Semantic Pyramids for Gender and Action Recognition

Fahad Shahbaz Khan, Joost van de Weijer, Rao Muhammad Anwer, Michael Felsberg, Carlo Gatta

Abstract—Person description is a challenging problem in computer vision. We investigate two major aspects of person description: gender and action recognition in still images. Most state-of-the-art approaches for gender and action recognition rely on the description of a single body part such as face or full-body. However, relying on a single body part is sub-optimal due to significant variations in scale, viewpoint and pose in real-world images.

This paper proposes a semantic pyramid approach for pose normalization. Our approach is fully automatic and based on combining information from full-body, upper-body and face regions for gender and action recognition in still images. The proposed approach does not require any annotations for upper-body and face of a person. Instead, we rely on pre-trained state-of-the-art upper-body and face detectors to automatically extract semantic information of a person. Given multiple bounding boxes from each body part detector, we then propose a simple method to select the best candidate bounding box which is used for feature extraction. Finally, the extracted features from the full-body, upper-body and face regions are combined into a single representation for classification.

To validate the proposed approach for gender recognition, experiments are performed on three large datasets namely: Human attribute, Head-Shoulder and Proxemics. For action recognition, we perform experiments on four datasets most used for benchmarking action recognition in still images: Sports, Willow, PASCAL VOC 2010 and Stanford-40. Our experiments clearly demonstrate that the proposed approach, despite its simplicity, outperforms state-of-the-art methods for gender and action recognition.

Index Terms—Gender Recognition, Action Recognition, Pyramid Representation, Bag-of-words

1 Introduction

Describing persons in images is one of the fundamental semantic problems in image analysis with many applications such as video surveillance, health care, image and video search engines and human-computer interaction etc. The problem is challenging since persons can appear in different poses and viewpoints in real-world scenarios, images can contain back-facing people, have low resolution, and can be taken under illumination and scale variations. In this paper, we focus on two challenging aspects of person description: gender and action recognition in still images.

In recent years, significant amount of work has been devoted to detect persons [7], [40], [16], [49] in real-world images. The part-based method of Felzenswalb et al. [16] is currently the state-of-the-art method for person detection [14]. The method works by modeling a person as a collection of parts, where each part is represented by a number of histograms of gradient orientations [7] over a number of cells. Other than full-body detection methods, several approaches exist in literature [48], [18], [60], [8] to detect the upper-body and face regions of a person. Combining different body part detectors for efficient person description is an open research problem. Here, we investigate the problem of combining semantic information from different body part detectors for gender and action recognition.

To solve the problem of gender recognition, most of the existing approaches [1], [52], [51], [37] rely only on face classification methods. These methods are generally applied on standard databases having high resolution aligned frontal faces. However, persons can appear in different scales and viewpoints in real-world images. In many cases gender recognition solely based on face cues could fail, and cues from clothes and hairstyle would be needed. The top row in Figure 1 shows exemplar images with persons from the different gender datasets used in this paper; they contain back-facing people, low resolution faces, and different clothing types. In this work, we aim at combining different body part detectors for robust gender classification.

In recent years, action recognition in static images has gained a lot of attention [43], [39], [10], [57], [36]. In action recognition, bounding boxes of humans performing actions are provided both at training and test time. The bottom row in Figure 1 shows exemplar images from different action recognition datasets used in this work. Most successful approaches to action recognition employ the bag-of-words (BOW) method popular in object and scene recognition [43], [9], [27]. The technique works by extracting local features such as color, shape and texture etc. on a dense grid of rectangular windows. These local features are then vector quantized into a fixed-size visual vocabulary. A histogram is constructed by

Fahad Shahbaz Khan and Michael Felsberg are with computer vision laboratory, dept. electrical engineering, Linköping University, Sweden.

Joost van de Weijer and Carlo Gatta are with Computer Vision Centre Barcelona, Spain.

Rao Muhammad Anwer is with Department of Information and Computer Science, Aalto University School of Science, Finland.

Fig. 1: Example images from the gender and action recognition datasets used in this work. Top row: images from the three gender recognition datasets. Factors such as back-facing people, scale and pose changes make it extremely difficult to rely on a single body part. Bottom row: example images from the different action recognition dataset. Conventional methods construct representations over the bounding box of a person. Combining semantic information from different body parts to improve the performance is still an open research problem.

counting the occurrence of each visual word in an image. Incorporating the part-based information within the bag-of-words framework is an active research problem [44], [26]. This paper investigates the problem of combining semantic information from different body parts within the bag-of-words approach for action recognition.

Both in gender recognition and in action recognition, the introduction of spatial information within the person bounding box has been primarily handled with spatial pyramids [44], [9], [27], [41], [6]. For deformable objects, like humans, spatial pyramids only provide a rough spatial description, because the pose of the object can vary significantly within the bounding box. To account for this, pose normalization has recently been proposed as a preprocessing step before performing feature extraction [3], [59]. Pose normalization would identify relevant parts, such as head and upper body, and subsequently describe these parts in the feature extraction phase. For gender recognition this has been studied by Bourdev et al. [3], who propose a computationally demanding part-based approach based on poselets. Other than existing work, we acknowledge that human part recognition is a much researched field, and tailored detectors exist for bodies [16], [49], [7], faces [48], [60], upper-bodies [12], [53], and even hands [38], [33]. In this paper we combine these existing detectors for pose normalization and use them to construct what we call *semantic pyramids*. To the best of our knowledge we are the first to investigate shape normalization based on existing part detectors.

Contributions: In this paper, we propose to combine different body part detectors for gender and action recognition. We combine information from full-body, upper-body and face regions of a person in an image. It is worth to mention that our approach does not require annotations for face and upper-body regions. Instead we use state-of-the-art upper-body and face detectors to automatically localize body parts in an image. Each detector fires at multiple locations in an image thereby providing multiple candidate bounding boxes. We propose a simple approach to select the best candidate bounding box from each body part detector. The selected bounding boxes are then used for feature extraction. For gender classification, we use a combination of visual descriptors. For action recognition, we employ the popular bag-of-words approach with spatial pyramid representation. Finally, the individual representations from the full-body, upper-body and face are combined into a single feature vector for classification.

We validate the proposed approach on three large datasets for gender classification namely: Human attribute, Head-Shoulder and Proxemics. For action recognition, we perform experiments on four benchmark datasets namely: Sports, Willow, PASCAL VOC 2010 and Stanford-40. For both gender and action recognition, our approach outperforms state-of-the-art methods in literature.

The paper is organized as follows. In Section 2 we discuss related work. In Section 3 we introduce our approach. The results on gender recognition are provided in Section 4. In Section 5 we provide a comprehensive evaluation of our approach for action recognition in still images. Section 6 finishes with a discussion and concluding remarks.

