and saturation channels:

$$H = \arctan\left(\frac{O1}{O2}\right), \tag{2}$$

$$S = \sqrt{O1^2 + O2^2}, \tag{3}$$

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 van de Weijer and Schmid (2006). 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:

$$ang_{\mathbf{x}}^{O} = \arctan\left(\frac{O1_{\mathbf{x}}}{O2_{\mathbf{x}}}\right), \tag{4}$$

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 (van de Weijer and Schmid, 2006). 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 (van de Weijer et al, 2009). 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 (Berlin and Kay, 1969).

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 (van de Weijer et al, 2009). 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 (Benavente et al, 2008). 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 (Khan et al, 2011) and object detection (Khan et al, 2012a). Finally, they have the additional advantage of being a very compact representation at only 11 dimensions.

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 (Sharma et al, 2012; Delaitre et al, 2010; Shapovalova et al, 2011). Here we discuss, within the context of action classification, different fu-

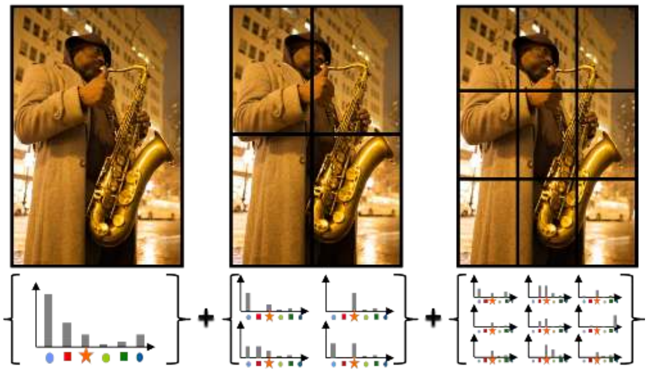

Fig. 2 We apply a three level pyramid on the bounding boxes of the action recognition datasets. Separate BOW histograms are constructed for each cell and are concatenated to form the final action descriptor. In this paper we use a pyramid representation with three levels, yielding a total of 14 cells.

sion approaches proposed in literature for combining color and shape cues within the bag-of-words framework. Throughout this paper we use the SIFT descriptor for describing the shape of local image patches (Lowe, 2004).

Figure 2 shows the bag-of-words action representation which is considered in this paper. 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 paper we will ignore background information and only describe the information within the bounding box of the person in action.[2] For all image representations, we incorporate spatial information via a spatial pyramid (Lazebnik et al, 2006). 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 (Delaitre et al, 2010).

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.[3] Pipelines for several fusion methods are illustrated in Figure 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 $H = [h_1, ..., h_C]$, where h_i corresponds to the histogram of visual words in spatial pyramid cell i. Visual vocabularies are denoted by $W^k = \{w_1^k, ..., w_{V^k}^k\}$, 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.[4] 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

$$h_i^k\left(w_n^k\right) \propto \sum_{j \in c_i} \delta\left(x_j^k, w_n^k\right), \tag{5}$$

where δ is the Dirac delta function:

$$\delta\left(x, y\right) = \begin{cases} 0 & \text{for} \quad x \neq y \\ 1 & \text{for} \quad x = y \end{cases} \tag{6}$$

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 (Khan et al, 2012b).

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 (Khan et al, 2012b).

4.3 Channel-based Early Fusion

Channel-based fusion for color and shape was first proposed by Bosch et al (2008) and later extensively investigated and tested by van de Sande et al (2010). First,

[2] Only in experiment 6.2.3 do we add additional information from the background.

[3] 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 (Lan et al, 2012).

[4] The combined vocabulary sc is constructed by concatenating the shape and color features before constructing the vocabulary in the combined feature-space.

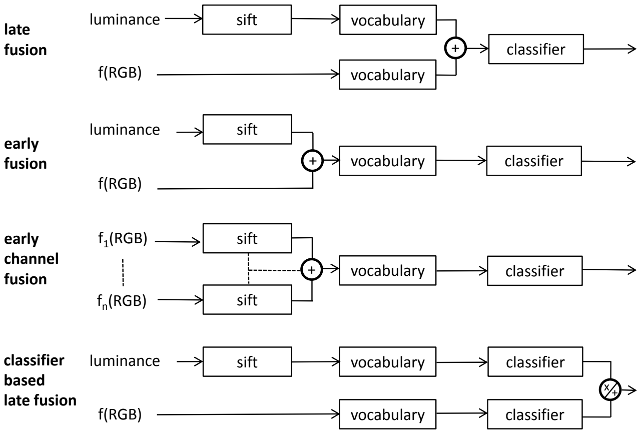

Fig. 3 Pipelines for four different fusion methods. The fusion between color and shape is indicated by a 'plus' in case of concatenation of vectors or vocabulary histograms. In the case of classifier based fusion, the encircled multiplication and sum symbols refer to the two methods of classifier fusion 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.

