Discriminative Dictionary Design for Action Classification in Still Images

Abhinaba Roy[1(✉)], Biplab Banerjee[2], and Vittorio Murino[1]

[1] Istituto Italiano di Technologia, Genova, Italy
{abhinaba.roy,vittorio.murino}@iit.it
[2] Indian Institute of Technology, Roorkee, India
bbanfcs@iitr.ac.in

Abstract. In this paper, we address the problem of action recognition from still images. Although used widespread, local features (SIFT, STIP) invariably engender two potential problems: the counts of such extracted features are not evenly distributed in different entities of a given category and many of such features are not paradigmatic of the visual concept the entities represent. In order to generate a discriminative dictionary taking the aforementioned issues into account, we propose a novel method for identifying robust and category specific local features which maximize the class separability to a possible extent. Specifically, we consider category independent region proposals to highlight local regions in still images. Further, the selection of potent local descriptors is cast as filtering based feature selection problem which ranks the local features per category based on a novel measure of distinctiveness. The underlying visual entities are subsequently represented based on the learned dictionary and this stage is followed by action classification using the random forest model. The framework is validated on the challenging Stanford-40 dataset and exhibits superior performances than the representative methods from the literature.

Keywords: Action recognition · Local features · Feature mining · Random forest

1 Introduction

Recognition of visual concepts is one of the most active research lines in computer vision. Other than the problem of generic object recognition from images [3], action recognition is another challenging research paradigm in this respect. With the growing amount of visual data available from various sources, intelligent analysis of human attributes and activities has gradually attracted the interest of the computer vision community. Action recognition based on local descriptors is another very popular paradigm which broadly depends upon three stages: (1) Extraction of local descriptors, (2) Codebook (dictionary) generation and feature encoding, and (3) Classification based on the encoded features. Although this framework has exhibited superlative performances for general

S. Battiato et al. (Eds.): ICIAP 2017, Part II, LNCS 10485, pp. 160–170, 2017.
https://doi.org/10.1007/978-3-319-68548-9_15

visual recognition tasks, the efficacy of such a model depends upon a number of factors, and effective codebook construction is undoubtedly the most noteworthy. The standard codebook construction process is based on vector quantizing local descriptors extracted from the available training data in which the cluster centroids define the codewords, the basic building blocks that are ultimately used to encode the underlying visual entities. Specifically, an entity is represented by a vector where the i^{th} component can be either the number of local descriptors that fall in the i^{th} cluster or a measure of proximity of local descriptors to the i^{th} cluster centroid. Needless to mention, the quality of the extracted local descriptors affect representation power of the codewords which, in turn, has direct impact on the recognition performance. For instance, the descriptors extracted from background regions or the ones shared by many visual categories add little to the discriminative capability of the codebook in comparison to the ones specifically extracted from the objects of interest. However, it is impossible to ensure the selection of potentially useful local descriptors in advance since such feature extraction techniques are typically engineered and ad hoc. In other words, there are certain immediate advantages if the most discriminative local descriptors are used for the purpose of a cogent codebook construction, though the process is intrinsically complex in general. Selection of the discriminative local descriptors for effective codebook generation coping with action recognition from images is the very core topic of this paper. We propose a simple algorithm which gradually filters out unrepresentative descriptors before constructing a compact global codebook. The proposed method is generic in the sense that it can work with different types of local features irrespective of the underlying visual entities they refer to. Specifically, we represent each still image by a large pool of category independent region proposals [1]. Each region proposal is represented by convolutional neural network (CNN) features (4096 dimensions) obtained from a pre-trained network. We propose a sequential method for codebook construction which first clusters the local descriptors of each entity using the non-parametric mean-shift (MS) technique [7]. The cluster centroids thus obtained represent the reduced set of non-repetitive local features for the entity from now onwards. Another round of MS clustering on the new set of local descriptors calculated from all the entities of a given category is followed and the centroids thus obtained are employed to build a temporary codebook specific to each category. Further, we propose a ranking criteria to highlight potentially discriminative codewords from each category specific codebook and the global dictionary is built by accumulating these reduced set of codewords from all the categories.

