Attributes and Action Recognition Based on Convolutional Neural Networks and Spatial Pyramid VLAD Encoding

Shiyang Yan[1], Jeremy S. Smith[2], Bailing Zhang[1]

[1]Xi'an Jiaotong-Liverpool University
{shiyang.yan, bailing.zhang}@xjtlu.edu.cn
[2]University of Liverpool
J.S.Smith@liverpool.ac.uk

Abstract. Determination of human attributes and recognition of actions in still images are two related and challenging tasks in computer vision, which often appear in fine-grained domains where the distinctions between the different categories are very small. Deep Convolutional Neural Network (CNN) models have demonstrated their remarkable representational learning capability through various examples. However, the successes are very limited for attributes and action recognition as the potential of CNNs to acquire both of the global and local information of an image remains largely unexplored. This paper proposes to tackle the problem with an encoding of a spatial pyramid Vector of Locally Aggregated Descriptors (VLAD) on top of CNN features. With region proposals generated by Edgeboxes, a compact and efficient representation of an image is thus produced for subsequent prediction of attributes and classification of actions. The proposed scheme is validated with competitive results on two benchmark datasets: 90.4% mean Average Precision(mAP) on the Berkeley Attributes of People dataset and 88.5% mAP on the Stanford 40 action dataset.

1 Introduction

Human attributes descriptions such as gender, clothing style, hair style and action categories such as using a computer, riding a horse or texting messages, are two popular yet challenging recognition problems in semantic computer vision. The tasks are particularly difficult in static images partly due to the lack of motion information. Large variances in illumination conditions, view point, human pose as well as occlusion add further obstacles to finding satisfactory solutions.

The description of human attributes and the classification of action categories all depends on local and global contextual information. On the one hand, the local regions that correspond to detailed, fine-grained appearance features may play critical roles in recognition; on the other hand, the global context of the surrounding objects and scenes is also instrumental to tackle the problem. As an example, human attributes like 'gender' not only depends on local features such as the face or hair style, but also relies on the global context, for example, the clothing style and body shape. As for action category classification, the pose, the objects a person interacts with, and the scene in which the action is performed, all contain useful information. This is better illustrated by the action types in sports. For example, for the action of 'playing basketball', the basketball and playground are both strong evidence for this action category.

A typical way for compactly representing image and incorporating global contextual information is to apply a patch feature encoding strategy such as the Bag-of-Visual-Words (BoVW) [1], Fisher Vectors (FV) [2], and Vector of Locally Aggregated Descriptors (VLAD) [3]. Among these, it is reported that the Fisher Vector outperforms many popular encoding methods previously published on benchmark image datasets. VLAD can be regarded as a simplified non-probabilistic version of FV and also shows comparable performance [3]. Recently, VLAD has continued to grow in popularity in computer vision, with many excellent demonstrations for problems including object detection, scene recognition and action recognition [4], [5], [6], [7]. In this paper, we will further explore the potential of VLAD for the human attribute prediction and action classification in still images.

Conventional patch feature encoding strategies largely depend on local features such as Scale Invariant Feature Transform (SIFT) [8]. Such a mid-level feature description can be considered as 'hallow features', as no deep training is involved. Recently, convolutional neural networks (CNN) achieved breakthrough performance in many vision tasks [9], [10], [11]. Yet, it is noteworthy that CNNs are originally trained for the classification of objects [9], with the typical goal of correctly identifying a single, predominant object in an image. Hence, existing CNN models alone are limited in their capability in the description of attributes and the classification of actions despite their powerful feature representation capabilities. For action and attribute recognition, a rational strategy is to take advantages from both of the CNN and patch feature encoding strategies. Based on this intuition, we encode the CNN features for sub-regions of an image to generate a compact representation. Our approach is similar to [12] in which the Fisher Vector encoding scheme is applied on CNN features and each image is represented as a bag of windows. Likewise, our method can also be considered as bag of patches as the image patches are first extracted using region proposal algorithms such as Edgeboxes [13] and then represented with VLAD encoding.

As previously explained, VLAD encoding combines local and global features by generating image representations that better reflect the local information of the image patches. A region proposal algorithm would be instrumental to generate object-like local patches. Through VLAD encoding, the CNN features of many regions will be encoded into a higher level description closer to the images inherent signature. Such a compact signature preserves most of the information from CNNs while reducing the dimensionality. More importantly, the fine-grained properties of an image could be retained with the proposed VLAD encoding.

