Scale Coding Bag-of-Words for Action Recognition

Fahad Shahbaz Khan[1], Joost van de Weijer[2], Andrew D. Bagdanov[2], Michael Felsberg[1]
[1]Computer Vision Laboratory, Linköping University, Sweden
[2]Computer Vision Center, CS Dept. Universitat Autonoma de Barcelona, Spain

***Abstract*—Recognizing human actions in still images is a challenging problem in computer vision due to significant amount of scale, illumination and pose variation. Given the bounding box of a person both at training and test time, the task is to classify the action associated with each bounding box in an image. Most state-of-the-art methods use the bag-of-words paradigm for action recognition. The bag-of-words framework employing a dense multi-scale grid sampling strategy is the *de facto* standard for feature detection. This results in a scale invariant image representation where all the features at multiple-scales are binned in a single histogram. We argue that such a scale invariant strategy is sub-optimal since it ignores the multi-scale information available with each bounding box of a person.**

This paper investigates alternative approaches to scale coding for action recognition in still images. We encode multi-scale information explicitly in three different histograms for small, medium and large scale visual-words. Our first approach exploits multi-scale information with respect to the image size. In our second approach, we encode multi-scale information relative to the size of the bounding box of a person instance. In each approach, the multi-scale histograms are then concatenated into a single representation for action classification. We validate our approaches on the Willow dataset which contains seven action categories: interacting with computer, photography, playing music, riding bike, riding horse, running and walking. Our results clearly suggest that the proposed scale coding approaches outperform the conventional scale invariant technique. Moreover, we show that our approach obtains promising results compared to more complex state-of-the-art methods.

I. Introduction

In recent years, recognizing human actions in still images has gained much attention [1], [2], [3], [4], [5], [6], [7]. In action recognition problems, bounding boxes of humans performing actions are provided both at training and test time. The task is then to assign an action category label to each bounding box at test time. The problem is challenging due to the significant amount of pose, viewpoint and illumination variation. Large scale changes, both within categories and across different action classes, complicate the problem further. Figure 1 shows example images from the *interacting with computer* and *running* action categories. These examples illustrate the changes in scale that can occur within an action class. In this work, we investigate the potential of exploiting multi-scale information for bag-of-words based action recognition.

Most state-of-the-art approaches employ the bag-of-words (BOW) model for object and action recognition [2], [3], [7], [8], [9], [10]. The bag-of-words approach begins with feature detection which involves detecting keypoint locations in an image. These keypoint locations are then used to extract visual features such as color, shape or texture. Generally, SIFT descriptors [11] are used to describe local appearance in intensity images. The feature extraction stage is followed by a vocabulary construction step after which the local features are vector quantized against a fixed-size visual vocabulary. Finally, a histogram-based image representation is constructed by counting the frequency of occurrence of each visual word in an image. Following this trend, we also use the bag-of-words approach for action recognition in still images.

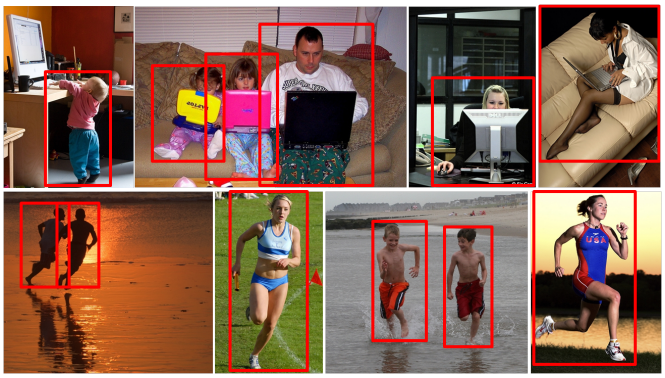

Fig. 1. Example images from the *interacting with computer* (top row) and *running* (bottom row) action categories from the Willow dataset. These examples demonstrate the variation in scale, especially with respect to the size of bounding boxes within each category. This suggests that alternative image representations may be desirable to incorporate multi-scale information.

State-of-the-art recognition pipelines using the bag-of-words model generally use dense multi-scale feature sampling in lieu of feature detection. This works by scanning the image at multiple scales at fixed locations on a grid of rectangular patches. Such a multi-scale dense sampling strategy is an integral component in virtually every bag-of-words based object recognition method [12], [13], [14], [15]. In object classification, invariance with respect to scale is crucial since a category instance can appear at different sizes. To achieve scale invariance, the feature descriptors are transformed into a common size which is then use to construct the visual vocabulary. As a result of this, feature points from all scales are encoded into a single scale-invariant histogram representation.

