Deep Semantic Pyramids for Human Attributes and Action Recognition

Fahad Shahbaz Khan[1](✉), Rao Muhammad Anwer[2], Joost van de Weijer[3], Michael Felsberg[1], and Jorma Laaksonen[2]

[1] Computer Vision Laboratory, Linköping University, Linköping, Sweden
fahad.khan@liu.se
[2] Department of Information and Computer Science,
Aalto University School of Science, Aalto, Finland
[3] Computer Vision Center, CS Department, Universitat Autonoma de Barcelona,
Barcelona, Spain

Abstract. Describing persons and their actions is a challenging problem due to variations in pose, scale and viewpoint in real-world images. Recently, semantic pyramids approach [1] for pose normalization has shown to provide excellent results for gender and action recognition. The performance of semantic pyramids approach relies on robust image description and is therefore limited due to the use of shallow local features. In the context of object recognition [2] and object detection [3], convolutional neural networks (CNNs) or deep features have shown to improve the performance over the conventional shallow features.

We propose deep semantic pyramids for human attributes and action recognition. The method works by constructing spatial pyramids based on CNNs of different part locations. These pyramids are then combined to obtain a single semantic representation. We validate our approach on the Berkeley and 27 Human Attributes datasets for attributes classification. For action recognition, we perform experiments on two challenging datasets: Willow and PASCAL VOC 2010. The proposed deep semantic pyramids provide a significant gain of 17.2 %, 13.9 %, 24.3 % and 22.6 % compared to the standard shallow semantic pyramids on Berkeley, 27 Human Attributes, Willow and PASCAL VOC 2010 datasets respectively. Our results also show that deep semantic pyramids outperform conventional CNNs based on the full bounding box of the person. Finally, we compare our approach with state-of-the-art methods and show a gain in performance compared to best methods in literature.

Keywords: Action recognition · Human attributes · Semantic pyramids

1 Introduction

Human attributes description such as gender, hair style, and clothing style and action category recognition such as playing music, riding bike, and taking photo are two of the most challenging problems in semantic computer vision. The two

R.R. Paulsen and K.S. Pedersen (Eds.): SCIA 2015, LNCS 9127, pp. 341–353, 2015.
DOI: 10.1007/978-3-319-19665-7_28

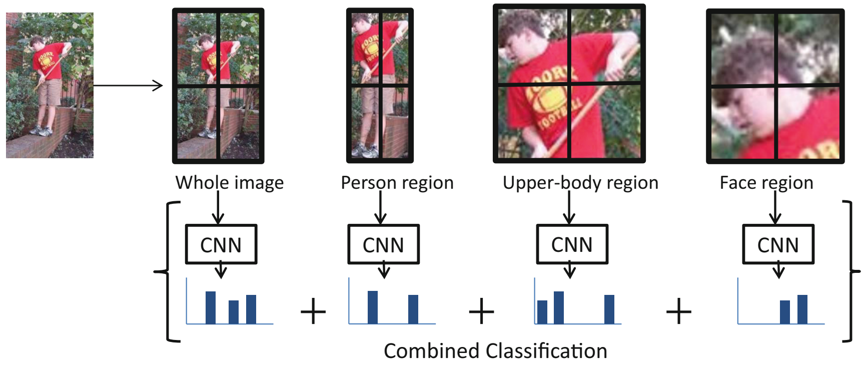

Fig. 1. Overview of deep semantic pyramids approach. We use whole image, full-body, upper-body and face regions for feature extraction. The upper-body and face regions are automatically localized using pre-trained state-of-the-art body part detectors. Each region is then used to construct a spatial pyramid representation of deep features. Finally, representations from all regions are concatenated into a single feature vector for classification. It is worth to mention that same pipeline is used for both human attributes classification and action recognition.

tasks are difficult since scale, viewpoint and pose varies significantly in real-world scenarios. Furthermore, images can appear in different illumination conditions, in low resolution and with back-facing people. These factors make the task of robust pose normalization and image description extremely challenging. In this paper, we focus on two problems namely: human attributes recognition and action category recognition.