We can summarize the main highlights of this paper as follows:

– The initial two level MS based clustering of the local descriptors on the entity and the category level largely reduces the effects of repetitive and uninteresting descriptors, yet selecting representative codewords from each locally dense region in the feature space. We further propose a novel measure to rank and select a subset of discriminative codewords per visual category under consideration The proposed ranking measure ensures that the selected set of

codewords are frequent in the entities of the same category while being sporadic in other visual categories.
– We evaluate the codebooks learned in this way for action categorization from still images. We observe that the learned codebooks, when used in conjunction with efficient feature encoding techniques, sharply outperform similar techniques from the literature.

The rest of the paper is organized as follows. We discuss a number of related works from the literature in Sect. 2. The proposed action recognition framework is described in Sect. 3. Experimental results are reported in Sect. 4, followed by concluding remarks and ideas of possible future endeavour.

2 Related Works

In this section, we primarily highlight two aspects of the proposed framework and discuss relevant techniques from the literature. First, action representation from images with a focus on local feature encoding based methods will be addressed and a discussion on the relevant codebook construction techniques will subsequently be followed.

2.1 Action Recognition from Still Images

Recognition of human actions and attributes [6] has been approached using traditional image classification methods [23]. In the standard dictionary learning based scenario, a typical framework extracts dense SIFT [16] from the training images and codebook is constructed by clustering the SIFT descriptors by k-means clustering. Further, efficient encoding techniques including bag of words (BoW), LLC, Fisher vector are used to represent the images before the classification stage is carried out in such a feature space [4].

Since the inherent idea of the BoW based frameworks is to learn recurring local patches, a different set of approaches directly models such object parts in images. Such techniques either initially define a template and try to fit it to object parts or iteratively learn distinctive parts for a given category.

Discriminative part based models (DPM) [8] are used extensively for this purpose and they served as the state of the art for a period. The hierarchical DPM model is used to parse human pose for action recognition in [22]. An efficient action and attributed representation based on sparse bases of local features is introduced in [24]. An expanded part model for human attribute and action recognition is proposed in [19]. The effects of empty cavity, ambiguity and pooling strategies are explored in order to design the optimal feature encoding for the purpose of human action recognition in still images in [25].

Very recently, the part learning paradigm has gained much attention because of its ability to represent mid-level visual features. Given a large pool of region proposals extracted from the images, such techniques iteratively learn part classifiers with high discriminative capabilities. Methods based on partness analysis [10], deterministic annealing for part learning [20] etc. are some of the representatives in this respect.

2.2 Efficient Codebook Construction for Classification

As already mentioned, the paper focuses on the identification of good local descriptors for building a better dictionary in order to enhance the action recognition performance. The dictionary learning strategies can be supervised or unsupervised in nature. A class of unsupervised dictionary learning strategies compute over-complete sparse bases considering the idea of alternate optimization. Such techniques iteratively update the dictionary components and sparse coefficients for the input samples using k-SVD and matching pursuit based methods. [21] proposes the LLC technique where a locality constraint is added to the loss function of sparse coding. [13] introduces an l_1-norm based sparse coding algorithm where feature-sign search is applied for encoding and Lagrange dual method for dictionary learning. Effective sampling strategies for the BoW model is the focus of discussion in [18] where several aspects including the codebook size, clustering techniques adopted etc. are exhaustively studied.

On the contrary, the supervised approaches include the class support in building the dictionary. Label consistent SVD [9], logistic regression based sparse coding [17] explicitly consider the class discrimination in designing the sparse bases for dictionary learning. Two different clustering based approaches for keypoints selection are introduced in [15] for the purpose of dictionary learning based generic scene recognition. Distance measures among the keypoints are modeled in an online fashion to filter out keypoints with low generalization capability.