Conventional bag-of-words based action recognition approaches employ the same object recognition pipeline by constructing a single histogram representing features extracted at all scales. Such a representation aims to achieve the same scale invariance so crucial for object recognition. However, in the case of action recognition, the bounding box information for each instance is available at both training and test time. Our hypothesis is that this bounding box information can be exploited in order to obtain multi- and relative-scale image representations for action recognition that relax the scale invariance normally used. Since such representations encode some scale information in the final histogram, we refer to them

as *scale coding image representations*.

This paper investigates alternative scale coding strategies that incorporate multi-scale information in the image representation. We propose two approaches to encoding multi-scale information using three different histograms: one for small-, one for medium-, and one for large-scale visual words. In the first approach, the multi-scale image representation is constructed with respect to the image size. The second approach takes into account the relative scale with respect to the size of the person bounding box. Here, the definition of a scale considered to be small, medium or large is dependent on the size of the bounding box. Instead of the conventional scale invariant approach, which puts all the scales in a single histogram, our representation preserves some multi-scale information of each feature relative to either the size of image or the bounding box. Our final image representation is obtained by concatenating the small-, medium- and large-scale histograms. To validate our proposed representations, we perform experiments on the Willow dataset which consists of seven action categories. The results of our experiments clearly suggest that the multi-scale methods outperform the conventional bag-of-words image representation.

The paper is organized as follows. In the next section we discuss related work. In Section III we introduce our scale coding methods. We report on a number of experiments in Section IV and conclude in Section V.

II. Related Work

Recognizing actions in still images is a difficult problem. The bounding boxes of humans performing actions are provided both at training and test time and the task is to associate an action category label to each person bounding box. The problem is difficult due to the lack of temporal information and due to large variations in human appearance, scale and pose. In recent years, several methods have been proposed which focus on finding human-object interactions [2], [4], [5], [6], [16]. Maji et al. [6] propose a poselet activation vector method that captures the pose in multi-scale manner. The technique works by capturing the 3D pose of a human and the corresponding action from the still images. A human-centric approach is proposed in [2] that works by first localizing a human and then finding an object and its relationship to it. Delaitre et al. [16] propose a discriminative approach where the model is constructed using spatial co-occurrences of objects and individual body parts. The problem of the large number of possible interaction pairs is handled using a discriminative learning procedure. A method based on attributes and parts is proposed in [5]. The approach works by learning a set of sparse attribute and part bases for action recognition.

Other than finding human-object interactions, several state-of-the-art methods are based on the bag-of-words (BOW) model [1], [3], [7]. The authors of [1] use a method based on a max margin classifier to learn the discriminative spatial saliency of images. Recently, Khan et al. [7] investigated the contribution of color for action recognition. In their evaluation, several color descriptors and color-shape fusion approaches are evaluated for both action classification and detection. Sharma et al. [17] propose an approach based on learning a model based on a collection of part templates learned discriminatively to select scale-space locations in the images.

Scale invariant bag-of-words based image representations are commonly used in for object and scene recognition [12], [13], [14], [15]. Several sampling strategies for BOW-based object recognition are evaluated by Nowak et al. [12]. In their evaluation, a random sampling strategy was shown to yield superior performance compared to sophisticated interest-point detectors. Bosch et al. [13] compute multiple dense color descriptors using different scales to allow for scale variation between images. Combining intensity-based and color interest point detectors together with dense multi-scale sampling [8] was shown to yield excellent results for object recognition. Vedaldi et al. [18] construct dense SIFT and colorSIFT features at four scales for object detection. The visual features are then encoded into a single histogram representation. The work in [19] uses an approach based on random forests with discriminative decision trees for feature mining to address the problem of fine-grained object recognition.

Despite the success of bag-of-words based action recognition, the principal state-of-the-art approaches all adopt the conventional technique of constructing a single histogram based on the occurrence of each visual-word independent of the original scale of the feature in an image [1], [3], [7]. In this paper, we take a different approach by exploiting multi-scale information for action recognition in still images. The first approach is dependent on the image size while the second approach takes into account the bounding box information available for each person instance both at training and test time. We construct three histograms: one for small-, one for medium- and one for large-scale visual-words. The scales are encoded into one of the three histograms depending on either the size of the image or the bounding box. Finally, the three histograms are concatenated into a single image representation for action classification.