The only downside to VLAD encoding is its lack of preservation of spatial information [14]. To compensate for this, we firstly build spatial pyramids of the image which are matched to region level CNN features, and then perform VLAD encoding on the separate pyramid. The generated VLAD codes are concatenated into a final representation, which is subsequently forwarded to a classifier, e.g. a SVM, for final classification. We conducted extensive experiments including various comparisons, and achieved promising results on the Berkeley Attributes of People dataset [15] and the Stanford 40 action dataset [16]. To the best of our knowledge, we are the first to apply a spatial pyramid VLAD encoding scheme for attribute and action recognition in still images.

2 Related works

2.1 Human Attributes and Action Recognition

Attributes as visual qualities of objects provide a bridge between lower level image representations and higher level semantic information. Accordingly, attribute learning has become important in many computer vision applications, for example, face verification [17], clothing description [18] and image retrieval [19]. At the same time, action recognition from still images has recently attracted more attention, because many action categories can be unambiguously defined in static images without motion information. The potential applications are obvious, for example, image annotation, image retrieval and human computer interaction.

The issue of the recognition of human attributes and actions has been researched for many years in computer vision and machine learning. A common practice is to apply the Bag-of-visual-words (BoVW) [20] [21] [22], which is advantageous in the global representation of an image. Vincent Delaitre et al. [23] handled the problem by applying a bag-of-features and part-based representation. Recently, as an extension of BoVW, Fisher Vector [2] and VLAD have been gaining ground in many vision problems, including video event classification [24] and action recognition [25].

For many vision problems, another influential train of thought is on part-based modelling, which has also witnessed some successes. Among them, the Deformable Part Model(DPM) [26] is a milestone in the development. Similarly, Poselets method [27]used a part-based template to interpolate a pose. Recently, Zhang et al. [28] proposed a Pose Aligned Networks Model (PANDA), which is a combination of deep learning model and Poselets, and demonstrated its capability in capturing human attributes.

For action recognition from still images, another common approach is centering on human-object interaction. Yao and Fei Fei [29] detected objects in cluttered scenes and estimated articulated human body parts in human object interaction activities. Alessandro Prest et al. [30] introduced a weakly supervised approach for learning human actions modeled as interactions between humans and objects. In these previous studies, the global scene information has not been taken into account, which is one of the challenges this paper addresses.

2.2 Deep Learning Powered Approaches

Recently, deep learning methods, especially deep convolutional neural networks, have dramatically improved the state-of-the-art in visual object recognition [9], object detection [10] [31], image segmentation [32] and many other vision tasks. For the task of action recognition, Oquab et al. [33] investigated transfer learning with a CNN model, showing that a CNN trained from a large dataset can be transferred to another visual recognition task when limited training data is available. Promising results were reported for action recognition [33], with the advantages of higher accuracy and a shorter training period. Gkioxari et al. [34] developed a part-based approach by leveraging trained deep body parts detectors for action and attribute classification. They showed state-of-the-art performance on the Berkeley Attributes of People dataset. Gkioxari et al. [11] proposed a joint training procedure for the automatic discovery of the top contributing contextual information for action recognition and attribute classification, which showed state-of-the-art results on several publicly available datasets. Recently, Ali Diba et al. proposed DeepCAMP [35], a scheme that utilizes CNN mid-level patterns for action recognition and attribute determination, which is also showing promising results.

Our work is different from these as we extract CNN features for post-processing using spatial pyramid VLAD coding. VLAD coding [36] [3], along with Fisher Vectors [37], is mostly applied in image classification or retrieval tasks [38][12]. Compared with BoVW and Fisher Vector, VLAD can be more balanced between memory usage and performance [3]. However, the main downside of VLAD, the lack of spatial information, has been less stressed. A well-known approach of encoding spatial information was proposed by Lazebnik et al. [39] by taking into account the spatial layout of keypoints in a pooling step, which divides the image into increasingly finer spatial sub-regions and creates histograms for each sub-region separately. In [40], the authors proposed to combine spatial pyramids and VLAD. Recently, Andrew Shin et al. [14] further examined the approach for image captioning. We applied the similar methods of [40][14] by extracting deep activation features from local patches at multiple scales, and coding them with VLAD. However, our approach extended much beyond the scene classification and object classification in [40][14], in which the significance of explicitly dealing with local objects and spatial information is less evident.

