GLOBAL FOR COARSE AND PART FOR FINE: A HIERARCHICAL ACTION RECOGNITION FRAMEWORK

Weiwei Liu[1], *Chongyang Zhang*[1,2*], *Jiaying Zhang*[1], *and Zhonghao Wu*[1]

[1]School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
[2]Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai 200240, China
*Corresponding email: sunny_zhang@sjtu.edu.cn

ABSTRACT

Action recognition is one significant yet challenging task in computer vision. Recent methods mainly model an end-to-end one-stage non-deep or deep learning networks to distinguish different action categories. In this paper we introduce one novel hierarchical action classification framework: Unlike existing one-stage recognition models, the proposed work improves the recognition accuracy by: 1) developing a hierarchical coarse-to-fine action classification framework by dividing the recognition processing into two stages: coarse-grained classification and fine-grained classification, and 2) representing actions in different stages with different granularity features representation: global features are utilized for coarse classifiers while more body parts patterns for fine-grained classifiers are aggregated. Experiments on two widely-tested benchmark datasets show that our method can achieve state-of-the-art or competitive performance compared with existing results using one-stage models, with advantages regarding the recognition accuracy and robustness.

Index Terms— Action recognition, coarse-to-fine, hierarchical framework, two stages, granularity

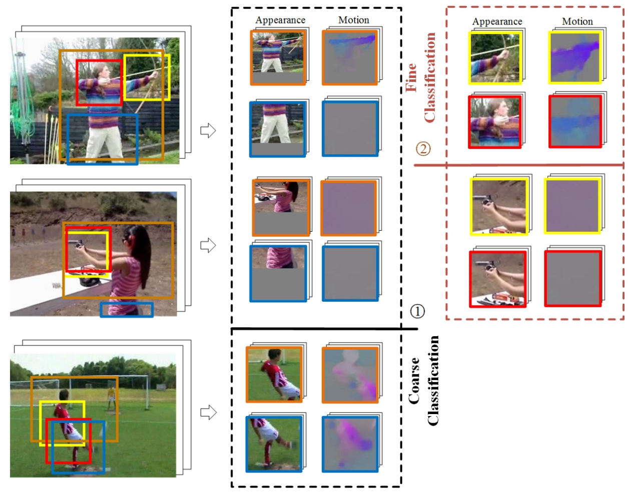

Fig. 1. Illustration of our motivation: In the first coarse-grained classification stage, shoot actions can be distinguished easily from kickball using whole body features; while in the second fine-grained classification stage, more attention should be paid to the parts of body to differentiate the shoot gun and shoot bow correctly.

1. INTRODUCTION

Action recognition has attracted much attention due to its importance in many applications. Different with image classification, the difficulties of action recognition include not only diverse interference factors, such as perspective, illumination, deformation, occlusion and background, but also the complexity in intra-class and between-classes induced by spatiotemporal 3D dimension information. Although the recent advances in deep convolutional networks (ConvNets) have brought some improvements on action recognition [1, 2, 3, 4], it remains challenging due to the large variation of video scenarios and the interferences from noisy contents unrelated to the video topic.

With the development of deep ConvNets[5], many ConvNet based methods were recently proposed for action recognition, which utilize ConvNets to automatically obtain the feature representation for actions. Ji, et al. and Tran, et al. utilize a 3D ConvNet to recognize actions in video[6, 7], Simonyan and Zisserman propose a two stream framework which uses two ConvNets to respectively extract features from two information streams (i.e., appearance and motion) and fuse them for recognition[4]. Based on these frameworks, recent researches further improve the effectiveness of ConvNet features by including additional information sources, such as pose or human part based CNN features proposed in [8] and [9]. Spatial-temporal attention model is also introduced into action recognition researches, such as recurrent attention convolutional neural network in [10], key volume mining in [11], and action tubes in [12]. Most of the existing works [3, 13, 14]are targeted at learning features for directly describing actions' individual action classes, while the shared