III. Scale Coding: Relaxing Scale Invariance

In this section we discuss several approaches to relaxing the scale invariance of local descriptors in the bag-of-words model. Originally, the BOW model was developed for image classification where the task was to determine the presence or absence of objects in images. In this case invariance with respect to scale was essential, since the object could be in the background of the image and thus appear small, or instead in the foreground and cover most of the image space. Therefore, extracted features were converted to a canonical scale — and from which point on the original feature scale was discarded — and mapped onto a visual vocabulary. When BOW was extended to object detection [18], [20] and later to action recognition [2], [3], [7] the same strategy was applied.

However, the invariance comes at the expense of discriminative power. A drawback of this representation is that the relative scale of features is lost; the representation is not suited to discriminate, for example, images which contain a large circle and a small circle from images which contain two circles of equal scale (in this example the circle can be thought of as one of the visual words). Especially, in the case of action recognition were we have bounding box information both at training and testing time, alternative scale coding strategies can be considered. The resulting representations still have a degree of scale invariance, however they do not suffer the same drop in discriminative power. Here we propose two strategies to

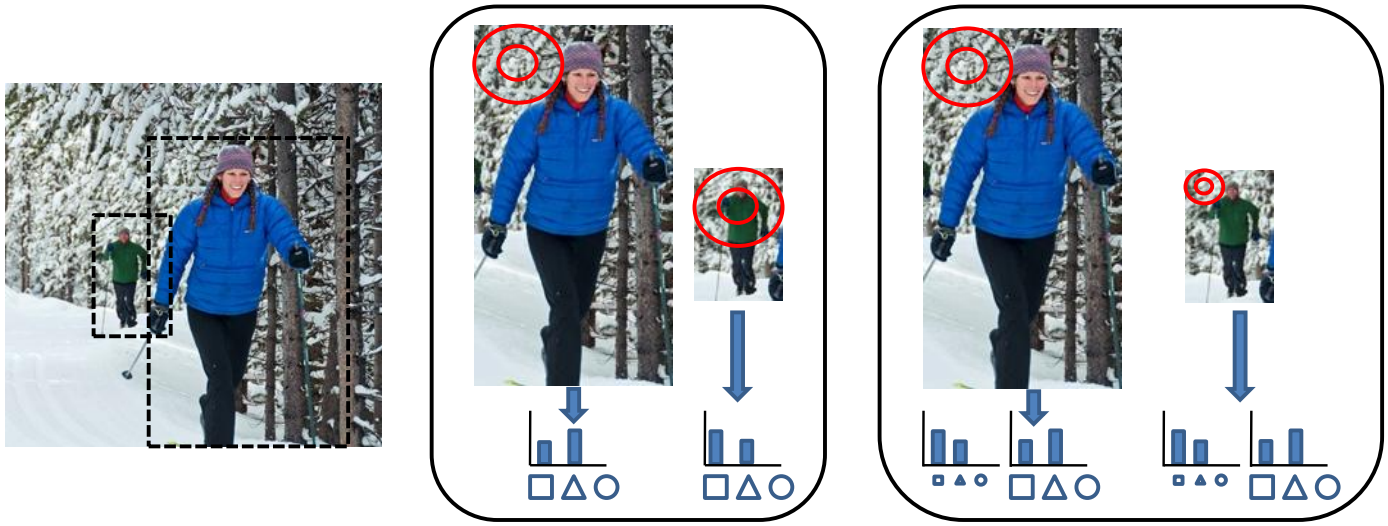

Fig. 2. Scale coding: (left) input image, superimposed bounding boxes indicate persons performing an action; (middle) in standard scale coding the scale is independent of the object size (red circles show the extracted feature scales), and they are all assembled in a single histogram per image; (right) our proposal of relative scale coding adapts to the bounding box of the object. This ensures that similar structures (such as hands and ski poles) are captured at the same scale independent of the bounding box size. The features are represented in several concatenated histograms which collect a range of feature scales.

handle multi-scale features for applications where bounding box information is provided.

A. Scale-invariant Image Representation

We first introduce some notation. Features are extracted from all bounding boxes using multiscale sampling on a dense grid. For each bounding box B, we extract a set of features:

$$F(B) = \{f_i^s | i \in \{1, \ldots, N\}, s \in \{1, \ldots M\}\}, \quad (1)$$

where $i \in \{1, \ldots, N\}$ indexes the N feature sites in B defined by the dense grid, and $s \in \{1, \ldots M\}$ indexes the M scales extracted at each site. Assume we have a visual vocabulary $W = \{\mathrm{w}_1, \ldots, \mathrm{w}_\mathrm{P}\}$ of P words. Every feature is quantized to its closest vocabulary word (in Euclidean sense); we denote the vocabulary word closest to f_i^s with w_i^s.

