Action Recognition from Still Images Based on Deep VLAD Spatial Pyramids

Shiyang Yan[a,b,*], Jeremy S. Smith[a], Bailing Zhang[b]

[a] *Electrical Engineering and Electronic, University of Liverpool, Liverpool, United Kingdom*
[b] *Department of Computer Science and Software Engineering, Xi'an Jiaotong-liverpool University, SuZhou, JiangSu Province, China*

Abstract

The recognition of human actions in images is a challenging task in computer vision. In many applications, actions can be exploited as mid-level semantic features for high level tasks. Actions often appear in fine-grained categorization, where the differences between two categories are small. Recently, deep learning approaches have achieved great success in many vision tasks, e.g., image classification, object detection, and attribute and action recognition. Also, the Bag-of-Visual-Words (BoVW) and its extensions, e.g., Vector of Locally Aggregated Descriptors (VLAD) encoding, have proved to be powerful in identifying global contextual information. In this paper, we propose a new action recognition scheme by combining the powerful feature representational capabilities of Convolutional Neural Networks (CNNs) with the VLAD encoding scheme. Specifically, we encode the CNN features of image patches generated by a region proposal algorithm with VLAD and subsequently represent an image by the compact code, which not only captures the more fine-grained properties of the images but also contains global contextual information. To identify the spatial information, we exploit the spatial pyramid representation and encode CNN features inside each pyramid. Experiments have verified that the proposed schemes are not only suitable for action recognition but also applicable

*Corresponding author
Email addresses: `Shiyang.Yan@xjtlu.edu.cn` (Shiyang Yan), `J.S.Smith@liverpool.ac.uk` (Jeremy S. Smith), `Bailing.Zhang@xjtlu.edu.cn` (Bailing Zhang)

to more general recognition tasks such as attribute classification. The proposed scheme is validated with four benchmark datasets with competitive mAP results of 88.5% on the Stanford 40 Action dataset, 81.3% on the People Playing Musical Instrument dataset, 90.4% on the Berkeley Attributes of People dataset and 74.2% on the 27 Human Attributes dataset.

Keywords: Actions, Convolutional Neural Networks, VLAD encoding, Spatial Pyramids

1. Introduction

In computer vision, many human actions such as ‘using a mobile phone’, ‘riding a bike’ or ‘reading a book’, provide a natural description for many still images, which could provide significant meta-data to many applications such as automatic scene description, and the indexing and searching of very large image repositories. Compared with more well-established video-based action recognition, these tasks are more difficult as there are a number of possible obstacles to find the satisfactory solutions, e.g., large variances in illumination conditions, the viewpoint, and the human pose, and more importantly, lack of motions.

Unlike the video-based action recognition which heavily relies on the spatial-temporal features, the solutions to human action classification from still images hinge on the acquisition of local and global contextual information. To be more specific, local information associated with discriminative parts provides detailed appearance features which would be particularly pertinent to fine-grained recognition. This is because human actions are often localized in space, e.g., the facial region for expressions and the wrist and hand regions for many common actions. Additionally, the global contextual information about the configuration of objects and scenes is also instrumental. For example, the articulation of body parts, the pose, the objects a person interacts with and the scene in which the action is performed, all contain useful information. This is well illustrated by the action types in sports. For example, for the action of ‘playing football’, the football itself and playground are both strong evidence for this action category.

To represent the contextual information of images, many methods have been proposed. Bangpeng et al. [1] proposed to use probabilistic graphical models, e.g, conditional random fields, to model the mutual contextual information. In this approach, the objects and humans or human body parts are described as nodes in conditional random fields. By modeling the conditional probabilities, the system can generate labels by discriminating not only on input features but also on the relationships between them.

Compared to holistic contextual features, local features or patches have the advantage of being more robust to misalignment and occlusions, and have been widely used for generic image classification. Popular local feature or patches encoding strategies include the Bag of Visual Words (BoVW) [2], Fisher Vectors (FV) [3], and Vector of Locally Aggregated Descriptors (VLAD) [4]. Among these, the FV often perform best on a number of benchmark image datasets. VLAD aggregates information of several features such as Scale-Invariant Feature Transform (SIFT) into a compact and fixed length descriptor, which can be regarded as a simplified non-probabilistic version of FV and also show comparable performance [5]. Another advantage of VLAD is its computational efficiency as it mainly involves primitive operations [6]. Recently, VLAD has been widely applied in computer vision, demonstrating an excellent performance in many tasks including object detection, scene recognition and action recognition [7], [8], [9], [10].

While the dominate patch encoding strategies are all based on hand-crafted features, deep neural networks, and Convolutional Neural Networks (CNN) in particular, emphasize the significance of learning robust feature representations from raw data. Krizhevsky et al. [11] shown that CNNs trained with large amounts of labeled data outperforms FV. Since then CNNs have consistently led the classification task in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [12]. Much of the published work considered the problem of incorporating contextual information in the CNN framework. For example, recurrent neural networks (RNNs) have been proposed to embed the contextual information into CNNs. Bell et al. [13] proposed a deep CNN structure by plugging in the RNNs to integrate contextual information for object detection. In [14], a conditional random field was formulated as RNNs and plugged into the CNN model, which was optimised using mean field for image semantic segmentation.

To date, convolutional neural networks (CNNs) have achieved a considerable success in many vision tasks [11], [15], [16]. Despite these achievements, deep CNN architectures meet with new challenges, which include the requirement for large amounts of training data, and the high computational cost with solutions relying on GPUs and other hardware acceleration techniques. Additionally, Convolutional Neural Networks still have some limitations, e.g., their lack of geometric invariance and their inability in conveying information on local elements. A promising direction for their improvement is to combine the CNN with traditional encoding approaches like VLAD to better express the local information of the images [17], [18], [19]. For example, Gong et al. [5] extracted CNN activations at multiple scale levels, and performed orderless VLAD pooling separately, which were then concatenated together to form a high dimensional feature vector which is more robust to global deformations.

In this paper, we follow that direction to further explore the potential of augmenting CNN with VLAD in the context of human action classification in still images. To take advantages from both CNN and the patch feature encoding strategy, we encode the CNN features upon sub-regions of the image for a compact representation. Our approach shares similarities with [19], in which the FV encoding scheme was applied on CNN features and each image was represented as a bag of windows. Our method can also be regarded as a bag of patches or windows as the image patches are extracted using region proposal algorithms such as Edgeboxes [20], which are subsequently encoded by VLAD for image representation.

Aiming to preserve crucial local features and identify contextual information from neighbouring objects and scenes, the proposed approach is more likely to capture the fine-grained properties of an image than the conventional approaches. To take account of the spatial information which is absent in VLAD [17], spatial pyramids of the image were generated and matched to region level CNN features. Then, VLAD encoding was applied on separate pyramids with the resulting VLAD codes concatenated and forwarded to a classifier for final classification. With extensive experiments, we achieved state-of-the-art results on the Stanford 40 action dataset [1] and People Playing Musical Instrument dataset (PPMI) [21].