Most state-of-the-art approaches rely on part-based representations [4–6] to counter the problem of pose normalization for human attributes classification. These approaches either use the deformable part models [7] or poselets [8] to obtain part locations. The part locations are then used to construct pose-normalized representations for classification. In the context of action recognition, conventional approaches either employ the bag-of-words framework [18,22,24] or focus on finding human-object interactions [15,20,21] to obtain improved performance. The bag-of-words based methods make use of local features such as shape, texture and color to represent local patches. On the other hand, approaches based on finding human-object interactions also use local features for patch description.

Recently, Khan et al. [1] have proposed an approach, semantic pyramids, for pose normalization. The method aims at fusing information from the full-body, upper-body and face regions of a person in an image. The parts of a person are automatically localized using state-of-the-art face and upper-body detectors. This ensures that no additional part annotations are required neither at training nor at test time. Spatial pyramid based image representations are then constructed for each part location. Consequently, the final image representation is obtained by combining the full-body, upper-body and face pyramids for each instance of a person. In this work, we also use pose-normalized semantic pyramids representation for human attributes and action recognition.

The performance of semantic pyramids approach heavily relies on the choice of the feature descriptors used to describe each part location in an image. Conventionally, local feature descriptors, also known as *shallow* features, have been employed for body part description [1]. Recently, convolutional neural networks (CNNs) [9], also known as *deep* features, have shown to provide excellent performance on large scale image classification tasks [10]. Deep features have also been applied successfully to many other applications such as object detection [3], pose estimation [11], and attributes classification [6]. However, substantial amount of training data is required for learning these robust networks. The work of [12] shows that off-the-shelf CNNs features, trained on ImageNet dataset, are generic and can be applied to any standard image classification dataset. Here, we investigate to what extent off-the-shelf CNNs features fare when used within the semantic pyramids method for the tasks of human attributes classification and action recognition.

Contributions: In this paper, we propose to augment semantic pyramids with deep features for attributes classification and action recognition. The approach combines information from the whole image, full-body, upper-body and face regions. Similar to [1], we employ state-of-the-art body part detectors to automatically localize face and upper-body regions. In this way, no extra annotations for body parts are required either at training or test time. The best candidate bounding boxes are selected for each part location for feature extraction. Instead of shallow features used in [1], we employ pre-trained deep features learned from the ImageNet dataset for each region description. Each body region is divided geometrically in various blocks to obtain a spatial pyramid representation. The deep features are then computed for each region block providing a rough spatial description. Finally, the deep spatial pyramid representations from the whole image, full-body, upper-body and face are combined into a single feature vector for classification. Figure 1 shows an overview of deep semantic pyramids approach.

For human attributes classification, we validate our approach on the Berkeley and 27 Human Attributes benchmark datasets. For action recognition, we perform experiments on two challenging datasets: Willow and PASCAL VOC 2010. Our results show that significant improvements are obtained by deep semantic pyramids over standard semantic pyamids for both attributes and action recognition. Furthermore, deep semantic pyramids approach improve the performance compared to conventional CNNs trained on the full bounding box of the person alone. Finally, we show deep semantic pyramids outperform state-of-the-art methods for both human attributes and action recognition tasks.

2 Our Approach

Our approach combines the advantages of semantic pyramids and deep features in a single framework for attributes classification and action recognition. We start by providing a brief introduction to conventional semantic pyramids followed by our proposed deep semantic pyramids.

2.1 Semantic Pyramids

The semantic pyramids approach has been recently introduced by Khan et al. [1] for pose normalization. Instead of relying on single body part, semantic pyramids approach aims at combining information from different part locations for gender and action recognition. In the work of [1], information is combined from full-body, upper-body and face regions of a person. The approach employs pre-trained state-of-the-art upper-body and face detectors to automatically extract semantic information. The use of pre-trained detectors ensures that no extra overhead to annotate part regions is required for each person instance. The method assumes that the bounding box information of a person is available. To obtain the upper-body part of a person, a pre-trained upper-body detector[1] based on part-based detection framework [7] is employed. To extract the face region of a person, a pre-trained face detector [13] built on top of part-based framework [7] is used. The face detector is trained using positive instances from the MultiPIE dataset and negative samples taken from the popular INRIA person dataset.