We focus on the supervised dictionary learning paradigm and follow a sequential approach in building the dictionary. In contrast to similar methods from the literature, our framework is flexible, scalable and non-parametric. The MS based clustering technique is inherently capable of detecting all possible clusters in the feature space based on data density. It better captures the local aspects of the entities than the traditional k-means based vector quantization techniques which require a number of hyper-parameters to be optimized. In addition, our ranking measure can efficiently highlight codewords which are highly discriminative by exploring their frequency distributions in all the underlying categories. Further, the proposed ranking measure is generic in the sense that it is applicable to the broader domains of feature selection, part learning, ranking in a retrieval system to name a few.

3 Proposed Algorithm

We detail the proposed action recognition framework in this section. As already mentioned, the proposed framework consists of four major stages: (1) Extraction of local features (2) Discriminative dictionary construction (3) Feature encoding (4) Action classification.

For notational convenience, let us consider that $TR = \{X_i, Y_i\}_{i=1}^{N}$ constitutes N training examples belonging to L action categories where each X_i represents an image and Y_i is the corresponding class label. Entities in TR are represented by a set of local descriptors $F_i = \{F_i^1, F_i^2, \ldots, F_i^{\alpha_i}\}$ where $F_i^k \in \mathbb{R}^d$ and $d = 4096$ or $d = 162$, respectively, depending on whether the underlying X_i is an

image. In addition, α_i represents the number of local descriptors extracted from X_i. Further, $\{C_1, C_2, \ldots, C_L\}$ represents the set of category specific codebooks learned by the proposed algorithm by exploiting the local features extracted from TR, whereas $C = [C_1 C_2 \ldots C_L]$ is the global codebook obtained by the concatenation of the local ones.

The framework is elaborated in the following sections.

3.1 Extraction of Local Features

We consider category independent region proposals to highlight local regions in still images.

Region proposal generation techniques highlight region segments in the image where the likelihood of the presence of an object part is high. This provides a structured way to identify interesting locations in the image and thus reduces the search space for efficient codeword generation. We specifically work with the objectness paradigm for region proposals generation from still images which is based on modeling several aspects regarding the characteristics of the objects in a Bayesian framework. Each region proposal is further represented by the CNN features. We prefer the ImageNet pre-trained VGG-F [5] model which has an architecture similar to AlexNet [11], and comprises of 5 convolutional layers and 3 fully-connected layers. The main difference of VGG-F and AlexNet is that VGG-F contains less number of convolutional layers and uses a stride of 4 pixels leading to better evaluation speed than the AlexNet architecture.

3.2 Discriminative Dictionary Learning

We first build category specific codebooks and then concatenate all the local codebooks to generate a global codebook.

Separate Dictionary Learning for Each Category. For a given $l \in \{1, 2, \ldots, L\}$, the dictionary learning process is summarized as follows:

1. For each training instance with the category label l, we first group the local descriptors using MS clustering and consider the cluster centroids as constituting the reduced set of local descriptors. MS is an iterative, non-parametric clustering method which does not require an estimation of the number of clusters as input. Instead, it relies on the kernel density estimate in the feature space to group samples which form dense clusters. Given $F_i = \{F_i^1, F_i^2, \ldots, F_i^{\alpha_i}\}$, the kernel density estimate at a point F_i^k is expressed as

$$f(F_i^k) = \frac{1}{\alpha_i h^d} \sum_{m=1}^{\alpha_i} K\left(\frac{F_i^k - F_i^m}{h}\right) \tag{1}$$

   where K is a radially symmetric kernel function and h defines the width of the Parzen window to highlight the neighbourhood around F_i^k. A cluster is identified as the region where the data density is locally maximum. This