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 (van de Sande et al, 2010) and evaluate five different channel-based descriptors: RGB-SIFT, RG-SIFT, OPP-SIFT, C-SIFT and HSV-SIFT.

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 $H^s = [h_1^s, \ldots, h_C^s]$ and $H^c = [h_1^c, ..., h_C^c]$. In the work of Gehler and Nowozin (2009) 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.

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 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 (Khan et al, 2012b) method follows the same pipeline as late fusion (see the first pipeline in Figure 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:

$$h_i = \left[h_i^{cl_1}, ..., h_i^{cl_m}\right], \tag{7}$$

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

$$h_i^{cl_t}\left(w_n^s\right) \propto \sum_{j \in c_i} p\left(cl_t | w_j^c\right) \delta\left(x_j^k, w_n^s\right), \tag{8}$$

where the only difference with respect to computing a shape histogram according to Eq. 5 is the modulation with $p\left(cl_t | w_j^c\right)$. 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 $p\left(cl_t | w_j^c\right)$. 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

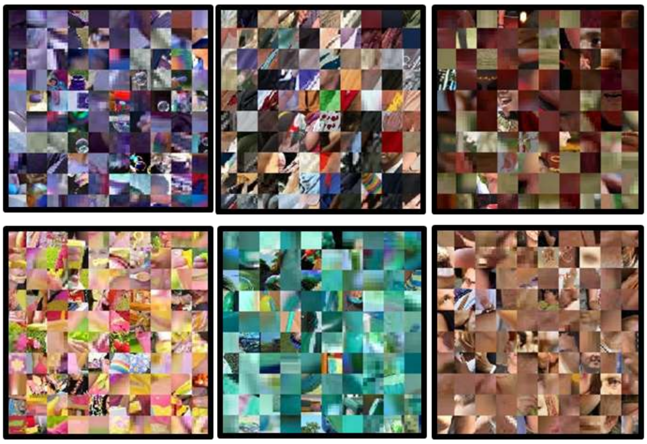

Fig. 4 Example portmanteau clusters from the Willow and Stanford-40 datasets. Note that each portmanteau cluster constitutes a distinct pattern of shape and color. Moreover, several clusters are representative of humans and specific actions such as gardening.

classes. For more details on color attention-based representations, see Khan et al (2012b).

The probability $p\left(cl_t|w_j^c\right)$ 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:

$$p_i\left(cl_t|w_n^c\right) = \frac{\sum\limits_{s\in cl_t}\sum\limits_{j\in c_i^s}\delta\left(x_j^c, w_n^c\right)}{\sum\limits_{s}\sum\limits_{j\in c_i^s}\delta\left(x_j^c, w_n^c\right)}, \tag{9}$$

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 $p\left(cl_t|w_j^c\right)$ for all cells of the object:

$$p\left(cl_t|w_n^c\right) = \frac{\sum\limits_{s\in cl_t}\sum\limits_{i}\sum\limits_{j\in c_i^s}\delta\left(x_j^c, w_n^c\right)}{\sum\limits_{s}\sum\limits_{i}\sum\limits_{j\in c_i^s}\delta\left(x_j^c, w_n^c\right)}, \tag{10}$$

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.

4.6 Portmanteau Vocabulary-based Fusion

A second approach to introducing feature binding into the late fusion representation is through portmanteau vocabularies (Khan et al, 2011). 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:

$$\begin{aligned} W &= \left\{w_1, w_2, ..., w_T\right\} \\ &= \left\{\left\{w_q^s, w_r^c\right\}|1 \leq q \leq V^s, 1 \leq r \leq V^c\right\}. \end{aligned} \tag{11}$$

The main drawback of this is that this leads to very large vocabularies of size $T = V^s \times V^c$. In Khan et al (2011) 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 $p\left(cl_t|w_n\right)$. 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 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 4, containing many patches with hands and plants, clearly encodes information about the gardening class.