2 Related Work

Describing person attributes is an active research problem in computer vision. Several methods exist in literature [1], [52], [51], [37], [3], [35], [59] to tackle the problem of gender recognition. An evaluation of gender classification methods using automatically detected and aligned faces is performed by [1]. Interestingly, the evaluation shows that using automatic face alignment methods did not increase the gender classification performance. Wu et al. [11] propose an approach for gender classification by using facial shape information to construct discriminating models. The facial shapes are represented using

2.5D fields of facial surface normals. The work of [3] propose to use a part-based approach based on poselets for describing human attributes. Recently, Zhang et al. [59] propose two pose-normalized descriptors based on deformable part models for attribute description. In this paper, we also focus on the problem of gender recognition in the wild using semantic information from different body parts.

Other than gender recognition, describing actions associated with humans is a difficult problem in computer vision. In action recognition, given the bounding box of a person both at train and test time, the task is to classify the action label associated with each bounding box. Several approaches exist in literature [24], [39], [10], [57], [36], [23], [44] to solve the problem of action recognition. The bag-of-words based approaches [43], [27], [9] have shown to obtain promising results for action recognition task. Sharma et al. [43] propose an approach based on learning a max margin classifier to learn the discriminative spatial saliency of images. A comprehensive evaluation of color features and fusion approaches is performed [27].

Besides the bag-of-words framework, several methods [39], [10], [57], [36] have recently been proposed to find human-object interactions for action recognition. A human-centric approach is proposed by [39] that works by first localizing a human and then finding an object and its relationship to it. Maji et al. [36] introduced an approach based on poselet activation vector that captures the pose in a multi-scale fashion. The work of [10] propose a method based on spatial co-occurrences of objects and individual body parts. A discriminative learning procedure is introduced to solve the problem of the large number of possible interaction pairs.

Recently, several methods [44], [26] look into combining part-based information within the bag-of-words framework. The work of [44] is based on learning a model based on a collection of part templates learnt discriminatively to select scale-space locations in the images. Similarly, our work also investigates how to combine the semantic part-based information within the bag-of-words framework for improved action recognition.

In recent years, significant progress has been made in the field of human detection [7], [40], [16], [49]. The part-based approach by Felzenszwalb et al. [16] has shown to provide excellent performance. Besides full-body person detection, localizing specific parts of human body such as face, upper-body and hand also exist in literature [48], [18], [60], [8]. Combining these different body part detectors for human attribute description is an open problem in computer vision. In this paper, we propose an approach by combining semantic information from full-body, upper-body and face for gender and action recognition problems.

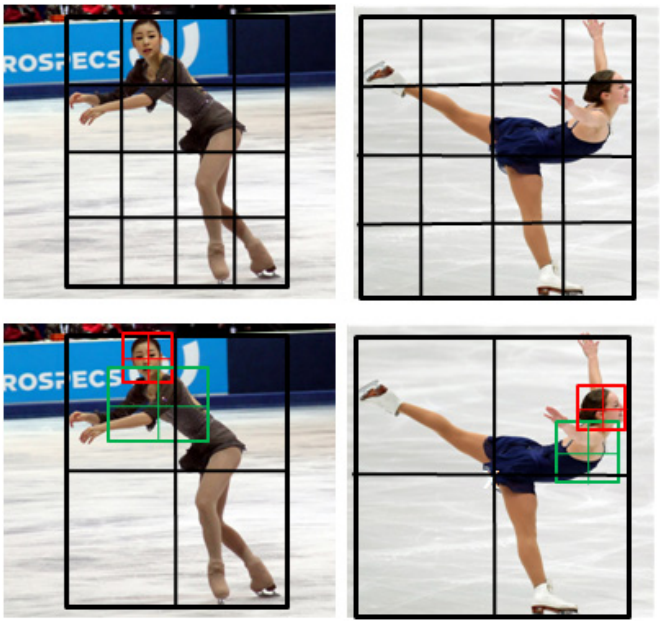

Fig. 2: Traditional spatial pyramid approach in top row where the person box is divided geometrically in various blocks. In the bottom row our approach of semantic pyramids which performs pose normalization by placing the representation on semantic parts (here we consider face detection and upper-body detection).

3 Semantic Pyramids for Person Description

In this section we outline our method of semantic pyramids for person description. We will apply the method to two cases of person description, namely gender recognition and action recognition.

Typically, both for gender recognition as well as for action recognition in still images, bounding boxes of persons are provided. This is done to decouple research in person detection from research in gender/action recognition. The task is then, given a bounding box of the person, to decide on the gender or/and the action of the person. Generally, this is approached by applying a spatial pyramid on the provided bounding box, similar as is done for image classification [32], [13] and object detection [7]. The pyramid encodes spatial information and allows the description of features dependent on their relative location in the bounding box. Following this strategy for several features (such as shape, texture, and color) was found to obtain competing results for gender classification [3], [35], [25] as well as action recognition [27].

The spatial pyramid allows to learn a rough spatial structure of the human outline. But because of the large variety of poses, i.e. people can e.g. be lying, sitting or being captured from the back, the discriminative power of such a representation remains inherently limited. This has recently been acknowledged by research in fine-grained object detection, where the task is to distinguish between hundreds of birds, flowers or airplane models

[15], [20]. The exact localization of semantic parts on these classes is considered an important step before going into the feature extraction phase [15], [59], see also Figure 2. The pre-localization of parts of the objects is also called the pose-normalization step. The method has also been applied to human attribute recognition in the poselet framework of Bourdez et al. [3].

Most existing methods to pose normalization are general and could be applied to a variety of objects [15], [20]. Here we take a different approach. Because of the importance of humans in many computer vision applications, the recognition of semantic human parts has been studied in great detail. There exists much work on person detection [16], [7], [49], face detection [48], [60], but also specialized work on upper-body detection [12], [53] and even hand detection [38], [33]. These detectors have been separately designed and optimized for their task. Therefore, in this paper, we focus on how to combine these tailored detectors for the task of gender and action recognition. Other than previous work we do not propose to relearn detectors for human parts [3], [59].

In the remainder of this section we outline our method of semantic pyramids for human attribute and action recognition. The main steps are indicated in Figure 3. First we run the part based detectors and obtain a set of hypotheses of possible part locations. Next we infer the most likely configuration given the person bounding box in the part selection step. In the next step, several features are extracted and represented in a histogram for all the elected part bounding boxes. Finally, a classifier is learned on the concatenated histograms. These steps are outlined in detail in the following subsections.

3.1 Human Parts Detection

In action recognition, the bounding box of a person is given both at training and test time. For gender recognition, the Human attribute dataset [3] has bounding box information given. The head-shoulder gender dataset [35] contains persons with only head-shoulder covering almost the entire image. Finally, the Proxemics dataset [54] also contains bounding box information of each person instance in an image. Therefore, in this work, we start with the assumption that the bounding box information of a person is available in prior. However, our approach can easily be extended for scenarios where no bounding box information is available in prior.

In order to automatically obtain the upper-body part of a person, we use a pre-trained upper-body detector.[1] The upper-body detector is based on the popular part-based object detection framework of Felzenswalb et al. [16]. In this work, we use a pre-trained model learned to detect near-frontal upper-body regions of a person. Given a cropped person image using the bounding box information, the upper-body detector returns bounding-boxes fitting the head and upper half of the torso of the person. In order to increase the robustness of the detector, the primary upper-body detections are regressed from the Viola and Jones face detector [47] to obtain secondary upper-body detections. The upper-body detection framework is successfully used for the problem of human pose estimation [12].

