Anwer, Rao Muhammad; Khan, Fahad Shahbaz; Laaksonen, Jorma

Two-stream part-based deep representation for human attribute recognition

Published in:
Proceedings - 2018 International Conference on Biometrics, ICB 2018

DOI:
10.1109/ICB2018.2018.00024

Published: 13/07/2018

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Anwer, R. M., Khan, F. S., & Laaksonen, J. (2018). Two-stream part-based deep representation for human attribute recognition. In *Proceedings - 2018 International Conference on Biometrics, ICB 2018* (pp. 90-97). IEEE. https://doi.org/10.1109/ICB2018.2018.00024

Two-Stream Part-based Deep Representation for Human Attribute Recognition

Rao Muhammad Anwer[1], Fahad Shahbaz Khan[2], Jorma Laaksonen[1]
[1]Department of Computer Science, Aalto University School of Science, Finland
[2]Computer Vision Laboratory, Linköping University, Sweden

***Abstract*—Recognizing human attributes in unconstrained environments is a challenging computer vision problem. State-of-the-art approaches to human attribute recognition are based on convolutional neural networks (CNNs). The de facto practice when training these CNNs on a large labeled image dataset is to take RGB pixel values of an image as input to the network. In this work, we propose a two-stream part-based deep representation for human attribute classification. Besides the standard RGB stream, we train a deep network by using mapped coded images with explicit texture information, that complements the standard RGB deep model. To integrate human body parts knowledge, we employ the deformable part-based models together with our two-stream deep model. Experiments are performed on the challenging Human Attributes (HAT-27) Dataset consisting of 27 different human attributes. Our results clearly show that (a) the two-stream deep network provides consistent gain in performance over the standard RGB model and (b) that the attribute classification results are further improved with our two-stream part-based deep representations, leading to state-of-the-art results.**

***Keywords*-Deep learning; Human attribute recognition; Part-based representation;**

I. Introduction

Recognizing human attributes such as gender, age, hair style, and clothing style in unconstrained environments is a challenging problem since humans can appear in different poses, under changing illumination and scale, and at low resolution. Human attribute recognition has many potential applications such as, including people search, person re-identification, and human identification. In case of visual surveillance recognizing fine-grained human attributes, to be used as soft-biometrics, has gained much importance due to many intelligent surveillance applications ranging from the monitoring of railway stations and airports to citizen-oriented applications such as monitoring assistants for the aged people. Initially, most approaches for human attribute recognition relied on face information with images having high resolution aligned frontal faces. However, humans appear in different scales and viewpoints in real-world situations, such as far-view video surveillance scenarios. In such scenarios, recognition solely based on facial cues could provide below-expected results, and cues from clothes and hairstyle are likely to provide valuable additional information. In this paper, we also investigate the problem of human attribute recognition in real-world images.

In recent years, convolutional neural networks (CNNs) have achieved an outstanding success, being the catalyst to

attribute	standard RGB	two-stream	ground-truth	attribute	standard RGB	two-stream	ground-truth
female	0.28	0.96	yes	female	0.97	0.98	yes
sitting	0.81	0.98	yes	sitting	0.17	0.14	no
armsbent	0.80	0.99	yes	armsbent	0.51	0.33	no
young	0.43	0.32	no	young	0.83	0.90	yes
short skirts	0.41	0.23	no	short skirts	0.17	0.80	yes
low cut top	0.29	0.18	no	low cut top	0.97	0.99	yes

Figure 1. Human attribute classification results on two images from the HAT-27 Dataset [1]. The probabilities from a certain attribute classifier is provided for both the standard RGB deep network and our two-stream deep network. The groun-truth labels are provided in blue text. An incorrect classification is shown in red. Our two-stream deep network based approach provides improved classification results compared to the baseline RGB deep network.

significant improvement in performance on a wide range of computer vision applications, including human attribute recognition [2], [3], [4]. CNNs consist of a series of convolution and pooling layers followed by one or more fully connected (FC) layers. CNNs or deep networks are trained on large amount of labeled training samples (i.e. ImageNet [5]). The de facto practice when training these deep networks is to take RGB pixel values of an image as input to the network. State-of-the-art human attribute recognition approaches [6], [7], [3], [4] either employ off-the-shelf pre-trained deep networks or fine-tuned them by transferring to the new domain of fine-grained human attribute data. Previous works [2], [8] have shown that deep features extracted from activations of the fully connected layers of the deep CNNs are general purpose image representations applicable to several visual recognition tasks. The resulting deep features are then used together with Support Vector Machines (SVMs) with linear kernel classifier.