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 Felzenszwalb et al (2010). 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 Felzenszwalb et al (2010), 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 Dalal and Triggs (2005). 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[5]. The first method, which we call channel fusion, computes HOGs separately on the three channels and concatenates the result:

$$D_i = \left[HOG_i^R, HOG_i^G, HOG_i^B\right], \tag{12}$$

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:

$$D_i = \left[HOG_i, C_i\right], \tag{13}$$

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 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 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. 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.

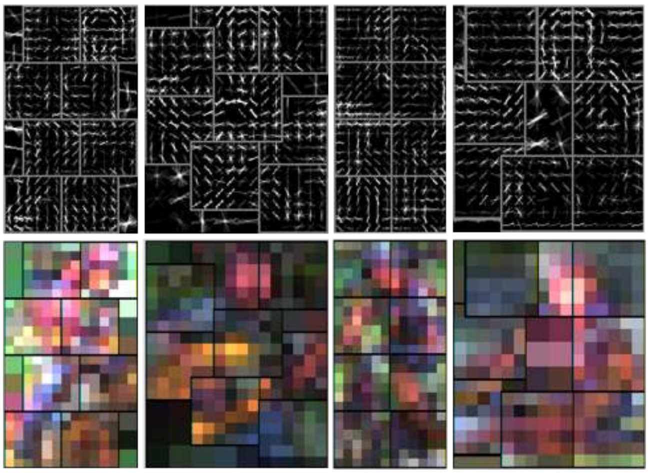

Fig. 5 Visualization of learned part-based models using CN-HOG on the Stanford-40 action dataset. Both the HOG and color names components of our trained models combined in a late fusion are shown. Each color cell is represented using the color obtained by multiplying the SVM weights for the 11 CN bins 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.

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.

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.[6]

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.[7]

Finally, we also present results on Stanford-40, which is one of the most challenging action recognition datasets currently available.[8] Stanford-40 consists of 9532 im-

[5] Due to the absence of a vocabulary stage, several of the fusion methods explained in Section 4 cannot be applied to part-based object detection.

[6] The Willow dataset is available at: `http://www.di.ens.fr/willow/research/stillactions/`

[7] PASCAL 2010 is available at: `http://www.pascal-network.org/challenges/VOC/voc2010/`

[8] The Stanford-40 dataset is available at `http://vision.stanford.edu/Datasets/40actions.html`

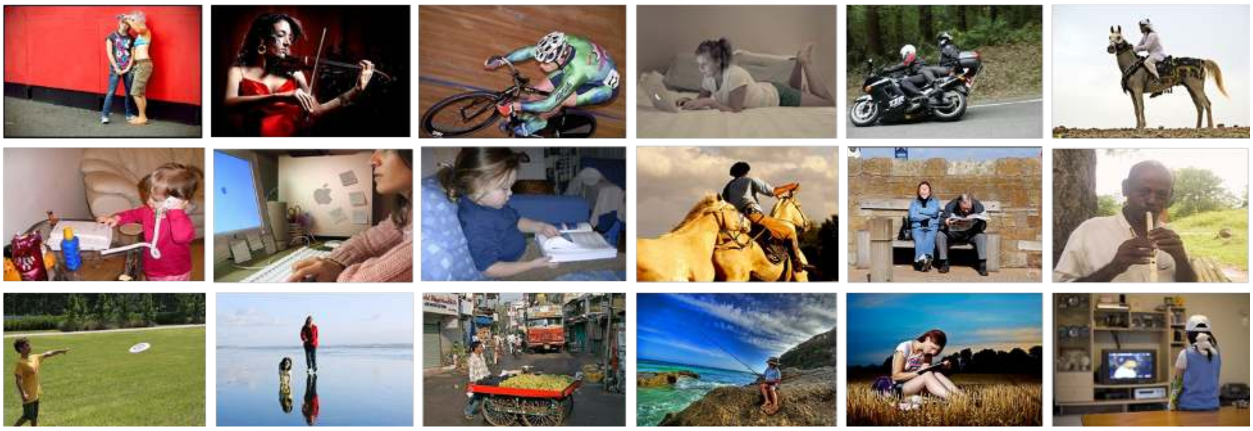

Fig. 6 Example images from the three datasets used to evaluate color descriptors and fusion techniques. Top row: images from the Willow dataset. Middle row: images from PASCAL VOC 2010 action recognition dataset. Bottom row: example images from the Stanford-40 dataset.

ages 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 6 shows some of example images from the three datasets.