In order to automatically extract the face of a person, we use a pre-trained face detector [60] constructed on top of part-based implementation of Felzenswalb et al. [16]. The method employs a tree-structured model with a shared pool of parts where every facial landmark is modeled as a part. The method efficiently captures global elastic deformations. In this work, we use a pre-trained detector learned using the positive samples from MultiPIE dataset[2] and the negative instances from the INRIA Person dataset [7].

3.2 Part Selection

Each of the part detectors fires at multiple location within the bounding box of the person, yielding a set of hypotheses for all of the parts. These detections come together with a detection score indicating the confidence of the detector. A straightforward method would be to select the highest scoring detector for each part. However, due to the difficulty of the problem - low resolution, strange body pose, partial occlusion - detectors give many false detections. Here we consider a method, related to the pictorial structure method of Felzenswalb and Huttenlocher [17], to improve part selection.

Here we outline our approach to find the optimal configuration of parts given a bounding box. We consider n part detectors. For a given person bounding box we run the detectors and obtain a number p_i detections for part i. Each detection j of detector i consists out of a location $\mathbf{x}_i^j$ and a detector score c_i^j. A possible combination of parts is represented by a configuration $L = \{l_1, ..., l_n\}$, where $l_i \in \{1, ..., p^i\}$ represents the index to one of the p^i detections of part i (e.g. $L = \{2, 4\}$ means that there are two parts, where in this configuration for the first part the second detection is considered and for the second part the fourth detection).

We consider the part's locations to be independent of each other given the person bounding box. Finding the optimal configuration can then be defined as an energy minimization problem [19]. We model the costs of the configuration to be dependent on the mismatch, m_i, of the appearance of the part and the deformation cost, d_i, based on the location of each part to the person bounding box. The optimal configuration L^* is then computed with:

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i^{l_i} + \sum_{i=1}^{n} \lambda_i d_i^{l_i} \right) \quad (1)$$

where λ_i is a weight which balances the relative strength of appearance mismatch versus deformation cost. Since

1. The upper-body detector is available at: http://groups.inf.ed.ac.uk/calvin/calvin_upperbody_detector/

2. The dataset is available at: http://multipie.org/

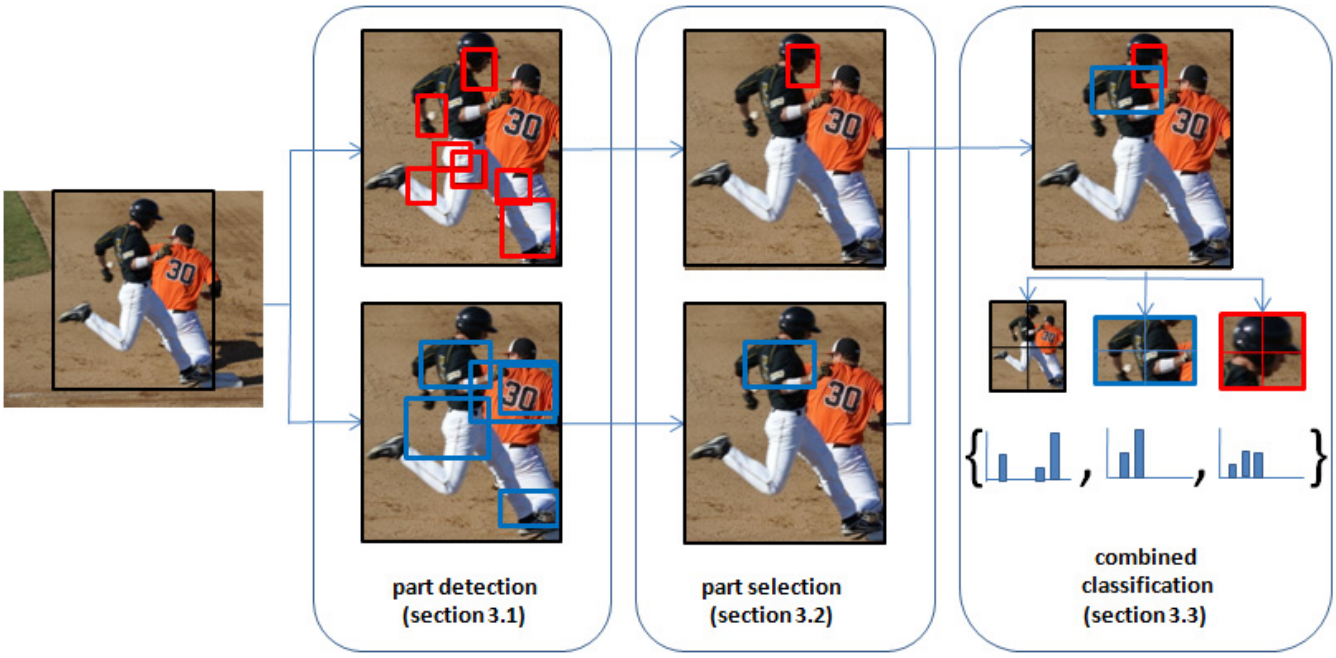

Fig. 3: Overview of our method. In the part detection stage the detectors for each of the parts are run and hypotheses for the locations are obtained. In the part selection stage a single optimal detection is selected for each part. In the last step, histogram representations for all parts are concatenated to form the final representation.

there are no dependencies between the part locations, Eq. 1 can be solved for each part separately:

$$l_i^* = \arg\min_{l_i} m_i^{l_i} + \lambda_i d_i^{l_i} \tag{2}$$

Given m_i, d_i and λ_i this is therefore straightforward to solve: the equation can be evaluated for all of the locations l_i where a part i was detected, after which the best location can be established. As the appearance mismatch cost we use minus the detector score $m_i^{l_i} = -c_i^{l_i}$. In the remainder we outline our method to compute d_i and λ_i.

First consider the case where we have a set of training images with ground truth bounding boxes for the parts. We transform all coordinates to relative coordinates by subtracting the upper left coordinate of the person bounding box, and dividing by the width and respectively height of the bounding box (we indicate relative coordinates with $\hat{.}$). Based on the ground truth bounding boxes we compute the mean location $\hat{\boldsymbol{\mu}}_i$ and its standard deviation $\hat{\boldsymbol{\sigma}}_i$ of part i. Now we can define the deformation cost of detection l_i of part i to be the normalized Euclidean distance, as given by:

$$d_i^{l_i} = \left(\hat{\mathbf{x}}_i^{l_i} - \hat{\boldsymbol{\mu}}_i\right)^{\mathrm{T}} M_i^{-1} \left(\hat{\mathbf{x}}_i^{l_i} - \hat{\boldsymbol{\mu}}_i\right) \tag{3}$$

where the matrix $M_i = diag\left(\hat{\boldsymbol{\sigma}}_i\right)$ contains the standard deviation of the coordinate. Hence, the deformation cost rises with the distance of the detection to the mean location of the part, and rises faster for parts which are well localized in space (i.e. which have low $\boldsymbol{\sigma}$).

