Machine Vision and Applications (2018) 29:55–71
https://doi.org/10.1007/s00138-017-0871-1

ORIGINAL PAPER

Scale coding bag of deep features for human attribute and action recognition

Fahad Shahbaz Khan[1] · **Joost van de Weijer**[2] · **Rao Muhammad Anwer**[3] · **Andrew D. Bagdanov**[4] · **Michael Felsberg**[1] · **Jorma Laaksonen**[3]

Received: 15 December 2016 / Revised: 19 June 2017 / Accepted: 9 August 2017 / Published online: 11 September 2017

Abstract Most approaches to human attribute and action recognition in still images are based on image representation in which multi-scale local features are pooled across scale into a single, scale-invariant encoding. Both in bag-of-words and the recently popular representations based on convolutional neural networks, local features are computed at multiple scales. However, these multi-scale convolutional features are pooled into a single scale-invariant representation. We argue that entirely scale-invariant image representations are sub-optimal and investigate approaches to scale coding within a bag of deep features framework. Our approach encodes multi-scale information explicitly during the image encoding stage. We propose two strategies to encode multi-scale information explicitly in the final image representation. We validate our two scale coding techniques on five datasets: Willow, PASCAL VOC 2010, PASCAL VOC 2012, Stanford-40 and Human Attributes (HAT-27). On all datasets, the proposed scale coding approaches outperform both the scale-invariant method and the standard deep features of the same network. Further, combining our scale coding approaches with standard deep features leads to consistent improvement over the state of the art.

✉ Fahad Shahbaz Khan
fahad.khan@liu.se

1 Computer Vision Laboratory, Linköping University, Linköping, Sweden

2 Computer Vision Centre Barcelona, Universitat Autonoma de Barcelona, Bellaterra, Spain

3 Department of Computer Science, Aalto University School of Science, Espoo, Finland

4 Media Integration and Communication Center, University of Florence, Florence, Italy

Keywords Action recognition · Attribute recognition · Bag of deep features

1 Introduction

Human attribute and action recognition in still images is a challenging problem that has received much attention in recent years [22,24,47,67]. Both tasks are challenging since humans are often occluded, can appear in different poses (also articulated), under varying illumination, and at low resolution. Furthermore, significant variations in scale both within and across different classes make these tasks extremely challenging. Figure 1 shows example images from different categories in the Stanford-40 and the Willow action datasets. The bounding box (in red) of each person instance is provided both at train and test time. These examples illustrate the inter- and intra-class scale variations common to certain action categories. In this paper, we investigate image representations which are robust to these variations in scale.

Bag-of-words (BOW) image representations have been successfully applied to image classification and action recognition tasks [24,31,52,56]. The first stage within the framework, known as feature detection, involves detecting keypoint locations in an image. The standard approach for feature detection is to use dense multi-scale feature sampling [23,43,52] by scanning the image at multiple scales at fixed locations on a grid of rectangular patches. Next, each feature is quantized against a visual vocabulary to arrive at the final image representation. A disadvantage of the standard BOW pipeline is that all scale information is lost. Though for image classification such an invariance with respect to scale might seem beneficial since instances can appear at different scales, it trades discriminative information for scale invariance. We distinguish two relevant sources of scale information: (i)

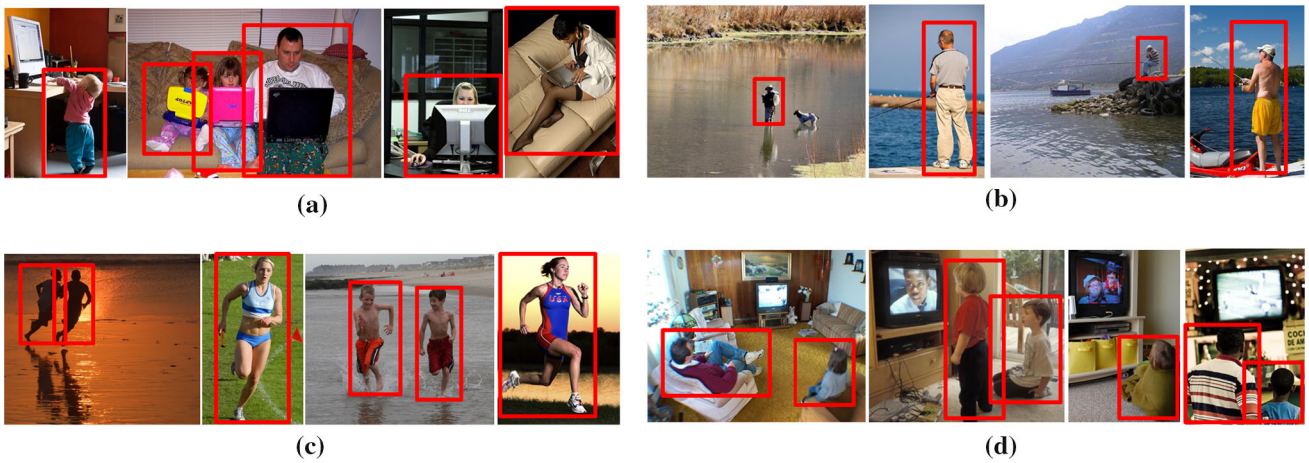


Fig. 1 Example images from the *interacting with computer*, *fishing*, *running* and *watching tv* action categories. These examples illustrate the scale variations present, especially with respect to the size of *bounding boxes* within each action category. This suggests that alternative image representations may be desirable that explicitly encode multi-scale information. **a** Class interacting with computer. **b** Class fishing. **c** Class running. **d** Class watching tv

dataset scale prior: due to the acquisition of the dataset some visual words could be more indicative of certain categories at a particular scale than at others scales (e.g., we do not expect persons of 15 pixels nor shoes at 200 pixels) and (ii) relative scale: in the presence of a reference scale, such as the person bounding box provided for action recognition, we have knowledge of the actual scale at which we expect to detect parts of the object (e.g., the hands and head of the person). Both examples show the relevance of scale information for discriminative image representations and are the motivation for our investigation into scale coding methods for human attribute and action recognition.

Traditionally, BOW methods are based on hand-crafted local features such as SIFT [37], HOG [7] or Color Names [64]. Recently, convolutional neural networks (CNNs) have had tremendous success on a wide range of computer vision applications, including human attribute [68] and action recognition [44]. Cimpoi et al. [6] showed how deep convolutional features (i.e., dense local features extracted at multiple scales from the convolutional layers of CNNs) can be exploited within a BOW pipeline for object and texture recognition. In their approach, a Fisher vector encoding scheme is used to obtain the final image representation (called FV-CNN). We will refer to this type of image representation as a *bag of deep features*, and in this work we will apply various scale coding approaches to a bag of deep features.

Contributions In this paper, we investigate strategies for incorporating multi-scale information in image representations for human attribute and action recognition in still images. Existing approaches encode multi-scale information only at the feature extraction stage by extracting convolutional features at multiple scales. However, the final image representation in these approaches is scale-invariant since all the scales are pooled into a single histogram. To prevent the loss of scale information, we will investigate two complementary scale coding approaches. The first approach, which we call *absolute scale coding*, is based on a multi-scale image representation with scale encoded with respect to the image size. The second approach, called *relative scale coding*, instead encodes feature scale relative to the size of the bounding box corresponding to the person instance. *Scale coding of bag of deep features* is performed by applying the coding strategies to the convolutional features from a pre-trained deep network. The final image representation is obtained by concatenating the small, medium and large scale image representations. We perform comprehensive experiments on five standard datasets: Willow, PASCAL VOC 2010, PASCAL VOC 2012, Stanford-40 and the Database of Human Attributes (HAT-27). Our experiments clearly demonstrate that our scale coding strategies outperform both the scale-invariant bag of deep features and the standard deep features extracted from fully connected layers of the same network. We further show that combining our scale coding strategies with standard features from the FC layer further improves the classification performance. Our scale coding image representations are flexible and effective, while providing consistent improvement over the state of the art.

In the next section, we discuss work from the literature related to our proposed scale coding technique. In Sects. 3 and 4, we describe our two proposals for coding scale information in image representations. We report on a comprehensive set of experiments performed on five standard benchmark datasets in Sect. 5, and in Sect. 6 we conclude with a discussion of our contribution.

2 Related work

Scale plays an important role in feature detection. Important work includes early research on pattern spectrums [39] based on mathematical morphology which provided insight into the existence of features at certain scales in object shapes. In the field of scale-space theory [28,65], the scale of features was examined by analyzing how images evolved when smoothed with Gaussian filters of increasing scale. This theory was also at the basis of the SIFT detector [37] which obtained scale-invariant features and showed these to be highly effective for detection of objects. The detection of scale-invariant features was much studied within the context of bag-of-words [41]. In contrast to the methods we describe in this paper, most of these works ignore the relative size of detected features.

In this section, we briefly review the state of the art in bag-of-words image recognition frameworks, multi-scale deep feature learning, and human action and attribute recognition.