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 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 (1×1, 2×2, and 3×3), yielding a total of 14 cells (Lazebnik et al, 2006). For classification, we use a nonlinear SVM with a χ^2 kernel (Zhang et al, 2007). 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.

6.2.1 Pure Color Descriptors for Action Classification

We compare six different color descriptors using the same experimental settings. Table 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

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	**44.7**	**34.4**	**17.6**

Table 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	**65.6**	53.7	39.4
RG-SIFT	256	1500	65.0	**54.6**	**39.6**
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 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.

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.

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 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. 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 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 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 ap-

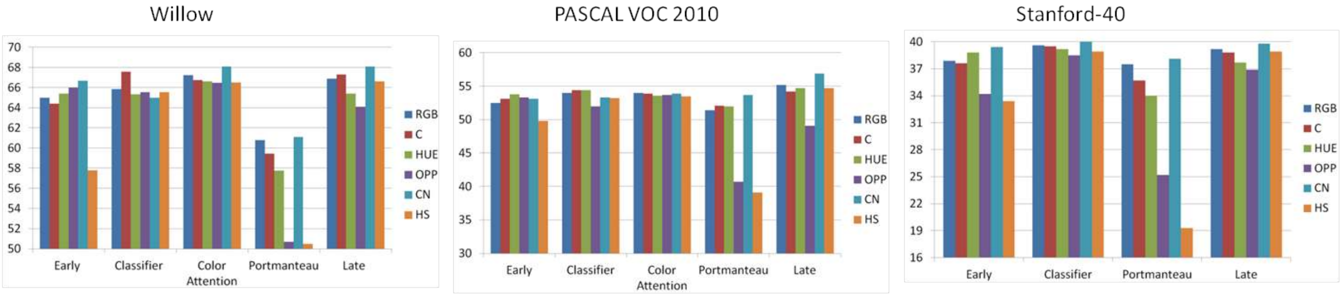

Fig. 7 Performance comparison of different approaches to fusing color and shape. The choice of color descriptor is crucial for portmanteau-based image presentations. On all three datasets, late fusion performs better than early fusion, and the best results are obtained using late fusion with color names.

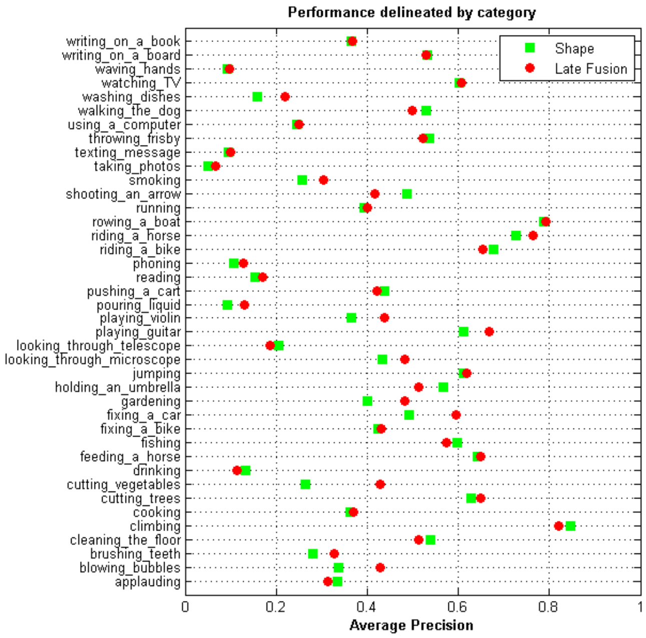

Fig. 8 Per-category comparison between late fusion and shape alone. Here late fusion refers to fusion of color names and shape. On many action categories combining color and shape improves performance over shape alone.

proaches, both late fusion and classifier-based methods yield the best performance. Figure 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. 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 Khan et al (2012b) 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 Elfiky et al (2012). 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.

6.2.3 Combining Fusion Techniques for Action Classification

In this section we analyze the potential of combining fusion approaches and determine if these strategies are

	int. computer	photographing	playingmusic	ridingbike	ridinghorse	running	walking	mean AP
Delaitre et al (2010)	58.2	35.4	73.2	82.4	69.6	44.5	54.2	59.6
Delaitre et al (2011)	56.6	37.5	72.0	90.4	75.0	59.7	57.6	64.1
Sharma et al (2012)	59.7	42.6	74.6	87.8	84.2	56.1	56.5	65.9
Sharma et al (2013)	**64.5**	40.9	75.0	**91.0**	**87.6**	55.0	59.2	67.6
Our approach	61.9	**48.2**	**76.5**	90.3	84.3	**64.7**	**64.6**	**70.1**