Above we considered the case where we have a training set with ground truth bounding boxes for the parts. In all the data sets we consider in this paper such ground truth is not present, and we need an automatic method to select good detections for each of the parts. We followed a simple procedure which was found efficient for the data sets we considered. For each bounding box in the training set we only consider the best detection, of these we take the 50% of detections with the highest classifier score. Based on this selection a first estimate of both $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ is made. We iteratively improve the estimates by selecting the 50% closest (in normalized Euclidean sense) and recompute $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ until convergence. This reduces the influence of outliers on the estimates. Finally, to determine the weights λ_i we use the following heuristic: the λ's are chosen in such a way that the average distance between the best and second best detection is equal for both deformation costs (times lambda) and classifier score. By doing so, we have chosen the influence of both cues to be equal. In Figure 4 several examples of the selection parts are shown. One can see that incorporating the deformation cost improves results. In the experiment section, we will compare our approach to the baseline method which ignores the deformation costs, and just picks the highest scoring detection.

3.3 Combining Detector Outputs

We combine the bounding boxes of upper-body and face selected using the approach proposed in section 3.2 with the conventional full-body person bounding box. For gender recognition, multiple features in a spatial pyramid representation are computed for each of the full-body, face and upper-body boxes. The three spatial pyramid representations are then concatenated into a single feature vector which is then input to the gender

Fig. 4: Results of detector output selection: in red the highest scoring detection and in blue the detection after taking into account the deformation cost. Top row: The method correctly identifies the head in these four examples using the boxes from the face detector. Bottom row: Our method accurately localizes the upper-body regions in these examples using the upper-body detector.

classifier.

In case of action recognition, We employ the same procedure by constructing bag-of-words based spatial pyramid representations each for full-body, face and upper-body boxes. The final image representation is then obtained by concatenation of the three feature vectors each coming from a different body part.

4 Gender Recognition

Here, we evaluate our approach for the problem of gender recognition. We start by introducing the datasets used in our experiments. Afterwards, we describe the details of our experimental setup followed by the features used in our evaluations. Finally, we provide the results of our experiments.

4.1 Datasets

We have evaluated our approach on three challenging gender recognition datasets namely: Human attribute, Head-Shoulder and Proxemics. These datasets pose the challenging problem of gender recognition "in the wild" since the images contain persons in different poses, viewpoints and scales.

The Human attribute [3] is the most challenging dataset[3] and comprises of 8035 images. The images are collected from the H3D [4] and PASCAL VOC 2010 [14] datasets. The dataset contains nine attributes where each has a label corresponding to absent, present and unspecified instances. We use the gender attribute from this dataset for our experiments.

The Head-Shoulder [35] dataset consists of 5510 images[4] of head-shoulder of men and women. This is the largest dataset for head-shoulder based gender recognition. The dataset contains 3856 training samples (2044 men and 1812 women) and 1654 test samples (877 men and 777 women).

Finally, we present results on the Proxemics dataset [54]. The Proxemics dataset[5] was recently introduced for the problem of recognizing proxemics in personal photos. We manually labeled the dataset with gender labels. The dataset consists of 1218 samples divided into 620 training and 598 test samples. The top row in Figure 1 shows example images from the three gender datasets.

4.2 Image Features

In this work, we use three visual features for image representation commonly used for gender recognition problem [35], [41], [6], [25].

CLBP[21]: Local binary patterns (LBP) is the most commonly used feature to extract texture information for image description. The LBP descriptor has shown to obtain state-of-the-art results for texture classification [29], [21], [22] and applied successfully for gender recognition task [41], [35]. In this paper, we employ the complete LBP (CLBP) approach [21] where a region in an image is represented by its center pixel and a local difference sign-magnitude transform. In our experiments, we compute the texture features at multiple pixel neighborhoods and radius spacings since it was shown to improve the performance.

PHOG[2]: To represent the shape information, we use the popular PHOG descriptor. The descriptor captures the local shape of an image along with its spatial layout and has been evaluated previously for gender recognition [6], [3]. In this work, we use 20 orientation bins in the range [0,360].

WLD[5]: The Weber Local Descriptor (WLD) has two components. The first component extract the gradient orientation of a pixel. The second component captures the ratio between the relative intensity differences of a pixel against its neighbors and the intensity of the current pixel. The WLD descriptor has shown to provide excellent results for texture classification. Based on its success for human face and gender recognition

3. Human attribute dataset is available at: http://www.cs.berkeley.edu/~lbourdev/poselets/

4. Head-Shoulder dataset is available at: http://limin81.cn/research.html/

5. Proxemics dataset is available at: http://www.ics.uci.edu/~dramanan/software/proxemics/

problems[5], [25], we also use this descriptor in our experiments.

4.3 Spatial Pyramid Image Representation

We use the conventional pyramid scheme by [32], which is a simple and computationally efficient method to capture the spatial information. The spatial pyramid scheme works by representing an image using multi-resolution histograms, which are obtained by repeatedly sub-dividing an image into increasingly finer sub-regions. The final image representation is a concatenation of the histograms of all the regions. The spatial pyramid representation has shown to provide excellent performance for object and action recognition [27], [31].

All three features mentioned above are computed in a spatial pyramid manner. In this work, we use a spatial pyramid representation with three levels, yielding a total of 14 sub-regions. Combining multiple visual features has shown to provide improved performance [35], [25]. For each body part, we also combine the spatial pyramid representations of all three features into a single representation which is then input to the classifier. In our experiments, we will show both the importance of multiple features and spatial pyramid representation for gender recognition.

4.4 Experimental Setup

We follow the same evaluation protocol as proposed with the respective datasets. For Human-attribute dataset, the performance is represented as average precision under the precision-recall curve. The results for The Head-Shoulder dataset are represented in terms of classification accuracies. We run separate classifiers for men and women. The final classification result is obtained as a mean recognition rate over the two categories. For the Proxemics dataset, we follow the same criteria used for Human-attribute dataset by measuring average precision under the precision-recall curve. The final performance is calculated by taking the mean average precision over the two categories. For classification, we use Support Vector Machines (SVM) with a χ^2 kernel [58].

4.5 Experimental Results

We start by evaluating the contribution of each Visual feature for gender classification. In the next experiment, we show the importance of spatial pyramid representation for improving the recognition performance. Afterwards, we demonstrate the effectiveness of our semantic pyramid representation together with a comparison with state-of-the-art approaches.

4.5.1 Experiment 1: Combining Visual Cues

In the first experiment, we evaluate to what extent combining multiple visual cues improve the performance of gender recognition. The experiments are performed on the full-body of the persons without the spatial pyramid representation.

Dataset	CLBP[21]	PHOG[2]	WLD[5]	Combine
Human-attribute	69.6	68.4	67.7	**73.4**
Head-Shoulder	76.0	81.0	70.5	**85.5**
Proxemics	66.2	65.2	64.7	**70.6**

TABLE 1: Comparison of different visual cues and their combination on the three gender datasets. For the Human-attribute and Proxemics dataset, the result are shown in average precision (AP). For the Head-Shoulder dataset, the performance is shown in terms of recognition rate. For all datasets, the best results are obtained by combining the three visual cues.