The bag-of-words framework In the last decade, the bag-of-words (BOW)-based image representation dominated the state of the art in object recognition [10] and image retrieval [20]. The BOW image representation is obtained by performing three steps in succession: feature detection, feature extraction and feature encoding. Feature detection involves keypoint selection either with an interest point detector [42] or with dense sampling on a fixed grid [2,62]. Several works [49,52] demonstrated the importance of using a combination of interest point and grid-based dense sampling. This feature detection phase, especially when done on a dense grid, is usually *multi-scale* in the sense that feature descriptors are extracted at multiple scales at all points.

Local descriptors, such as SIFT and HOG, are extracted in the feature extraction phase [7,37]. Next, several encoding schemes can be considered [11,46,70]. The work of [11] investigated soft assignment of local features to visual words. Zhou et al. [70] introduced super-vector coding that performs a nonlinear mapping of each local feature descriptor to construct a high-dimensional sparse vector. The Improved Fisher vectors, introduced by Perronnin et al. [46], encode local descriptors as gradients with respect to a generative model of image formation (usually a Gaussian mixture model (GMM) over local descriptors which serves as a visual vocabulary for Fisher vector coding).

Regardless of the feature encoding scheme, most existing methods achieve scale invariance by simply quantizing local descriptors to a visual vocabulary *independently* of the scale at which they were extracted. Visual words have no associated scale information, and scale is thus marginalized away in the histogram construction process. In this work, we use a Fisher vector encoding scheme within the BOW framework and investigate techniques to relax scale invariance in the final image representation. We refer to this as *scale coding*, since scale information is preserved in the final encoding.

Deep features Recently, image representations based on convolutional neural networks (CNNs) [32] have demonstrated significant improvements over the state of the art in image classification [44], object detection [12], scene recognition [29], action recognition [33], and attribute recognition [68]. CNNs consist of a series of convolution and pooling operations followed by one or more fully connected (FC) layers. Deep networks are trained using raw image pixels with a fixed input size. These networks require large amounts of labeled training data. The introduction of large datasets (e.g., ImageNet [50] and the parallelism enabled by modern GPUs have facilitated the rapid deployment of deep networks for visual recognition.

It has been shown that intermediate, hidden activations of fully connected layers in trained deep network are general-purpose features applicable to visual recognition tasks [1,44]. Several recent methods [6,14,36] have shown superior performance using convolutional layer activations instead of fully connected ones. These convolutional layers are discriminative, semantically meaningful and mitigate the need to use a fixed input image size. Gong et al. [14] proposed a multi-scale orderless pooling (MOP) approach by constructing descriptors from the fully connected (FC) layer of the network. The descriptors are extracted from densely sampled square image windows. The descriptors are then pooled using the VLAD encoding [21] scheme to obtain final image representation.

In contrast to MOP [14], Cimpoi et al. [6] showed how deep convolutional features (i.e., dense local features extracted at multiple scales from the convolutional layers of CNNs) can be exploited within a BOW pipeline. In their approach, a Fisher Vector encoding scheme is used to obtain the final image representation. We will refer to this type of image representation as a *bag of deep features*, and in this work we will apply various scale coding approaches to a bag of deep features. Though FV-CNN [6] employs multi-scale convolutional features, the descriptors are pooled into a single Fisher vector representation. This implies that the final image representation is scale-invariant since all the scales are pooled into a single feature vector. We argue that such a representation is sub-optimal for the problem of human attribute and action recognition and propose to explicitly incorporate multi-scale information in the final image representation.

Action recognition in still images Recognizing actions in still images is a difficult problem that has gained a lot of attention recently [24,44,47,67]. In action recognition, bounding box information of each person instance is provided both at train and test time. The task is to associate an action category label with each person bounding box. Several approaches have addressed the problem of action recognition by finding human–object interactions in an image [38,47,67]. A poselet-based approach was proposed in [38] where poselet activation vectors capture the pose of a person. Prest et al. [47]

proposed a human-centric approach that localizes humans and objects associated with an action. Yao et al. [67] propose to learn a set of sparse attribute and part bases for action recognition in still images. Recently, a comprehensive survey was performed by Ziaeefard et al. [72] on action recognition methods exploiting semantic information. In their survey, it was shown that methods exploiting semantic information yield superior performance compared to their nonsemantic counterparts in many scenarios. Human action recognition in still images is also discussed within the context of fuzzy domain in a recent survey [34].

Other approaches to action recognition employ BOW-based image representations [24,25,56]. Sharma et al. [56] proposed the use of discriminative spatial saliency for action recognition by employing a max margin classifier. A comprehensive evaluation of color descriptors and color-shape fusion approaches was performed by Khan et al. [24] for action recognition. Khan et al. [25] proposed pose-normalized semantic pyramids employing pre-trained body part detectors. A comprehensive survey was performed by Guo and Lai [15] where existing action recognition methods are categorized based on high-level cues and low-level features.

Recently, image representations based on deep features have achieved superior performance for action recognition [13,18,44]. Oquab et al. [44] proposed mid-level image representations using pre-trained CNNs for image classification and action recognition. The work of [13] proposed learning deep features jointly for action classification and detection. Hoai et al. [18] proposed regularized max pooling and extract features at multiple deformable sub-windows. The aforementioned approaches employ deep features extracted from activations of the fully connected layers of the deep CNNs. In contrast, we use dense local features from the convolutional layers of networks for image description.

The incorporation of scale information has been investigated in the context of action recognition in videos [53,71]. The work of [53] proposes to construct multiple dictionaries at different resolutions in a final video representation. The work of [71] proposes multi-scale spatio-temporal concatenation of local features resulting in a set of natural action structures. Both these methods do not consider relative scale coding. In addition, our approach is based on recent advancements of deep convolutional neural networks (CNNs) and Fisher vector encoding scheme. We revisit the problem of incorporating scale information for the popular CNNs-based deep features. To the best of our knowledge, we are the first to investigate and propose scale-coded bag of deep feature representations applicable for both human attribute and action recognition in still images.

Human attribute recognition Recognizing human attributes such as, age, gender and clothing style is an active research problem with many real-world applications. State-of-the-art approaches employ part-based representations [3,25,68] to counter the problem of pose normalization. Bourdev et al. [3] proposed semantic part detection using poselets and constructing pose-normalized representations. Their approach employs HOGs for part descriptions. Later, Zhang et al. [68] extended the approach of [3] by replacing the HOG features with CNNs. Khan et al. [25] proposed pre-trained body part detectors to automatically construct pose-normalized semantic pyramid representations.

In this work, we investigate scale coding strategies for human attribute and action recognition in still images. This paper is an extended version of our earlier work [26]. Instead of using the standard BOW framework with SIFT features, we propose scale coding strategies within the emerging bag of deep features paradigm that uses dense convolutional features in classical BOW pipelines. We additionally extend our experiments with results on the PASCAL VOC 2010, PASCAL 2012, Standord-40 and Human Attribute (HAT-27) datasets.

3 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 is to determine the presence or absence of objects in images. In such situations, invariance with respect to scale is important since the object could be in the background of the image and thus appear small, or instead appear in the foreground and cover most of the image space. Therefore, extracted features are converted to a canonical scale—and from that point on the original feature scale is discarded—and mapped onto a visual vocabulary. When BOW was extended to object detection [16,62] and later to action recognition [8,24,47] this same strategy for ensuring scale invariance was applied.

However, this invariance comes at the expense of discriminative power through the loss of information about relative scale between features. In particular, we distinguish two sources of scale information: (i) dataset scale prior: the acquisition and/or collection protocol of a dataset results in a distribution of the object-sizes as a function of the size in the image, e.g., most cars are between 100–200 pixels, and (ii) relative scale: in the presence of a reference scale, such as the person bounding box, we have knowledge of the actual scale at which we expect to detect parts or objects (e.g., the size at which the action-defining object such as the mobile phone or musical instrument should be detected). These sources of information are lost in scale-invariant image representations. We propose two strategies to encode scale information of features in the final image representation.

3.1 Scale-invariant image representation

We first introduce some notation. Features are extracted from the person bounding boxes (available at both training and testing time) using multi-scale sampling at all feature locations. For each bounding box B, we extract a set of features:

$$F(B) = \left\{ \mathbf{x}_i^s \mid i \in \{1, \ldots, N\}, s \in \{1, \ldots M\} \right\},$$

where $i \in \{1, \ldots, N\}$ indexes the N feature sites in B, and $s \in \{1, \ldots M\}$ indexes the M scales extracted at each site.

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

$$h(B) \propto \sum_{i=1}^{N} \sum_{s=1}^{M} c\left(\mathbf{x}_i^s\right) \tag{1}$$

where $c : \Re^p \rightarrow \Re^q$ denotes a coding scheme which maps the input feature space of dimensionality p to the image representation space of dimensionality q.