Fusing body part detector outputs: Each body part detector provides a set of hypotheses by firing at multiple locations. In the work of [1], a simple method is proposed to select the optimal part locations. The method works by defining the task of finding the optimal part location as an energy minimization problem. It was shown to provide improved results compared to the baseline approach of selecting the part location with the highest scoring confidence. Finally, the full-body, upper-body and face bounding boxes are combined to obtain a semantic representation.

Image representations: For gender recognition, multiple feature descriptors (HOG, WLD and CLBP) are extracted in a spatial pyramid manner for each of the body parts. The different pyramid representations are then combined in a single feature vector for classification. Similarly, for action recognition, the bag-of-words framework with multiple visual features (SIFT, Color names, Color-SIFT) is employed to construct semantic pyramids.

2.2 Deep Semantic Pyramids

We combine part-based semantic pyramids and deep features to obtain a robust pose-normalized deep representation. To this end, our objective is to use CNNs for learning powerful features and the simplicity of semantic pyramids to obtain robust pose-normalized representation. We use deep features [14] pre-trained on the ImageNet dataset while demonstrating best performance on ImageNET 2014 challenge. The network takes fixed size 224x224 RGB image as input. The depth of the network varies from 11 (8 convolution and 3 FC) weight layers to 19 weight layers (16 convolution and 3 FC). Inside these two deep networks, the number of channels start from 64 in the first layer and increased by a factor of 2 after each max-pooling layer.

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

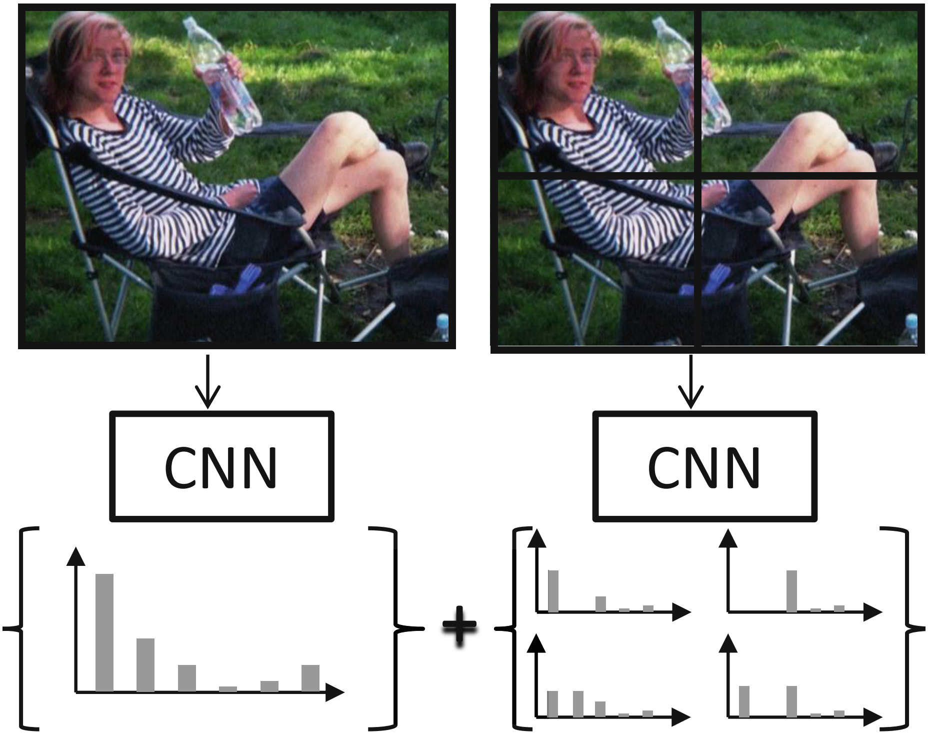

Fig. 2. Overview of deep spatial pyramid representation used for attributes classification and action recognition. We apply a two-level pyramid scheme where deep features are extracted separately for each image partition. The final representation is obtained by concatenating deep features from all the region partitions.