Dataset	Level 1	level 2	Level 3
Human-attribute	73.4	76.1	**77.7**
Head-Shoulder	85.5	88.0	**89.5**
Proxemics	70.6	72.4	**73.6**

TABLE 2: Evaluation of spatial pyramid representation on the three gender datasets. Level 1 corresponds to standard image representation with no spatial pyramids. On all three datasets, a significant gain in performance is obtained by using the spatial pyramid representation.

Table 1 shows the results of combining visual cues on the three gender datasets. On the Human-attribute dataset, the single best feature (CLBP) provides an average precision score (AP) of 69.6. The results are significantly improved on all datasets when the three visual cues are combined. The performance is improved by 3.8% when combining multiple visual cues. Similarly, on the Head-Shoulder and Proxemics datasets a gain of 4.5% and 4.4% is obtained when using a combination of visual cues compared to the single best feature.

The results clearly suggest that combining multiple visual cues always provide better performance compared to using a single visual feature for gender classification. This further shows that the visual features used in this paper contain complementary information and should be combined for improve gender classification.

4.5.2 Experiment 2: Spatial Pyramid Representation

Here, we evaluate the impact of using the spatial pyramid representation for gender classification. The experiments are performed on the full-body of the persons using a combination of multiple visual cues. In this work, we use a 3 level spatial pyramid: level 1 corresponds to standard image-wide representation, level 2 contains the 2x2 division of the image and level 3 comprises of 9 sub-window as a results of 3x3 division. For each higher level, the histogram representations of previous levels are concatenated (i.e. level 3 is obtained as a result of histogram concatenation of 14 sub-windows).

Table 2 shows the results obtained by using different level of pyramids on the three gender datasets. A significant improvement in performance is obtained by using spatial pyramid representation on all three datasets. A

gain of 4.3%, 4.0% and 3.0% is obtained on Human-attribute, Head-Shoulder and Proxemics datasets respectively by using the spatial pyramid scheme compared to the standard image representation.

4.5.3 Experiment 3: Semantic Pyramid Representation

We provide the results of our semantic pyramid representation for gender classification. We combine the full-body, upper-body and face pyramid histograms into a single image representation. As a baseline, we use two approaches. In the first approach, called Horizontal pyramids, we divide the full-body of a person into three horizontal regions. This provides a rough estimation of body parts without any alignment. A spatial pyramid histogram is then computed for each horizontal region. The three horizontal pyramid histograms are then concatenated into a single representation for classification. As a second baseline, called Maximum pyramid scheme, we directly use the bounding box with maximum confidence from the detector output directly (and ignore the deformation cost).

Table 3 shows the results using semantic pyramid representations on the three gender datasets. Using only the full-body based representation (FB) provides an average precision (AP) of 77.7% and 73.6% respectively on Human-attribute and Proxemics datasets. The baseline, Horizontal Pyramids, provide inferior results which clearly suggest that alignment of different body parts is crucial to obtain robust semantic information. Using the bounding box selection based on the detector confidence, Maximum pyramids, provides improved performance for both face (FB) and upper-body (UP) parts. Our detector output selection method, described in Section 3.2, obtains the best results. Finally, a significant gain in classification performance is obtained by combining the semantic pyramids of full-body (FB), face (FC) and upper-body (UP) based representations. This clearly shows that a single body part is sub-optimal and different body parts should be combined for robust gender recognition.

State-of-the-art Comparison: we compare our semantic pyramids with state-of-the-art approaches in literature. Table 4 shows a comparison of state-of-the-art methods with our semantic pyramids based approach on the Human-attribute and Head-Shoulder datasets. On the Human-attribute dataset, Cognitec which is one of the leading face and gender recognizing tool provides an average precision (AP) of 75.0%. The poselet based approach [3] obtains a classification score of 82.4%. Our semantic pyramids approach, despite its simplicity, outperforms the existing methods by obtaining a score of 84.8% on this dataset. On the Head-Shoulder dataset, combining LBP, HOG and gabor features (MC) provides a classification accuracy of 87.8%. Combining multiple cues with PLS based classification approach [35] provides a recognition score of 88.6%. Our approach outperforms state-of-the-art methods by obtaining a classification accuracy of 92.0% on this dataset.

Dataset	Cognitec [3]	Poselet [3]	MC [35]	MC-PLS [35]	Ours
Human-attribute	75.0	82.4	-	-	**84.8**
Head-Shoulder	-	-	87.8	88.6	**92.0**

TABLE 4: Comparison of our approach with state-of-the-art methods in literature. On the Human-attribute dataset, our approach outperforms both professional face-based software Cognitec and the Poselet methods. Similarly, on the Head-Shoulder dataset, our method outperform the previous best results obtained using multiple cues and PLS based classification.

5 Action Recognition

We now evaluate our approach for the task of action recognition. In action recognition the bounding boxes of humans performing actions are provided both at training and test time. The task is to classify the action associated with each bounding box. In this work, we follow the successful bag-of-words (BOW) framework which has shown to provide state-of-the-art performance in literature [43], [27], [9]. Conventional pyramid representation is used with these BOW based approaches for action recognition. Similar to [27], we follow the BOW approach with multiple visual cues and compare our semantic pyramids with conventional pyramids for action recognition.

5.1 Datasets

We evaluate our approach on four challenging benchmark action recognition datasets: Sports, Willow, PASCAL VOC 2010 and Stanford-40. The Sports action dataset consists of 6 action categories of humans doing different sports. The action categories in this dataset are: cricket batting, cricket bowling, croquet, tennis forehand, tennis backhand and volleyball smash.[6] Each of the action category in this dataset consists of 30 training images and 20 test images. The Willow dataset consists of 7 action categories: interacting with computer, photographing, playing music, riding bike, riding horse, running and walking.[7] The second dataset which we use is the PASCAL VOC 2010 dataset, which comprises of 9 action categories: phoning, playing instrument, reading, riding bike, riding horse, running, taking photo, using computer and walking.[8] Lastly, we also validate our approach on the challenging Stanford-40 dataset, which is the largest and most challenging action recognition datasets currently available.[9] This dataset contains out of 9532 images of 40 different action categories such as jumping, repairing a car, cooking, applauding, brushing teeth, cutting vegetables, throwing a frisbee, etc.

6. The Sports dataset is available at: http://www.cs.cmu.edu/~abhinavg/Downloads.html/

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

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

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

Dataset	FB	Horizontal Pyramids				Maximum Pyramids			**Semantic Pyramids**		
		H1	H2	H3	H1+H2+H3	FC	UP	FB+FC+UP	**FC**	**UP**	**FB+FC+UP**
Human-attribute	77.7	75.9	71.5	67.7	77.1	79.0	77.1	82.4	81.2	79.4	**84.8**
Head-Shoulder	89.5	83.5	84.5	81.0	85.5	87.0	87.5	90.5	88.5	89.0	**92.0**
Proxemics	73.6	65.4	68.8	61.7	71.3	67.1	76.3	77.9	69.9	78.1	**80.5**

TABLE 3: Classification performance of different methods using full-body (FB), face (FC) and upper-body (UP) representations. For all representations, we use the same feature set. We compare our semantic pyramids with FB, Horizontal and Maximum score based pyramid methods. Our semantic pyramids outperforms other methods on all three datasets. Furthermore, the best results are obtained by combining our semantic FB, FC and UP pyramids.