Let us first consider the case of standard bag-of-words with nearest neighbor assignment to the closest vocabulary word. Assume we have a visual vocabulary $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_P\}$ of P words. Every feature is quantized to its closest (in the Euclidean sense) vocabulary word:

$$w_i^s = \underset{k \in \{1, \ldots, P\}}{\operatorname{argmin}} \, d(\mathbf{x}_i^s, \mathbf{w}_k),$$

where $d(\cdot, \cdot)$ is the Euclidean distance. Index w_i^s corresponds to the vocabulary word to which feature $\mathbf{x}_i^s$ is assigned. Letting $e(i)$ be the one-hot column vector of length q with all zeros except for the index i where it is one, we can write the standard hard assignment bag-of-words by plugging in:

$$c_{\text{BOW}}\left(\mathbf{x}_i^s\right) = e(w_i^s)$$

as the coding function in Eq. 1.

For the case of Fisher vector encoding [45], a Gaussian Mixture Model (GMM) is fitted to the distribution of local features $\mathbf{x}$:

$$u_\lambda(\mathbf{x}) = \sum_{1}^{K} w_k u_k(\mathbf{x}),$$

where $\lambda = \left\{ w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \right\}_{k=1}^{K}$ are the parameters defining the GMM, respectively, the mixing weights, the means, and covariance matrices for the K Gaussian mixture components, and

$$u_k(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} \left|\boldsymbol{\Sigma}_k\right|^{1/2}} \times \exp\left\{ -\frac{1}{2} \left(\mathbf{x} - \boldsymbol{\mu}_k\right)' \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_k\right) \right\}.$$

The coding function is then given by the gradient with respect to all of the GMM parameters:

$$c_{\text{Fisher}}\left(\mathbf{x}_i^s\right) = \nabla \log u_\lambda\left(\mathbf{x}_i^s\right).$$

and plugging this encoding function into Eq. 1. For more details on the Fisher vector encoding, please refer to [51]. Since the superiority of Fisher coding has been shown in several publications we apply Fisher coding throughout this paper [4].

3.2 Absolute scale coding

The first scale preserving scale coding method we propose uses an absolute multi-scale image representation. Letting $S = \{1, \ldots, M\}$ be the set of extracted feature scales, we encode features in groups of scales:

$$h^t(B) \propto \sum_{i=1}^{N} \sum_{s \in S^t} c(\mathbf{x}_i^s). \tag{2}$$

Instead of being marginalized completely away as in Eq. 1, feature scales are instead divided into several subgroups S^t that partition the entire set of extracted scales (i.e., $\bigcup_t S^t = \{1, \ldots, M\}$). In this work, we consider a split of all scales into three groups with $t \in \{\text{s}, \text{m}, \text{l}\}$ for small, medium and large scale features. For absolute scale coding, these three-scale partitions are defined as:

$$\begin{aligned} S^{\text{s}} &= \left\{ s \mid s \leq s^{\text{s}}, s \in S \right\} \\ S^{\text{m}} &= \left\{ s \mid s^{\text{s}} < s \leq s^{\text{l}}, s \in S \right\} \\ S^{\text{l}} &= \left\{ s \mid s^{\text{l}} < s, s \in S \right\}, \end{aligned}$$

where the two cutoff thresholds s^{s} and s^{l} are parameters of the encoding. The final representation is obtained by concatenating these three encodings of the box B and thus preserves coarse scale information about the originally extracted features, and it exploits what we refer to as the *dataset scale prior* or absolute scale. However, note that this representation does not exploit the relative scale information.

3.3 Relative scale coding

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

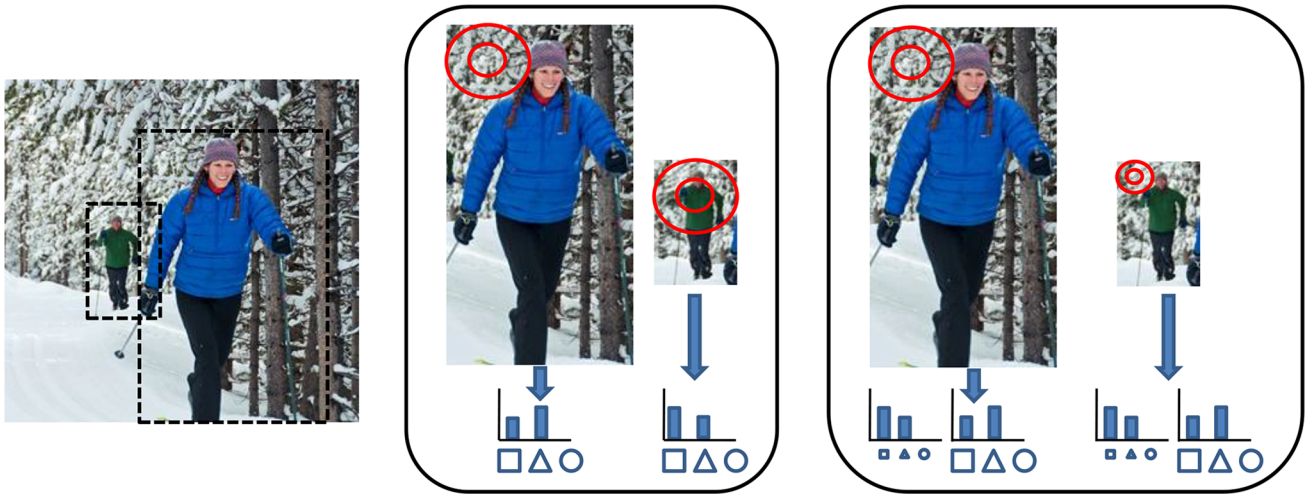


Fig. 2 Scale coding (*left*) input image, superimposed *bounding boxes* indicate persons performing an action; (*middle*) in standard scale coding the scale is independent of the object size (*red circles* show the extracted feature scales), and they are all assembled in a single histogram per image; (*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 absolute *bounding box* size. The features are represented in several concatenated histograms which collect a range of feature scales (color figure online)

$$h^t(B) \propto \frac{1}{|\hat{S}^t|} \sum_{i=1}^{N} \sum_{s \in \hat{S}^t} c(\mathbf{x}_i^s) \tag{3}$$

The difference between Eqs. 2 and 3 is that the scale of each feature s is reparameterized relative to the size of the bounding box B in which it was observed:

$$\hat{s} = \frac{B_w + B_h}{\bar{w} + \bar{h}} s$$

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. Taking into account, the boundary length ensures that elongated objects have large scales.

As for absolute scale coding, described in the previous section, we group relative scales into three groups. The relative scale splits $\hat{S}^t$ are defined with respect to relative scale:

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

Since the number of scales which fall into the small, medium and large scale range image representation now varies with the size of the bounding box, we introduce a normalization factor $|\hat{S}^t|$ in Eq. 3 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 image representations for small, medium and large scale visual words, which are then concatenated together to form the final representation for B. 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 given in Fig. 2. In contrast to the standard approach, this method preserves the relative scale of visual words without completely sacrificing the scale invariance of the original representation.

3.4 Scale partitioning

Until now, we have considered partitioning the features into three scale-groups: small, medium and large. Here, we evaluate this choice and compare it with other partitioning of the scales.

To evaluate the partitioning of scales, we extracted features at $M = 21$ different scales on Stanford-40 and the PASCAL VOC 2010 datasets. For this evaluation, we performed absolute scale encoding and varied the number of scale partitions from one (equivalent to standard scale-invariant coding) to 21 in which case every scale is represented by a single image representation. In Fig. 3, we plot the mean average precision (mAP) on Stanford-40 and PASCAL 2010 as a function of the number of scale partitions. The curve clearly shows that absolute scale coding outperforms the generally applied representation based on scale-invariant coding (which collects all scales in a single partition). Furthermore, it shows that after three-scale partitions, the gain of increasing the number of partitions is negligible. Throughout this paper, we use three-scale partitions for all scale coding experiments.

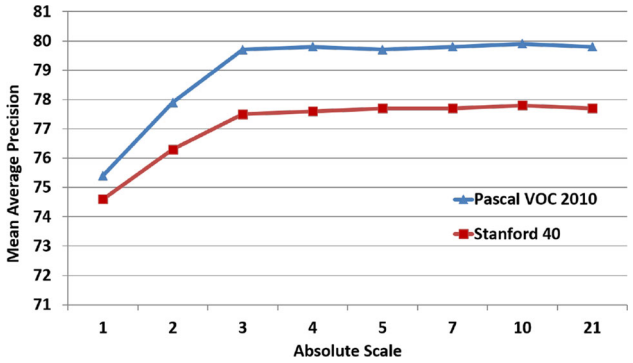


Fig. 3 Mean average precision as a function of number of scale partitions for absolute scale coding (Sect. 3.2. Performance is shown on the Stanford-40 and PASCAL VOC 2010 validation sets. On both, absolute scale coding improves performance compared to scale-invariant coding (which groups all scales to single representation). A consistent improvement is achieved when using three-scale partitions for absolute coding

4 The bag of deep features model

Inspired by the recent success of CNNs, we use deep features in our scale coding framework.

Deep convolutional features Similar to [6], we use the VGG-19 network proposed by Simonyan and Zisserman [60], pre-trained on the ImageNet dataset. It was shown to provide the best performance in a recent evaluation [5,6] for image classification tasks. In the VGG-19 network, input images are convolved with 3×3 filters at each pixel at a stride of 1 pixel. The network contains several max-pooling layers which perform spatial pooling over 2×2 pixel windows at a stride of 2 pixels. The VGG-19 network contains 3 fully connected (FC) layers at the end. The width of the VGG-19 network starts from 64 feature maps in the first layer and increases by a factor of 2 after each max-pooling layer to reach 512 feature maps at its widest (see [60] for more details).