Table 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
Maji et al (2011)	49.6	43.2	27.7	**83.7**	89.4	85.6	31.0	59.1	67.9	59.7
Shapovalova et al (2011)	45.5	54.5	31.7	75.2	88.1	76.9	32.9	64.1	62.0	59.0
Delaitre et al (2011)	48.6	53.1	28.6	80.1	**90.7**	85.8	33.5	56.1	69.6	60.7
Yao et al (2011)	42.8	60.8	41.5	80.2	90.6	87.8	**41.4**	**66.1**	**74.4**	**65.1**
Prest et al (2012)	**55.0**	**81.0**	**69.0**	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	**88.1**	37.3	59.9	66.5	62.4

Table 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.

Method	Object Bank	LLC	Sparse Bases	EPM	Ours
mAP	32.5	35.2	45.7	42.2	**51.9**

Table 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.

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 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 (Delaitre et al, 2011; Prest et al, 2012; Delaitre et al, 2010; Sharma et al, 2013). A mean AP of 64.1 is reported by Delaitre et al (2011) with an approach that models complex interactions between persons and objects. The interactions are modeled using external data to train body part detectors. Sharma et al (2012) 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 4. Most state-of-the-art approaches rely on detection techniques to find human-object relationships. Maji et al (2011) report a mean AP of 59.7 using a poselet detector that captures the pose in multiscale manner. A mean AP of 62.0 is reported by Prest et al (2012) using a human-centric approach to localize humans and find object-human relationships. The best result of 65.1 is reported by Yao et al (2011) 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 shows a comparison with state-of-the-art performance reported on the Stanford-40 dataset (Yao et al, 2011; Li et al, 2010; Wang et al, 2010). In order to improve the overall performance on this large dataset we increase the vocabulary size for shape to 4000. Recently Sharma et al (2013) 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 (Yao et al, 2011).

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 (Everingham et al, 2010). 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[9]. 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 descriptors described in Section 3.

In Table 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) which gives it an advantage over luminance-based detection.

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	**27.5**

Table 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.

Figure 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.

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 10 shows the confusion matrix obtained on the Willow dataset using late fusion and the color name descriptor[10]. 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.

To illustrate further the contribution of color information for action classification we generated heat maps

[9] We also performed experiments replacing HOG with pure color descriptors but significantly inferior results were obtained.

[10] The confusion matrix is constructed by assigning each image to the class for which it gets the highest classification score.

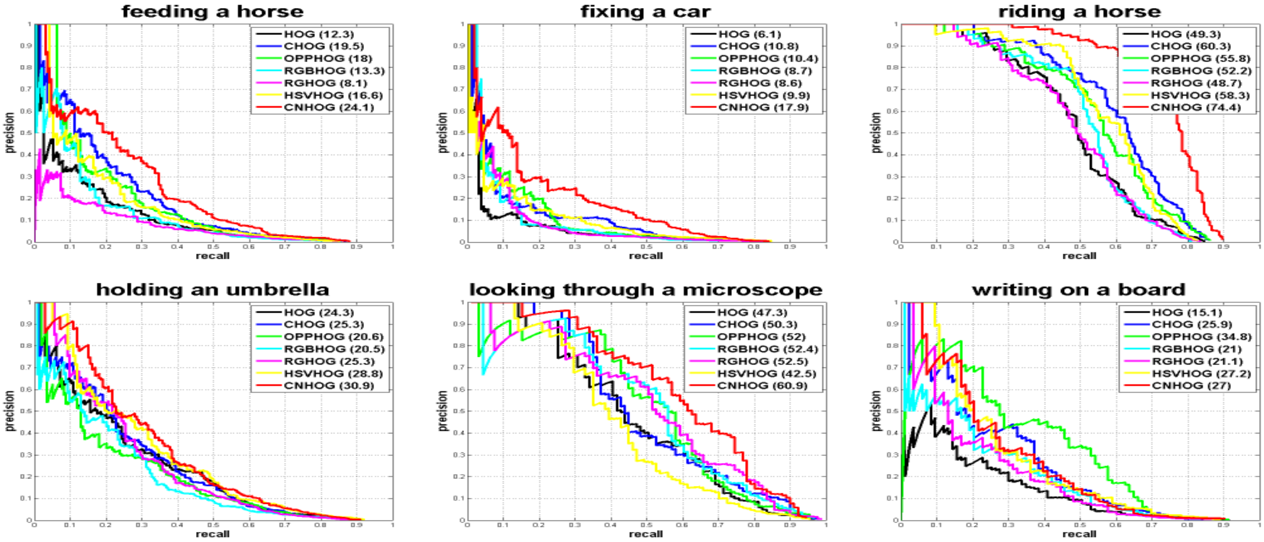

Fig. 9 Precision/recall curves of the various channel based approaches, HOG and CN-HOG on six different action categories from the Stanford-40 dataset. Other than the writing on a board action category, CN-HOG provides significantly improved performance over channel based methods.