Unlike previous network architectures which employed large receptive fields in the first layers, the deep networks [14] employ small 3x3 receptive fields throughout the entire network. The respective fields are convolved with every pixel input with a stride of 1. The use of small receptive fields enables the incorporation of three non-linear rectification layers instead of a single one. This helps to obtain a more discriminative decision function. In our work, we use the fully connected layers from both 11 and 19 weight layer networks for image description. This provides us with a feature vector of size 4096x2. It is worth to mention that the dimensionality of deep features is significantly lower than the shallow representations commonly employed in the bag-of-words framework. For more details, we refer to [14].

In this work, we use whole image, full-body, upper-body and face regions. The full bounding box information for each person instance is provided both at training and test time with all the datasets. As discussed earlier, the bounding boxes of upper-body and face regions are automatically extracted using the approach presented by [1]. We construct a spatial pyramid representation using deep features for each region as illustrated in Figure 1. A spatial pyramid upto

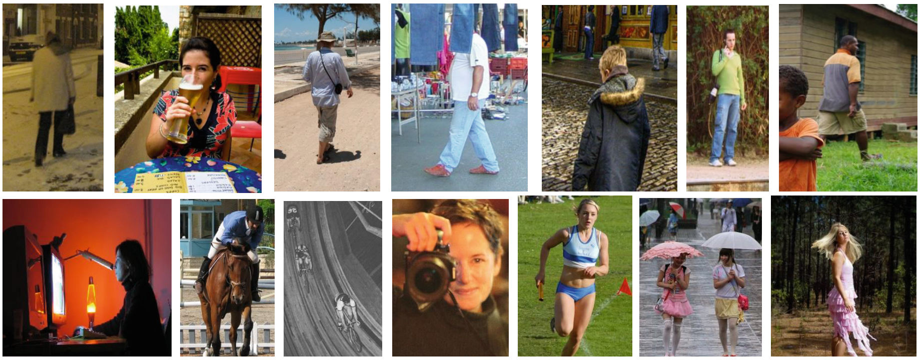

Fig. 3. Example images from the datasets used in our experiments. Top row: images from the Berkeley Attributes dataset used for attributes classification task. Bottom row: images from the Willow action dataset used for action recognition task.

level 2 is used in our work. Figure 2 shows an overview of our deep spatial pyramid representation approach applied for each of the body parts.

Since the deep network takes a fixed-size input, we crop each image partition and resize it to 224x224 pixels. Each partition is then represented by a 4096 dimensional feature vector. The final representation is obtained by concatenating all the feature vectors from each image partition resulting in a 5x4096 dimensional feature vector. The spatial pyramids of the whole image, full-body, upper-body and face regions are normalized and concatenated, resulting in a 4x5x4096 size feature vector which is then input to the classifier for classification. The same procedure is applied for both deep networks described above. We employ non-linear SVM with intersection kernel for classification.

3 Attributes Classification

We start with an introduction of the datasets used in our experiments followed by our experimental results. Finally, we compare deep semantic pyramids with state-of-the-art methods in literature.

3.1 Dataset

We perform experiments on the Berkeley and 27 Human Attributes benchmark datasets. The Berkeley dataset consists of 4013 training and 4022 test instances. The images are collected from PASCAL and H3D datasets. The dataset consists of nine attributes: *male, long hair, glasses, hat, tshirt, longsleeves, shorts, jeans and long pants.*[2] The images in the dataset are very challenging since persons appear in different poses, viewpoints and scales with only 60% of the persons in

[2] The Berkeley dataset is available at: http://www.cs.berkeley.edu/~lbourdev/poselets/

the photos have both eyes visible. The 27 Human Attributes (HAT) dataset consists of 9344 images of 27 different human attributes such as *crouching, casual jacket, wedding dress, young and female.*[3] Figure 3 (top row) shows some example images from the Berkeley dataset.

3.2 Experimental Results

We first compare the performance of deep semantic pyramids with the standard shallow semantic pyramids. Afterwards, we provide a comparison with state-of-the-art approaches. For attributes classification, the performance is represented by average precision as the area under the precision-recall curve.