For many tasks in computer vision such as video surveillance, image search and human-computer interaction, objects can often be conveniently identified by a set of mid-level, nameable descriptions termed as semantic attributes or attributes [22]. For example, a human object can be described by hair-length, eye color, clothing style, gender, ethnicity and age. Therefore, recognition of visual attributes often directly leads to many high-level tasks. To give an intuition that our proposed approach can also be generalized to attribute classification, we conducted experiments on Berkeley Attributes of People dataset [22] and the 27 Human Attributes dataset (HAT) [23], with promising results.

The rest of the paper is organized as follows. In section 2, we briefly introduce previous research in action classification, which is followed by our proposed approach explained in section 3. Section 4 provides our experimental procedure and presents results to prove the effectiveness of the proposed approach on attributes classification, with the conclusions presented in section 5.

2. Related works

2.1. Action Recognition

Still image-based human action recognition has been much addressed in recent years [24], [16], [25] due to the potential for providing useful meta-data to many applications such as image understanding, human-computer interaction and the indexing and searching of large-scale image archives.

The most popular conventional method for the task is the BoVW [26], [18], [27], which is capable of achieving a global representation of an image. Delaitre et al. [28] applied a BoVW for image representation and an SVM classifier for action recognition in still images. Later on, two extensions of BoVW, namely, FV [3] and VLAD, have attracted wide attention due to their advantages. Sun et al. [29] utilized FV in large-scale web video event classification. Jain et al. [30] combined the dense trajectory descriptors with new features computed from optical flow, and encoded them using VLAD for final action recognition. However, a significant problem with FV and VLAD is the absence of spatial layout information. A number of methods have been proposed to overcome the problem by incorporating spatial information into the BoVW representation. For example, the issue was addressed by Savarese et al. [31] with a BoVW encoding over spatially neighbouring image regions. A related problem to learn discriminative spatial representation for image classification, action and attributes recognition was emphasized by Sharma et al. [23]. Fahad et al. [32] directly utilized CNN features and semantic pyramids for action and attribute recognition, achieving impressive results on several datasets.

A special feature of action recognition is the modelling of a human-object interaction. Yao and FeiFei [33] exploit both pose information and the objects people interact with in the context of object-action interaction. Prest et al. [34] proposed a weakly supervised learning scenario for learning the relationship between humans and objects. Though some satisfactory results have been achieved, ignorance of the background or scene information limits the approaches to human-object interaction.

Also, when a person is interacting with objects, it is often termed activity recognition [35], [36], [37]. This is normally addressed in egocentric videos. For instance, with the aid of the saliency-based object recognition and contextual information incorporation, Diaz et al.[35] recognized activities in egocentric videos in the instrumental activities of daily living for medical research. Crispim-Junior et al. [36] proposed a hybrid framework with a concept-based knowledge framework and a probabilistic inference method for activity recognition in egocentric videos, with promising results. Karaman et al. [37] also worked on this domain with a Hierarchical Hidden Markov Model for the purpose of dementia studies.

For more general action recognition, part-based modeling has been one of the mainstream paradigms, with the Deformable Part Model (DPM) [38] as the most influential one. The Poselets model [39], which employs key points to build an ensemble model of human body parts, achieves improved performance in some vision tasks. The model proposed by Gkioxari et al. [25] combines CNNs and Poselets, for human action and attributes classification. However, Poselets need strong supervision and extensive annotations on key body parts are necessary which is time-consuming and labor intensive.

2.2. Deep Learning Powered Approach

In the last two years, visual object classification, detection and many other vision tasks have advanced quickly with the application of deep learning and CNNs [11], [15] [40], [41]. For action recognition, Oquab et al. [42] investigated the transfer learning [43] capability from a pre-trained CNN model. Transfer learning, allows the domains, tasks, and distributions involved in training and testing to be different [43]. Oquab et al. [42] showed that the pre-trained CNN parameters can be adapted to new domains of data by only retraining the classifier. Gkioxari et al. [25] emphasised the importance of parts for the tasks of action and attribute classification and developed a part-based approach by leveraging convolutional network features, with the effectiveness being experimentally confirmed on the Berkeley Attributes of People dataset. Gkioxari et al. [16] also used a scheme similar to R-CNN [15], by combining context with deep networks for two tasks, namely, action classification and detection. Recently, Diba et al. [44] proposed a method for action recognition and attribute determination by mining CNN mid-level patterns, which also showed promising results.

Compared with the previous approaches, we emphasize the importance of spatial pyramid VLAD coding on CNN features for action recognition. VLAD [45], [4], and FV [46], have been mainly applied in image classification or retrieval [47], [19]. With the accumulation of residuals on each visual word concatenated into a single vector, VLAD achieves reasonable trade-offs on both search accuracy and memory usage [4]. Also, VLAD coding is ignorant of spatial information, which has not been sufficiently stressed. The conventionally popular approach of encoding spatial information is spatial pyramid matching (SPM) by Lazebnik et al. [49] which was leveraged by Zhou et al. [50] in their proposal of spatial pyramid VLAD. The methodology was further developed by Shin et al. [17] for image captioning. [48] proposed a unified deep CNN model by implementing VLAD encoding as a layer for a weakly-supervised place recognition. However, their system performance largely depends on the initialization value of clusters. Hence, in this paper, instead of developing a homogeneous system, following a similar train of thought of [17], we extracted deep activation features from local patches at multiple scales, and then coded them with VLAD. While the emphasis of [50] and [17] was on scene classification and object classification, our focus is on the explicit abstraction of local objects and their corresponding spatial information, which was not obviously evident in [50], [17].

3. Methods

In this section, the main components of the proposed method will be described, which include patch generation, deep feature extraction and Spatial Pyramid VLAD encoding. The system pipeline is illustrated in Fig.1.

3.1. Deep Feature Extraction

Region proposals have become a standard practice for many vision tasks involving object detection as a component. In our proposed scheme, a set of image regions are generated using a bottom-up object proposal algorithm. From the recently published work, we applied Edgeboxes [20] because of its computational efficiency and high-level performance [51].

Different from Shin et al. [17], in which the pre-trained ImageNet model [52] was directly applied for feature extraction, we further fine-tuned the CNN model with the labelled candidate regions provided by Edgeboxes, this is beneficial to

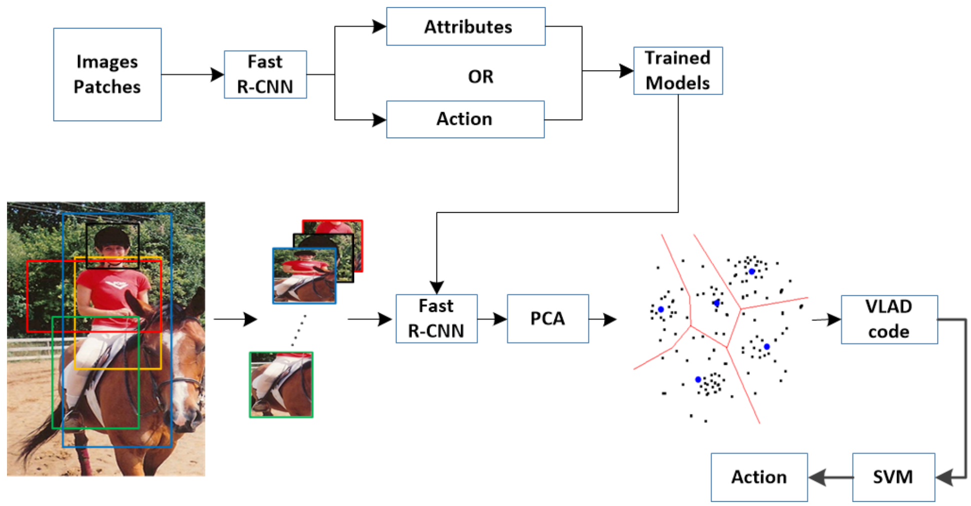


Figure 1: Full pipeline of the proposed method: Each window is generated by a region proposal algorithm and represented by FC6 features, Principle Component Analysis(PCA) is applied for dimension reduction, followed by K-means for centroid learning(the larger blue dots). Actions can thus be classified with VLAD code and a SVM classifier.