Dataset	SIFT		CN		Early Fusion		Late Fusion		C-SIFT		OPP-SIFT	
	SP	SM-SP	SP	SM-SP	SP	SM-SP	SP	SM-SP	SP	SM-SP	SP	SM-SP
Actions	83.3	**85.8**	70.1	**72.3**	85.8	**87.5**	87.4	**90.0**	89.1	**90.0**	88.2	**89.3**
Willow	64.9	**67.3**	44.7	**47.6**	66.6	**67.3**	68.1	**69.2**	62.6	**64.3**	62.9	**64.0**
PASCAL VOC 2010	54.1	**56.5**	34.4	**36.6**	53.0	**55.1**	56.9	**58.5**	52.7	**53.9**	49.8	**51.7**
Stanford-40	40.6	**44.2**	17.6	**18.9**	39.4	**43.2**	41.7	**44.4**	37.6	**41.9**	35.3	**41.6**

TABLE 5: Comparison of our semantic pyramids with conventional spatial pyramid approach on the four action recognition dataset. We evaluate our approach on a variety of visual features. In all cases, our approach outperforms the conventional spatial pyramid method on all four datasets.

5.2 Experimental Setup

As mentioned earlier, we use the popular bag-of-words (BOW) framework with multiple visual cues. For feature detection, we use the dense sampling strategy at multiple scales. To extract the shape features, we use the SIFT descriptor, commonly used for shape description in BOW models. For color feature extraction, we use the color names [46] descriptor which has shown to provide excellent results for action recognition [27], object detection [28] and texture classification [29]. We use a visual vocabulary of 1000 and 500 words for SIFT and color names respectively. Due to the large size of Stanford-40 dataset, we use a visual vocabulary of 4000 words.

To combine color names and SIFT, we use early and late fusion approaches. In case of early fusion, a joint visual vocabulary of color-shape words is constructed. This results in a joint color-shape histogram representation. In early fusion, separate visual vocabularies are constructed for both color names and SIFT. Afterwards, the two histograms are concatenated into a single image representation. For early fusion, we use a larger vocabulary of 1500 visual words. Besides early and late fusion, we also use the colorSIFT descriptors by [45]. Similar to early fusion, we also use a visual vocabulary of 1500 words for colorSIFT descriptors. The image representations are then input to a nonlinear SVM with a χ^2 kernel [58] classifier. The performance is measured as mean average precision under the precision-recall curve over all action categories.

For each representation, we use the conventional spatial pyramid of three levels (1×1, 2×2, and 3×3), yielding a total of 14 regions [32] over the bounding box of a person. For our semantic pyramid representation, we use the same three level pyramids for the bounding box of a person (full-body), face and upper-body region respectively. The spatial pyramids from the three body parts are concatenated into a single image representation.

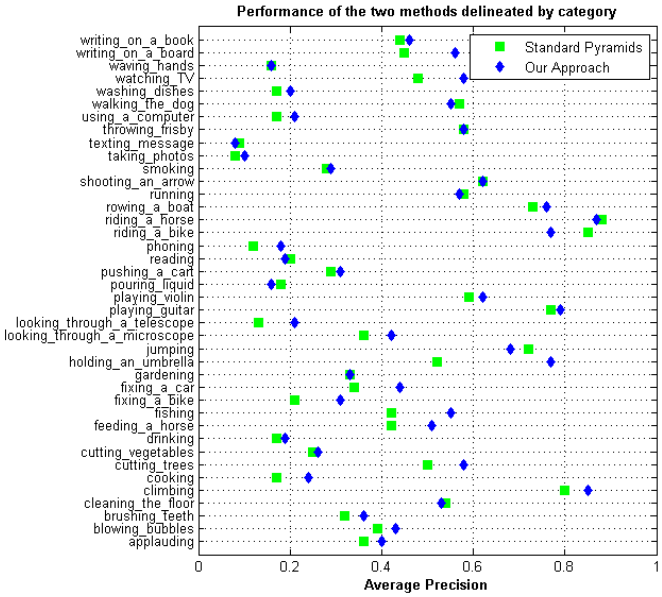

Fig. 5: Per-category performance comparison of our approach compared to the conventional pyramid method on the Stanford-40 action recognition dataset. Note that our approach improves the performance on 25 out of 40 action categories on this dataset.

tion.

5.3 Experimental Results

We first compare our semantic pyramid approach with the conventional pyramids commonly employed in action recognition frameworks. Afterwards, we provide a comparison with state-of-the-art approaches on the three action recognition datasets.

Method	HMI [39]	SFC [24]	MMC [56]	HOI[57]	Ours
Accuracy	83.0	79.0	83.0	87.0	**92.5**

TABLE 6: Comparison of our semantic pyramids method with state-of-the-art results on Sports action dataset. On this dataset, our approach outperforms the best reported results in the literature.

5.3.1 Conventional Pyramids vs Semantic Pyramids

Here, we compare our semantic pyramid representation with the conventional spatial pyramid scheme used in the vast majority of action recognition frameworks. The conventional scheme is based on constructing spatial pyramids on the bounding box (full-body) of the person. Similar to gender recognition, our semantic representation constructs the spatial pyramids on the full-body, upper-body and face bounding boxes which are later concatenated into a single image representation.

Table 5 shows the comparison of our semantic pyramids (SM-SP) with conventional spatial pyramid scheme (SP) for color, shape and color-shape features. On the Action dataset, our approach improves the performance from 83.3% to 85.8% for shape alone. On the Willow dataset the spatial pyramids obtain a mean AP of 64.9% when using shape alone. Our approach improves the performance by 2.4% mean AP on this dataset. Similarly, on the PASCAL VOC 2010 validation set, our method provides a mean AP 56.5% compared to 54.1% obtained using the conventional spatial pyramids. On the challenging Stanford-40 dataset, we obtain a significant gain of 3.6% mean AP using shape alone. Finally, in all cases, our approach improves the performance compared to the conventional method.

Figure 5 shows a per-category performance comparison of conventional spatial pyramids with our semantic pyramids approach on the challenging Stanford-40 dataset. Our approach improves the performance on 25 out of 40 action categories on this dataset. Especially relevant performance gains are obtained for holding-an-umbrella (+25%), fishing (+13%), writing-on-a-board (+11%), fixing-a-car (+10%), and watching-tv (+10%) compared to conventional spatial pyramid approach.

5.3.2 Comparison with State-of-the-art

We compare our approach with state-of-the-art methods in literature. To this end, we combine all the feature representations based on semantic pyramids by adding the individual classifier outputs. Table 6 shows a state-of-the-art comparison on the Sports dataset. Our approach achieves a recognition accuracy of 92.5%, which is the best result reported on this dataset [39], [57], [56], [24]. The work of [39] obtains a recognition rate of 83.0% by modeling interactions between humans and objects. The approach of [57] achieves an action recognition accuracy of 87.0% using a mutual context model that jointly models objects and human poses. Our approach, on this dataset, provides a gain of 5.5% compared to the second best method.