Typically, the activations from the FC layer(s) are used as input features for classification. For VGG-19 this results in a 4096-dimensional representation. In contrast, we use the output of the last convolutional layer of the network since it was shown to provide superior performance compared to other layers [6]. This layer returns dense convolutional features at a stride of eight pixels. We use these 512-dimensional descriptors as local features within our scale coding framework. To obtain multi-scale samples, we rescale all images over a range of scales and pass them through the network for feature extraction. Note that the number of extracted local convolutional patches depend on the size of the input image.

Vocabulary construction and assignment In standard BOW all features are quantized against a scale-invariant visual vocabulary. The local features are then pooled in a single scale-invariant image representation. Similar to [6], we use the Fisher vector encoding for our scale coding models. For vocabulary construction, we use the Gaussian mixture model (GMM). The convolutional features are then pooled via the Fisher encoding that captures the average first- and second-order differences. The 21 different scales are pooled into the three-scale partitions to ensure that the scale information is preserved in the final representation. It is worth mentioning that our scale coding schemes can also be used with other encoding schemes such as hard assignment, soft assignment, and VLAD [21].

5 Experimental results

In this section, we present the results of our scale coding strategies for the problem of human attribute and action recognition. First we detail our experimental setup and datasets used in our evaluation, and then present a comprehensive comparison of our approach with baseline methods. Finally, we compare our approach with the state of the art in human attribute and action recognition.

5.1 Experimental setup

As mentioned earlier, bounding boxes of person instances are provided at both train and test time in human attribute and action recognition. Thus, the task is to predict the human attribute or action category for each person bounding box. To incorporate context information, we extend each person bounding box by 50% of its width and height.

In our experiments, we use the pre-trained VGG-19 network [60]. Similar to Cimpoi et al. [6], we extract the convolutional features from the output of the last convolutional layer of the VGG-19 network. The convolutional features are not de-correlated by using PCA before employing Fisher Vector encoding, since it has been shown [6] to deteriorate the results. The convolutional features are extracted after rescaling the image at 21 different scales $s \in \{0.5 + 0.1n \mid n = 0, 1, \ldots, 20\}$. This results in 512-dimensional dense local features for each scaled image. On an image of size 300×300, the feature extraction on multi-core CPU takes about 5 seconds. For our scale coding approaches, we keep a single, constant threshold for all datasets.

For each problem instance, we construct a visual vocabulary using a Gaussian mixture model (GMM) with 16 components. In Fig. 4, we plot the mean average precision (mAP) on Willow and PASCAL 2010 datasets as a function of the number of Gaussian components. We observed no significant gain in classification performance by increasing the number of Gaussian components beyond 16. The parameters of this model are fit using a set of dense descriptors sampled from descriptors over all scales on the training set. We randomly sample 100 descriptor points from each training image. The resulting sampled feature descriptors from the

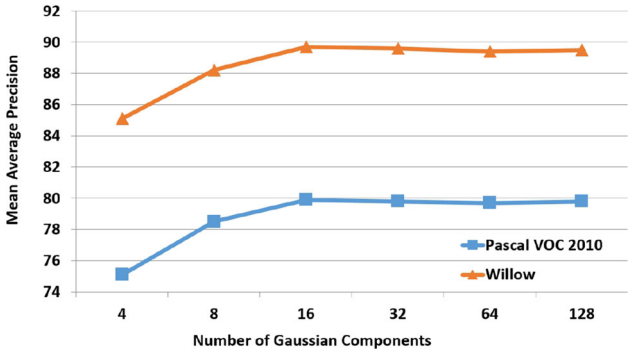


Fig. 4 Mean average precision as a function of number of Gaussian components for relative scale coding. Performance is shown on the Willow and PASCAL VOC 2010 sets. On both, our scale coding provides best results when using a Gaussian mixture model (GMM) with 16 components

whole training set are then used to construct a GMM-based dictionary. We also perform experiments by varying the number of feature samples per image. However, no improvement in performance was observed with increased feature samples per image. We employ a GMM with diagonal covariances. Finally, the Fisher vector representations discussed in Sect. 3 are constructed for each image. The Fisher vector encoding is performed with respect to the Gaussian mixture model (GMM) with means, diagonal covariances, and prior probabilities. The dense image features have dimensionality 512, and so our final scale-coded Fisher vector representation has dimensionality $3 \times (2 \times 16 \times 512 + 16) = 49200$ (i.e., it is a concatenation of the Fisher vector encoding the three-scale categories). In our experiments, we use the standard VLFeat library [61] commonly used to construct GMM-based vocabulary and the improved Fisher vector-based image representations. For classification, we employ SVMs with linear kernels on the concatenated Fisher vectors of each scale coding groups described above.

5.2 Datasets

We perform experiments on five datasets to validate our approach:

- The *Willow Action Dataset* consisting of seven action categories *interacting with computer, photographing, playing music, riding bike, riding horse, running and walking*.[1]
- The *Stanford-40 Action Dataset* consisting of 9532 images of 40 different action categories such as *gardening, fishing, applauding, cooking, brushing teeth, cutting vegetables, and drinking*.[2]
- The *PASCAL VOC 2010 Action Dataset* consisting of 9 action categories *phoning, playing instrument, reading, riding bike, riding horse, running, taking photo, using computer and walking*.[3]
- The *PASCAL VOC 2012 Action Dataset* consisting of 10 different action classes *phoning, playing instrument, reading, riding bike, riding horse, running, taking photo, using computer, walking and jumping*.[4]
- The *27 Human Attributes Dataset (HAT-27)* consisting of 9344 images of 27 different human attributes such as *crouching, casual jacket, wedding dress, young and female*.[5]

The test sets for both the PASCAL VOC 2010 and 2012 datasets are withheld by the organizers and results must be submitted to an evaluation server. We report the results on the test sets in Sect. 5.2.2 and provide a comparison with state-of-the-art methods. For the Willow [8], Stanford-40 [67] and HAT-27 [55] datasets we use the train and test splits provided by the respective authors.

Evaluation criteria We follow the same evaluation protocol as used for each dataset. Performance is measured in average precision as area under the precision-recall curve. The final performance is calculated by taking the mean average precision (mAP) over all categories in each dataset.

5.2.1 Baseline scale coding performance analysis

We first give a comparison of our scale coding strategies with the baseline scale-invariant coding. Our baseline is the FV-CNN approach [6] where multi-scale convolutional features are pooled in a single scale-invariant image representation. The FV-CNN approach is further extended with spatial information by employing spatial pyramid pooling scheme [31]. The spatial pyramid scheme is used with two levels (1×1 and 2×2), yielding a total of 5 cells. We also compare our results with standard deep features obtained from the activations of the first fully connected layer of the CNN. Additionally, we compare our approach with multi-scale orderless pooling (MOP) [14] by extracting FC activations at three levels: 4096-dimensional CNN activation from the entire image patch (the person bounding box), 128×128 patches of 4096 dimensions pooled using VLAD encoding with 100 visual words, and the same VLAD encoding but with 64×64 patches. The

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

[2] Stanford-40 is at http://vision.stanford.edu/Datasets/40actions.html.

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

[4] PASCAL 2012 is at: http://www.pascal-network.org/challenges/VOC/voc2012/.

[5] HAT-27 is available at: https://sharma.users.greyc.fr/hatdb/.

Table 1 Comparison (in mAP) of the standard deep features (FC, for "fully connected"), the MOP approach, the baseline scale-invariant approach (FV-CNN), the scale-invariant spatial pyramid approach (FV-CNN-SP), and our proposed relative and absolute scale coding schemes

	Willow	PASCAL 2010	PASCAL 2012	Stanford-40	HAT-27
VGG-19 FC [60]	87.1	72.0	74.0	70.3	61.2
MOP [14]	87.6	74.8	75.3	74.2	64.1
FV-CNN [6]	87.9	75.4	75.6	74.6	64.5
FV-CNN-SP	88.4	78.1	77.3	76.9	66.6
Absolute scale coding	89.3	79.7	78.1	77.5	67.3
Relative scale coding	89.7	79.9	78.4	77.8	67.4
Absolute + relative + FC	92.1	82.7	80.3	80.0	70.6

Scale coding yields consistent improvements on all 5 datasets

Table 2 Comparison of our approach with the state of the art on the willow dataset

	Int. computer	Photographing	Playingmusic	Ridingbike	Ridinghorse	Running	Walking	mAP
BOW-DPM [8]	58.2	35.4	73.2	82.4	69.6	44.5	54.2	59.6
POI [9]	56.6	37.5	72.0	90.4	75.0	59.7	57.6	64.1
DS [56]	59.7	42.6	74.6	87.8	84.2	56.1	56.5	65.9
CF [24]	61.9	48.2	76.5	90.3	84.3	64.7	64.6	70.1
EPM [57]	64.5	40.9	75.0	91.0	87.6	55.0	59.2	67.6
SC [26]	67.2	43.9	76.1	87.2	77.2	63.7	60.6	68.0
SM-SP [25]	66.8	48.0	77.5	93.8	87.9	67.2	63.3	72.1
EDM [33]	86.6	**90.5**	89.9	98.2	92.7	46.2	58.9	80.4
NSP [40]	88.6	61.8	93.4	98.8	98.4	69.4	62.3	81.7
DPM-VR [59]	84.9	72.0	91.2	96.9	93.6	73.4	61.0	81.9
This paper	**96.6**	89.2	**98.2**	**99.8**	**99.3**	**83.0**	**78.7**	**92.1**

Our proposed approach achieves best results on 6 out of 7 action categories
The bold numbers in the tables indicate best results (highest classification scores)

three representations are concatenated into a single feature vector for classification. Note that we use the same VGG-19 network for all of these image encodings.