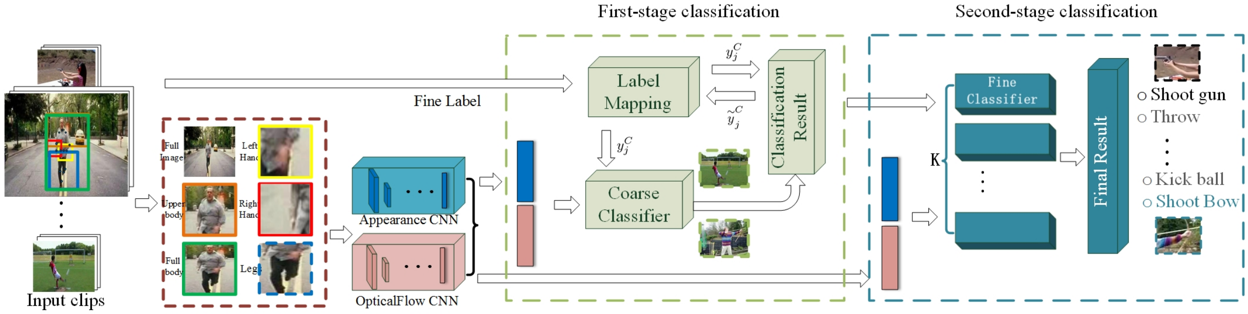

Fig. 2. Hierarchical action classification structure. Firstly body part patches are cropped based on human joints and different part patches appearance and motion features are aggregated for both two stage classifiers. For first-stage classification we pay more attention on global information like global body, while for each second-stage classifier we focus on body part features.

characteristics in different action class granularities are less studied [15]. This restrains them from precisely distinguishing the subtle difference among ambiguous actions. Although some methods [16] obtain different levels of generality by integrating features in multi-ConvNet layers, they still focus on directly representing the individual action classes and do not consider the more precise feature representation for actions at different class granularities.

In this paper we introduce one novel hierarchical action classification framework: Unlike existing one-stage recognition models, the proposed work improves the recognition accuracy by: 1) developing a hierarchical coarse-to-fine action classification framework by dividing the recognition processing into two stages: coarse-grained classification and fine-grained classification, and 2) representing actions in different stages with different granularity features: global features are utilized for coarse-grained classification while more body parts patterns are aggregated for fine-grained classification.

2. HIERARCHICAL TWO-STAGE ACTION RECOGNITION

The framework of our approach is shown in Fig.2. Similar to[8], Human left/right wrist joint are used to cropping left/right hand body part image. Human neck, belly, face, shoulder, hip and elbow body joints are taken to defined the upper body part patch. Lower body part patch is constructed based on hip, knee and ankle body joint. Body parts' appearance (RGB) and optical flow deep features are obtained firstly. In the first-stage, each sample is identified as one coarse action category by coarse-grained classifier; Secondly, the shared appearance & motion deep features and the coarse label obtained in the first stage are combined to decide the final action class of the input video by one fine-grained classifier. Note that different body parts of the input video will be aggregated in different stages in the proposed framework.

Coarse-grained classification can be divided into two steps. Firstly action videos are roughly separated into K coarse categories. JHMDB action dataset can be separated into $UpperBody\ Actions$, $LowerBody\ Actions$ and $FullBody\ Actions$ based on the main activity-executing parts of each action category. For j_{th} action category we give a coarse label y_j^C with $Map : [1, T] \to [1, K]$ where T is the number of fine categories. Details about roughly separation on JHMDB dataset are show in Tabel1.

Coarse Category	Actions
LowerBody	climb_stairs, jump, kick_ball, run, walk
UpperBody	brush_hair, catch, clap, golf, pour, shoot_ball, shoot_bow, shoot_gun swing_baseball, throw, wave
FullBody	pick, pullup, push, sit, stand

Table 1. Roughly separation of JHMDB dataset in 3 coarse categories

After roughly separation, confusion matrix result of coarse-grained classification is analyzed. The j_{th} fine action category, where coarse-grained classification error exceeds threshold Thr, will be modified to the corresponding coarse category. Our goal of changing the j_{th} action coarse label y_j^C is to achieve better initial classify result :

$$y_j^{C*} = \arg\max_{y_j^C} \sum_{j=1}^{T} \sum_{i=1}^{M_j} L_{ij}(y_{ij}^C, \tilde{y}_{ij}^C), y_j^C \in [1, K] \quad (1)$$

where M_j denotes number of j_{th} action category testing sample. y_j^C means coarse-grained classification label and $\tilde{y}_j^C$ is the predicted coarse label. Here L is function:

$$L(y_1, y_2) = \begin{cases} 1 & y_1 = y_2, \\ 0 & y_1 \neq y_2 \end{cases} \quad (2)$$

In our experiments on JHMDB dataset, action category $shoot\_ball$ should be grouped into $FullBody\ Action$ coarse category where $Thr = 0.3$. After coarse label modification, coarse-grained classifier are re-trained with updated category separation. The coarse-grained classification accuracy came to 91.3% with 5.6% gains. Good classification result of first-stage implies video dataset is split to coarse categories well.