In the scale-invariant representation a single histogram $h(\cdot|B)$ is constructed for each bounding box B:

$$h(\mathrm{w}_n|B) \propto \sum_{i=1}^{N} \sum_{s=1}^{M} \delta(\mathrm{w}_\mathrm{i}^\mathrm{s}, \mathrm{w}_n), \quad (2)$$

where δ is the Dirac delta function. The final histogram contains the frequency of each visual word independent of the original scale of the feature. This scale invariant representation is by far the most applied approach to handling multiple scales in the BOW model [2], [3], [7]. In the next section we propose two new scale coding approaches to multi-scale image representations.

B. Absolute Scale Coding

The first scale preserving scale coding method we propose uses an absolute multi-scale image representation:

$$h^t(\mathrm{w}_n|B) \propto \sum_{i=1}^{N} \sum_{s \in S^t} \delta(\mathrm{w}_\mathrm{i}^\mathrm{s}, \mathrm{w}_n) \quad (3)$$

where the scales, instead of being marginalized completely away as in equation (2), are divided into several subgroups S^t that partition the entire set of extracted scales (i.e. $\bigcup_t S^t = S$). In this work we consider a split of all extracted scales into three groups with $t \in \{\mathrm{s}, \mathrm{m}, \mathrm{l}\}$ for small, medium and large scale features. These three scale partitions are defined as:

$$\begin{aligned} S^\mathrm{s} &= \{s \mid s \leq s^\mathrm{s}, s \in S\} \\ S^\mathrm{m} &= \{s \mid s^\mathrm{s} < s \leq s^\mathrm{l}, s \in S\} \\ S^\mathrm{l} &= \{s \mid s^\mathrm{l} < s, s \in S\}, \end{aligned} \quad (4)$$

where the two cutoff thresholds s^s and s^l are input parameters. This representation thus preserves coarse scale information about the originally extracted features, however note that they are not relative with respect to the bounding box of the object.

C. Relative Scale Coding

In relative scale coding features are represented relative to the bounding box size of the object (in our case the person bounding box). The representation is computed with:

$$h^t(\mathrm{w}_n|B) \propto \frac{1}{|\hat{S}^t|} \sum_{i=1}^{N} \sum_{s \in \hat{S}^t} \delta(\mathrm{w}_\mathrm{i}^\mathrm{s}, \mathrm{w}_n) \quad (5)$$

The difference between Eq. 5 and 3 is that the scale of each feature s is re-parameterized relative to the size of the bounding box B in which it was observed:

$$\hat{s} = \frac{B_w + B_h}{\bar{w} + \bar{h}} s \quad (6)$$

where B_w and B_h are the width and height of bounding box B and $\bar{w}$ and $\bar{h}$ are the mean width and hight of all bounding boxes in the training set.

As for absolute scale coding, described in the previous section, we group relative scales into three groups. The relative

Fig. 3. Example images from the Willow action recognition dataset. The dataset contains seven action categories namely: interacting with computer, photographing, playing music, riding bike, riding horse, running and walking.

scale splits $\hat{S}^t$ are defined with respect to relative scale:

$$\begin{aligned} \hat{S}^{\mathrm{s}} &= \{\hat{s} \mid \hat{s} \leq s^{\mathrm{s}}, s \in S\} \\ \hat{S}^{\mathrm{m}} &= \{\hat{s} \mid s^{\mathrm{s}} < \hat{s} \leq s^{\mathrm{m}}, s \in S\} \\ \hat{S}^{\mathrm{l}} &= \{\hat{s} \mid s^{\mathrm{m}} < \hat{s}, s \in S\}\,. \end{aligned} \tag{7}$$

Since the number of scales which falls into the small, medium and large scale range histogram now varies with the size of the bounding box we introduce a normalization factor $|\hat{S}^t|$ to counter this. Here $|\hat{S}^t|$ is the cardinality of the set $\hat{S}^t$.