Table 1 gives the baseline comparison on all five datasets. Since the PASCAL VOC 2010 and 2012 test sets are withheld by the organizers, performance is measured on the validation sets for the baseline comparison. The standard multi-scale-invariant approach (FV-CNN) improves the classification performance compared to the standard FC deep features. The spatial pyramid-based FV-CNN further improves over the standard FV-CNN method. Our absolute and relative scale coding approaches provide a consistent gain in performance on all datasets, compared to baselines using features from the same deep network. Note that the standard scale-invariant (FV-CNN) and our scale coding schemes are constructed using the same visual vocabulary (GMM) and set of local features from the convolutional layer. Finally, a further gain in accuracy is obtained by combining the classification scores of our two scale coding approaches with the standard FC deep features. This combination is done by simply adding the three classifier outputs. On the Stanford-40 and HAT-27 datasets, this approach yields a considerable gain of 6.5 and 4.8% in mAP, respectively, compared to the MOP approach employing FC features from the same network (VGG-19). These results suggest that the FC, absolute scale, and relative scale encodings have complementary information that when combined yield results superior to each individual representation.

5.2.2 *Comparison with the state of the art*

We now compare our approach with the state of the art on the five benchmark datasets. In this section, we report results for the combination of our relative and absolute scale coding strategies with the FC deep features. The combination is done by simply adding the three classifier outputs.

Willow Table 2 gives a comparison of our combined scale coding approach with the state of the art on the Willow dataset. Our approach achieves the best performance reported on this dataset, with an mAP of 92.1%. The shared part detectors approach of Mettes et al. [40] achieves an mAP of 81.7%, while the part-based deep representation approach [59] obtains an mAP of 81.9%. Our approach, without exploiting any part information, yields the best results on 6 out of 7 action categories, with an overall gain of 10.2% in mAP compared to [59].

PASCAL VOC 2010 Table 3 compares our combined scale coding approach with the state of the art on the PASCAL VOC 2010 Action Recognition *test* set. The color fusion

Table 3 Comparison with the state-of-the-art results on the PASCAL VOC 2010 *test* set

	Phoning	Playingmusic	Reading	Ridingbike	Ridinghorse	Running	Takingphoto	Usingcomputer	Walking	mAP
Poselets [38]	49.6	43.2	27.7	83.7	89.4	85.6	31.0	59.1	67.9	59.7
IaC [54]	45.5	54.5	31.7	75.2	88.1	76.9	32.9	64.1	62.0	59.0
POI [9]	48.6	53.1	28.6	80.1	90.7	85.8	33.5	56.1	69.6	60.7
LAP [67]	42.8	60.8	41.5	80.2	90.6	87.8	41.4	66.1	74.4	65.1
WPOI [47]	55.0	81.0	69.0	71.0	90.0	59.0	36.0	50.0	44.0	62.0
CF [24]	52.1	52.0	34.1	81.5	90.3	88.1	37.3	59.9	66.5	62.4
SM-SP [25]	52.2	55.3	35.4	81.4	91.2	89.3	38.6	59.6	68.7	63.5
This paper	64.3	94.5	65.1	96.9	96.8	93.4	77.1	87.7	78.9	83.7

Our scale coding-based approach provides consistent improvements compared to existing methods

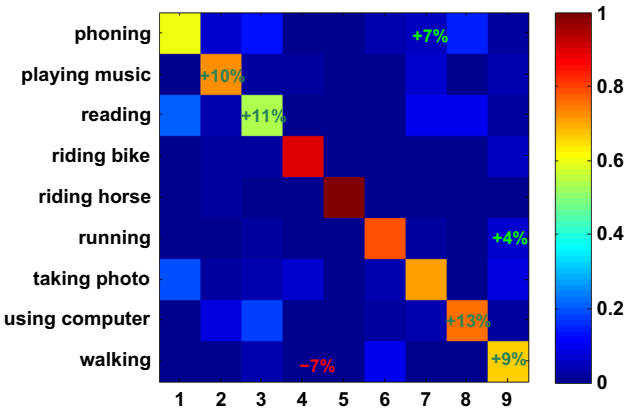


Fig. 5 Confusion matrix for our approach, combining both absolute and relative scale coding, on PASCAL VOC 2010. We superimposed the differences with the confusion matrix for the scale-invariant FV-CNN approach for confusions where the absolute change is at least 4%. Our approach provides consistent improvements, in general, but improves significantly the performance for playing music (10%), reading (11%) and using computer (13%) categories

approach of Khan et al. [24] achieves an mAP of 62.4%, the semantic pyramid approach by Khan et al. [25] obtains a mAP of 63.5%, and the method of Yao et al. [67] based on learning a sparse basis of attributes and parts achieves an mAP of 65.1%. Our approach yields consistent improvement over the state of the art with an mAP of 83.7% on this dataset. Figure 5 shows the confusion matrix for our scale coding-based approach. The differences with the confusion matrix based on the standard scale-invariant FV-CNN approach are superimposed for confusions where the absolute change is at least 4%. Overall, our approach improves the classification results with notable improvements for playing music (10%), reading (11%) and using computer (13%) action categories. Further, our approach reduces confusion all categories except for walking.

PASCAL VOC 2012 In Table 4, we compare our approach with state of the art on the PASCAL VOC 2012 Action Recognition *test* set. Among existing approaches, Regularized Max Pooling (RMP) [18] obtains a mAP score of 76.4%.

Table 4 Comparison of our proposed approach with the state of the art on the PASCAL VOC 2012 *test* set

	Phoning	Playingmusic	Reading	Ridingbike	Ridinghorse	Running	Takingphoto	Usingcomputer	Walking	Jumping	mAP
Stanford	44.8	66.6	44.4	93.2	94.2	87.6	38.4	70.6	75.6	75.7	69.1
Oxford	50.0	65.3	39.5	94.1	95.9	87.7	42.7	68.6	74.5	77.0	69.5
Action poselets [38]	32.4	45.4	27.5	84.5	88.3	77.2	31.2	47.4	58.2	59.3	55.1
MDF [44]	46.0	75.6	45.3	93.5	95.0	86.5	49.3	66.7	69.5	78.4	70.2
WAB [19]	49.5	67.5	39.1	94.3	96.0	89.2	44.5	69.0	**75.9**	79.6	70.5
Action R-CNN [13]	47.4	77.5	42.2	94.9	94.3	87.0	52.9	66.5	66.5	76.2	70.5
RMP [18]	52.9	84.3	53.6	95.6	96.1	89.7	60.4	76.0	72.9	82.3	76.4
TL [27]	62.4	91.3	61.1	93.3	95.1	84.1	59.8	84.5	53.0	84.9	77.0
VGG-19 + VGG-16 + Full image [60]	**71.3**	**94.7**	**71.3**	97.1	**98.2**	**90.2**	73.3	**88.5**	66.4	89.3	**84.0**
This paper	69.7	92.4	70.8	**97.2**	98.0	89.8	**73.8**	88.4	69.4	**89.5**	83.9

The best existing results are obtained by combining FC features from two CNNs (VGG-16 and VGG-19). The features are extracted both from the full image and bounding box of a person. Our approach, based only on VGG-19 network and without using full image information, obtains comparable performance with best results on 3 out of 10 action categories

The bold numbers in the tables indicate best results (highest classification scores)

The best results on this dataset are obtained by combining the FC features of the VGG-16 and VGG-19 networks. These FC features are extracted both from the full image and the provided bounding box of the person. Our combined scale coding-based approach provides the best results on 3 out of 10 action categories, and achieves an mAP of 83.9% on the PASCAL 2012 test set. It is worth mentioning that our scale coding-based approach employs a single network (VGG-19) and does not exploit the full image information. Combining our scale coding-based approaches using multiple deep networks is expected to further improve performance.