Deep vs Shallow Semantic Pyramids Here, we validate deep semantic pyramids approach with the conventional shallow semantic pyramids. In all cases, we use spatial pyramid representations for attributes classification. We use 2 level spatial pyramid: level 1 corresponds to standard image-level representation and level 2 comprises of 2x2 partitioning of an image. In case of level 2, feature representation from the previous level is also concatenated.

Table 1(a) shows a comparison between deep pyramids and shallow pyramids using different regions of an image for attributes classification. In case of the whole image (WI), using deep semantic pyramids improve the performance by 22.3% and 11.2% in mean AP on Berkeley and 27 Human Attributes (HAT) datasets respectively. Overall, deep semantic pyramids provide a performance boost of 17.2% and 13.9% in mean AP over the conventional shallow pyramids on the to datasets respectively.

The results clearly suggest that combining deep pyramids based on different body part regions improve the performance compared to using only the full bounding box of a person. Moreover, deep semantic pyramids significantly improve the performance over standard semantic pyramids for attributes classification. It is worth to mention that the deep features used in the semantic pyramids are generic and not trained for the task of attributes classification.

State-of-the-art Comparison: We compare deep semantic pyramids with state-of-the-art approaches in literature. Table 2 shows a comparison of state-of-the-art approaches with deep semantic pyramids on the Berkeley Attributes dataset. The conventional poselets approach [4] provides a mean AP of 65.2% on this dataset. The DLPoselets approach which employs the same poselets to train an attribute classifier provides a mean AP of 69.2%. The only difference between poselets and DLPoselets is that the latter uses deep features which improves the performance by 4.0% over the traditional poselets.

The approach of [6] provides a mean AP of 78.9% on this dataset. The method employs poselets to obtain part locations and train a poselet-level deep network on an additional large dataset of human attributes. Moreover, the method uses

[3] The 27 Human Attributes dataset is available at: https://sharma.users.greyc.fr/hatdb/

Table 1. Classification performance of deep semantic pyramids (DP) compared to standard shallow pyramids (SP) for attributes classification and action recognition tasks. The results are shown for whole image (WI), full-body (FB), upper-body (UB), face (FC) and combined representations. The deep pyramids approach significantly outperforms the standard shallow pyramids on all datasets.

(a) Attributes Classification

	WI	FB	UB	FC	Combine
Berkeley (SP)	51.6	57.2	53.7	55.0	62.1
Berkeley (DP)	73.9	75.2	68.6	65.6	**79.3**
27 HAT (SP)	44.3	46.2	43.0	38.4	57.6
27 HAT (DP)	55.5	66.8	59.1	55.9	**71.5**

(b) Action Recognition

	WI	FB	UB	FC	Combine
Willow (SP)	62.4	63.7	51.1	52.7	66.7
Willow (DP)	87.9	88.6	57.8	56.0	**91.0**
PASCAL (SP)	51.6	52.8	47.8	48.6	55.8
PASCAL (DP)	71.9	81.2	59.8	58.8	**85.3**

Table 2. Comparison of deep semantic pyramids approach with state-of-the-art on the Berkeley dataset. Deep semantic pyramids, despite their simplicity, achieve the best performance on 5 out of 9 categories while providing competitive performance compared to state-of-the-art methods.

	male	long hair	glasses	hat	tshirt	longsleeves	shorts	jeans	long pants	mean AP
Poselets [4]	82.4	72.5	55.6	60.1	51.2	74.2	45.5	54.7	90.3	65.2
DLPoselets [4]	92.1	82.3	76.3	65.6	44.8	77.3	43.7	52.5	87.8	69.2
DPD [5]	83.7	70.0	38.1	73.4	49.8	78.1	64.1	78.1	93.5	69.9
RAD [16]	88.0	80.1	56.0	75.4	53.5	75.2	47.6	69.3	91.1	70.7
PANDA [6]	**91.7**	**82.7**	**70.0**	74.2	49.8	86.0	79.1	**81.0**	96.4	79.0
This Paper	88.8	79.8	47.6	**84.2**	**66.4**	**88.0**	**83.3**	79.1	**96.5**	**79.3**

pre-trained deep features trained on the ImageNet to describe the full body of a person. In this way, the classifier exploits the complementarity in the deep features of parts and holistic regions since they are trained on different image data. Different to [6], our approach, while only using off-the-shelf deep features trained on the ImageNet, provides comparable performance to the previous best method.