Relative scale coding represents visual words at a certain relative scale with respect to the bounding box size. Again, it consists of three histograms for small, medium and large scale visual words. However, depending on the size of the bounding box the scales which are considered small, medium and large change. An illustrative overview of this approach is provided in Figure 2. In contrast to the standard approach this method does allow the relative scale of visual words to be maintained in the representation, without completely sacrificing the scale invariance of the original representation.

IV. Experimental Results

We evaluate our two new scale coding strategies for the problem of action recognition in still images. We first detail our experimental setup, then present a baseline comparison of scale invariant and our new scale coding schemes. Finally, we compare with state-of-the-art action recognition methods.

A. Experimental Setup

All experiments are performed on the Willow action dataset. The dataset consists of seven action categories: interacting with computer, photographing, playing music, riding bike, riding horse, running and walking.[1] We densely sample features at nine different scales. To describe each local feature we use the SIFT descriptor [11], commonly used for shape description in BOW models. A visual vocabulary of 1000 visual-words is obtained using the K-means algorithm. The histogram-based multi-scale representations, as discussed in Section III, are constructed for final image representations. We extract nine scales: $S = \{\sqrt{2}, 2, 2\sqrt{2}, ..., 16\}$ and set $s_s = 4\sqrt{2}$ and $s_l = 8\sqrt{2}$. Finally, the image representations are input to a nonlinear SVM with a χ^2 kernel [21]. Performance is evaluated using the PASCAL criteria as average precision (AP) under the precision-recall curve. The final score is the mean AP over all seven action categories.

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

B. Baseline Comparison of Scale Coding Schemes

We first compare our scale coding approaches with conventional scale invariant coding. Note that the same visual vocabulary is used in all experiments and only the image representation varies depending on the method of constructing the final histogram. Table I compares our scale coding approaches to the conventional method of constructing a single histogram ignoring the scale of visual words. The conventional scale-invariant coding approach yields a mean AP of 64.9%. Absolute scale coding improves on this with mean AP of 66.7%. The performance improves significantly for action categories such as *interacting with computer*, *playing music* and *running*. Finally, the second multi-scale representation based on relative scale coding further improves the performance with a mean AP of 67.4%. On the *interacting with computer* action category, the relative scale representation yields a gain of 13.8%. Similarly, the performance also improves for the *riding a bike* category. Overall, the two multi-scale representations improve the performance on 5 out of 7 action categories.

C. Comparison with the State-of-the-art

We now compare our scale coding representations with state-of-the-art methods from the literature. To obtain the best possible final results, we combine the representations based on absolute and relative coding by combining the classifier outputs. As a single feature is used in our experiments (SIFT), we only compare with approaches also using a single visual cue. Table II compares our approach with state-of-the-art methods. Our approach yields the best results on 5 out of 7 action categories on this dataset. We achieve a mean AP of 68.0%, which is the best result reported on this dataset [1], [3], [16], [17] using a single visual cue. Delaitre et al. [16] obtain a mean AP of 64.1% with an approach that models complex

	int. computer	photographing	playingmusic	ridingbike	ridinghorse	running	walking	mean AP
Scale Invariant Coding	52.9	**44.7**	73.3	86.1	**77.9**	59.3	60.4	64.9
Absolute Scale Coding	59.6	43.5	**77.0**	86.4	77.2	**63.5**	60.0	66.7
Relative Scale Coding	**66.7**	43.0	75.2	**87.4**	77.2	62.2	**60.5**	**67.4**

TABLE I. COMPARISON OF THE CONVENTIONAL SCALE INVARIANT APPROACH WITH OUR PROPOSED SCALE CODING IMAGE REPRESENTATIONS. THE SCALE CODING METHODS OUTPERFORM THE CONVENTIONAL APPROACH ON 5 OUT OF 7 ACTION CATEGORIES.

	int. computer	photographing	playingmusic	ridingbike	ridinghorse	running	walking	mean AP
Delaitre et al.[3]	58.2	35.4	73.2	82.4	69.6	44.5	54.2	59.6
Delaitre et al.[16]	56.6	37.5	72.0	90.4	75.0	59.7	57.6	64.1
Sharma et al.[1]	59.7	42.6	74.6	87.8	84.2	56.1	56.5	65.9
Sharma et al.[17]	64.5	40.9	75.0	**91.0**	**87.6**	55.0	59.2	67.6
Our approach	**67.2**	**43.9**	**76.1**	87.2	77.2	**63.7**	**60.6**	**68.0**

TABLE II. COMPARISON OF OUR SCALE CODING APPROACH WITH THE STATE-OF-THE-ART USING SINGLE CUE ON THE WILLOW DATASET. OUR APPROACH PROVIDES THE BEST RESULTS ON 5 OUT OF 7 ACTION CATEGORIES ON THIS DATASET.

interactions between persons and objects. In their method, the interactions are modeled using external data to train body part detectors. Sharma et al. [1] achieve a mean AP of 65.9% using a spatial saliency based approach. The work of [17] reports a mean AP of 67.6% by learning part-based representations and using a bag-of-words based framework.