Table 5 Comparison of the proposed approach with the state-of-the-art methods on Stanford-40 dataset

	SB	CF	SM-SP	Place	D-EPM	TL	Ours
mAP	45.7	51.9	53.0	55.3	72.3	75.4	**80.0**

Our approach yields a significant gain over the best reported results in the literature

Stanford-40 dataset In Table 5, we compare scale coding with state-of-the-art approaches: SB [67], CF [24], SM-SP

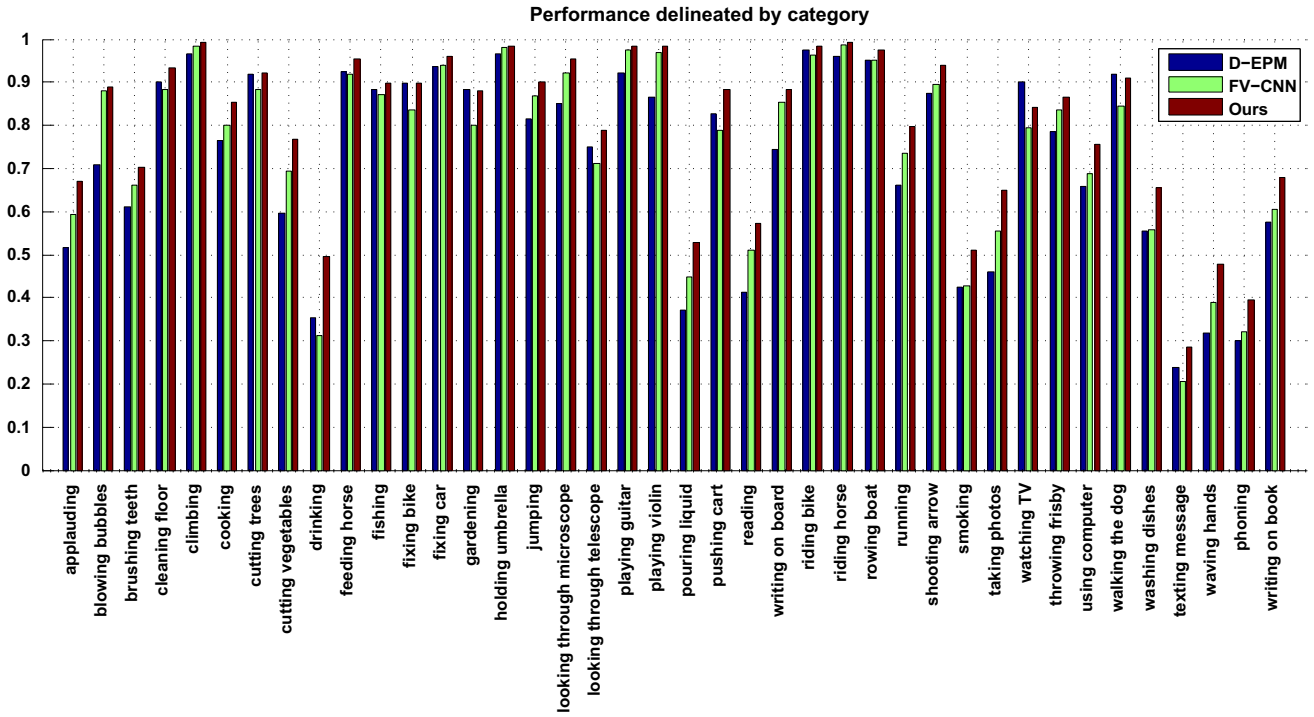
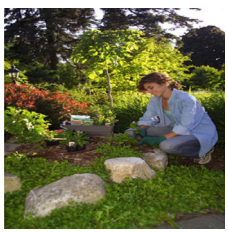


Fig. 6 Per-category performance comparison (in AP) of our approach with the D-EPM method [58] and the scale-invariant FV-CNN approach [6]. Our approach improves the results on 37 out 40 action classes compared to these two methods

Table 6 Images from pouring liquid, gardening, using computer and fishing action categories from the Stanford-40 dataset

Ranking of different action categories				
Method				
VGG-19 FC [60]	186 (98)	51 (5)	104 (24)	56 (5)
FV-CNN [6]	144 (62)	57 (4)	98 (22)	52 (5)
Our approach	32 (5)	17 (1)	30 (1)	38 (2)

The number indicates the absolute rank of corresponding image in the list of all test images sorted by the probability for the corresponding class. The number in parentheses after each rank is the number of false positives appearing *before* the example test image in the ranked list. Lower absolute rank reflects higher confidence in the class label. The action category list contains 5532 test instances. Our approach outperforms both VGG-19 FC and FV-CNN methods on these images demonstrating the importance of coding multiple scales in the final image representation

Table 7 Comparison of our approach with the state of the art on the 27 Human Attributes (HAT-27) dataset. Our method, without using part-based information, achieves the best performance compared the state-of-the-art D-EPM method [58] exploiting the part-based information

	Female	Frontalpose	Profilepose	Turnedback	Upper-body	Standing	Runwalk	Crouching	Sitting	Armsbent	Elderly	Middleaged	Young	Teen
EPM [57]	85.9	93.6	67.3	77.2	97.9	98.0	74.6	24.0	62.7	94.0	38.9	68.9	64.2	36.2
RAD [22]	91.4	**96.8**	**77.2**	**89.8**	96.3	97.7	63.5	12.3	59.3	95.4	32.1	70.0	65.6	33.5
SM-SP [25]	86.1	92.2	60.5	64.8	94.0	96.6	76.8	23.2	63.7	92.8	37.7	69.4	67.7	36.4
D-EPM [58]	**93.2**	95.2	72.6	84.0	**99.0**	98.7	75.1	**34.2**	77.8	95.4	46.4	72.7	70.1	36.8
This paper	92.0	95.7	62.9	86.9	95.1	**98.8**	**80.3**	31.6	**87.0**	95.5	**54.7**	**74.6**	**72.9**	**39.3**

	Kid	Baby	Tanktop	Tshirt	Casualjacket	Mensuit	Longskirt	Shortskirt	Smallshorts	Lowcuttop	Swimsuit	Weddingdress	Bermudashorts	mAP
EPM [57]	49.7	24.3	37.7	61.6	40.0	57.1	44.8	39.0	46.8	61.3	32.2	64.2	43.7	58.7
RAD [22]	53.5	16.3	37.0	67.1	42.6	64.8	42.0	30.1	49.6	66.0	46.7	62.1	42.0	59.3
SM-SP [25]	55.9	18.3	40.6	65.6	40.6	57.4	33.3	38.9	44.0	67.7	46.7	46.3	38.6	57.6
D-EPM [58]	62.5	**39.5**	48.4	75.1	**63.5**	**75.9**	**67.3**	52.6	56.6	84.6	**67.8**	**79.7**	53.1	69.6
This paper	**70.5**	31.3	**56.5**	**80.4**	62.8	69.2	62.0	**52.9**	66.4	**84.7**	63.5	72.5	**65.2**	**70.6**

The bold numbers in the tables indicate best results (highest classification scores)

[25], Place [69], D-EPM [58] and TL [27]. Stanford-40 is the most challenging action dataset and contains 40 categories. The semantic pyramids of Khan et al. [25] achieve an mAP of 53.0%. Their approach combines spatial pyramid representations of full-body, upper-body and face regions using multiple visual cues. The work of [69] uses deep features trained on ImageNet and a recently introduced large scale dataset of place scenes. Their hybrid deep feature-based approach achieves a mAP of 55.3%. The D-EPM approach [58] based on expanded part models and deep features achieves a mAP score of 72.3%. The transfer learning (TL)-based approach [27] with deep features obtains a mAP score of 75.4%. Our combined scale coding approach achieves state-of-the-art results with a gain of 4.6% in mAP compared to the TL-based approach [27].

In Figure 6, we compare the per-category performance of our approach with two state-of-the-art approaches: D-EPM [58] and FV-CNN [6]. Our scale coding-based approach achieves the best performance on 37 out of 40 action categories on this dataset. A significant gain in performance is achieved especially for drinking (+14%), washing dishes (+9%), taking photos (+9%), smoking (+8%), and waving hands (+8%) action categories, all compared to the two state-of-the-art methods. Table 6 shows example images from pouring liquid, gardening, using computer and fishing categories. The corresponding ranks are shown for the standard VGG-19 FC, FV-CNN and our scale coding-based approach. The number indicates the absolute rank of corresponding image in the list of all test images sorted by the probability for the corresponding class. A lower number implies higher confidence in the action class label. We also show rank with respect to the number of false positives appearing before the example test image in the ranked list. Our approach obtains improved rank on these images compared to the two standard approaches.