On the right side of Figure2, similar to first-stage classification, CNN features for each coarse category are extracted and aggregated. Fine-grained classifier are trained with fine label and extracted features descriptor.

We refer $p_k(y_i^F = j|x_i)$ as the k_{th} coarse category classifier prediction probability. Combined with the first-stage prediction result $p(y_i^C = k|x_i)$, the final prediction[17] for each x_i:

$$p\left(y_i = j|x_i\right) = \frac{\sum_{k=1}^{K} p(y_i^C = k|x_i) p_k(y_i^F = j|x_i)}{\sum_{k=1}^{K} p(y_i^C = k|x_i) I_k\left(x_i\right)} \tag{3}$$

Here $I_k(x_i)$ denotes whether video clip x_i's coarse label is k:

$$I_k(x_i) = \begin{cases} 1 & y_i^C = k \\ 0 & y_i^C \neq k \end{cases} \tag{4}$$

Softmax function[18] is used to compute prediction probability for coarse and each fine classifier.

3. STAGED PATCHES AGGREGATION: GLOBAL FOR COARSE AND PARTS FOR FINE

Considering the fact that different action class granularities may need different representation features , one staged part patches aggregation mechanism is developed in this work.

For each action video clip, RGB image and optical flow image were cropped into part patches including $lefthand$, $righthand$, $upperbody$, $fullbody$ and $fullimage$ parts based on poses information where $legs$ part is appended to $LowerBody\ Action$ coarse category individually. Each part patch was resized to standard CNN input size: $224 * 224$. And CNN features for each part patch were extracted with finetuned *VGG-f*[19] network.

Features from optical flow image are computed with motion network provided by[12]. Normalization and non-linear pooling over multi-frames processing are added for both spatial and temporal features for final feature vector.

For first-stage classifier and each classifier of second-stage, different combinations of the human body part patches were investigated for finding discriminative part patches. Feature vector V_i of i_{th} video clip is composed by feature descriptor f_i^p where p denotes different part patch.

$$V_i = [f_i^{p_1}, f_i^{p_2}, ... f_i^{p_n}]^{\mathrm{T}} \tag{5}$$

We went through and concatenated all possible combinations of part patches descriptor for each classifier and try to maximize the number of true positive sample:

$$V_i^* = \arg\max_{V_i} TP(V_i) \tag{6}$$

Here TP means true positive and the evaluation metric is average accuracy of three splits:

$$Average\ accuracy = \tfrac{1}{3} \sum_{i=1}^{3} \frac{TP(V^*)}{N_i} \tag{7}$$

N_i denotes number of testing sample in i_{th} split. This metric was used to find discriminative part patches for each classifier and detailed results are discussed in Section4.2.

We use linear kernel one-vs-all SVM[20]to classify video and note that feature vector dimension was reduced compared with[8] when we selected one or several patches rather than all of them.

4. EXPERIMENTS RESULT

Dataset: Experiments are performed on two datasets: JHMDB [9], which contains 928 clips in 21 action categories and sub-MPII Cooking, which is a subset of fine-grained MPII Cooking[21]. MPII cooking action dataset has 12 subjects and each subject completed some activities continuously. Several action categories have fewer sample number than others. We choose action category which the number of samples is greater than 50 and refer it as sub-MPII Cooking dataset which comprising 2976 video clips distributed in 27 action categories. Same as common action datasets, the evaluation metric is the average accuracy of 3 splits.