Before the recent revolution of CNNs, hand-crafted texture features such as local binary pattern (LBP) [9] were shown to provide excellent performance for face [10], texture recognition [11], and gender recognition [12]. In a recent study [13], LBP and its variant descriptors were shown to provide texture classification performance similar to deep CNN features. The classification performance of LBP variants were especially competitive in the presence of rotations and different noise types. Due to their success, the problem of integrating LBP within deep learning archi-

tecture has been recently studied in the context of emotion recognition [14] and texture classification [15]. Motivated by these observations, we evaluate the impact of learning robust texture description within deep learning architectures for human attribute recognition. The resulting architecture is a two-stream deep network where texture is used as a second stream and fuse it with the standard RGB stream.

The standard paradigm for human attribute recognition assumes that person bounding boxes are provided both at training and test time. In such a paradigm, deep features are extracted from human bounding boxes. Beside the bounding box, many approaches [16], [12], [6] rely on part-based representations to counter the problem of pose normalization for human attribute classification. These approaches either use the deformable part models [17] or poselets [18] to obtain part locations. The work of [12] proposes semantic pyramids where parts of a person are automatically localized using state-of-the-art face and upper-body detectors. In this paper, we integrate the part-based deep representations, from the deformable part models (DPMs) [17], in our two-stream deep architecture. The DPMs based approach models the person as a structured constellation of parts. The resulting semantic part-based deep representations enable pooling across pose and viewpoint.

As discussed earlier, deep features extracted from activations of the fully connected (FC) layers of the deep CNNs are typically used for image representation. Instead of FC layers, activations from the last convolutional layer of the deep networks have been shown to provide excellent performance in recent works [19], [20], [21]. The convolutional layers are known to be discriminative and semantically meaningful. Further, they mitigate the need to use a fixed input image size. Generally, deep convolutional features are extracted at multiple scales from the convolutional layers of the deep network. These dense local features are then pooled either by VLAD [22] or Fisher vector (FV) [23] encoding schemes. Most notably, the work of [19] extract multi-scale features from the last convolutional layer of the deep network and pool the resulting dense local features using Fisher vector encoding to obtain an image representation (FV-CNN). Despite employing multi-scale convolutional features, the FV-CNN approach pool all the descriptors into a single scale-invariant image representation.

A. Our Approach

We propose a two-stream part-based deep representation for the problem of human attribute recognition in still images. Our two-stream deep architecture combines the standard RGB stream with a texture stream. The texture stream is obtained by first extracting LBP based codes from an image. The unordered LBP code values are then mapped to points in a 3D metric space [14] to obtain texture coded mapped images. The texture network stream is then separately trained using these texture coded mapped images as input to the deep network. Both the standard RGB stream and the texture network streams are trained on the ImageNet dataset.

To integrate the part-based knowledge, we employ the deformable part models (DPMs) [17] to obtain human body parts. Given an image, our approach then extract dense multi-scale convolutional features from the last convolutional layer of the RGB and texture deep networks. The features are extracted for both the whole person and the associated body parts. We then construct separate scale-coded FV-CNN representations for both the whole person and associated body parts. The scale-coded FV-CNN representations from the two-streams whole person and body parts are concatenated into a single feature vector for classification (see figure 2).

Experiments are performed on the challenging Human Attributes (HAT-27) Dataset [1] consisting of 27 different human attributes such as crouching, young, elderly, bermuda shorts, wedding dress, young, and female. Our results clearly suggest that our two-stream based image representations provide significant improvements compared to the standard RGB deep network. Moreover, our two-stream part-based deep representation provides further gain in classification performance, leading to consistent improvement over the state-of-the-art. Figure 1 shows attribute classification results from the standard RGB network and our approach on two example images.

II. Related Work

Human attribute recognition is an active research problem with several real-world applications [16], [3], [7], [24], [25]. State-of-the-art human attribute recognition approaches employ deep features and part-based representations. The work of [16] proposed pose-normalized representations by using deep features and semantic part detection using poselets [18]. Khan et al. [3] proposed deep semantic pyramids by employing deep features and pre-trained body part detectors to construct pose-normalized image representations. The work of [7] proposed a class activation map deep framework by designing a new exponential loss function to measure the appropriateness of the attention heat map which is an intermediate result in the class activation map. The work of [7] introduced an expanded part model by mining parts and learning corresponding discriminative templates together with their respective locations.