Human Attributes (HAT-27) dataset Finally, Table 7 shows a comparison of our scale coding-based approach with state-of-the-art methods on the Human Attributes (HAT-27) dataset. The dataset contains 27 different human attributes. The expanded part-based approach by Sharma et al. [57] yields an mAP of 58.7%, and semantic pyramids [25], combining body part information in a spatial pyramid representation, an mAP of 57.6%. The approach of [22] is based on learning a rich appearance part-based dictionary and achieves an mAP of 59.3%. Deep FC features from the VGG-19 network obtains a mAP score of 62.1%. The D-EPM method [58] based on deep features and expanded part-based models achieves the best results among the existing methods with a mAP of 69.6%. On this dataset, our scale coding-based approach outperforms the D-EPM method with a mAP score of 70.6%. Scale coding yields the best classification perfor-

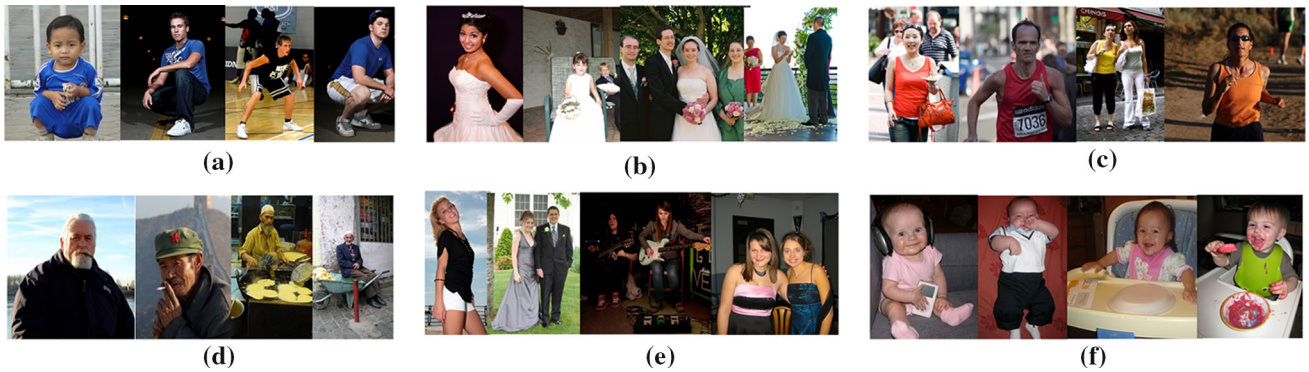


Fig. 7 Attribute classification performance of our approach on the HAT-27 dataset. We show top correct predictions of six attribute categories: 'crouching,' 'wedding dress,' 'tank top,' 'elderly,' 'young' and 'baby.' **a** Class crouching. **b** Class wedding dress. **c** Class tank top. **d** Class elderly. **e** Class young **f** Attribute category baby

mance on 15 out of 27 attribute categories compared to the state of the art.

Figure 7 illustrates the top four predictions of six attribute categories from the HAT-27 dataset. These examples show inter- and intra-class variations among different categories. The variations in scale and pose of persons make the problem of attribute classification challenging. Our scale coding-based approach consistently improves the performance on this dataset.

5.2.3 Generality of our approach

We have validated our approach on two challenging problems: human attribute and action classification. However, our scale coding approach is generic and is more broadly applicable to other recognition tasks. To validate the generality of our approach, we perform additional experiments on the popular MIT indoor scene 67 dataset [48] for scene recognition task. The dataset contains 15620 images of 67 indoor scene classes. The training and test configurations are provided by the original authors, where each category has around 80 images for training and 20 for testing. The performance is measured in terms of mean classification accuracy computed over all the categories in the dataset. Most existing methods [17,30,35,63] report results using VGG16 model, pre-trained on either ImageNet or Places dataset. For fair comparison, we also validate our absolute scale coding approach using the VGG16 model and only compare with approaches pre-trained on ImageNet dataset.

Table 8 shows a comparison of our absolute scale coding-based approach with state-of-the-art methods on the MIT indoor scene 67 dataset. Among existing approaches, the work of [17] also investigated multi-scale CNN architecture by training scale-specific networks on the ImageNet dataset, focusing on the CNN models. Several scale-specific networks are combined by concatenating the FC7 features of all networks, yielding a mean accuracy score of 79.0%. Instead, our approach proposes multi-scale image representations by using a single pre-trained deep network and preserving scale information in the pooling method, obtaining a mean classification score of 81.9%. The results are further improved when combining the standard FC features with our scale coding approach. It is worth to mention that a higher recognition score of 86.0% is obtained by [17], when combining scale-specific networks trained on both ImageNet and Places scene dataset. However, when using the same deep model architecture (VGG16) and only ImageNet dataset for network training, our results of 83.1% are superior compared to 79.0% obtained by the multi-scale scale-specific networks [17].

Table 8 Comparison of our approach with the state of the art on the MIT indoor scene 67 dataset

Method	Accuracy
DAG-CNN [66]	77.5
Deep spatial pyramid [35]	78.3
B-CNN [35]	79.0
FV-CNN [6]	79.2
SPLeap [30]	73.5
Standard VGG16 [60]	69.6
Standard VGG16 FT [17]	76.4
Multi-scale network [17]	79.0
Our approach	81.9
Our approach + standard VGG16	**83.1**

Our method achieves superior performance compared to existing approaches based on the same VGG16 model
The bold numbers in the tables indicate best results (highest classification scores)

6 Conclusions

In this paper, we investigated the problem of encoding multi-scale information for still images in the context of

human attribute and action recognition. Most state-of-the-art approaches based on the BOW framework compute local descriptors at multiple scales. However, multi-scale information is not explicitly encoded as all the features from different scales are pooled into a single scale-invariant histogram. In the context of human attribute and action recognition, we demonstrate that both absolute and relative scale information can be encoded in final image representations and that relaxing the traditional scale invariance commonly employed in image classification can lead to significant gains in recognition performance.

We proposed two alternative scale coding approaches that explicitly encode scale information in the final image representation. The absolute scale of local features is encoded by constructing separate representations for small, medium and large features, while the relative scale of the local features is encoded with respect to the size of the bounding box corresponding to the person instance in human action or attribute recognition problems. In both cases, the final image representation is obtained by concatenating the small, medium and large scale representations.

Comprehensive experiments on five datasets demonstrate the effectiveness of our proposed approach. The results clearly demonstrate that our scale coding strategies outperform both the scale-invariant bag of deep features and the standard deep features extracted from the same network. An interesting future direction is the investigation of scale coding strategies for object detection and fine-grained object localization. We believe that our scale coding schemes could be very effective for representing candidate regions in object detection techniques based on bottom-up proposal of likely object regions.

Acknowledgements This work has been funded by the projects TIN2013-41751, TIN2016-79717-R and of the Spanish Ministry of Science, the Catalan project 2014 SGR 221, the CHISTERA project PCIN-2015-251, SSF through a grant for the project SymbiCloud, VR (EMC2), VR starting Grant (2016-05543), through the Strategic Area for ICT research ELLIIT, the Grant 251170 of the Academy of Finland. The calculations were performed using computer resources within the Aalto University School of Science "Science-IT" project and NSC. We also acknowledge the support from Nvidia.

References

1. Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: CVPRW, pp. 512–519 (2014)
2. Bosch, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. PAMI **30**(4), 712–727 (2008)
3. Bourdev, L., Maji, S., Malik, J.: Describing people: a poselet-based approach to attribute classification. In: ICCV, pp. 1543–1550 (2011)
4. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC (2011)
5. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: BMVC (2014)
6. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: CVPR, pp. 3828–3836 (2015)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
8. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: BMVC (2010)
9. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: NIPS, pp. 1503–1511 (2011)
10. Everingham, M., Gool, L.J.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010)
11. van Gemert, J., Veenman, C., Smeulders, A., Geusebroek, J.M.: Visual word ambiguity. PAMI **32**(7), 1271–1283 (2010)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587 (2014)
13. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: R-CNNS for pose estimation and action detection. arXiv preprint arXiv:1406.5212 (2014)
14. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: ECCV, pp. 392–407 (2014)
15. Guo, G., Lai, A.: A survey on still image based human action recognition. PR **47**(10), 3343–3361 (2014)
16. Harzallah, H., Jurie, F., Schmid, C.: Combining efficient object localization and image classification. In: ICCV (2009)
17. Herranz, L., Jiang, S., Li, X.: Scene recognition with cnns: objects, scales and dataset bias. In: CVPR (2016)
18. Hoai, M.: Regularized max pooling for image categorization. In: BMVC (2014)
19. Hoai, M., Ladicky, L., Zisserman, A.: Action recognition from weak alignment of body parts. In: BMVC (2014)
20. Jegou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. IJCV **87**(3), 316–336 (2010)
21. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR, pp. 3304–3311 (2010)
22. Joo, J., Wang, S., Zhu, SC.: Human attribute recognition by rich appearance dictionary. In: ICCV, pp. 721–728 (2013)
23. Khan, F.S., van de Weijer, J., Vanrell, M.: Modulating shape features by color attention for object recognition. IJCV **98**(1), 49–64 (2012)
24. Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A., Lopez, A., Felsberg, M.: Coloring action recognition in still images. IJCV **105**(3), 205–221 (2013)
25. Khan, F.S., van de Weijer, J., Anwer, R.M., Felsberg, M., Gatta, C.: Semantic pyramids for gender and action recognition. TIP **23**(8), 3633–3645 (2014a)
26. Khan, FS., van de Weijer, J., Bagdanov, A., Felsberg, M.: Scale coding bag-of-words for action recognition. In: ICPR, pp. 1514–1519 (2014b)