the performance improvement. During training, all boxes extracted from the original image using the Edgeboxes algorithm acted as candidate regions for the fine-tuning of the fast R-CNN framework. In our work, the VGG16 model from [53] was applied for action classification. Further details of the model architecture are outlined in Table.1. For the task of action classification, as it is essentially a multi-class classification problem, Softmax Loss layer(Softmax activation with cross-entropy loss) from the Caffe platform [54] is suitable for the task as Softmax activation transfers the model outputs to a probability value for all categories. To prove that our method can also be applied to more general recognition tasks, we further tested the methods for attribute recognition. As attribute classification is a multiple two-class classification problem, [16] applied a Sigmoid Cross Entropy Loss layer as the cost layer for attribute recognition. When the Softmax Loss layer is replaced by a Sigmoid Cross Entropy Loss layer, each input can have multiple label probabilities [55]. Hence, it is applicable for attribute classification, we also set this layer as the cost function for attribute

Table 1: Architecture of the CNN Model

Number	Layer	Kernel Size	Output Number
1	Conv1_1	3	64
2	Conv1_2	3	64
3	Conv2_1	3	128
4	Conv2_2	3	128
5	Conv3_1	3	256
6	Conv3_2	3	256
7	Conv3_3	3	256
8	Conv4_1	3	512
9	Conv4_2	3	512
10	Conv4_3	3	512
11	Conv5_1	3	512
12	Conv5_2	3	512
13	Conv5_3	3	512
14	RoI_Pooling	7X7	512
15	FC6	Fully-Connection	4096
16	FC7	Fully-Connection	4096
17	Cls_Score	Fully-Connection	Class Categories

prediction.

After fine-tuning, the CNN features for the top 1000 boxes produced by Edgeboxes for each image were extracted from the first fully connected layer (FC6). From our experiments, we found that 1000 regions are sufficient for the representation of an image. Empirically, as the Edgeboxes algorithm provides ranking for the generated boxes with confidence values, the top ranked 1000 boxes have higher probabilities which implies they contain objects. For the same reason as [17], we do not apply non-maximum suppression. However, feature extraction of multiple regions in a CNN is time-consuming. Consequently, we implemented our algorithm on top of a fast R-CNN [40] in which the RoI projection and RoI pooling scheme enable the completion of feature extraction of one image in only one feed forward process, thus significantly reducing the computational cost and running time. The final dimension of the VLAD code is the number of clusters times the dimension of CNN features after PCA dimensionality reduction.

3.2. VLAD Encoding

VLAD is a type of global discriminative feature descriptor generated on a set of local features (say, SIFT) extracted from an image. The basic principles are as follows:

Let $X = \{x_i\}_{i=1}^{N}$ be a set of local descriptors. Then a codebook $C = \{c_1, ..., c_k\}$ of k visual words can be learnt by the k-means algorithm. Each local descriptor x_i can be quantized to the nearest visual word. For each visual word, the sum of the differences between the center and each local descriptor assigned to this center can be subsequently obtained. This can be expressed as

$$\delta_j(X) = \sum_{i=1}^{N} a_j^i (c_j - x_i) \tag{1}$$

where a_j^i is a binary assignment weight indicating if the local descriptors belongs to this visual word, and N is the number of local descriptors. Then the VLAD code is a concatenation vectors of cumulated differences δ_j of each cluster:

$$v(X) = [\delta_1^T(X), \delta_2^T(X), \delta_3^T(X), ..., \delta_k^T(X)] \tag{2}$$

The overall dimension of the VLAD code d×k, where d is the dimension of local descriptors and k is the number of dictionary entries (clusters).

3.3. Spatial Pyramid VLAD

Although VLAD encoding performs well in preserving local features, spatial information is largely ignored. To compensate for this, recent papers [50], [17] have proposed spatial pyramid VLAD. In this paper, we apply it to CNN features following the same train of thought described in [19] and demonstrate the significance of the scheme in the explicit abstraction of local objects and their corresponding spatial information for action and attribute recognition. Also, Lazebnik et al. [49] also applied spatial pyramid scheme for recognizing natural scene categories. They extracted conventional image features and place them inside corresponding spatial grids whilst we fully made use of CNN features and assigned candidate regions into the spatial pyramids. Fig. 2 provides an illustration of the spatial pyramid VLAD approach. More specifically, we implemented a 3 level spatial pyramid: 1×1, 2×2, and 4×1 as shown in Fig. 2. Regions are allocated into each spatial grid, with assignments determined by the distribution of the centers of the regions.

With the CNN features (4096 dimensions), VLAD encoding is performed for each spatial pyramid separately. As has being pointed out in [56] , appropriate dimension reduction on original features would further improve the performance of the VLAD encoding. Subsequently, we apply PCA on the CNN features of each region. However, as the number of features is large, training conventional PCA on all of the features would be unrealistic. As an effective alternative, we first randomly select a number of features for training, and then perform PCA on all of the remaining features. This method may poorly generalize as only limited samples are applied for PCA training. In our implementation, an incremental PCA [57] was utilized due to its merit of high efficiency in memory usage. We perform PCA on all the features to reduce from 4096 dimensions to 256.

Following the steps of VLAD, codeword learning with k-means clustering is subsequently performed, with the number of clusters set at 12, 16, 24, and 64. The efficient k-means++ algorithm [58] was chosen to improve the performance of the conventional k-means as the random initialization of it often result in poor performance. The final dimensionality of the VLAD codes is the number of clusters multiplied by the CNN features after PCA dimension reduction.