Recent years have seen deep convolutional neural networks or deep networks being the catalyst to significant performance gains in many computer vision applications. Deep CNNs are generally trained on a large amount of labeled data and consist of a series of convolution and pooling operations followed by fully connected (FC) layers. Combining multiple feature streams within the deep learning architecture has been recently studied. The work of [26]

proposed a two-stream deep architecture for action recognition where the spatial (RGB) stream is combined with the motion (optical flow) stream. The work of [14] proposed an approach for emotion recognition based on the standard RGB stream and texture stream using texture coded mapped images. The texture coded mapped images are obtained by mapping the unordered LBP codes of an image to points in a 3D metric space. The mapping is performed by applying the approximation of Earth Mover's Distance (EMD), resulting in a three channel mapped image. The deep network is then trained by taking this three channel mapped image as input. To the best of our knowledge, such a two-stream deep architecture, combining RGB and texture information, is yet to be investigated for human attribute recognition problem.

Our work is inspired by recent two-stream based works [14], [15], [27] on emotion and texture recognition. In this work, we evaluate the impact of the two-stream deep architecture, combining RGB and texture information, for human attribute recognition problem. We train a second CNN stream using texture coded mapped images together with the standard RGB CNN stream and construct the scale-coded FV-CNN representations from the two-stream deep network. Additionally, we also integrate the human body part information by employing deformable part-based models together with our two-stream deep network. To the best of our knowledge, we are the first to propose a two-stream part-based deep representation for human attribute recognition.

III. Two-Stream Part-based Deep Features

Here, we describe the construction of our two-stream part-based deep image representation. We start by describing our two-stream deep network. Afterwards, we describe our two-stream scale-coded FV-CNN full person body representation. Finally, we present how to integrate body part information to construct scale-coded FV-CNN body part representation.

A. Texture Coded Two-Stream Deep Architecture

Here, we describe our texture coded two-stream deep architecture for investigating to what extent texture coded deep networks complement the standard RGB based deep networks. Our two-stream deep architecture uses RGB pixel values to train RGB stream and texture coded mapped images to train texture stream. The two streams are trained on the ImageNet ILSVRC-2012 dataset [5]. Our two-stream network is based on the VGG-M architecture [28], which is similar to the Zeiler and Fergus network [29]. The VGG-M architecture takes as input an image of 224×224 pixels and consists of five convolutional layers followed by three FC layers. The first convolutional layer of the deep network employs smaller stride (1) and receptive field (or the filter size) with 96 convolution filters. The second convolutional layer uses a larger stride (2 compared to 1) with 256 convolution filters. The third, fourth and fifth convolutional layers comprises of 512 convolution filters.

To train the texture stream, texture coded mapped images are constructed as in [14], [15]. The underlying texture representation of texture coded mapped images is based on Local Binary Patterns (LBP) [9]. The LBP descriptor captures the local gray-scale distribution by describing the neighborhood of a pixel in an image by its binary derivatives, resulting in a short code. The short LBP codes are binary numbers (0 or 1) depending on a threshold, where each LBP code can be considered as a micro-texton. LBP codes can be computed over any neighborhood size with typical computations performed over a 8 pixel neighborhood, resulting in a binary string of eight-bit numbers between 0 and 255. Given an image $Im(a_{ct}, b_{ct})$ of size $H \times W$, with $(a_{ct} \in \{0, ..., H-1\}, b_{ct} \in \{0, ..., W-1\})$. Here, (a_{ct}, b_{ct}) represent the coordinates of the center pixel of a circular local neighborhood (Pt, R), where Pt denotes the number of sampling points. Here, $R > 0$ is the radius of the circular local neighborhood. The LBP code (a Pt-bit word) is then computed as:

$$LBP_{Pt,R}(a_{ct}, b_{ct}) = \sum_{pt=0}^{Pt-1} s(Im(a_{pt}, b_{pt}) - Im(a_{ct}, b_{ct}))2^{pt}, \tag{1}$$

where the thresholding function $s(t)$ is defined as:

$$s(t) = \begin{cases} 0 & \text{for } t < 0 \\ 1 & \text{for } t \geq 0. \end{cases} \tag{2}$$

In the standard texture classification task, the final image representation is obtained by constructing a histogram as a LBP code distribution over the whole image region. When training CNNs using LBP codes, a straight-forward strategy is to train the deep network by using LBP code values as input to the network. However, such a straightfroward strategy is infeasible since the unordered LBP code values are unsuitable for the convolution operations, equivalent to a weighted average of the input values, performed within CNN models. To counter this issue, the work of [14] proposes to train CNN models by mapping the LBP code values to points in a 3D metric space using Multi Dimensional Scaling for emotion recognition. Within the 3D metric space, the Euclidean distance approximates the distance between LBP code values. The resulting transformation allows the LBP code values to be used within CNN models, since the code values can now be averaged together using convolution operations, while approximately preserving the original code-to-code distances. In [14], Earth Mover's Distance (EMD) was used as a measure of the difference between two LBP codes to account for differences in spatial locations of pixel codes. For more details, we refer to [14].