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 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 Hoiem et al (2012) 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[11] 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.

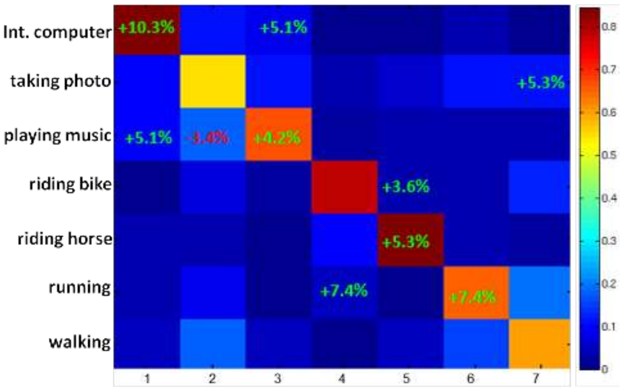

Fig. 10 Confusion matrix for late fusion of shape and color names on the Willow dataset. Superimposed are differences with the confusion matrix based on luminance alone for confusions where the absolute change is at least 3%. Late fusion reduces the confusion among different categories in general, but particularly so in the interacting with computer and playing music categories.

Figure 12 shows the results of this analysis on the Standford-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

[11] 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.

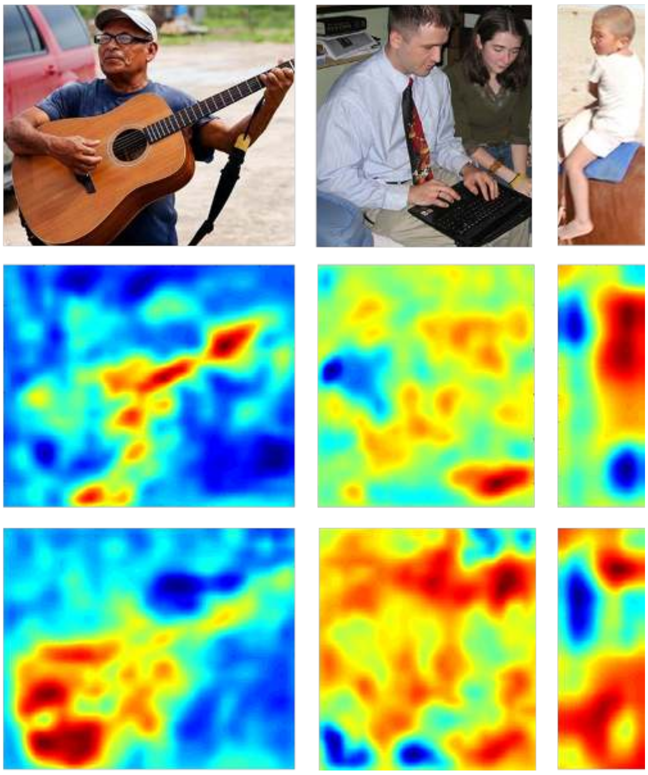

Fig. 11 Heat maps of classifier responses for playing guitar, using computer and riding horse categories. The top row contains the original images. The second row shows heat maps using shape alone. The third row contains heat maps using fused color-shape features. Despite their being no location information coded into the classifiers, adding color information helps to localize the horse and guitar in the images.

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 then 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 (Mullen, 1985), implying that for precise localization the human vision system relies on luminance.

7 Discussion and Conclusion

In this article 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 paper, 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.

Willow	PASCAL 2010	Stanford-40
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 7 The best performing color descriptors on the three datasets used for action classification in this paper. Note that for all three datasets the color name descriptor is the best choice.