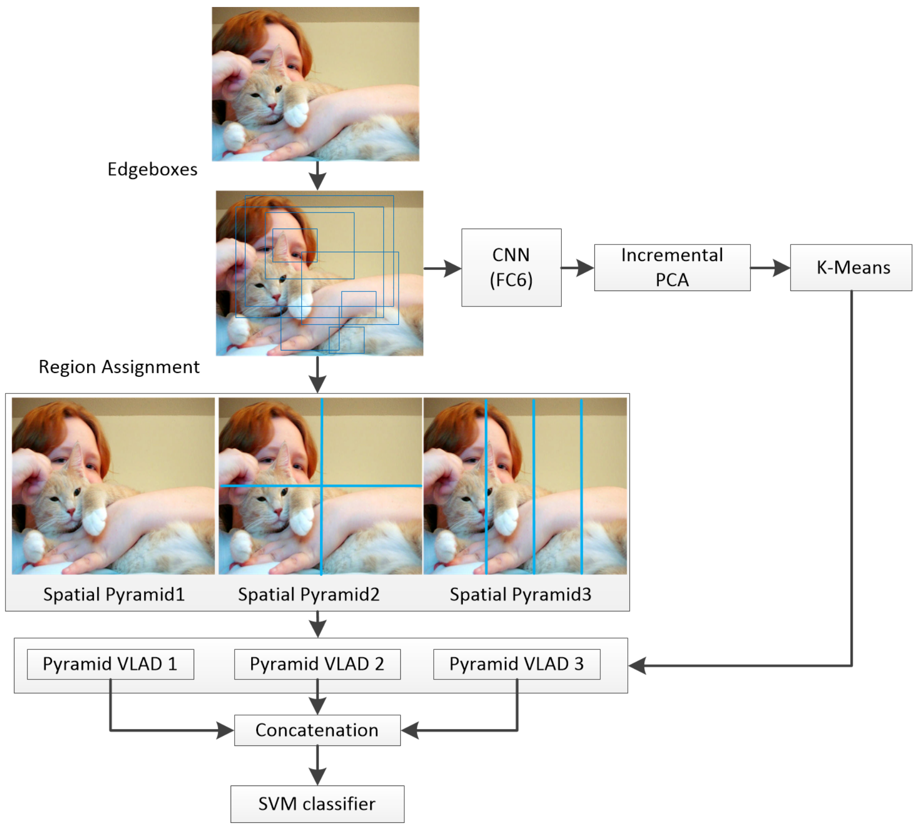


Figure 2: VLAD encoding with a spatial pyramid: The image was divided with a 3 level spatial pyramid: 1×1, 2×2 and 4×1. Each pyramid is encoded separately with VLAD.

4. Experiments and Results

In this section, the experimental set up will be briefly described, followed by the details of the experiments on the four benchmark datasets: the Stanford 40 Action dataset, the People Playing Musical Instrument dataset (PPMI) for action recognition, the Berkeley Attributes of People dataset and the 27 Human Attributes dataset (HAT) for attribute classification.

4.1. Deep Learning Model

All of the models have been implemented on the Caffe deep learning framework. The VGG16 from [53] was employed with the network pre-trained on

ImageNet and then fine-tuned on specific datasets. As pointed out by Girshick [40], it is not necessary to fine-tune weights from all layers in VGG16. Hence, during fine-tuning, we kept the weights of the first two convolutional layers unchanged and adjusted the other layers. The maximum training iterations and learning rate were chosen as 40000 and 0.001 respectively. During training, we set all the boxes generated from Edgeboxes as candidate regions for training. As action recognition is a general multi-class classification problem, we used the widely applied Softmax Loss function in deep convolutional neural networks. However, as attribute classification is a multiple independent two class classification problem, another loss function, namely Sigmoid Cross Entropy Loss would be preferable. The other parameters are the same as the fast R-CNN [40].

The reasons for choosing VGG16 as the CNN model are as follows:

1. In terms of the system efficiency, VGG16 model is more GPU demanding compared with some other shadow network structures. In practice, the VGG16 is more straightforward to use than those complicated structures such as GoogleNet [59] or ResNet [60]. On the other hand, the RoI pooling in our VGG16 model inherits the advantage of fast R-CNN [40] to efficiently extract the CNN features from candidate regions.
2. Another reason for using VGG16 is to compare the proposed spatial pyramid VLAD encoding scheme with previous state-of-the-art methods which employed VGG16 as their basic model, e.g., R*CNN [16] and Action parts [25] for action and attribute classification.

4.2. VLAD Encoding

We completed our experiments under the Linux operating system, with the incremental PCA and k-means++ implemented using the scikit-learn machine learning package [61]. VLAD encoding was realized in Matlab using the VLFeat toolbox [62]. Action recognition is a multi-class classification problem in which the data can only belong exclusively to one class. For such a multi-class problem, a multinomial classifier implemented by logistic regression using a Softmax

Figure 3: Some examples of the Stanford 40 action dataset, each image corresponds to one action type of the 40 actions.

classifier and its MLP variant is better than an SVM implemented as multiple binary classifiers. As noted in [40], Softmax , unlike one-vs-rest SVMs, introduces competition between classes and shows better results than SVMs [40]. Hence, this task was achieved with the aid of a multi-layer perceptron (MLP) neural network provided in the Matlab Neural Network toolbox. As for attribute prediction, it can be considered as a multiple of the two-class classification problem. SVM is a superior two-class classifier as it directly optimizes the decision boundaries from the data [63]. Hence, a SVM linear classifier was used from the LIBSVM toolbox [64] for attribute classification.

4.3. Stanford 40 Action Dataset

To evaluate the system performance on action recognition, we experimented using the Stanford 40 Action dataset [1], which has 9532 images in total corresponding to 40 classes of actions. The dataset was split into training and testing sets of 4000 and 5532 images respectively. There are 180-300 images for

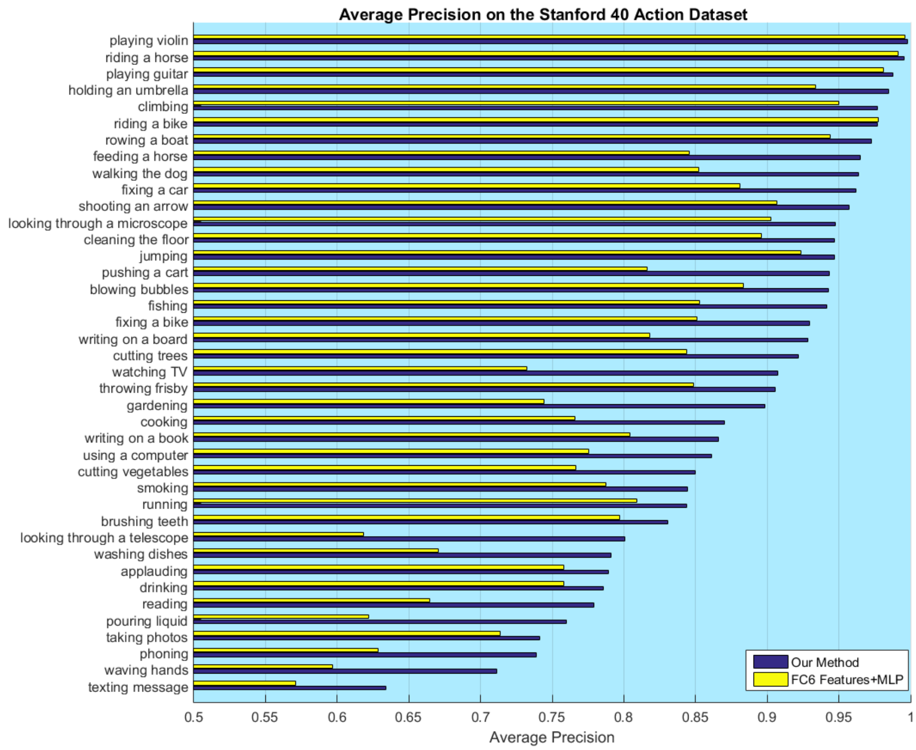


Figure 4: Results on the Stanford 40 Action dataset and comparison with the baseline approach.