B. Scale-coded Two-Stream Deep Representation

The texture coded two-stream deep architecture, described above, is trained on the large scale ImageNet dataset. The

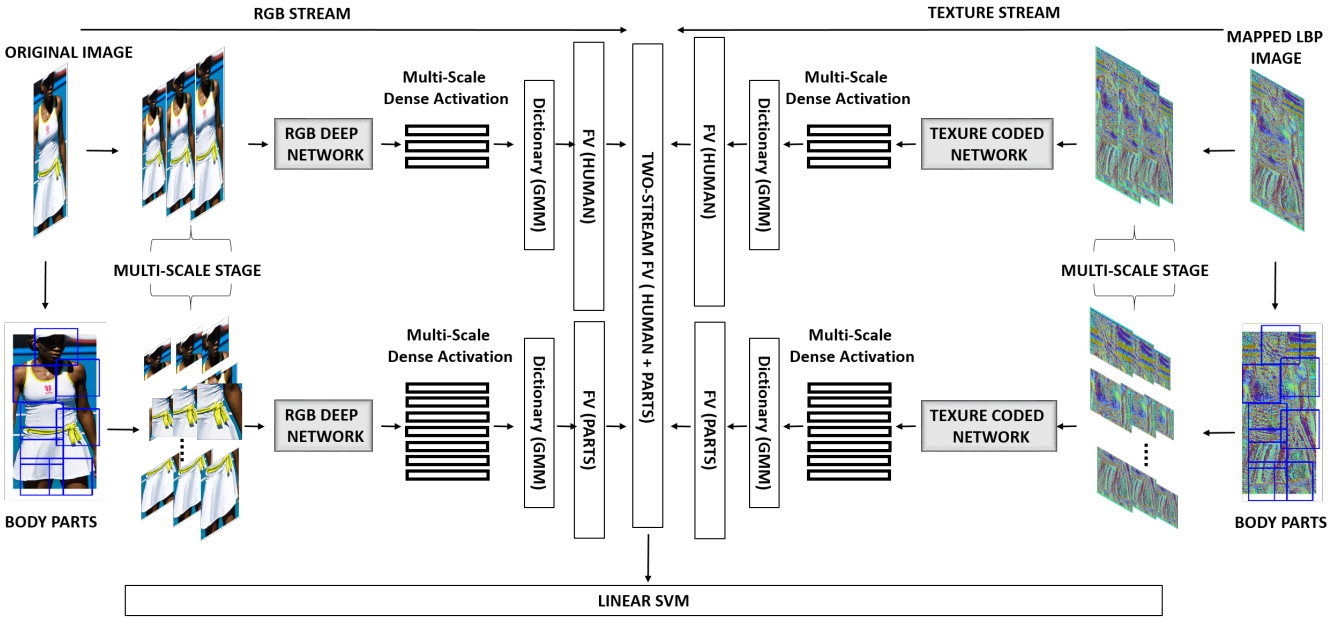

Figure 2. An overview of our proposed two-stream part-based deep image representation. The RGB network stream takes RGB image as input and is used to construct scale-coded FV-CNN representations from full body and human body parts. The texture network stream takes mapped coded texture image as input and is also used to construct scale-coded FV-CNN representations from full body and human body parts. The four scale-coded FV-CNN representations are concatenated into a long feature vector which is then input to linear SVMs.

pre-trained RGB and texture network streams are then used to construct image representations for human attribute recognition. Given an image, multi-scale convolutional features are extracted from the last (fifth) convolutional layer of the RGB and texture networks. The multi-scale features are extracted from person bounding boxes (available at both training and testing time). Afterwards, a Gaussian Mixture Model (GMM) is fitted to the distribution of dense multi-scale convolutional features. For each person bounding box PB, we extract a set of features:

$$Ft(PB) = \{\mathbf{x}_i^s \mid i \in \{1,\dots,K\}, s \in \{1,\dots,M\}\},$$

where $i \in \{1,\dots,K\}$ indexes the K feature locations in person box PB, and $s \in \{1,\dots M\}$ indexes the M scales extracted at each location. Similar to [21], we construct a scale-coded representation $h^t(PB)$ for each person bounding box PB by encoding features in group of extracted feature scales ($S = \{1,\dots,M\}$):

$$h^t(PB) \propto \sum_{i=1}^{K} \sum_{s \in S^t} cd(\mathbf{x}_i^s). \quad (3)$$