can alternatively be interpreted as the local regions where $\nabla f \approx 0$. ∇f can efficiently be calculated by iteratively shifting the centroids of the Parzen windows until the locally dense regions are reached [7].
Since all the descriptors in a dense region in the feature space highlight near similar local features, the mean-shift clustering is able to select one unique representative for all of them. Further, since mean-shift implicitly estimates the number of clusters present in the dataset, hence, the problem of over-merging is greatly reduced. On the other hand, spherical clustering techniques like k-means and fuzzy c-means create suboptimal codebooks as most of the cluster centroids fall near high density regions, thus underrepresenting equally discriminant low-to-medium density regions. MS resolves such problem by focusing on locally dense regions in the feature space. Let $\widehat{F}_i = \{\widehat{F}_i^1, \widehat{F}_i^2, \ldots, \widehat{F}_i^{\widehat{\alpha_i}}\}$ represents the new set of local descriptors for the i^{th} training instance where each $\widehat{F}_i^k$ represents a cluster centroid.

2. Once $\widehat{F}_i$s are constructed for all the training instances with category label l, we vector quantize all such $\widehat{F}_i$s using MS clustering to build a temporary codebook $C_l = \{C_l^1, C_l^2, \ldots, C_l^{\beta_l}\}$ for the category with each C_l^k representing a codeword (cluster centroid). Similar to the previous stage, it is guaranteed that C_l is ensured to capture all the potential local features for the l^{th} category.

$\{C_1, C_2, \ldots, C_L\}$ are constructed in the similar fashion for $l \in \{1, 2, \ldots, L\}$. It is to be noted that the labels of the codewords depend upon the action categories they refer to. Further, the sizes of the C_ls may differ from each other. The C_ls thus obtained are not optimal in the sense that they contain many codewords with low discriminative property. Such codewords need to be eliminated in order to build robust category specific codebooks. However, we need a measure to rank the descriptors based on their discriminative ability. In this respect, the following observations can be made:

- A potentially discriminative codeword is not frequent over many of the categories constituting the dataset.
- Most of its nearest neighbors in $\{C_1, C_2, \ldots, C_L\}$ share the same class label with the codeword under consideration.

We model the first observation in terms of the idea of conditional entropy whereas the second observation is replicated by the tf-idf score.

For a given codeword C_l^k, we find out the labels of its T nearest neighbours over the entire set of codewords in $\{C_1, C_2, \ldots, C_L\}$ and subsequently define the conditional entropy measure as:

$$H(Y|C_l^k) = -\sum_{l'=1}^{L} p(l'|C_l^k) \log_2 p(l'|C_l^k) \tag{2}$$

where $p(l'|C_l^k)$ represents the fraction of the retrieved codewords with label l'. For discriminative codewords, i.e. the ones which do not span many categories,

H is small whereas the value of H grows with the selection of codewords shared by many categories.

In addition to the H score, we also expect the nearest neighbours to be populated from the same category as of C_l^k. In order to impose this constraint, we define the tf-idf score for C_l^k as follows:

$$TI(C_l^k) = \frac{|C_{l'}^{k'}|C_{l'}^{k'} \in knn(C_l^k)\ AND\ l' = l|}{|C_{l'}^{k'}|C_{l'}^{k'} \in knn(C_l^k)|} \tag{3}$$

Both the measures are further combined in a convex fashion to define the ranking measure as follows:

$$Rank(C_l^k) = w_1\ \frac{1}{H(Y|C_l^k)} + (1 - w_1)\ TI(C_l^k) \tag{4}$$

We repeat this stage for all the codewords in $\{C_1, C_2, \ldots, C_L\}$. As already mentioned, the $Rank(C_l^k)$ has high values for potentially discriminative and category specific codewords. We rank the codewords on the basis of the *Rank* scores and select top B codewords in a greedy fashion in order to define the final codebook $\widehat{C}_l$ for category l.

Global Dictionary Construction. The local codebooks obtained in the previous stage are concatenated in order to obtain a global codebook $\widehat{C} = [\widehat{C_1}\widehat{C_2}\ldots\widehat{C_L}]$ of size $L * B$.