Table 2: The Mean AP results on the Stanford 40 dataset using different pre-trained models

Methods	Mean AP(%)
FC6 features(VGG-M-1024 [65])	43.8
FC6 features(VGG16 [53])	61.3

each class. The images within each class have large variations in human pose, appearance, and background clutter. Fig.3 presents 40 examples corresponding to the 40 action categories in this dataset.

Details on the experiments on this dataset are explained as follows:

1. CNN features

Table 3: The Mean AP results on Stanford 40 Action dataset and comparison with different approaches.

Methods	Mean AP(%)
FC6 features(pre-trained model)	61.3
FC6 features(fine-tuned model)	81.2
PCA256+16clusters(No Spatial Pyramid)	84.9
PCA256+16clusters(With Spatial Pyramid)	85.9
PCA256+16clusters+FC6 features(No Spatial Pyramid)	86.6
PCA256+16clusters+FC6 features(With Spatial Pyramid)	**88.5**

Table 4: Mean AP results on the Stanford 40 Action dataset and comparison with previous results.

Method	Mean AP(%)
Object bank [66]	32.5
LLC [67]	35.2
EPM [68]	40.7
DeepCAMP [44]	52.6
Khan et al. [24]	75.4
Semantic parts [69]	80.6
(Ours)PCA256+24clusters+FC6 features	81.5
(Ours)PCA256+64clusters+FC6 features	81.8
(Ours)PCA256+12clusters+FC6 features	87.7
(Ours)PCA256+16clusters+FC6 features	**88.5**

Before selecting the model for subsequent experiments, we first evaluated the performance from different models. The VGG-M-1024 [65] and VGG16

Table 5: The Mean AP results on the Stanford 40 Dataset using different PCA reduced dimensions.

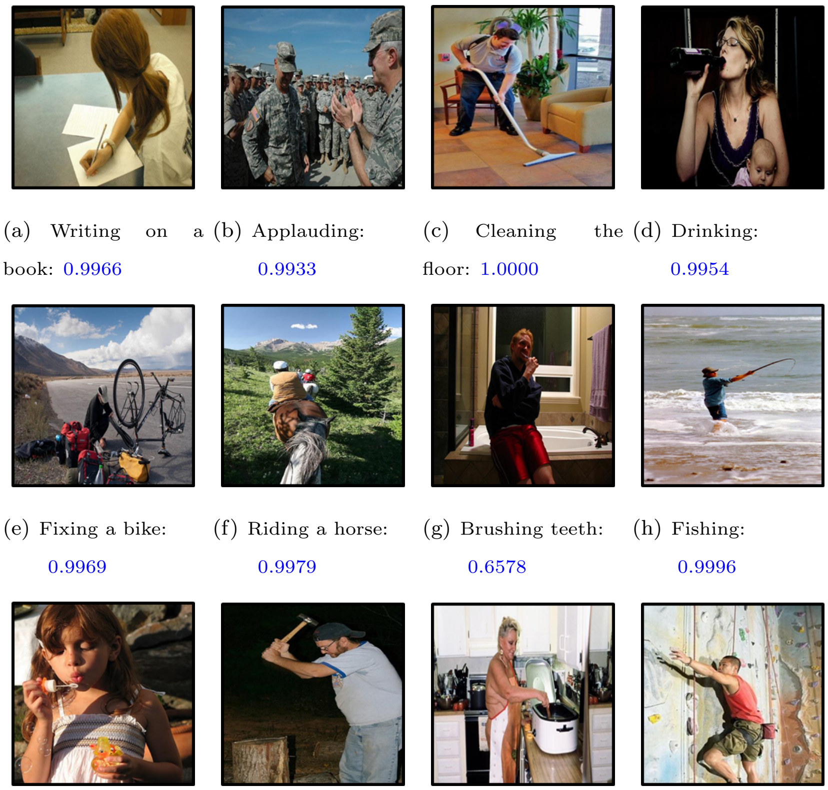
Methods	Mean AP(%)
PCA512+16clusters+FC6 features	88.4
PCA256+16clusters+FC6 features	**88.5**

Table 6: Comparative study of the Stanford 40 dataset on the different number of patches to form VLAD code.

Method	Mean AP(%)
PCA256+16clusters(3000 regions)+FC6 features	87.8
PCA256+16clusters(2000 regions)+FC6 features	88.1
PCA256+16clusters(1000 regions)+FC6 features	**88.5**

model [53] were selected for comparison. VGG16 turns out to be much better than the VGG-M-1024 model in terms of recognition rates as shown in Table.2. Hence, we chose the VGG16 model for subsequent experiments. Also, to prove that fine-tuning of the CNN model can significantly improve the feature representation capability, we extracted FC6 features from both the pre-trained CNN model and the fine-tuned model. As can be seen in Table.3, with the same experimental setting, the fine-tuned model gains about a 20% increase in recognition performance, from 61.3% to 81.2%.

2. VLAD coding with different learnt clusters
   To select the best number of centroids learnt with k-means, we performed extensive comparative experiments. From Table.4, the best performance was achieved when the cluster number of CNN features is 16. This is an interesting result which matches the findings in [4] that only a small number of clusters can generate promising results. The advantage of small number of feature clusters also stem from the characteristics of VLAD, which,



(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

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

unlike traditional BoVW, is based on the accumulation of the differences between a local descriptor and each learnt cluster. Also, VLAD can be considered as a simplified version of FV, which is more efficient than FV and more powerful than traditional BoVW. As noted in [56], dimension reduction plays a significant role in VLAD encoding. The same procedure was repeated with the setting up of dimensionality-reduced CNN features of 512 dimensionality, with slightly poorer mAP results (Table.5). Hence, the CNN features of 256 dimensionality will be the focus in most of the experiments.

3. VLAD coding without CNN features
   As can be seen from Table.3, to evaluate the stand-alone performance of the VLAD encoding scheme, each image was represented by a VLAD code from 256 dimension features and 16 learnt clusters. Adding the spatial pyramid boosted the performance from 84.9% to 85.9%.
4. VLAD coding with CNN features
   The ground-truth region was provided to indicate the target person within the image, hence it is instrumental for recognition. Adding the CNN features of the ground-truth region further raised the performance to 88.5%.
5. VLAD coding from different number of regions
   To validate that 1000 regions per image is sufficient for the VLAD encoding scheme, recognition results from 2000 and 3000 boxes per image were also provided. It is clear from Table.6 that 1000 boxes yields the best performance. This is partly because regions generated from the Edgeboxes algorithm are ranked and the top 1000 boxes include most of the important patches in the images. Including more regions may add noise to the final representation.
6. Standard Deviation of AP results from different methods
   As there are 40 action categories in our task, it is important to see whether the proposed methods have improved robustness over different categories. Hence, we calculated the Standard Deviation (SD) values on AP results from different methods. The SD on AP values from method only using CNN FC6 features is 11.5 while the SD on AP results from the proposed methods (PCA256+16clusters+FC6features) is 9.2 which indicates our approach has improved robustness over different categories.

The comparisons with previously published methods are shown in Table.4, which demonstrates that our method has the highest mean AP. It is noteworthy that Khan et al. [24] did not utilize a ground-truth bounding box during action recognition. In our configuration (PCA256+16clusters), the proposed method yields a 10.5% increase in mean AP even without ground truth. This results

Figure 6: Some examples in the PPMI dataset, the images in the first row correspond with the action of 'Playing Instrument' while the images from second row correspond with 'With Instrument'.

further demonstrate the suitability of spatial pyramid VLAD encoding in action recognition. Fig.4 shows the AP value of each categories of our approach and a comparison with results from CNN features.