3 Methods

In this section, we describe the main components of our proposed pipeline, starting with region-based feature extraction after EdgeBoxes, and ending with Spatial Pyramid VLAD encoding for the attributes and actions to be classified. Pictorially, the overall workflow of our model can be described by Fig.1.

3.1 Feature extraction

Inspired by the recent successes of CNN-based object detection, which relies on category independent region proposals, we also start from a set of region

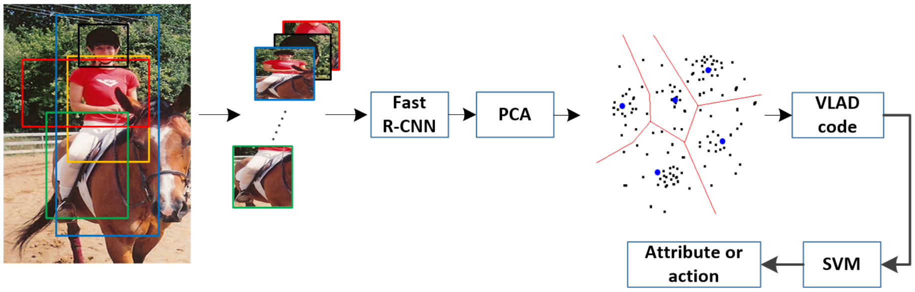


Fig. 1. Full pipeline of the proposed methods: Each window is generated by region proposal algorithm and represented by fc6 features, Principle Component Analysis (PCA) is applied for dimension reduction, followed by k-means for centroid learning(the blue dots). Attributes and action can thus by classified with VLAD codes and a classifier.

proposals from images to pursue accuracy with an affordable computation cost [13].

Among the recently published off-the-shelf technologies that follow this paradigm, we empirically choose EdgeBoxes [13] as the first component in our workflow to produce high-quality region proposals, mainly based on the computational efficiency and high-level performance in terms of localizing objects [41]. We consider each image as a bag, the object proposals from an image are considered as a bag of patches or windows. Thus a compact method for representing the information can be obtained through VLAD encoding. For the sake of efficiency, we simply extract 1000 patches per image.

While [14] directly exploited a CNN pre-trained on ImageNet as a generic extractor, we added an extra step of fine-tuning the pre-trained VGG16 model [42], which was conducted on top of the fast R-CNN [31] with candidate regions from EdgeBoxes. We annotated the target bounding boxes and corresponding labels, then performed training based on fast R-CNN. As verified by our experiments, this process yields a better category dependent feature representation capability.

We then use the CNN model as a generic feature extractor for 1000 image patches generated by EdgeBoxes. The top 1000 boxes have higher possibilities of containing objects. For the same reason in [14], we do not apply non-maximum suppression. The generated region proposals have arbitrary size, a common way to extract features using a CNN would be resizing them and forwarding them to a CNN model one by one. However, feature extraction of multiple regions in a CNN can be really time-consuming. Hence, we implemented our algorithm on top of the fast R-CNN in which the RoI projection and RoI pooling scheme [31] enable feature extraction of arbitrary size windows of one image in only one feed forward process, thus the computational cost can be much reduced.

3.2 Spatial Pyramid VLAD

While it is well-recognized that the VLAD encoding scheme performs well in preserving local features, it discards the spatial information. To tackle this problem, several recent papers [40] [14] proposed spatial pyramid VLAD as a solution. In this paper, we followed the methodology with an efficient implementation for attribute prediction and action classification. More specifically, as shown in Fig.2, a 3 level (1X1, 2X2, and 4X1) spatial pyramid is exploited. To allocate regions into each spatial grid, we assign each region according to the distribution of their centers. This simple yet effective approach can avoid overlapping and is discriminatively powerful.

For the 4096 dimensional features extracted from the CNN, one possibility is to perform VLAD encoding for each spatial pyramid separately. However, the 4096 dimensions would be too large to encode. As pointed out in [43], dimension reduction marginally affects the overall representation of VLAD. Hence, we perform dimensionality reduction with Principle Component Analysis (PCA) on the CNN features of each region. As the number of features is large, training a conventional PCA on all the features would be unrealistic. An alternative would be randomly selecting a certain number of features for training, and then performing PCA on all features. Here, we chose to implement incremental PCA [44] on all the features because of its high efficiency in terms of memory usage. We perform PCA on all the features with dimension of 256. Also, to study the influence of dimension on overall performance, we also reduced the features to 512 for comparison.