3.3 Feature Encoding Using $\widehat{C}$

We represent each visual entity with respect to $\widehat{C}$ using efficient LLC based encoding technique due to its ability to deal with CNN features in case of action recognition in still images.

3.4 Classification

The final classification is performed using random forest ensemble classifier [2]. The decision tree learning algorithm used is information gain and bootstrap aggregation is employed to learn the ensemble model. Thus the forest reduces classifier variance without increasing bias. Random subspace splitting is used for each tree split and we consider $\sqrt{d}$ features for each split given d original feature dimensions. The generalization is performed by applying majority voting on the outcomes of the learned trees.

4 Experimental Details

4.1 Dataset

We consider the Stanford-40 [24] still image action recognition database to evaluate the effectiveness of the proposed framework.

Stanford-40 actions is a database of human actions with 40 diverse action types, e.g. brushing teeth, reading books, blowing bubbles, etc. The number of images per category ranges between 180 to 300 with a total of 9352 images. We use the suggested [24] train-test split with 100 images per category as training and remaining for testing.

4.2 Experimental Setup

The following experimental setup is considered in order to evaluate the performance of the proposed framework.

- MS clustering is used in conjunction with the Gaussian kernel. The adaptive bandwidth parameter (h) is fixed empirically as $\frac{D}{m}$ ($1 \leq m \leq 10$), where D is the average pairwise distance of all the local descriptors extracted from all the visual entities of each category. The same setup is repeated for MS clustering in the entity and the category levels (Sect. 3.2).
- We extract 500 region proposals per image for the Stanford-40 dataset. Figure 1 depicts the extracted region proposals for a pair of images from the dataset. We further discard proposals which are largely overlapping to each other (overlap of $\geq 50\%$) in order to highlight potentially discriminative local patches in the images.
- The number of final distinctive codewords selected for each class based on the proposed ranking measure in Sect. 3.2 is set between $100 \leq B \leq 500$ and the best classification performances are reported. In LLC encoding, 100–200 nearest neighbors per local descriptor are considered to encode the images. We select the optimal hyper-parameters by cross-validation. Each image in

Fig. 1. Extraction region proposals from images of Stanford-40 using objectness

the Stanford-40 dataset is optimally represented by a sparse vector of length 8000×1 (100 neighbors in LLC).
- Each component tree in the random forest model is essentially a classification and regression tree (CART) [2]. We conduct experiments with random forest of different sizes (500–2000) and find that a random forest with 1000 CART trees exhibits superior performance.
- We compare the overall classification performance of the proposed technique with the representative techniques from the literature. All the experiments are repeated multiple times and the average performance measures are reported.

4.3 Performance Evaluation

Table 1 mentions the accuracy assessment of different techniques for the Stanford-40 dataset. The performances of the methods based on hand-crafted SIFT-like features are comparatively less ($\approx$35.2%) [21] since the differences in human attributes for many of the action classes are subtle. Part learning based strategies obtain better recognition performance in this respect by explicitly modeling category specific parts. An overall classification accuracy of 40.7% is obtained with the generic expanded part models (EPM) of [19] which is further enhanced to 42.2% while the contextual information is incorporated in EPM. The best performance with shallow features obtained for this dataset is 45.7% by [24] which performs action recognition by combining bases of attributes, objects and poses. Further they derive their bases by using large amount of external information. The deep CNN models further enhance the state of the art performance in this respect thanks to the high level features they encode and the ImageNet pre-trained AlexNet reports a classification accuracy of 46% [11].