It can seen from Fig.4 that the spatial pyramid VLAD encoding scheme outperforms plain CNN features in all action classes except 'riding a bike', in which the performances are similar. More importantly, VLAD performs significantly better in the more fine-grained action classes, for instance, 'writing on a board'. This is because VLAD encoding preserves local information from small patches, and the important spatial information is retained with the spatial pyramid VLAD. Fig.5 provides some examples of correct recognition in the Stanford 40 Action dataset.

4.4. People Playing Musical Instruments Dataset

PPMI [21] is a dataset emphasizing subtle difference in interactions between humans and objects (fine grained classification). PPMI consists of 12 different musical instruments. Each class includes 150 PPMI+ images (humans playing instruments) and 150 PPMI- images (humans holding the instruments). Fig.6 provides some examples of the PPMI dataset. Hence, there are 24 categories to classify. We evaluated our approaches on the 24 categories classification task.

The dataset did not provide a ground-truth region for each person. Hence, different from Standford 40 Action dataset, we fine-tuned the pre-trained VGG16

Table 7: The Mean AP results on PPMI dataset and comparison with different approaches.

Methods	Mean AP(%)
FC6 features	80.7
PCA256+16clusters(No Spatial Pyramid)	74.3
PCA256+16clusters(With Spatial Pyramid)	76.6
PCA256+16clusters+FC6 features(No Spatial Pyramid)	80.8
PCA256+16clusters+FC6 features(With Spatial Pyramid)	**81.3**

Table 8: Comparison with other published methods on PPMI dataset.

Methods	SPM [49]	Grouplet [21]	LLC [67]	Spatial Saliency [70]	Ours
Mean AP(%)	35.6	36.7	39.8	49.4	**81.3**

model following the common image classification procedure in the Caffe platform [54]. The learning rate is set as 0.0001 and the batch size as 128. We set the maximum iterations as 40000. Once the model was trained, FC6 features were extracted from top the 1000 regions generated from Edgeboxes. The VLAD encoding was accomplished after PCA dimensionality reduction and codeword learning with k-means++.

From Table.7, the following results can be observed: On this dataset, Image-level CNN features alone provide satisfactory results. However, with CNN features combined with VLAD spatial pyramid, the performance increased to 81.3% which proves the VLAD and CNN features are complementary. The SD of AP results on image-level features are 9.7 while the SD of AP results from our methods is 9.5 which indicates the proposed method has good robustness over different categories. Also, We also achieved state-of-the-art results on this dataset when compared with other approaches as shown in Table.8.

4.5. Berkeley Attributes of People Dataset

Classification of people's attributes is an important task in computer vision as semantic attributes can often bridge the gap between low-level and high-level features in computer vision tasks. The main task of human attribute recognition is to recognize a person's multiple features such as gender, hair style and type of clothes for the purpose of describing a person under realistic viewpoints, pose and occlusion.

To see if our method can be applied to attribute classification, we evaluated our method on the Berkeley Attributes of People Dataset [22], which 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. Fig.7 shows some examples from this dataset. Compared with the many other benchmark computer vision datasets, only limited research has been published on experiments using it [71] [22].

We followed the Spatial Pyramid VLAD encoding of CNN features previously explained, and applied an SVM classifier for the final prediction. Specifically, the pre-trained VGG16 model [53] was utilized for subsequent fine-tuning. The training process was implemented in the fast R-CNN [40] framework. The region proposal algorithm Edgeboxes was applied on each image, and FC6 features were then extracted for each region. The VLAD encoding was accomplished after PCA dimensionality reduction and codeword learning with k-means++.

More details about the experiment procedure and three comparative settings are described as follows:

1. CNN features

   As shown in Table.9, CNN features from the first fully connected layers (FC6) corresponding to the ground truth region were extracted, and directly applied for attribute classification as a comparative baseline. Despite the effective representational capability of VGG16, the mean AP is

Figure 7: Some examples of the Berkeley Attributes of People dataset.

only 78.1%, which implies that CNN feature alone are insufficient.

2. VLAD coding without CNN features
To evaluate the stand-alone performance of VLAD encoding, each image was represented by CNN features of 256 dimensionality and then VLAD was applied to the 16 learnt codewords. The mAP from this configuration is 78.3%. There is no ground-truth region in this scenario and the spatial pyramid has not been taken into account.

3. CNN features combined with VLAD coding
In this configuration, CNN features of the ground-truth region are combined with VLAD coding. The concatenated features yield a mAP performance increase of up to 8.5%, which suggests that the local features (ground-truth region) and compact global representation (image-level VLAD code) are complementary. A ground-truth region specifies a target person in an image. Subsequently, the combination of features for ground-truth regions and the image level VLAD coding introduces the contextual information associated with the target person, which is beneficial to the improvement in action classification. The increase in performance agrees with our intuition that global contextual information is helpful for the recognition task.

4. CNN features combined with the spatial pyramid VLAD coding
Finally, to test the influence on overall performance of the spatial pyramid

Table 9: The AP results of the Berkeley Attributes of People dataset and comparison of different approaches.

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
PCA256+16clusters(No Spatial Pyramid)	88.9	76.4	74.7	68.2	68.5	88.5	73.3	71.8	94.2	78.3
PCA256+16clusters+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
PCA256+16clusters+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**

VLAD encoding, we added spatial pyramid encoding, and concatenated the VLAD codes of each pyramid into one representation, respectively, with CNN features with and without ground-truth regions. Experimental results showed that adding the spatial pyramid does improve the overall mAP performance, by 3.6%. The SD of the AP values is 6.3 while the SD of AP values from CNN features is 11.5 which proves our method's improved robustness on different categories. More interestingly, as can be seen from Table.9, the AP values from all categories increased by adding a spatial pyramid which proves that the spatial information is very important for recognition. Fig.8 provides some examples of recognition results on this dataset. The precision-recall figure of the proposed approach can be seen in Fig.9. It is clearly seen from the figure that our method on all categories has higher AP values than the method purely based on CNN features.

We also evaluated the influence of the number of k-means clusters by performing VLAD encoding with 12, 16, 24 and 64 centroids separately. The results show that 16 clusters works the best from the comparative experiments. Additionally, when comparing with other published methods, our approach generates competitive results as shown in Table.10.

4.6. 27 Human Attributes Dataset(HAT)

This human attributes dataset was collected by Sharma et al. [23]. The dataset contains 9344 images, split into 7000 training images and 2344 test images. A total of 27 attribute annotations are presented in the dataset toolkit. As explained in [23], the dataset contains a wide variety of human images in

(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

Figure 8: Examples of attribute classification: the probabilities of certain attributes are provided, the blue text are the ground truth labels. the red text show an incorrect classification example.

different poses, with different ages, wearing different clothing and with diverse accessories. Also, there might be more than one person in an image for attribute query, thus increasing the difficulties in recognizing attributes.

Fig.10 illustrates some examples from the HAT dataset. As can be seen from the figure, there exist large variations in the viewpoint, people's clothing style and illumination. Also, people in the image are performing various activities with different poses, which make attribute recognition more challenging.