where $cd : \Re^p \rightarrow \Re^q$ represents a feature coding scheme which maps the input feature space of p dimensions to the final image representation (person bounding box) space of q dimensions. In the scale-coded image representation, feature scales are divided into several scale subgroups S^t that partition the whole set of extracted scales (i.e. $\bigcup_t S^t = \{1,\dots,M\}$). The multi-scale convolutional features are divided in three scale groups ($t \in \{\mathrm{sm}, \mathrm{md}, \mathrm{lg}\}$): small, medium and large scale features. The three scales are partitioned as in [21]. The multi-scale convolutional features are then pooled using Fisher vector (FV) encoding scheme. The final representation preserves the coarse scale information and is obtained by concatenating these three encodings of the person bounding box. Separate scale-coded image representations are obtained for both RGB and texture streams (see figure 2).

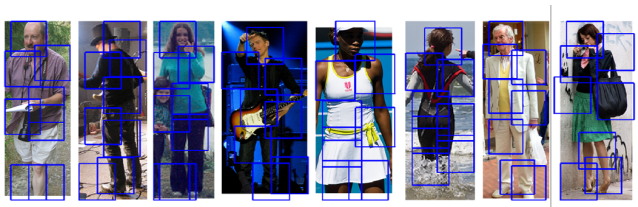

Figure 3. Visualization of body parts (in blue) on example images from the HAT-27 dataset.

Extension to Part-based Representation: To incorporate body part information, we employ the deformable part based framework [17]. The DPM based approach has been previously used to automatically detect parts for fine-grained classification [30], scene recognition [31], painting classification [32]. The DPM approach comprises of root filter and a deformable collection of moveable parts. The DPM framework represents both the root and parts by a dense grid

Figure 4. Example images from the HAT-27 dataset. The dataset consists of 27 different human attribute categories such as *crouching, young, elderly, bermuda shorts, wedding dress, young, and female*.

of non-overlapping cells. A 31-dimensional HOG [33] histogram is constructed for each non-overlapping cell. Within the DPM framework, the detection score for each window is computed by concatenating the root filter, the part filters and the configuration deformation cost of all corresponding parts.

The standard DPM framework exploits the bounding box information available during the training stage. However, no ground-truth part locations are available during the training stage and the part locations are therefore treated as latent information. The learning is performed using Latent SVM (LSVM) formulation. To obtain the human body parts, we train the DPM detector on the human class of the PASCAL VOC dataset [34]. We employ 8 parts for trained human DPM model. The trained human DPM model is then applied to human attribute images. We first crop each person instance using the provided bounding box information and then apply the trained human DPM model on the whole human. Figure 3 shows visualization of body parts (in blue) on example images from the HAT-27 dataset. The discriminative part regions, obtained using the trained human DPM model, are cropped from an image and rescaled over a range of scales before passing through our RGB and texture streams (see figure 2). Similar to whole human body, the multi-scale convolutional features from the human body parts are pooled into a scale-coded Fisher vector (FV-CNN) representation. Separate scale-coded body part representations are obtained for both RGB and texture streams (see figure 2).

IV. Experiments

In this section, we provide results of our approach for human attribute recognition. As discussed earlier, bounding boxes of person instances are provided at both train and test time in human attribute recognition. Thus the task is to predict the human attribute class associated with each person bounding box. We first provide details about our experimental setup and the human attribute dataset used in our evaluation. Afterwards, we present a comprehensive comparison of our approach with the baseline followed by a comparison with state-of-the-art methods in literature.

Experimental Setup: To train our two-stream deep network, we employ the Matconvnet library [35] and train the CNN models on the ImageNet ILSVRC-2012 dataset [5]. The ImageNet dataset consists of 1000 object classes and 1.2 million training images. During the training of two network streams, the learning rate is set to 0.001, a weight decay that acts as a regularizer and helps reducing the training error of the model is set to 0.0005. The momentum rate is associated with the gradient descent method used to minimize the objective function and is set to 0.9 during the training. The pre-trained two-stream deep network is then used as feature extractors for human attribute recognition. We compare our two-stream part-based deep approach with the baseline standard RGB deep network based on (a) features from the FC layers (FC-CNN) and (b) scale-coded Fisher vector (FV-CNN) representations. To construct FC-CNN representations, we remove the last FC layer (FC8) of the network which performs 1000-way ImageNet classification, and instead use 4096 dimensional activations from the FC7 (second last) layer as image features. The resulting image features are L_2-normalised and input to a linear kernel. Throughout our experiments, we fixed the weights (no fine-tuning) of both the baseline RGB network and our two-stream deep network for fair comparison.