27. Khan, F.S., Xu, J., van de Weijer, J., Bagdanov, A., Anwer, R.M., Lopez, A.: Recognizing actions through action-specific person detection. TIP **24**(11), 4422–4432 (2015)
28. Koenderink, J.: The structure of images. Biol. Cybern. **50**(5), 363–370 (1984)
29. Koskela, M., Laaksonen, J.: Convolutional network features for scene recognition. In: ACM Multimedia, pp. 1169–1172 (2014)
30. Kulkarni, P., Jurie, F., Zepeda, J., Perez, P., Chevallie, L.: Spleap: soft pooling of learned parts for image classification. In: ECCV (2016)
31. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. 2169–2178 (2006)
32. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. In: NIPS, pp. 396–404 (1989)
33. Liang, Z., Wang, X., Huang, R., Lin, L.: An expressive deep model for human action parsing from a single image. In: ICME, pp. 1–6 (2014)
34. Lim, C.H., Vats, E., Chan, C.S.: Fuzzy human motion analysis: a review. PR **48**(5), 1773–1796 (2015)
35. Lin, TY., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: ICCV (2015)
36. Liu, L., Shen, C., van den Hengel, A.: The treasure beneath convolutional layers: cross-convolutional-layer pooling for image classification. In: CVPR, pp. 4749–4757 (2015)
37. Lowe, D.: Distinctive image features from scale-invariant points. IJCV **60**(2), 91–110 (2004)
38. Maji, S., Bourdev, LD., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR, pp. 3177–3184 (2011)
39. Maragos, P.: Pattern spectrum and multiscale shape representation. PAMI **11**(7), 701–716 (1989)
40. Mettes, P., van Gemer, J., Snoek, C.: No spare parts: sharing part detectors for image categorization. CVIU **152**, 131–141 (2016)
41. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV **60**(1), 63–86 (2004a)
42. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV **60**(1), 63–86 (2004b)
43. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: ECCV, pp. 490–503 (2006)
44. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR, pp. 1717–1724 (2014)
45. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR, pp. 1–8 (2007)
46. Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV, pp. 143–156 (2010)
47. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. PAMI **34**(3), 601–614 (2012)
48. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR (2009)
49. Rojas, D., Khan, FS., van de Weijer, J., Gevers, T.: The impact of color on bag-of-words based object recognition. In: ICPR, pp. 1549–1553 (2010)
50. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. arXiv preprint, arXiv:1409.0575 (2014)
51. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: theory and practice. Int. J. Comput. Vis. **105**(3), 222–245 (2013)
52. van de Sande, K., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. PAMI **32**(9), 1582–1596 (2010)
53. Shabani, A.H., Zelek, J., Clausi, D.: Multiple scale-specific representations for improved human action recognition. PRL **34**(15), 1771–1779 (2013)
54. Shapovalova, N., Gong, W., Pedersoli, M., Roca, FX., Gonzalez, J.: On importance of interactions and context in human action recognition. In: IbPRIA, pp. 58–66 (2011)
55. Sharma, G., Jurie, F.: Learning discriminative spatial representation for image classification. In: BMVC (2011)
56. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: CVPR, pp. 3506–3513 (2012)
57. Sharma, G., Jurie, F., Schmid, C.: Expanded parts model for human attribute and action recognition in still images. In: CVPR, pp. 652–659 (2013)
58. Sharma, G., Jurie, F., Schmid, C.: Expanded parts model for semantic description of humans in still images. arXiv preprint, arXiv:1509.04186 (2015)
59. Sicre, R., Jurie, F.: Discriminative part model for visual recognition. CVIU **141**, 28–37 (2015)
60. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
61. Vedaldi, A., Fulkerson, B.: Vlfeat: an open and portable library of computer vision algorithms. In: ACM MM, pp. 1469–1472 (2010)
62. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV, pp. 606–613 (2009)
63. Wei, XS., Gao, BB., Wu, J.: Deep spatial pyramid ensemble for cultural event recognition. In: ICCV Workshop (2015)
64. van de Weijer, J., Schmid, C., Verbeek, J.J., Larlus, D.: Learning color names for real-world applications. TIP **18**(7), 1512–1524 (2009)
65. Witkin, A.: Scale-space filtering: a new approach to multi-scale description. In: ICASSP (1984)
66. Yang, S., Ramanan, D.: Multi-scale recognition with dag-cnns. In: ICCV (2015)
67. Yao, B., Jiang, X., Khosla, A., Lin, AL., Guibas, LJ., Li ,FF.: Human action recognition by learning bases of action attributes and parts. In: ICCV, pp. 1331–1338 (2011)
68. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: pose aligned networks for deep attribute modeling. In: CVPR, pp. 1637–1644 (2014)
69. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS, pp. 487–495 (2014)
70. Zhou, X., Yu, K., Zhang, T., Huang, T.: Image classification using super-vector coding of local image descriptors. In: ECCV, pp. 141–154 (2010)
71. Zhu, X., Li, M., Li, X., Yang, Z., Tsien, J.: Robust action recognition using multi-scale spatial-temporal concatenations of local features as natural action structures. PLoS One **7**(10), e46686 (2012)
72. Ziaeefard, M., Bergevin, R.: Semantic human activity recognition: a literature review. PR **48**(8), 2329–2345 (2015)

Fahad Shahbaz Khan is a research fellow at Computer Vision Laboratory, Linköping University, Sweden. He received the M.Sc. degree in Intelligent Systems Design from Chalmers University of Technology, Sweden, and a Ph.D. degree in Computer Vision from Autonomous University of Barcelona, Spain. From 2012 to 2014, he was postdoctoral fellow at Computer Vision Laboratory, Linköping University, Sweden. His research interests include a wide range of topics within computer vision: object detection, tracking, and recognition, semantic description of humans, and vision for autonomous systems. He has published articles in high-impact computer vision journals and conferences in these areas.

Joost van de Weijer is a senior scientist at the Computer Vision Center Barcelona. He leads the Learning and Machine Perception team. He received a Ph.D. degree in 2005 from the University of Amsterdam. From 2005 to 2007, he was a Marie Curie Intra-European Fellow in the LEAR Team, INRIA Rhone-Alpes, France. From 2008 to 2012, he was a Ramon y Cajal Fellow at the Universidad Autonoma de Barcelona. His main research is on the usage of color information in computer vision application, and machine learning for computer vision.

Rao Muhammad Anwer received the masters degree in Intelligent Systems Design from the Chalmers University of Technology, Sweden, and the Ph.D. degree in Computer Vision from the Autonomous University of Barcelona, Spain. He is a Postdoctoral Research Fellow with the Department of Information and Computer Science, Aalto University School of Science, Finland. His research interests are in object detection, pedestrian detection, and action recognition.

Andrew D. Bagdanov is currently Associate Professor at the University of Florence, Italy. He received his Ph.D. in Computer Science in 2004 from the University of Amsterdam, after which he held a postdoctoral position at the University of Florence. Dr. Bagdanov held a senior development position at the FAO of the United Nations in Rome where he worked on developing large-scale, multilingual ontologies for cross-language retrieval of agricultural information in over twenty languages. In 2014 he became Senior Researcher and Ramón y Cajal Fellow at the Computer Vision Center, Barcelona, and in December 2015 began his tenure as Associate Professor at the University of Florence. His research spans a broad spectrum of computer vision, image processing and machine learning.

Michael Felsberg received the Ph.D. degree in engineering from the University of Kiel, Kiel, Germany, in 2002. Since 2008, he has been a Full Professor and the Head of the Computer Vision Laboratory, Linköping University, Linköping, Sweden. His current research interests include signal processing methods for image analysis, computer and robot vision, and machine learning. He has published more than 100 reviewed conference papers, journal articles, and book contributions. He was a recipient of awards from the German Pattern Recognition Society in 2000, 2004, and 2005, from the Swedish Society for Automated Image Analysis in 2007 and 2010, from Conference on Information Fusion in 2011 (Honorable Mention), and from the CVPR Workshop on Mobile Vision 2014. He has achieved top ranks on various challenges (VOT: 3rd 2013, 1st 2014, 2nd 2015; VOT-TIR: 1st 2015; OpenCV Tracking: 1st 2015; KITTI Stereo Odometry: 1st 2015, March). He has coordinated the EU projects COSPAL and DIPLECS, he is an Associate Editor of the Journal of Mathematical Imaging and Vision, Journal of Image and Vision Computing, Journal of Real-Time Image Processing, Frontiers in Robotics and AI. He was Publication Chair of the International Conference on Pattern Recognition 2014 and Track Chair 2016, he was the General Co-Chair of the DAGM symposium in 2011, and he will be general Chair of CAIP 2017.

Jorma Laaksonen received his Dr. of Science in Technology degree in 1997 from Helsinki University of Technology, Finland, and is presently a permanent senior university lecturer at the Department of Computer Science, Aalto University School of Science. He is an author of 30 journal and 150 conference papers on pattern recognition, statistical classification, machine learning and neural networks. His research interests are in content-based multimodal information retrieval and computer vision. Dr. Laaksonen has been an Associate Editor of Pattern Recognition Letters, IEEE senior member, and a founding member of the SOM and LVQ Programming Teams and the PicSOM Development Group.