The proposed method further enhances the recognition performance to 49% while considering $B = 200$ codewords per visual category. We further observe a minor enhancement ($\leq$0.75%) in the classification accuracy for $B \leq 275$ after which the performance degrades to some extent before being saturated to 47%. In this case, our method encapsulates the advantages of deep and shallow models effectively in a single framework. The CNN based region proposals are capable of encoding high level abstractions from the local regions. Since the images are captured in unconstrained environments, the backgrounds are uncorrelated in different images of a given category. The per category dictionary learning strategy reduces the effects of such background patches and the proposed ranking measure further boosts the proposals corresponding to the shared human attributes, human-objects interaction etc. for a given action category. In contrast to other techniques which are based on SVM classifier, our framework relies on the random forest model which does not explicitly require any cross-validation. Further, the ensemble nature of the classifier reduces the number of misclassified actions to some extent. We further observe that performance of the random forest model gradually improves with growing number of CART trees within the range 500–1000 and a random forest model with 1000 trees outputs the best performance.

Table 1. A summary of the performance of our classification framework for the Stanford-40 data in comparison to the literature

Method	Classification accuracy
ObjectBank [14]	32.5%
LLC with SIFT features [21]	35.2%
Spatial pyramid matching kernel [12]	34.9%
Expanded parts model [19]	40.7%
CNN AlexNet [11]	46%
Proposed framework (with a per class codebook with 200 codewords and LLC encoding)	49%

5 Conclusion

We introduce a novel supervised discriminative dictionary learning strategy for the purpose of action recognition from still images. We leverage the available training samples to optimally rank local features which are robust and discriminative. Further, we initially cluster the local features at the entity as well as category levels to eliminate the effects of features corresponding to non-recurrent or background locations. The proposed ranking paradigm holds wider applications in areas including feature selection, ranked set generation for retrieval etc. The effectiveness of the proposed dictionary learning is validated on challenging dataset (Stanford-40), on which, superior performance measures can be observed in comparison to popular techniques from the literature. We currently focus on the *learning to rank* paradigm for effective dictionary learning.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2189–2202 (2012)
2. Bishop, C.M.: Pattern recognition. Mach. Learn. **128** (2006)
3. Campbell, R.J., Flynn, P.J.: A survey of free-form object representation and recognition techniques. Comput. Vis. Image Underst. **81**(2), 166–210 (2001)
4. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC, vol. 2, p. 8 (2011)
5. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
6. Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K., Buckles, B.P.: Advances in human action recognition: a survey. arXiv preprint arXiv:1501.05964 (2015)
7. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
9. Jiang, Z., Lin, Z., Davis, L.S.: Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1697–1704. IEEE (2011)

10. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: distinctive parts for scene classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 923–930 (2013)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 2, pp. 2169–2178. IEEE (2006)
13. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems, pp. 801–808 (2006)
14. 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, pp. 1378–1386 (2010)
15. Lin, W.C., Tsai, C.F., Chen, Z.Y., Ke, S.W.: Keypoint selection for efficient bag-of-words feature generation and effective image classification. Inf. Sci. **329**, 33–51 (2016)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
17. Mairal, J., Ponce, J., Sapiro, G., Zisserman, A., Bach, F.R.: Supervised dictionary learning. In: Advances in Neural Information Processing Systems, pp. 1033–1040 (2009)
18. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006). doi:10.1007/11744085_38
19. 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, pp. 652–659 (2013)
20. Sicre, R., Jurie, F.: Discriminative part model for visual recognition. Comput. Vis. Image Underst. **141**, 28–37 (2015)
21. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3360–3367. IEEE (2010)
22. Wang, Y., Tran, D., Liao, Z., Forsyth, D.: Discriminative hierarchical part-based models for human parsing and action recognition. J. Mach. Learn. Res. **13**(Oct), 3075–3102 (2012)
23. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2030–2037. IEEE (2010)
24. 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: 2011 International Conference on Computer Vision, pp. 1331–1338. IEEE (2011)
25. Zhang, L., Li, C., Peng, P., Xiang, X., Song, J.: Towards optimal VLAD for human action recognition from still images. Image Vis. Comput. **55**, 53–63 (2016)