To construct the scale-coded FV-CNN representations, we extract the convolutional features from the output of the last convolutional layer (conv5) of the deep network. The 512-dimensional dense convolutional features are extracted

	female	frontalpose	profilepose	turnedback	upperbody	standing	runwalk	crouching	sitting	armsbent	elderly	middleaged	young	teen
Standard RGB (FC-CNN)	84.1	93.0	53.1	78.8	98.3	97.0	70.8	25.3	66.5	92.1	31.0	62.3	53.8	23.5
Two-Stream (FC-CNN)	88.6	95.0	63.1	82.4	98.7	97.8	75.5	24.5	76.4	94.4	36.9	63.8	59.2	28.2
Two-Stream + Parts (FC-CNN)	93.2	96.3	69.6	90.7	98.6	98.1	76.6	22.2	73.3	95.0	57.2	73.0	69.6	32.4
Standard RGB (FV-CNN)	90.5	95.4	62.5	85.9	97.4	98.4	79.2	33.2	77.9	95.1	52.4	72.0	71.3	37.0
Two-Stream (FV-CNN)	91.6	95.7	68.3	88.4	97.9	98.6	81.1	33.5	80.3	95.8	55.1	75.0	73.4	39.1
Two-Stream + Parts (FV-CNN)	94.1	96.7	72.1	91.8	98.4	98.8	81.7	31.9	79.3	95.8	58.7	76.7	74.1	38.1

	kid	baby	tanktop	tshirt	casualjacket	mensuit	longskirt	shortskirt	smallshorts	lowcuttop	swimsuit	weddingdress	bermudashorts	mAP
Standard RGB (FC-CNN)	46.0	21.0	36.0	63.0	43.1	65.1	53.3	36.6	38.9	72.1	56.4	77.9	53.0	59.0
Two-Stream (FC-CNN)	54.1	26.2	41.7	70.0	47.8	69.8	54.8	40.0	50.0	78.3	57.2	71.1	55.0	63.0
Two-Stream + Parts (FC-CNN)	64.2	29.6	48.4	74.5	50.2	73.9	60.4	48.0	56.3	84.7	60.4	75.5	58.3	67.7
Standard RGB (FV-CNN)	63.5	23.5	50.5	78.2	59.1	65.9	52.4	49.6	57.6	78.9	56.1	67.6	50.8	66.7
Two-Stream (FV-CNN)	66.2	26.0	52.3	79.7	61.5	67.5	55.6	54.0	58.8	80.9	57.7	69.7	54.3	68.8
Two-Stream + Parts (FV-CNN)	67.2	28.9	53.0	80.0	62.8	74.0	58.5	55.1	60.3	85.3	58.6	76.3	55.7	**70.5**

Table I

BASELINE COMPARISON (IN MAP) OF OUR TWO-STREAM PART BASED APPROACH WITH THE STANDARD RGB DEEP NETWORK. THE COMPARISON IS PRESENTED FOR BOTH THE FC BASED FEATURES (FC-CNN) AND SCALE-CODED FV-CNN REPRESENTATIONS. OUR TWO-STREAM APPROACH CONSISTENTLY OUTPERFORMS THE STANDARD RGB NETWORK. FURTHER GAIN IN PERFORMANCE IS OBTAINED BY INTEGRATING THE BODY PART INFORMATION. FOR FAIR COMPARISON, WE USE THE SAME NETWORK ARCHITECTURE TOGETHER WITH THE SAME SET OF PARAMETERS FOR BOTH THE STANDARD RGB AND OUR TWO-STREAM DEEP NETWORKS.

after rescaling the image at 21 different scales $s \in \{0.5 + 0.1n \mid n = 0, 1, \ldots, 20\}$. For vocabulary construction, we employ a Gaussian Mixture Model (GMM) with 16 components. Consequently, Finally, the scale-coded Fisher vector representations (FV-CNN) discussed in section III-B are constructed for both whole body and body parts and using both RGB and texture network streams. The resulting scale-coded FV-CNNs for whole body and body parts and from RGB and texture network streams are concatenated into a single final representation which is then input to the linear kernel SVM classifier.

We follow the same evaluation protocol proposed by the authors of the dataset [1]. The classification performance is measured in average precision (AP) as area under the precision-recall curve. The overall final performance is then measured by taking the mean average precision (mAP) over all human attribute categories in the dataset.