The last step of the encoding is similar to BoVW, i.e., code word learning implemented by unsupervised learning like k-means clustering. The number of clusters was set to 12, 16, 24, and 64 for testing. We also followed the practice of [14] by exploiting k-means++ [45] due to its improved speed and the accuracy from the randomized seeding technique. After obtaining the code words from k-means++, VLAD coding with L2-normalization can thus be implemented. The final dimensionality of a VLAD code is the number of clusters times the length of PCA reduced CNN features.

4 Experiments

4.1 Deep Learning Model

Experiments were conducted using the Caffe CNN platform [46]. A CNN model was first pre-trained using the ImageNet dataset which was subsequently fine-tuned using our specific datasets for the different tasks, as described in the previous section. We set the max training iteration at 40000, the other parameters were the same as for the original fast R-CNN.

4.2 VLAD Encoding

The incremental PCA and k-means++ are all implemented on top of the Scikit-learn Python machine learning package [47]. With the obtained dimensionality-

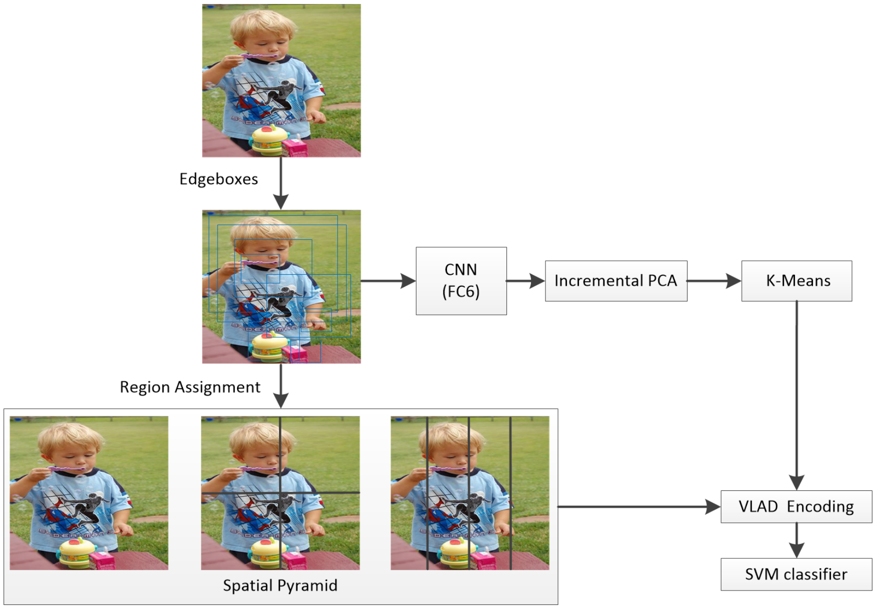


Fig. 2. VLAD encoding with a spatial pyramid: The image was divide with a 3 level spatial pyramid: 1X1, 2X2 and 4X1. Each pyramid is encoded separately with VLAD.

reduced features and codewords, VLAD encoding was implemented in Matlab using the VLfeat toolbox [48]. For attribute classification, a SVM linear classifier was utilized using the LIBSVM toolbox [49], while for the action classification, we implemented a multi-layer perceptron based on the Matlab Neural Network Toolbox.

4.3 Attribute Recognition

We evaluated our method using the Berkeley Attributes of People Dataset [15]. The dataset includes 4013 images for training, and 4022 test images collected from the PASCAL and H3D datasets. This is a very challenging dataset as the people in the images often have large appearance variance and occlusion. Few reported methods worked well with this dataset [15][28].

As previous explained above, we first applied the spatial pyramid VLAD encoding, and then employed a SVM classifier for the final prediction. Specifically, the pre-trained VGG16 model [42] was used for fine-tuning. The training process was implemented in a fast R-CNN [31] framework. We then run Edgeboxes on each image, and extracted the features of the first fully connected layers (fc6) of each sub-region. VLAD encoding is then accomplished after PCA dimensionality reduction and codeword learning with k-means++.