Table 3 shows a comparison of state-of-the-art approaches with deep semantic pyramids on 27 Human Attributes (HAT) dataset. The approach of [15] based on expanded part based models (EPM) obtain a mean AP of 58.7%. The rich appearance part dictionary of humans approach (RAD) by [16] achieves a mean AP of 59.3%. The standard semantic pyramids approach (SP) provides a mean AP of 57.6%. Deep semantic pyramids outperform best reported results in literature by achieving a mean AP of 71.5% on this dataset.

4 Action Recognition

Here, we evaluate the performance of deep semantic pyramids for the task of action recognition in still images. In case of action recognition, the bounding box of each person instance is provided both at training and test time. The task is to recognize the action category label associated with the bounding box. We use the same pipeline as was used for the task of attributes recognition earlier. For action recognition, the performance is again represented by average precision as area under the precision-recall curve.

Table 3. Comparison of deep semantic pyramids approach with state-of-the-art on 27 Human Attributes (HAT) dataset. Deep semantic pyramids obtain the best performance on 22 out of 27 categories compared to state-of-the-art methods.

	female	frontalpose	profilepose	turnedback	upperbody	standing	runwalk	crouching	sitting	armsbent	elderly	middleaged	young	teen
EPM [15]	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 [16]	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
SP [1]	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
This Paper	**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**

	kid	baby	tanktop	tshirt	casualjacket	mensuit	longskirt	shortskirt	smallshorts	lowcuttop	swimsuit	weddingdress	bermudashorts	mean AP
EPM [15]	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 [16]	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
SP [1]	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
This Paper	**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**

4.1 Datasets

To validate deep semantic pyramids, we use two challenging action recognition datasets: Willow and PASCAL VOC 2010. The willow dataset comprises of seven action classes: *interacting with computer, photographing, playing music, riding bike, riding horse, running and walking.*[4] We also validate our approach on the PASCAL VOC 2010 dataset. The PASCAL VOC dataset consists of nine action classes: *phoning, playing instrument, reading, riding bike, riding horse, running, taking photo, using computer and walking.*[5] Both these datasets are extremely challenging due to significant amount of scale, illumination, pose and viewpoint variations. Figure 3 (bottom row) shows some example images from these datasets.

4.2 Deep vs Shallow Semantic Pyramids

Here, we compare deep semantic pyramids with conventional semantic pyramids for the task of action recognition. In the work of [1], the bag-of-words framework with multiple features have been employed for each part location in an image. As a baseline, we use the bag-of-words framework with SIFT features to construct shallow semantic pyramids. Table 1(b) shows a comparison between deep pyramids and shallow pyramids on the two action recognition datasets. On the Willow action dataset, deep semantic pyramids improve the overall performance by 24.3% in mean AP. Similarly, on the PASCAL VOC 2010 validation set the conventional shallow pyramids provide a mean AP of 55.8%. Deep semantic pyramids improve the classification performance by providing a mean AP of 85.3%. The results obtained on both action datasets clearly suggest that deep semantic pyramids significantly improve the performance compared to standard semantic pyramids for action classification.

Figure 4 shows top correct (top-row) and incorrect predictions (bottom-row) for the phoning class from the PASCAL VOC 2010 dataset. Three out of four