Dataset: We perform comprehensive experiments on the challenging Human Attributes Dataset (HAT-27) [1]. The dataset comprises of 9344 images of 27 different human attributes such as *long skirt, armsbent, crouching, frontal pose, casual jacket, wedding dress, young, and female*.[1]

We follow the same evaluation protocol proposed by the authors of the dataset [1]. The classification performance is measured in average precision (AP) as area under the precision-recall curve. The overall final performance is then measured by taking the mean average precision (mAP) over all human attribute categories in the dataset. We employ the train and test splits provided by the respective authors [1]. Figure 4 shows example images from the HAT-27 dataset.

A. Baseline Comparison

Table I shows the baseline comparison on the HAT-27 dataset. Our baseline is the standard RGB VGG-M deep network. It is worth to mention that we use the same VGG-M network architecture together with the same set of parameters for both the standard RGB and our two-stream deep networks. The FC-CNN approach from the standard RGB deep network obtains a mAP score of 59.0%. Our FC-CNN approach from the two-stream deep network provides a significant gain of 4.0% over the Standard RGB (FC-CNN), with a mAP score of 63.0%. Our Two-Stream (FC-CNN) approach improves the results on 25 out of 27 human attribute categories. A significant gain in performance is achieved especially for small shorts (+11%), profile pose (+9%), sitting (+9%), kid (+8%), and tshirt (+6%) action categories, all compared to the Standard RGB (FC-CNN) approach. Furthermore, integrating the body part information in our Two-Stream (FC-CNN) approach improves the classification results with a mAP score of 67.7%.

All scale-coded FV-CNN image representations provide improved results compared to their respective FC-CNN counterparts. The Standard RGB (FV-CNN) approach achieves a mAP score of 66.7%. Our Two-Stream (FV-CNN) approach outperforms the Standard RGB (FV-CNN) method by achieving a gain of 2.1% in mAP. The results are further improved by integrating part-based information with a mAP score of 70.5%. In conclusion, our baseline experiments clearly suggest that the proposed two-stream approach provides consistent improvements over the baseline standard RGB deep network. A further gain in classification performance is achieved by integrating body part knowledge together with our two-stream approach.

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

Finally, we compare our approach with state-of-the-art results reported in literature on the HAT-27 dataset. Most state-of-the-art approaches report classification results based on very deep networks (VGG-16 or VGG-19) [39]. We therefore also combine our Two-Stream + Parts (FV-CNN) approach with the pre-trained VGG-16 deep network features. Table II shows the state-of-the-art comparison on the HAT-27 dataset. The expanded part-based model (EPM) approach [36] that learns a collection of discriminative

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

	female	frontalpose	profilepose	turnedback	upperbody	standing	runwalk	crouching	sitting	armsbent	elderly	middleaged	young	teen
EPM [36]	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 [37]	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 [12]	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 [4]	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
SC-BODF [21]	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
Deep SMP [3]	93.7	95.6	67.0	85.2	96.0	98.4	83.6	32.1	86.6	95.1	55.1	76.6	75.3	44.8
Deep VLAD [38]	97.5	97.4	83.0	96.6	98.6	99.1	80.0	30.8	87.9	97.3	69.0	80.1	73.9	38.4
this paper	95.4	97.1	75.1	93.0	98.8	98.9	83.9	41.6	85.2	96.4	66.4	78.6	77.7	44.5

	kid	baby	tanktop	tshirt	casualjacket	mensuit	longskirt	shortskirt	smallshorts	lowcuttop	swimsuit	weddingdress	bermudashorts	mAP
EPM [36]	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 [37]	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 [12]	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 [4]	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
SC-BODF [21]	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
Deep SMP [3]	74.9	39.8	55.9	81.5	62.2	74.1	59.7	53.1	62.4	85.8	63.0	75.7	58.3	71.5
Deep VLAD [38]	71.0	31.9	65.5	88.8	60.7	75.6	62.5	50.0	69.2	89.1	55.0	75.1	77.7	74.2
this paper	75.1	37.2	58.1	83.2	66.3	80.7	64.6	57.0	64.7	89.9	68.3	83.5	64.7	**75.0**

Table II
COMPARISON OF OUR APPROACH WITH STATE-OF-THE-ART METHODS ON THE 27 HUMAN ATTRIBUTES (HAT-27) DATASET. OUR APPROACH IMPROVES THE STATE-OF-THE-ART BY ACHIEVING A MAP SCORE OF 75.0%.

templates appearing at specific scale-space positions obtains a mAP score of 58.7%. The D-EPM approach [4] combines the expanded part model with deep features and achieves a mAP score of 69.6%. The semantic pyramid (SM-SP) approach of [12] employing body part detectors together with spatial pyramids achieves a mAP score of 57.6%. The deep features variant of the semantic pyramid (Deep SMP) approach [3] obtains a mAP score of 71.5%. The appearance part based dictionary approach (RAD) [37] achieves a mAP score of 59.3%. The scale-coded bag of deep feature approach (SC-BODF) of [21] combines the scale-coded deep representations with FC features of Very deep (VGG-19) network. The Deep VLAD approach of [38] combines the deep features with the VLAD encoding scheme and obtains a mAP score of 74.2%. Our approach improves the state-of-the-art by achieving a mAP score of 75.0%.