Table 7 shows a state-of-the-art comparison on the Willow dataset. Our approach provides the best results on 4 out of 7 action categories on this dataset. On this dataset, we achieve a mean AP of 72.1%, which is the best result reported on this dataset [9], [10], [43], [27], [44]. The work of [27] obtains a mean AP of 70.1% by combining color fusion methods for action recognition. Recently, Sharma et al. [44] report a mean AP of 67.6% by learning part-based representations with the bag-of-words based framework. Our approach provides a gain of 2.0% mean AP on this dataset compared to the second best method.

Table 8 shows a state-of-the-art comparison on the PASCAL VOC 2010 test set. The method of [36] based on poselets activation vectors obtain a mean Ap 59.7%. The work of [39] employing a human-centric approach to localize humans and and object-human relationships achieve a mean AP 62.0%. The color fusion method Khan et al. [27] obtains a mean AP 62.4%. Our approach provides the best performance on two categories on this dataset. The best results [55] on this dataset is achieved by learning a sparse basis of attributes and parts. It is worthy to mention that our approach is complementary to this method [55] since semantic pyramids is not designed to capture the human-object interactions explicitly. It would be interesting to combine the two approaches in order to obtain further gain in performance.

Finally, Table 9 shows the results on the most challenging Stanford-40 action recognition dataset. Sharma et al. [44] obtains a mean AP 42.2% using an expanded part model (EPM) approach based on learning a discriminative collection of part templates. The sparse basis (SB) approach [55] based on using attributes and parts, where attributes represent human actions and parts are model objects and poselets. The color fusion (CF) method by Khan et al. [27] achieves a mean AP 51.9. On this dataset, our approach provides a mean AP 53.0 outperforming existing results [34], [50], [55], [27], [44] on this dataset.

6 Discussion

In this paper, we have proposed a semantic pyramid approach for pose normalization evaluated on two tasks, namely gender and action recognition. Our approach combines information from the full-body, upper-body and face regions of a person in an image. State-of-the-art upper-body and face detectors are used to automatically localize respective body parts of a person. Each body part detector provides with multiple bounding boxes by firing at multiple locations in an image. We then proposed a simple approach to select the best candidate bounding box for each body part. Image representation based on spatial pyramids is then constructed for each body part. The final representation is obtained by concatenating the full-body, upper-body and face pyramids for each instance of a person.

Our approach for gender recognition is evaluated on three challenging datasets: Human-attribute, Head-Shoulder and Proxemics. We show that relying on single

	int. computer	photographing	playingmusic	ridingbike	ridinghorse	running	walking	mean AP
Delaitre et al.[9]	58.2	35.4	73.2	82.4	69.6	44.5	54.2	59.6
Delaitre et al.[10]	56.6	37.5	72.0	90.4	75.0	59.7	57.6	64.1
Sharma et al.[43]	59.7	42.6	74.6	87.8	84.2	56.1	56.5	65.9
Khan et al.[27]	61.9	**48.2**	76.5	90.3	84.3	64.7	**64.6**	70.1
Sharma et al.[44]	64.5	40.9	75.0	91.0	87.6	55.0	59.2	67.6
Khan et al.[30]	**67.2**	43.9	76.1	87.2	77.2	63.7	60.6	68.0
Our approach	66.8	48.0	**77.5**	**93.8**	**87.9**	**67.2**	63.3	**72.1**

TABLE 7: Comparison of our semantic pyramids approach with state-of-the-art results on the Willow dataset. Our approach provides best results on 4 out of 7 action categories on this dataset. Moreover, we achieve a gain of 2.0 mean AP over the best reported results.

	phoning	playingmusic	reading	ridingbike	ridinghorse	running	takingphoto	usingcomputer	walking	mean AP
Maji et al.[36]	49.6	43.2	27.7	**83.7**	89.4	85.6	31.0	59.1	67.9	59.7
Shapovalova et al.[42]	45.5	54.5	31.7	75.2	88.1	76.9	32.9	64.1	62.0	59.0
Delaitre et al.[10]	48.6	53.1	28.6	80.1	90.7	85.8	33.5	56.1	69.6	60.7
Yao et al.[55]	42.8	60.8	41.5	80.2	90.6	87.8	**41.4**	**66.1**	**74.4**	**65.1**
Prest et al.[39]	**55.0**	**81.0**	**69.0**	71.0	90.0	59.0	36.0	50.0	44.0	62.0
Khan et al.[27]	52.1	52.0	34.1	81.5	90.3	88.1	37.3	59.9	66.5	62.4
Our approach	52.2	55.3	35.4	81.4	**91.2**	**89.3**	38.6	59.6	68.7	63.5

TABLE 8: Comparison of our semantic pyramids approach with state-of-the-art methods on the PASCAL VOC 2010 test set. Our approach, despite its simplicity, achieves the best performance on two categories while providing competitive performance compared to state-of-the-art methods.

Method	OB [34]	LLC [50]	SB [55]	CF[27]	EPM[44]	Ours
mAP	32.5	35.2	45.7	51.9	42.2	**53.0**

TABLE 9: Comparison of our semantic pyramids method with state-of-the-art results on Stanford-40 dataset. On this dataset, our approach outperforms the best reported results in the literature.

body part for recognizing gender is sub-optimal especially in real-world datasets where the images contain back-facing people, low resolution faces, different clothing types and body proportions. This is validated by our results, which found that faces provide the best performance on Human-attribute dataset, full-bodies are the best choice for the Head-Shoulder dataset and the upper-body region is the best for the Proxemics dataset. Our approach, that combines the semantic information from these three body parts, provides significant improvement on all three challenging real-world gender datasets. The results clearly demonstrate the effectiveness of combining different semantic regions obtained using a detector output selection strategy. This is further validated by our results on the Human-attribute dataset where our approach significantly outperforms a leading professional face and gender recognizing software, Cognitec, which uses careful alignment and advanced proprietary biometric analysis.

We have also evaluated our semantic pyramid approach for the task of action recognition in still images. We validate the performance of our method on three challenging action recognition datasets: Willow, PASCAL VOC 2010nd Stanford-40. On all three datasets, our results clearly suggest that combining full-body, face and upper-body regions improves the performance compared to the conventional approaches relying on full-body only. Our semantic approach significantly outperforms the conventional pyramid based method on all three datasets, thereby showing the importance of pose normalization.

Most of the action categories such as phoning, taking photo, playing guitar, feeding horse etc. contain objects associated with the action in the upper region of the person. The explicit use of an upper-body detector can better capture these associated objects. This is especially demonstrated on the challenging Stanford-40 dataset, where our approach when using shape alone improves the performance on 25 out of 40 action categories compared to conventional scheme based on full-body only. The results clearly suggest that pose normalization by means of semantic pyramids improves action recognition; in most cases leading to state-of-the-art performance.

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, CADICS and The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170). The work of Carlo Gatta is supported by MICINN under a Ramon y Cajal Fellowship.