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

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

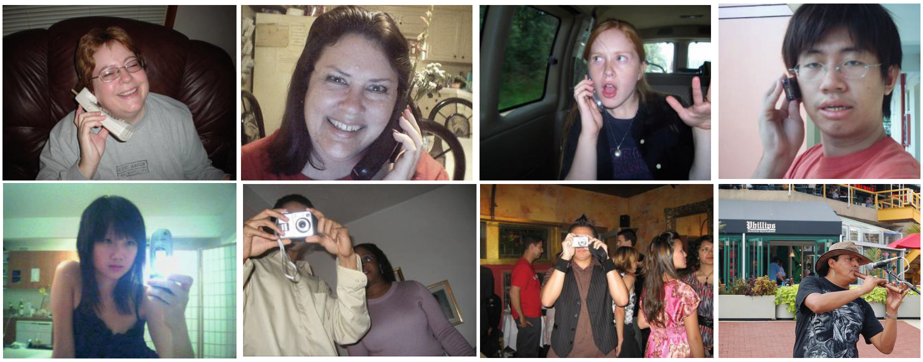

Fig. 4. Images from the PASCAL VOC 2010 dataset. Top row: top correct predictions for phoning class. Bottom row: top incorrect predictions for phoning class.

misclassified examples are from taking photo category which has certain degree of visual similarity with the phoning class.

State-of-the-art Comparison: we compare deep semantic pyramids with state-of-the-art methods in literature. Table 4 shows a comparison with state-of-the-art methods on the Willow dataset. Our approach provides the best performance on 6 out of 7 action categories on this dataset. Deep semantic pyramids approach obtains a mean AP of 91.0%, which is the best results reported on this dataset [1,17,18,22–24]. The work of [17] based on using manually labeled data for enhancing the efficiency of the pre-training and fine-tuning stages of the deep feature training obtains a mean AP of 80.4%. Khan et al. [18] propose to fuse color and shape features and obtain 70.1% mean AP. The work of [1] based on multi-cue semantic pyramids obtains a mean AP of 72.1%. Our approach which augments the semantic pyramids with deep features significantly improves the performance from 72.1% to 91.0% mean AP.

Table 5 shows a comparison with state-of-the-art methods on the PASCAL VOC 2010 test set. The method of [19] based on poselets vectors achieves a mean AP of 59.7%. The color and shape fusion approach by [18] provides a mean AP of 62.4%. The work of [20] based on localizing humans and human-object relationships achieves a recognition performance of 62.0%. Learning a sparse basis of attributes and parts framework by [21] obtains a mean AP of 65.1%. The multi-cue semantic pyramids approach [1] provides a mean AP of 63.5%. On this dataset, deep semantic pyramids achieve a mean AP of 86.1%, which is the best results reported on this dataset [1,18–21]. It is worthy to mention that deep semantic pyramids method does not take into account the human-object interactions. Such approaches [20,21] are complementary and could be combined with the proposed method to further improve the results.

Table 4. Comparison of deep semantic pyramids approach with state-of-the-art results on the Willow dataset. Deep semantic pyramids provide the best results on 6 out of 7 action classes on this dataset.

	int. computer	photographing	playingmusic	ridingbike	ridinghorse	running	walking	mean AP
BOW-DPM [22]	58.2	35.4	73.2	82.4	69.6	44.5	54.2	59.6
POI [23]	56.6	37.5	72.0	90.4	75.0	59.7	57.6	64.1
DS [24]	59.7	42.6	74.6	87.8	84.2	56.1	56.5	65.9
CF [18]	61.9	48.2	76.5	90.3	84.3	64.7	64.6	70.1
EPM [15]	64.5	40.9	75.0	91.0	87.6	55.0	59.2	67.6
SC [25]	67.2	43.9	76.1	87.2	77.2	63.7	60.6	68.0
SM-SP [1]	66.8	48.0	77.5	93.8	87.9	67.2	63.3	72.1
EDM [17]	86.6	**90.5**	89.9	98.2	92.7	46.2	58.9	80.4
Our approach	**96.6**	87.0	**99.4**	**99.7**	**99.6**	**79.4**	**75.0**	**91.0**

Table 5. Comparison of deep semantic pyramids approach with state-of-the-art methods on the PASCAL VOC 2010 test set.