More details about the experiments are explained in the following section. First of all, the first fully connected layers(fc6) CNN features of the ground

(a)		(b)	
Male:0.0064,	No	Male:0.8289,	Yes
Long-hair:0.7897,	Yes	Long-hair:0.0748,	No
Glasses:0.0643,	No	Glasses:not-certain,	Not-certain
Hat:0.0001,	No	Hat:0.9430,	Yes
T-shirt:0.6533,	No	T-shirt:0.0416,	No
Long-sleeves:0.0054,	No	Long-sleeves:0.2946,	No
Shorts:not-certain,	Not-certain	Shorts:0.0030,	No
Jeans:0.0276,	No	Jeans:0.0126,	No
Long-pants:0.0088	No	Long-pants:0.9929,	Yes

Fig. 3. Examples of attribute classification: the probabilities of certain attributes are provided, the blue text are the ground truth label. The red text shows an incorrect classification example. The threshold 0.5 was applied as a standard for classification.

truth region are extracted, and directly applied for attribute classification as a comparison baseline. As shown in Table.1, the mean average precision (mAP) is 78.1%, which means it is not very accurate in representing the attributes associated with the image based on the primitive CNN feature. Secondly, to evaluate the stand-alone performance of VLAD encoding, we encoded each images with CNN features of 256 dimensions and 16 learnt clusters into the VLAD code. The mean average precision is 78.3%. In this settings, we did not use the ground truth region for classification, and spatial pyramid are also not applied. Concatenating the VLAD code with CNN features yields an 8.5% rise in performance, which implies that the combination of local patches features and compact global representation all contribute to attribute recognition. Finally, to examine the influence of spatial pyramid coding, experiments were organized with results confirming that adding spatial pyramid coding does improve the overall performance, by 3.6% in the mean average precision.

To evaluate the influence of the number of k-means clusters, we performed VLAD encoding with 12, 16, 24, 64 centroids separately. It can be observed that 16 clusters works the best in our experiments as shown in Table 2. Also, we repeated the experiments with dimensionality-reduced CNN features of 512, which produced similar results. As noted in [43], dimension reduction plays a signifi-

Table 1. The average precision results on the Berkeley Attributes of People test Dataset and comparison with different approaches. The results show that with the combination of the original CNN fc6 features, there is a gain of 8.5%. Moreover, with the spatial pyramid, the mean average precision is improved by 3.6%.

Attribute	male	long hair	glasses	hat	tshirt	longsleeves	shorts	jeans	long pants	Mean AP
fc6 features of ground truth region	90.1	80.8	77.6	80.6	57.4	84.2	64.9	71.1	96.5	78.1
PCA 256+16 clusters (No Spatial Pyramid)	88.9	76.4	74.7	68.2	68.5	88.5	73.3	71.8	94.2	78.3
PCA 256+16 clusters+fc6 features (No Spatial Pyramid)	92.5	87.4	85.2	90.4	68.3	89.7	85.5	83.9	98.0	86.8
PCA 256+16 clusters+fc6 features (With Spatial Pyramid)	**94.1**	**90.4**	**89.4**	**94.0**	**74.0**	**92.5**	**91.9**	**88.6**	**98.5**	**90.4**

cant role in VLAD encoding, the lower dimensional CNN features may improve the performance. Therefore, we set the CNN features as 256 dimensionality in most of the experiments.

We also compared our results with existing methods. As illustrated in Table 2, our methods outperformed the published results listed in the table. In Fig.3, we provide some examples of recognized attributes on the datasets.

4.4 Action Recognition

To evaluate the system performance on action recognition, we experimented on the Stanford 40 action dataset [16]. which contains 9532 images corresponding to 40 classes of actions. The dataset was split into a training set with 4000 images, and a testing set with 5532 instances. There are 180-300 images for each class. The images from each class have large variations in human pose, appearance, and background clutter.

We extract the CNN features(fc6) and forward them directly to a Multi-layer Perceptron (MLP) classifier as the baseline for comparison. As the parameters of 256 dimensionality of the CNN features and 16 clusters achieved the best results in the previous experiments for attribute recognition, we directly took the same parameters in action recognition. As shown in Fig.4, our proposed spatial pyramid VLAD encoding scheme outperforms the primitive CNN features in all action classes except the ‘riding a bike’ class. In this action class, the performance results are similar. More importantly, VLAD performs better for the more fine-grained action classes, for instance, ‘writing on a board’. This is because VLAD encoding can retain the local information from small patches, and spatial pyramid coding can form a more compact representation of an image.