In our experiment, we applied PCA dimension reduction on the FC6 features from the trained VGG16 CNN model, following the similar procedure used with the Berkeley Human Attributes Dataset. PCA, clustering with k-means++

Table 10: The AP results of Berkeley Attributes of People dataset and comparison with previous methods.

Attribute	male	long hair	glasses	hat	tshirt	longsleeves	shorts	jeans	long pants	Mean AP(%)
Poselets [22]	82.4	72.5	55.6	60.1	51.2	74.2	45.5	54.7	90.3	65.0
PANDA [71]	91.7	82.7	70.0	74.2	49.8	86.0	79.1	81.0	96.4	79.0
R*CNN [16]	92.8	88.9	82.4	92.2	**74.8**	91.2	92.9	89.4	97.9	89.2
Gkioxari et al. [25]	92.9	90.1	77.7	93.6	72.6	**93.2**	**93.9**	**92.1**	**98.8**	89.5
Ours (PCA256+12clusters+FC6 features)	93.8	90.0	88.5	93.4	72.9	92.2	90.8	87.7	98.4	89.7
Ours (PCA256+64clusters+FC6 features)	93.8	**92.2**	89.1	93.8	73.1	92.1	91.4	87.8	98.4	90.0
Ours (PCA256+24clusters+FC6 features)	**94.1**	90.4	**89.5**	**94.0**	73.8	92.5	91.9	88.5	98.4	90.3
Ours (PCA256+16clusters+FC6 features)	**94.1**	90.4	89.4	**94.0**	74.0	92.5	91.9	88.6	98.5	**90.4**

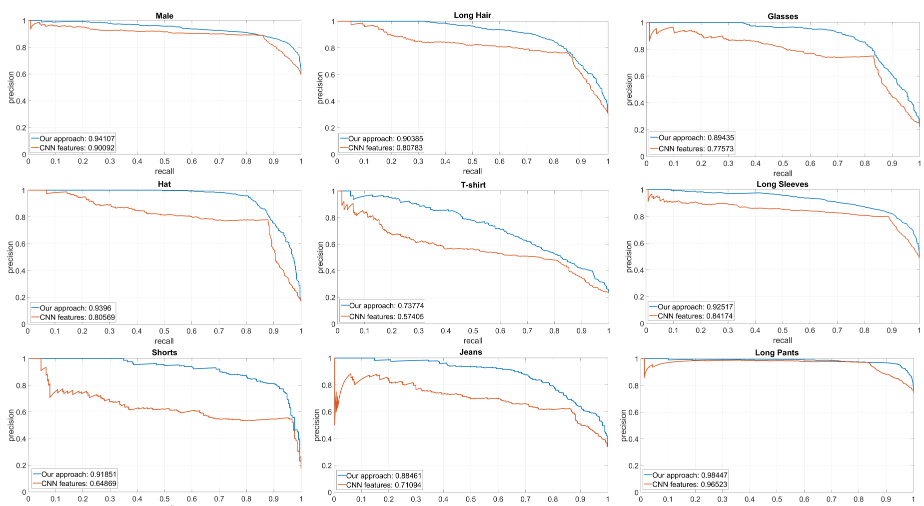


Figure 9: The precision recall curve of Berkeley Attributes of People Dataset. The red curves indicate results only based on CNN features while the blue curves show the results based on the proposed method.

and VLAD encoding with the spatial pyramid were performed consecutively to generate the concatenated features for final classification.

We treated the prediction of each attribute as an independent two-class classification problem. The final results on Average Precision (AP) are presented in Table.11. Specifically, we achieved 74.2% mean AP with SD 20.1 on this

Figure 10: Some examples from the HAT dataset.

Table 11: AP results on the 27 Human Attributes Dataset(HAT).

Attributes	AP(%)	Attributes	AP(%)	Attributes	AP(%)	Attributes	AP(%)
Female	97.5	Crouching/bent	30.8	Small kid	71.0	Female short skirt	50.0
Frontal pose	97.4	Sitting	87.9	Small baby	31.9	Wearing short shorts	69.2
Side pose	83.0	Arms bent/crossed	97.3	Wearing tank top	65.5	Low cut top	89.1
Turned back	96.6	Elderly	69.0	Wearing tee shirt	88.8	Female in swim suit	55.0
Upper body	98.6	Middle aged	80.1	Wearing casual jacket	60.7	Female wedding dress	75.1
Standing straight	99.1	Young (college)	73.9	Formal mens suit	75.6	Bermuda/beach shorts	77.7
Running/walking	80.0	Teen aged	38.4	Female long skirt	62.5	Mean AP	74.2

dataset. In Table.12, a comparison with previously published results is also presented. The Deep Semantic Pyramid (DSP) proposed in [32] also utilized Deep Convolutional Neural Networks and Spatial Pyramid, which shows a better performance than other published methods.

Table 12: Comparison with previous methods on the HAT dataset.

Methods	DSR [23]	SPM [49]	EPM [68]	DSP [32]	Ours
Mean AP(%)	53.8	55.5	59.7	71.5	**74.2**

5. Conclusion

Action recognition in static images is a challenging task, partly due to the fine-grained property and the absence of motion information. Our study indicates that information from local patches and the global contextual information are critically important contributing factors to improve the performance of action recognition. This is validated by our re-implementation of Vector of Locally Aggregated Descriptors (VLAD) on top of a spatial pyramid for CNN features to identify local information and global spatial information simultaneously. Experiments were conducted not only with ground-truth regions but also with images without ground-truth annotations where the neighboring objects and scenes are comprehensively coded into compact representations. Our experiments revealed that the combination of CNN features and VLAD codes brings performance gains for both action recognition and general recognition tasks such as attribute prediction from still images. The beneficial effect of spatial pyramids has also been confirmed by demonstrating the performance enhancement. Four different datasets have been tested, namely, the Stanford 40 Action dataset, the People Playing Musical Instrument dataset (PPMI), the Berkeley Attributes of People dataset and the 27 Human Attributes dataset (HAT) with the results all demonstrating the advantages of our proposed deep Spatial Pyramids VLAD coding scheme. We will develop a prototype system in future works by implementing the proposed scheme in a more homogeneous way.

References

[1] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, L. Fei-Fei, Human action recognition by learning bases of action attributes and parts, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 1331–1338.

[2] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2, IEEE, 2005, pp. 524–531.

[3] G. Csurka, F. Perronnin, Fisher vectors: Beyond bag-of-visual-words image representations, in: Computer Vision, Imaging and Computer Graphics. Theory and Applications, Springer, 2010, pp. 28–42.

[4] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3304–3311.

[5] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: European Conference on Computer Vision, Springer, 2014, pp. 392–407.

[6] M. Harandi, M. Salzmann, F. Porikli, When vlad met hilbert, arXiv preprint arXiv:1507.08373.

[7] G. Sharma, F. Jurie, C. Schmid, Discriminative spatial saliency for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 3506–3513. doi:10.1109/CVPR.2012.6248093.

[8] V. Delaitre, I. Laptev, J. Sivic, 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, pp. 97.1–97.11, doi:10.5244/C.24.97.

[9] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, International journal of computer vision 103 (1) (2013) 60–79.