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

Willow	PASCAL 2010	Stanford-40
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 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 8 we order the different fusion approaches evaluated for action classification in this paper. 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

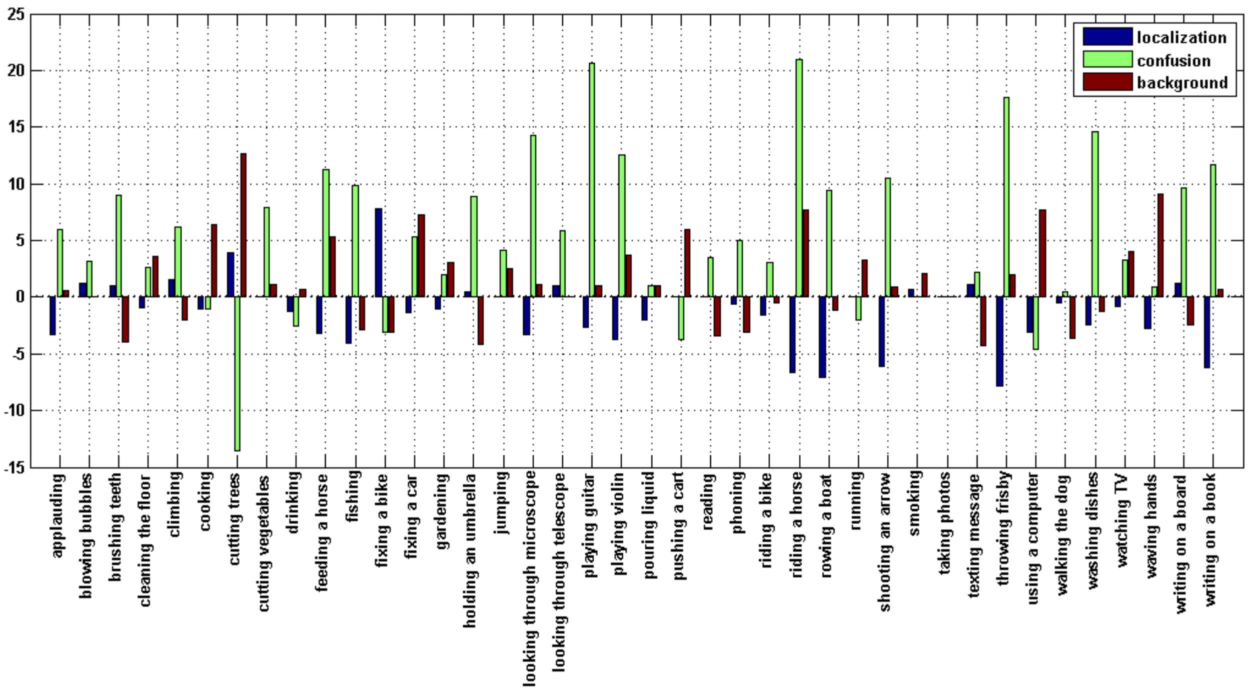

Fig. 12 Analysis of detection errors for the Standford-40 dataset. The graph shows the decrease in errors which occur when going from standard HOG to CN-HOG (negative changes in the graph signify an increase in errors). Errors are split into errors caused by misaligned localization, confusion with other classes, and detections on the background.

shape further improve performance for action classification. Note that nowhere in this paper 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.

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.

Acknowledgements

We gratefully acknowledge the support of MNEMOSYNE (POR-FSE 2007-2013, A.IV-OB.2), Collaborative Unmanned Aerial Systems (within the Linnaeus environment CADICS), ELLIIT, the Strategic Area for ICT research, funded by the Swedish Government, and Spanish projects TRA2011-29454-C03-01 and TIN2009-14173..

References

Benavente R, Vanrell M, Baldrich R (2008) Parametric fuzzy sets for automatic color naming. JOSA 25(10):2582–2593

Berlin B, Kay P (1969) Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley, CA

Bosch A, Zisserman A, Munoz X (2006) Scene classification via plsa. In: ECCV

Bosch A, Zisserman A, Munoz X (2008) Scene classification using a hybrid generative/discriminative approach. PAMI 30(4):712–727

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: CVPR

Delaitre V, Laptev I, Sivic J (2010) Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: BMVC

Delaitre V, Sivic J, Laptev I (2011) Learning person-object interactions for action recognition in still images. In: NIPS

Desai C, Ramanan D (2012) Detecting actions, poses, and objects with relational phraselets. In Proc. ECCV

Elfiky N, Khan FS, van de Weijer J, Gonzalez J (2012) Discriminative compact pyramids for object and scene recognition. PR 45(4):1627–1636

Everingham M, Gool LV, Williams CKI, JWinn, Zisserman A (2009) The pascal visual object classes challenge 2009 results.