The comparison with previous reported methods is shown in Table 3, which demonstrates that our method outperforms all of the previously published work as listed in the table. It is noteworthy that F.S. Khan et al. [53] did not utilize a ground truth bounding box during action recognition, to impartially compare with their results, we also experimented without the ground truth region. With a setting of 256 dimensions from PCA and 16 word codes from clustering, our

Table 2. The average precision results of the Berkeley Attributes of People Dataset and comparison with previous methods. We provide results on 256 dimensionality of CNN features after PCA with 12, 16, 24 and 64 clusters of k-means. We also perform VLAD encoding on 512 dimensionality, there is not much differences in terms of performance. Performing VLAD encoding on 256 dimensionality CNN features with 16 clusters yields the best results.

Attribute	male	long hair	glasses	hat	tshirt	longsleeves	shorts	jeans	long pants	Mean AP
Poselets[15]	82.4	72.5	55.6	60.1	51.2	74.2	45.5	54.7	90.3	65.0
PANDA[28]	91.7	82.7	70.0	74.2	49.8	86.0	79.1	81.0	96.4	79.0
R*CNN[11]	92.8	88.9	82.4	92.2	**74.8**	91.2	92.9	89.4	97.9	89.2
Gkioxari et al.[34]	92.9	90.1	77.7	93.6	72.6	**93.2**	**93.9**	**92.1**	**98.8**	89.5
Ours (PCA 256+12 clusters+fc6 features)	93.8	90.0	88.5	93.4	72.9	92.2	90.8	87.7	98.4	89.7
Ours (PCA 256+64 clusters+fc6 features)	93.8	**92.2**	89.1	93.8	73.1	92.1	91.4	87.8	98.4	90.0
Ours (PCA 256+24 clusters+fc6 features)	**94.1**	90.4	**89.5**	**94.0**	73.8	92.5	91.9	88.5	98.4	90.3
Ours (PCA 256+16 clusters+fc6 features)	**94.1**	90.4	89.4	**94.0**	74.0	92.5	91.9	88.6	98.5	**90.4**

Table 3. Mean average precision results on the Stanford 40 action dataset and comparison with previous results.

Method	Mean AP
Object bank [50]	32.5
LLC [51]	35.2
EPM [52]	40.7
DeepCAMP[35]	52.6
F.S. Khan et al. [53]	75.4
Ours(fc6 features of ground truth region)	81.2
Ours(PCA 256+16 clusters)	85.9
Ours(PCA 256+16 clusters+fc6 features)	**88.5**

proposed method yields a 10.5% increase in the mean average precision compared to [53]. Some of the correctly recognized examples are shown in Fig.5.

5 Conclusion

Human attribute and action recognition in static images are challenging tasks, with the main challenges being fine-grained recognition without motion information. How to efficiently exploit both the global features and local features is key to solving these problems. In this paper, we applied Vector of Locally Aggregated Descriptors on top of spatial pyramids to detect local information,