[10] X. Peng, C. Zou, Y. Qiao, Q. Peng, Action recognition with stacked fisher vectors, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 581–595.

[11] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[13] S. Bell, C. L. Zitnick, K. Bala, R. Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, arXiv preprint arXiv:1512.04143.

[14] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1529–1537.

[15] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587. doi:10.1109/CVPR.2014.81.

[16] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with r* cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1080–1088.

[17] A. Shin, M. Yamaguchi, K. Ohnishi, T. Harada, Dense image representation with spatial pyramid vlad coding of cnn for locally robust captioning, arXiv preprint arXiv:1603.09046.

[18] D. Oneata, J. Verbeek, C. Schmid, Action and event recognition with fisher vectors on a compact feature set, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1817–1824.

[19] T. Uricchio, M. Bertini, L. Seidenari, A. D. Bimbo, 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, pp. 1020–1026. `doi:10.1109/ICCVW.2015.134`.

[20] C. L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 391–405.

[21] B. Yao, L. Fei-Fei, Grouplet: A structured image representation for recognizing human and object interactions, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 9–16.

[22] L. Bourdev, S. Maji, J. Malik, Describing people: A poselet-based approach to attribute classification, in: 2011 International Conference on Computer Vision, 2011, pp. 1543–1550. `doi:10.1109/ICCV.2011.6126413`.

[23] G. Sharma, F. Jurie, Learning discriminative spatial representation for image classification, in: BMVC 2011-British Machine Vision Conference, BMVA Press, 2011, pp. 1–11.

[24] F. S. Khan, J. Xu, J. van de Weijer, A. D. Bagdanov, R. M. Anwer, A. M. Lopez, Recognizing actions through action-specific person detection, Image Processing, IEEE Transactions on 24 (11) (2015) 4422–4432.

[25] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2470–2478.

[26] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, arXiv preprint arXiv:1405.4506.

[27] M. M. Ullah, S. N. Parizi, I. Laptev, Improving bag-of-features action recognition with non-local cues., in: BMVC, Vol. 10, Citeseer, 2010, pp. 95–1.

[28] V. Delaitre, I. Laptev, J. Sivic, 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/.

[29] C. Sun, R. Nevatia, Large-scale web video event classification by use of fisher vectors, in: Applications of Computer Vision (WACV), 2013 IEEE Workshop on, IEEE, 2013, pp. 15–22.

[30] M. Jain, H. Jégou, P. Bouthemy, Better exploiting motion for better action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2555–2562.

[31] S. Savarese, J. Winn, A. Criminisi, Discriminative object class models of appearance and shape by correlatons, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, IEEE, 2006, pp. 2033–2040.

[32] F. S. Khan, R. M. Anwer, J. van de Weijer, M. Felsberg, J. Laaksonen, Deep semantic pyramids for human attributes and action recognition, in: Scandinavian Conference on Image Analysis, Springer, 2015, pp. 341–353.

[33] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 17–24. doi:10.1109/CVPR.2010.5540235.

[34] A. Prest, C. Schmid, V. Ferrari, Weakly supervised learning of interactions between humans and objects, IEEE Transactions on Pattern Analysis and

Machine Intelligence 34 (3) (2012) 601–614. doi:10.1109/TPAMI.2011.158.

[35] I. González Díaz, V. Buso, J. Benois-Pineau, G. Bourmaud, R. Megret, Modeling instrumental activities of daily living in egocentric vision as sequences of active objects and context for alzheimer disease research, in: Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare, ACM, 2013, pp. 11–14.

[36] C. Crispim-Junior, K. Avgerinakis, V. Buso, G. Meditskos, A. Briassouli, J. Benois-Pineau, Y. Kompatsiaris, F. Bremond, Semantic event fusion of different visual modality concepts for activity recognition.

[37] S. Karaman, J. Benois-Pineau, V. Dovgalecs, R. Mégret, J. Pinquier, R. André-Obrecht, Y. Gaëstel, J.-F. Dartigues, Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia, Multimedia tools and applications 69 (3) (2014) 743–771.

[38] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (9) (2010) 1627–1645.

[39] L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3d human pose annotations, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1365–1372.

[40] R. Girshick, Fast r-cnn, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448. doi:10.1109/ICCV.2015.169.

[41] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. S. Torr, Conditional random fields as recurrent neural networks, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1529–1537. doi:10.1109/ICCV.2015.179.

[42] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1717–1724.

[43] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (10) (2010) 1345–1359.

[44] A. Diba, A. M. Pazandeh, H. Pirsiavash, L. Van Gool, Deepcamp: Deep convolutional action & attribute mid-level patterns.

[45] R. Arandjelovic, A. Zisserman, All about vlad, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 1578–1585. doi:10.1109/CVPR.2013.207.

[46] J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: Theory and practice, International journal of computer vision 105 (3) (2013) 222–245.

[47] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, N. Vasconcelos, Scene classification with semantic fisher vectors, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2974–2983. doi:10.1109/CVPR.2015.7298916.

[48] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, Netvlad: Cnn architecture for weakly supervised place recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.

[49] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, Vol. 2, IEEE, 2006, pp. 2169–2178.

[50] R. Zhou, Q. Yuan, X. Gu, D. Zhang, Spatial pyramid vlad, in: Visual Communications and Image Processing Conference, 2014 IEEE, IEEE, 2014, pp. 342–345.

[51] J. Hosang, R. Benenson, B. Schiele, How good are detection proposals, really?, in: 25th British Machine Vision Conference, BMVA Press, 2014, pp. 1–12.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (3) (2015) 211–252. `doi:10.1007/s11263-015-0816-y`.

[53] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[54] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.

[55] S. Shankar, V. K. Garg, R. Cipolla, Deep-carving: Discovering visual attributes by carving deep neural nets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3403–3412.

[56] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: IEEE Conference on Computer Vision & Pattern Recognition, 2010.
URL `http://lear.inrialpes.fr/pubs/2010/JDSP10`

[57] D. A. Ross, J. Lim, R. Lin, M. Yang, Incremental learning for robust visual tracking, International Journal of Computer Vision 77 (1-3) (2008) 125–141.

[58] D. Arthur, S. Vassilvitskii, 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, pp. 1027–1035.

[59] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[60] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[62] A. Vedaldi, B. Fulkerson, VLFeat: An open and portable library of computer vision algorithms (2008).

[63] M. M. Adankon, M. Cheriet, Support vector machine, Encyclopedia of biometrics (2015) 1504–1511.

[64] C.-C. Chang, C.-J. Lin, 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`.

[65] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, arXiv preprint arXiv:1405.3531.

[66] L.-J. Li, H. Su, L. Fei-Fei, E. P. Xing, Object bank: A high-level image representation for scene classification & semantic feature sparsification, in: Advances in neural information processing systems, 2010, pp. 1378–1386.

[67] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Computer Vision and Pattern

Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3360–3367.

[68] G. Sharma, F. Jurie, C. Schmid, 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, pp. 652–659.

[69] Z. Zhao, H. Ma, X. Chen, Semantic parts based top-down pyramid for action recognition, Pattern Recognition Letters 84 (2016) 134–141.

[70] G. Sharma, F. Jurie, C. Schmid, Discriminative spatial saliency for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 3506–3513.

[71] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, Panda: Pose aligned networks for deep attribute modeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1637–1644.