The best result on this dataset is 70.1% by Khan et al. [7] obtained by combining multiple color-shape fusion strategies. Our approach, despite its simplicity, outperforms the more complex approaches using a single visual cue on this dataset. It is also worth mentioning that our scale coding approach is generic and could be incorporated in any of the state-of-the-art methods [1], [3], [7], [17].

V. CONCLUSION

In the traditional bag-of-words approach scale information is ignored and the representation is invariant with respect to scale. With this invariance the representation loses discriminative power. In this article we have proposed two alternative approaches that encode scale information in the final BOW histograms representing images. In the first scale coding approach, the absolute scale of the feature is coded in the representation by constructing different histograms for small, medium and large features. In the second scale coding approach, the relative scale of features with respect to the person bounding box is used to separate features into different histograms. Results on action recognition show that scale coding image representations obtain superior results compared to the scale invariant baseline, especially in the case of relative scale coding where a gain of 2.5% is obtained. Furthermore, the method compares favorably with respect to other state-of-the-art methods based on a single cue.

In this work we have exploited available bounding box information to compute the relative scale of features. Using other cues such as depth information (or estimation) to compute relative scale would extend the applicability of the proposed approach to image classification applications where bounding box information is not present. It would also be interesting to extend the work to object detection based on BOW. Especially exciting would be a combination with recent approaches which selectively evaluate object presence only for a limited set of bounding boxes [22].

ACKNOWLEDGEMENTS

This work has been supported by SSF through a grant for the project CUAS, by VR through a grant for the project ETT, through the Strategic Area for ICT research ELLIIT, and CADICS. Andrew D. Bagdanov acknowledges the support of a Ramon y Cajal Fellowship.

REFERENCES

[1] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *CVPR*, 2012.

[2] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *PAMI*, vol. 34, no. 3, pp. 601–614, 2012.

[3] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," in *BMVC*, 2010.

[4] B. Yao and F.-F. Li, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *PAMI*, vol. 34, no. 9, pp. 1691–1703, 2012.

[5] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and F.-F. Li, "Human action recognition by learning bases of action attributes and parts," in *ICCV*, 2011.

[6] S. Maji, L. D. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *CVPR*, 2011.

[7] F. S. Khan, R. M. Anwer, J. van de Weijer, A. Bagdanov, A. Lopez, and M. Felsberg, "Coloring action recognition in still images," *IJCV*, vol. 105, no. 3, pp. 205–221, 2013.

[8] D. Rojas, F. S. Khan, J. van de Weijer, and T. Gevers, "The impact of color on bag-of-words based object recognition," in *ICPR*, 2010.

[9] N. Elfiky, F. S. Khan, J. van de Weijer, and J. Gonzalez, "Discriminative compact pyramids for object and scene recognition," *PR*, vol. 45, no. 4, pp. 1627–1636, 2012.

[10] F. S. Khan, J. van de Weijer, A. D. Bagdanov, and M. Vanrell, "Portmanteau vocabularies for multi-cue image representations," in *NIPS*, 2011.

[11] D. G. Lowe, "Distinctive image features from scale-invariant points," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[12] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *ECCV*, 2006.

[13] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *PAMI*, vol. 30, no. 4, pp. 712–727, 2008.

[14] F. S. Khan, J. van de Weijer, and M. Vanrell, "Modulating shape features by color attention for object recognition," *IJCV*, vol. 98, no. 1, pp. 49–64, 2012.

[15] K. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *PAMI*, vol. 32, no. 9, pp. 1582–1596, 2010.

[16] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in *NIPS*, 2011.

[17] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *CVPR*, 2013.

[18] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *ICCV*, 2009.

[19] B. Yao, A. Khosla, and F.-F. Li, "Combining randomization and discrimination for fine-grained image categorization," in *CVPR*, 2011.

[20] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *ICCV*, 2009.

[21] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object catergories: A comprehensive study," *IJCV*, vol. 73, no. 2, pp. 213–218, 2007.

[22] K. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *ICCV*, 2011.