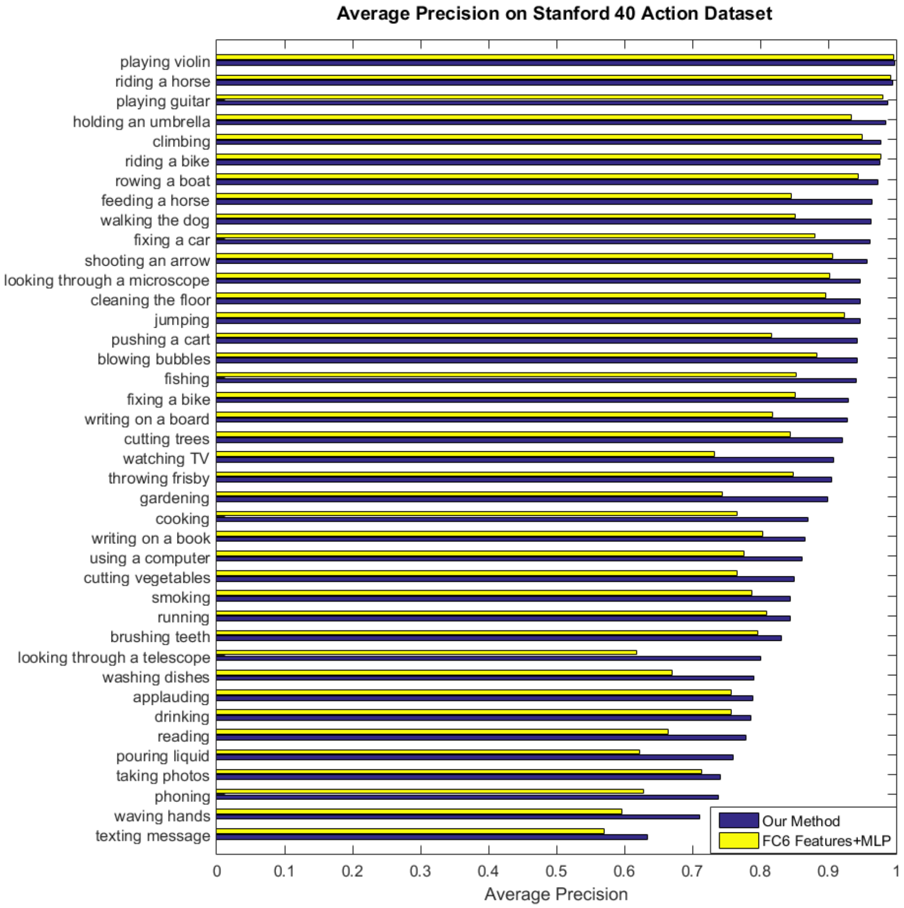
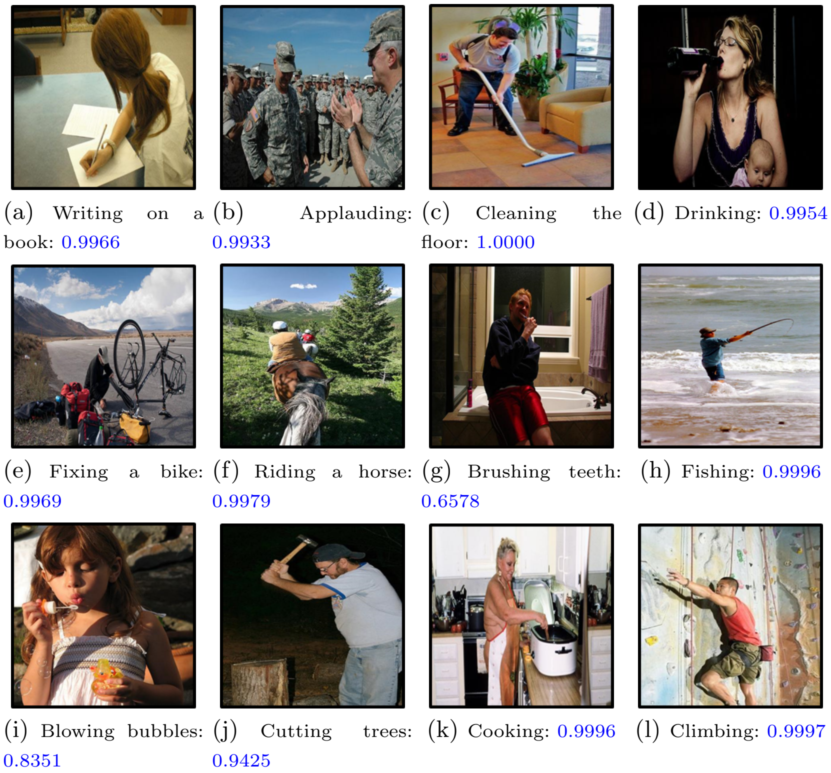


Fig. 4. Our best average precision results on the Stanford 40 action dataset and comparison with the baseline approach.

not only from the ground truth region but also from nearby objects and scenes. Experiments confirmed that the combination of CNN features and VLAD codes is very effective in retaining both local and global information. As we encode CNN features on the first fully connected layer, the next step is to explore the possibility of directly encoding the original CNN features.

References

1. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Volume 2. (2005) 524–531 vol. 2



(a) Writing on a book: 0.9966
(b) Applauding: 0.9933
(c) Cleaning the floor: 1.0000
(d) Drinking: 0.9954
(e) Fixing a bike: 0.9969
(f) Riding a horse: 0.9979
(g) Brushing teeth: 0.6578
(h) Fishing: 0.9996
(i) Blowing bubbles: 0.8351
(j) Cutting trees: 0.9425
(k) Cooking: 0.9996
(l) Climbing: 0.9997

Fig. 5. Some examples of correct recognition in the Stanford 40 action dataset: The predicted label and corresponding confidence values are provided.