C. Conclusions

In this paper, we evaluated the impact of two-stream deep architecture for the problem of human attribute recognition. In our two-stream deep architecture, we trained a second stream using texture coded mapped images together with the standard RGB stream. Afterwards, a scale-coded Fisher vector representation is constructed from the two-stream deep network. Furthermore, we also integrated the body part information by employing deformable part-based models together with our two-stream deep network.

Experiments on the challenging HAT-27 dataset clearly demonstrate that our two-stream deep architecture provides complementary information to standard RGB deep model of the same network architecture. The integration of part based information together with our two-stream network leads to further gain in classification performance, leading to consistent improvements over the state-of-the-art.

Acknowledgments: This work has been funded by the grant 251170 of the Academy of Finland, H2020-ICT project MeMAD (780069), SSF through a grant for the project SymbiCloud, VR starting grant (2016-05543). 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] G. Sharma and F. Jurie, "Learning discriminative spatial representation for image classification," in *BMVC*, 2011.

[2] H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *CVPRW*, 2014, pp. 512–519.

[3] F. S. Khan, R. M. Anwer, J. van de Weijer, M. Felsberg, and J. Laaksonen, "Deep semantic pyramids for human attributes and action recognition," in *SCIA*, 2015.

[4] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for semantic description of humans in still images," *PAMI*, vol. 39, no. 1, pp. 87–101, 2017.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[6] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *ICCV*, 2015.

[7] H. Guo, X. Fan, and S. Wang, "Human attribute recognition by refining attention heat map," *PRL*, vol. 94, pp. 38–45, 2017.

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*, 2014, pp. 1717–1724.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[10] T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with local binary patterns." in *ECCV*, 2004.

[11] F. S. Khan, R. M. Anwer, J. van de Weijer, M. Felsberg, and J. Laaksonen, "Compact color-texture description for texture classification," *PRL*, vol. 51, pp. 16–22, 2015.

[12] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta, "Semantic pyramids for gender and action recognition," *TIP*, vol. 23, no. 8, pp. 3633–3645, 2014.

[13] L. Liu, P. Fieguth, X. Wang, M. Pietikainen, and D. Hu, "Evaluation of lbp and deep texture descriptors with a new robustness benchmark," in *ECCV*, 2016.

[14] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *ICMI*, 2015.

[15] R. M. Anwer, F. S. Khan, J. van de Weijer, and J. Laaksonen, "Tex-nets: Binary patterns encoded convolutional neural networks for texture recognition," in *ICMR*, 2017.

[16] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *CVPR*, 2014, pp. 1637–1644.

[17] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[18] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *ICCV*, 2009.

[19] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, "Deep filter banks for texture recognition, description, and segmentation," *IJCV*, vol. 118, no. 1, pp. 65–94, 2016.

[20] L. Liu, C. Shen, and A. van den Hengel, "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification," in *CVPR*, 2015, pp. 4749–4757.

[21] F. S. Khan, J. van de Weijer, R. M. Anwer, A. Bagdanov, M. Felsberg, and J. Laaksonen, "Scale coding bag of deep features for human attribute and action recognition," *MVA*, 2017.

[22] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.

[23] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, pp. 222–245, 2013.

[24] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," in *ICB*, 2015.

[25] Y. Deng, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *ACM MM*, 2014.

[26] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[27] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *arXiv preprint arXiv:1705.03428*, 2017.

[28] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[29] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[30] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *ICCV*, 2013.

[31] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *ICCV*, 2011.

[32] R. M. Anwer, F. S. Khan, J. van de Weijer, and J. Laaksonen, "Combining holistic and part-based deep representations for computational painting categorization," in *ICMR*, 2016.

[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.

[34] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge." *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[35] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *ACM Multimedia*, 2015.

[36] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *CVPR*, 2013, pp. 652–659.

[37] J. Joo, S. Wang, and S.-C. Zhu, "Human attribute recognition by rich appearance dictionary," in *ICCV*, 2013, pp. 721–728.

[38] S. Yan, J. Smith, and B. Zhang, "Action recognition from still images based on deep vlad spatial pyramids," *SPIC*, vol. 54, pp. 118–129, 2017.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.