	phoning	playingmusic	reading	ridingbike	ridinghorse	running	takingphoto	usingcomputer	walking	mean AP
Poselets [19]	49.6	43.2	27.7	83.7	89.4	85.6	31.0	59.1	67.9	59.7
IaC [26]	45.5	54.5	31.7	75.2	88.1	76.9	32.9	64.1	62.0	59.0
POI [23]	48.6	53.1	28.6	80.1	90.7	85.8	33.5	56.1	69.6	60.7
LAP [21]	42.8	60.8	41.5	80.2	90.6	87.8	41.4	66.1	74.4	65.1
WPOI [20]	55.0	81.0	69.0	71.0	90.0	59.0	36.0	50.0	44.0	62.0
CF [18]	52.1	52.0	34.1	81.5	90.3	88.1	37.3	59.9	66.5	62.4
SM-SP [1]	52.2	55.3	35.4	81.4	91.2	89.3	38.6	59.6	68.7	63.5
Our approach	**65.1**	**94.0**	**71.9**	**97.6**	**97.7**	**93.8**	**83.3**	**93.4**	**77.2**	**86.1**

5 Conclusion

This paper combines pose-normalized semantic pyramids and deep features representation. Semantic pyramids combine information from the whole image, full-body, upper-body and face regions. We employ pre-trained body part detectors that automatically localize upper-body and face regions in an image. The use of pre-trained detectors ensures that no extra annotations are required either at training or test times. We propose to use a spatial pyramid based deep feature representation to describe each of these image regions. The final representation is obtained by combining the pyramidal feature vectors from al regions. The proposed approach is evaluated on two challenging tasks: human attributes classification and action recognition, demonstrating promising performance compared to state-of-the-art methods in literature.

Currently our approach employs pre-trained deep features from ImageNet. Future work involves learning deep features on large attributes and action datasets with a more careful optimization of network topology, choice of activation and pooling functions.

Acknowledgments. This work has been supported by SSF through a grant for the project CUAS, by VR through a grant for the projects ETT, by EU's Horizon 2020 Program through a grant for the project CENTAURO, through the Strategic Area for ICT research ELLIIT, and CADICS, project TIN2013-41751 of Spanish Ministry of Science and the Catalan project 2014 SGR 221, grants 255745 and 251170 of the Academy of Finland and *Data to Intelligence (D2I)* DIGILE SHOK project. The calculations were performed using computer resources within the Aalto University School of Science "Science-IT" project.

References

1. Khan, F.S., van de Weijer, J., Anwer, R.M., Felsberg, M., Gatta, C.: Semantic pyramids for gender and action recognition. TIP **23**(8), 3633–3645 (2014)
2. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR (2014)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
4. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: ICCV (2011)
5. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: ICCV (2013)
6. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: CVPR (2014)
7. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI **32**(9), 1627–1645 (2010)
8. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV (2009)
9. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. In: NIPS (1989)
10. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
11. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: CVPR (2014)
12. Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPRW (2014)
13. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR (2012)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
15. Sharma, G., Jurie, F., Schmid, C.: Expanded parts model for human attribute and action recognition in still images. In: CVPR (2013)
16. Joo, J., Wang, S., Zhu, S.C.: Human attribute recognition by rich appearance dictionary. In: ICCV (2013)
17. Liang, Z., Wang, X., Huang, R., Lin, L.: An expressive deep model for human action parsing from a single image. In: ICME (2014)
18. Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A., Lopez, A., Felsberg, M.: Coloring action recognition in still images. IJCV **105**(3), 205–221 (2013)
19. Maji, S., Bourdev, L.D., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR (2011)
20. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. PAMI **34**(3), 601–614 (2012)
21. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L.J., Li, F.F.: Human action recognition by learning bases of action attributes and parts. In: ICCV (2011)
22. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: BMVC (2010)

23. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: NIPS (2011)
24. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: CVPR (2012)
25. Khan, F.S., van de Weijer, J., Bagdanov, A., Felsberg, M.: Scale coding bag-of-words for action recognition. In: ICPR (2014)
26. Shapovalova, N., Gong, W., Pedersoli, M., Roca, F.X., Gonzàlez, J.: On importance of interactions and context in human action recognition. In: Vitrià, J., Sanches, J.M., Hernández, M. (eds.) IbPRIA 2011. LNCS, vol. 6669, pp. 58–66. Springer, Heidelberg (2011)