2. Csurka, G., Perronnin, F.: Fisher vectors: Beyond bag-of-visual-words image representations. In: Computer Vision, Imaging and Computer Graphics. Theory and Applications. Springer (2010) 28–42
3. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 3304–3311
4. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. (2012) 3506–3513
5. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: Proceedings of the British Machine Vision Conference, BMVA Press (2010) 97.1–97.11 doi:10.5244/C.24.97.
6. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision **103** (2013) 60–79
7. Peng, X., Zou, C., Qiao, Y., Peng, Q.: Action recognition with stacked fisher vectors. In: Computer Vision–ECCV 2014. Springer (2014) 581–595
8. Lowe, D.G.: Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on.

Volume 2., IEEE (1999) 1150–1157

9. Alex Krizhevsky, Ilya Sutskever, G.E.: Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems (2012)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. (2014) 580–587
11. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with r* cnn. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1080–1088
12. Uricchio, T., Bertini, M., Seidenari, L., Bimbo, A.D.: Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). (2015) 1020–1026
13. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014. Springer (2014) 391–405
14. Shin, A., Yamaguchi, M., Ohnishi, K., Harada, T.: Dense image representation with spatial pyramid vlad coding of cnn for locally robust captioning. arXiv preprint arXiv:1603.09046 (2016)
15. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: 2011 International Conference on Computer Vision. (2011) 1543–1550
16. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 1331–1338
17. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and Simile Classifiers for Face Verification. In: IEEE International Conference on Computer Vision (ICCV). (2009)
18. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Computer Vision–ECCV 2012. Springer (2012) 609–623
19. Cai, J., Zha, Z.J., Zhou, W., Tian, Q.: Attribute-assisted reranking for web image retrieval. In: Proceedings of the 20th ACM international conference on Multimedia, ACM (2012) 873–876
20. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. arXiv preprint arXiv:1405.4506 (2014)
21. Oneata, D., Verbeek, J., Schmid, C.: Action and event recognition with fisher vectors on a compact feature set. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 1817–1824
22. Ullah, M.M., Parizi, S.N., Laptev, I.: Improving bag-of-features action recognition with non-local cues. In: BMVC. Volume 10., Citeseer (2010) 95–1
23. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. (2010) updated version, available at http://www.di.ens.fr/willow/research/stillactions/.
24. Sun, C., Nevatia, R.: Large-scale web video event classification by use of fisher vectors. In: Applications of Computer Vision (WACV), 2013 IEEE Workshop on, IEEE (2013) 15–22
25. Jain, M., Jégou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 2555–2562

26. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32** (2010) 1627–1645
27. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 1365–1372
28. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1637–1644
29. Yao, B., Fei-Fei, L.: Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 1691–1703
30. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 601–614
31. Girshick, R.: Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV). (2015) 1440–1448
32. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). (2015) 1529–1537
33. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1717–1724
34. Gkioxari, G., Girshick, R., Malik, J.: Actions and attributes from wholes and parts. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2470–2478
35. Diba, A., Pazandeh, A.M., Pirsiavash, H., Van Gool, L.: (Deepcamp: Deep convolutional action & attribute mid-level patterns)
36. Arandjelovic, R., Zisserman, A.: All about vlad. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. (2013) 1578–1585
37. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. International journal of computer vision **105** (2013) 222–245
38. Dixit, M., Chen, S., Gao, D., Rasiwasia, N., Vasconcelos, N.: Scene classification with semantic fisher vectors. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 2974–2983
39. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 2169–2178
40. Zhou, R., Yuan, Q., Gu, X., Zhang, D.: Spatial pyramid vlad. In: Visual Communications and Image Processing Conference, 2014 IEEE, IEEE (2014) 342–345
41. Hosang, J., Benenson, R., Schiele, B.: How good are detection proposals, really? In: 25th British Machine Vision Conference, BMVA Press (2014) 1–12
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
43. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision & Pattern Recognition. (2010)
44. Ross, D.A., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. International Journal of Computer Vision **77** (2008) 125–141

45. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics (2007) 1027–1035
46. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, ACM (2014) 675–678
47. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12** (2011) 2825–2830
48. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008)
49. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27 Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
50. Li, L.J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Advances in neural information processing systems. (2010) 1378–1386
51. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 3360–3367
52. Sharma, G., Jurie, F., Schmid, C.: Expanded parts model for human attribute and action recognition in still images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 652–659
53. Khan, F.S., Xu, J., van de Weijer, J., Bagdanov, A.D., Anwer, R.M., Lopez, A.M.: Recognizing actions through action-specific person detection. Image Processing, IEEE Transactions on **24** (2015) 4422–4432