Everingham M, Gool LJV, Williams CKI, Winn JM, Zisserman A (2010) The pascal visual object classes (voc) challenge. IJCV 88(2):303–338

Felsberg M, Hedborg J (2007) Real-time view-based pose recognition and interpolation for tracking initialization. J Real-Time Image Processing 2(3):103–115

Felzenszwalb PF, Girshick RB, McAllester DA, Ramanan D (2010) Object detection with discriminatively trained part-based models. PAMI 32(9):1627–1645

Gaidon A, Harchaoui Z, Schmid C (2011) Actom sequence models for efficient action detection. In Proc. CVPR

Gehler PV, Nowozin S (2009) On feature combination for multiclass object classification. In Proc. ICCV

Geusebroek JM, van den Boomgaard R, Smeulders AWM, Geerts H (2001) Color invariance. PAMI 23(12):1338–1350

Hoiem D, Chodpathumwan Y, Dai Q (2012) Diagnosing error in object detectors. In: ECCV

Hu Y, Cao L, Lv F, Yan S, Gong Y, Huang TS (2009) Action detection in complex scenes with spatial and temporal ambiguities. In Proc. ICCV

Khan FS, van de Weijer J, Bagdanov AD, Vanrell M (2011) Portmanteau vocabularies for multi-cue image representations. In: NIPS

Khan FS, Anwer RM, van de Weijer J, Bagdanov AD, Vanrell M, Lopez AM (2012a) Color attributes for object detection. In: CVPR

Khan FS, van de Weijer J, Vanrell M (2012b) Modulating shape features by color attention for object recognition. IJCV 98(1):49–64

Lan ZZ, Bao L, Yu SI, Liu W, Hauptmann AG (2012) Double fusion for multimedia event detection. In: MMM

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proc. CVPR

Lenz R, Bui TH, Hernandez-Andres J (2005) Group theoretical structure of spectral spaces. Journal of Mathematical Imaging and Vision 23(3):297–313

Li LJ, Su H, Xing EP, Li FF (2010) Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: NIPS

Lowe DG (2004) Distinctive image features from scale-invariant points. IJCV 60(2):91–110

Maji S, Bourdev LD, Malik J (2011) Action recognition from a distributed representation of pose and appearance. In: CVPR

Mullen KT (1985) The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. The Journal of Physiology (359):381–400

Pagani A, Stricker D, Felsberg M (2009) Integral p-channels for fast and robust region matching. In: ICIP

Prest A, Schmid C, Ferrari V (2012) Weakly supervised learning of interactions between humans and objects. PAMI 34(3):601–614

van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. PAMI 32(9):1582–1596

Shapovalova N, Gong W, Pedersoli M, Roca FX, Gonzalez J (2011) On importance of interactions and context in human action recognition. In: IbPRIA

Sharma G, Jurie F, Schmid C (2012) Discriminative spatial saliency for image classification. In: CVPR

Sharma G, Jurie F, Schmid C (2013) Expanded parts model for human attribute and action recognition in still images. In: CVPR

Tran D, Yuan J (2012) Max-margin structured output regression for spatio-temporal action localization. In Proc. NIPS

Vedaldi A, Gulshan V, Varma M, Zisserman A (2009) Multiple kernels for object detection. In: ICCV

Vigo DAR, Khan FS, van de Weijer andTheo Gevers J (2010) The impact of color on bag-of-words based object recognition. In: ICPR

Wang J, Yang J, Yu K, Lv F, Huang TS, Gong Y (2010) Locality-constrained linear coding for image classification. In: CVPR

van de Weijer J, Schmid C (2006) Coloring local feature extraction. In: ECCV

van de Weijer J, Schmid C (2007) Applying color names to image description. In: ICIP

van de Weijer J, Schmid C, Verbeek JJ, Larlus D (2009) Learning color names for real-world applications. IEEE Transaction in Image Processing (TIP) 18(7):1512–1524

Yao B, Li FF (2012) Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. PAMI 34(9):1691–1703

Yao B, Jiang X, Khosla A, Lin AL, Guibas LJ, Li FF (2011) Human action recognition by learning bases of action attributes and parts. In: ICCV

Yuan J, Liu Z, Wu Y (2011) Discriminative video pattern search for efficient action detection. PAMI 33(9):1728–1743

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

Zhang J, Huang K, Yu Y, Tan T (2010) Boosted local structured hog-lbp for object localization. In: CVPR