REFERENCES

[1] L. A. Alexandre. Gender recognition: A multiscale decision fusion approach. *PRL*, 31(11):1422–1427, 2010.
[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.
[3] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011.
[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.
[5] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao. Wld: A robust local image descriptor. *PAMI*, 32(9):1705–1720, 2010.

[6] M. Collins, J. Zhang, P. Miller, and H. Wang. Full body image feature representations for gender profiling. In *ICCV Workshop*, 2009.
[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
[8] M. Dantone, J. Gall, G. Fanelli, and L. J. V. Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012.
[9] 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.
[10] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011.
[11] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012.
[12] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *IJCV*, 99(2):190–214, 2012.
[13] N. Elfiky, F. S. Khan, J. van de Weijer, and J. Gonzalez. Discriminative compact pyramids for object and scene recognition. *PR*, 45(4):1627–1636, 2012.
[14] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.
[15] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, 2011.
[16] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.
[17] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.
[18] V. Ferrari, M. J. Marn-Jimnez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.
[19] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, 100(1):67–92, 1973.
[20] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *ICCV*, 2013.
[21] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *TIP*, 19(6):1657–1663, 2010.
[22] Z. Guo, L. Zhang, and D. Zhang. Rotation invariant texture classification using lbp variance (lbpv) with global matching. *PR*, 43(3):706–719, 2010.
[23] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007.
[24] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *PAMI*, 31(10):1775–1789, 2009.
[25] M. Hussain, S. Al-Otaibi, G. Muhammad, H. Aboalsamh, G. Bebis, and A. M. Mirza. Gender recognition using nonsubsampled contourlet transform and wld descriptor. In *SCIA*, 2013.
[26] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.
[27] F. S. Khan, R. M. Anwer, J. van de Weijer, A. Bagdanov, A. Lopez, and M. Felsberg. Coloring action recognition in still images. *IJCV*, 105(3):205–221, 2013.
[28] F. S. Khan, R. M. Anwer, J. van de Weijer, A. Bagdanov, M. Vanrell, and A. Lopez. Color attributes for object detection. In *CVPR*, 2012.
[29] F. S. Khan, J. van de Weijer, S. Ali, and M. Felsberg. Evaluating the impact of color on texture recognition. In *CAIP*, 2013.
[30] F. S. Khan, J. van de Weijer, A. Bagdanov, and M. Felsberg. Scale coding bag-of-words for action recognition. In *ICPR*, 2014.
[31] F. S. Khan, J. van de Weijer, A. D. Bagdanov, and M. Vanrell. Portmanteau vocabularies for multi-cue image representations. In *NIPS*, 2011.
[32] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
[33] C. Li and K. Kitani. Pixel-level hand detection for ego-centric videos. In *CVPR*, 2013.
[34] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.
[35] M. Li, S. Bao, W. Dong, Y. Wang, and Z. Su. Head-shoulder based gender recogntion. In *ICIP*, 2013.
[36] S. Maji, L. D. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.
[37] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *PAMI*, 30(3):541–547, 2008.
[38] A. Mittal, A. Zisserman, and P. Torr. Hand detection using multiple proposals. In *BMVC*, 2011.
[39] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *PAMI*, 34(3):601–614, 2012.
[40] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *ICCV*, 2009.
[41] C. Shan. Learning local binary patterns for gender classification on real-world face images. *PRL*, 33(4):431–437, 2012.
[42] N. Shapovalova, W. Gong, M. Pedersoli, F. X. Roca, and J. Gonzalez. On importance of interactions and context in human action recognition. In *IbPRIA*, 2011.
[43] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*, 2012.
[44] G. Sharma, F. Jurie, and C. Schmid. Expanded parts model for human attribute and action recognition in still images. In *CVPR*, 2013.
[45] K. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010.
[46] J. van de Weijer, C. Schmid, J. J. Verbeek, and D. Larlus. Learning color names for real-world applications. *TIP*, 18(7):1512–1524, 2009.
[47] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.
[48] P. A. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.
[49] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *CVPR*, 2010.
[50] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.
[51] J. Wu, W. Smith, and E. Hancock. Facial gender classification using shape-from-shading. *IVC*, 28(6):1039–1048, 2010.
[52] J. Wu, W. Smith, and E. Hancock. Gender discriminating models from facial surface normals. *PR*, 44(12):2871–2886, 2011.
[53] J. Xing, H. Ai, and S. Lao. Multiple human tracking based on multi-view upper-body detection and discriminative learning. In *ICPR*, 2010.
[54] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012.
[55] 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.
[56] B. Yao and F.-F. Li. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.
[57] B. Yao and F.-F. Li. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *PAMI*, 34(9):1691–1703, 2012.
[58] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–218, 2007.
[59] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.
[60] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.

Fahad Shahbaz Khan is a post doctoral fellow at Computer Vision Laboratory, Linköping University, Sweden. He has a master in Intelligent Systems Design from Chalmers University of Technology, Sweden and a PhD degree in Computer Vision from Autonomous University of Barcelona, Spain. His research interests are in color for computer vision, object recognition, action recognition and visual tracking. 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. Joost van de Weijer has a master in applied physics at Delft University of Technology and a PhD degree at the University of Amsterdam. He obtained a Marie Curie Intra-European scholarship in 2006, which was carried out in the LEAR team at INRIA Rhne-Alpes. From 2008-2012 he was a Ramon y Cajal fellow at the Universitat Autonma de Barcelona. His research interest are in color for computer vision, object recognition, and color imaging. He has published in total over 60 peer reviewed papers. He has given several postgraduate tutorials at mayor ventures such as ICIP 2009, DAGM 2010, and ICCV 2011.

Rao Muhammad Anwer is a post doctoral research fellow at Department of Information and Computer Science, Aalto University School of Science, Finland. He has a master in Intelligent Systems Design from Chalmers University of Technology, Sweden and a PhD degree in Computer Vision from Autonomous University of Barcelona, Spain. His research interests are in object detection, pedestrian detection and action recognition.

Michael Felsberg received PhD degree (2002) in engineering from University of Kiel. Since 2008 full professor and head of CVL. Research: signal processing methods for image analysis, computer vision, and machine learning. More than 80 reviewed conference papers, journal articles, and book contributions. Awards of the German Pattern Recognition Society (DAGM) 2000, 2004, and 2005 (Olympus award), of the Swedish Society for Automated Image Analysis (SSBA) 2007 and 2010, and at Fusion 2011 (honourable mention). Coordinator of EU projects COSPAL and DIPLECS. Associate editor for the Journal of Real-Time Image Processing, area chair ICPR, and general co-chair DAGM 2011.

Carlo Gatta obtained the degree in Electronic Engineering in 2001 from the Università degli Studi di Brescia (Italy). In 2006 he received the Ph.D. in Computer Science at the Università degli Studi di Milano (Italy), with a thesis on perceptually based color image processing. In September 2007 he joined the Computer Vision Center at Universitat Autónoma de Barcelona as a postdoc researcher working mainly on medical imaging. He is a co-founder and member of the LAMP team. He is currently a senior researcher at the Computer Vision Center, under the Ramon y Cajal program. His main research interests are medical imaging, computer vision, machine learning, contextual learning and unsupervised deep learning.