4.1. Experiments on coarse-grained classification

Based on the main activity-executing parts of each action category, JHMDB dataset was divided into three coarse categories: *UpperBody* Actions, *LowerBody* Actions and *FullBody* Actions. Another reason for that separation is that part patches were cropped from videos based on human poses information. Similar separation is conducted on sub-MPII Cooking dataset. VGG16[19] network structure is experimented as feature extractor. In Table2 we reported the accuracy for our first-stage classification accuracy.

datasets	JHMDB	sub-MPII Cooking
VGG-f	85.7	82.5
VGG-f+LM	91.3	91.0
finetuned VGG-f+LM	92.3	91.2
finetuned VGG-f+LM+SP	**92.6**	**92.7**

Table 2. Accuracy of first-stage classification. LM: re-Label Mapping. SP: Staged Part patches (% average accuracy)

Re-label mapping according to Formula1 led to accuracy increases of 5.6% and 9.5% on JHMDB and sub-MPII Cooking dataset, respectively. Finetuning the network with basic two stream method improved the results to some extent. Overfitting when finetuned network might limit performance due to the few amount of training data. Staged part patches descriptors increased result slightly than all part patches and detailed discussions are in the section4.2.

4.2. Experiments on staged patches aggregation

Firstly the influence of single part patch was investigated for both two stage classifiers. As showed in Figure3, for coarse-grained classification $fullbody$ and $upperbody$ part which have global information get the competitive result accuracy. For the $UpperBody\ Actions$ category classification, only upper part of body patch $upperbody$ have good performance with a margin of $\sim 5\%$ than other single part patch.

Comparing two part patches combination result in Figure3, for coarse-grained classification the accuracy differentiate small from best result where any of $upperbody$, $fullbody$ and $fullimage$ part patches feature are aggregated. In Figure3(b), the performances become competitive only when the $upperbody$ part patch is chosen, denoting that $upperbody$ part patch is discriminative for $UpperBody\ Actions$ classification.

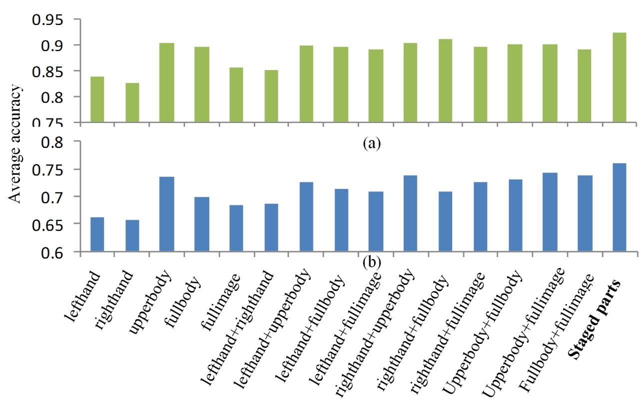

Fig. 3. Accuracy of coarse and fine classifier with single or two part patches on JHMDB dataset. (a)Coarse-grained classifier result (b)$UpperBody\ Actions$ category classifier result. Staged Parts means best part patches combination with highest performance.

Experiments on all combination of patches are implemented and table3 shows details. Here patch$_d$ means discriminative part patch. The above experimental results show that for coarse-grained classification. global information is necessary. For each fine-grained classification, we need to aggregate more part patch information where some global information like $FullBody$ is not important.

4.3. Comparison with the state-of-the-art methods

To prove our coarse-to-fine hierarchy classification structure, performances on two datasets are showed in Table4. For sub-MPII Cooking dataset, we have implemented HLPF[9]and P-CNN[8] methods where poses are provided by[8]. The metric

	Lowerbody	Upperbody	FullBody
Full image	73.4	68.4	82.4
Single patch$_d$	73.4	73.5	85.5
All parts	67.1	73.3	88.7
Staged Parts	**75.8**	**76.0**	**90.0**

Table 3. Accuracy of each fine-grained classifier with different patches combination on JHMDB dataset. (% average accuracy)

is average accuracy of three splits, in which each category is split with the ratio of 2:1 of training and testing.

datasets	JHMDB	sub-MPII Cooking
HLPF	**76.0**	28.7
iDT+FV	65.9	-
P-CNN$_{MatConvNet}$	73.7	52.2
Ours $_{MatConvNet}$	75.2	**54.7**

Table 4. Average accuracy of the state-of-the-art methods on JHMDB and sub-MPII Cooking dataset(% average accuracy)

Comparing with HLPF and iDT+FV methods, our work benefits from part body information while HLPF is heavily dependent on pose correctness and not suitable for fine-grained action classification task on sub-MPII Cooking dataset since the pose variation is small. Especially when we use pose estimation result rather than ground-truth pose. Different from P-CNN, our work analyzes the influence of different part patches combination and owing to hierarchical structure, we choose the corresponding staged parts for each classifier and achieve better performance on two datasets.

5. CONCLUSION

This paper presents a novel framework for action recognition. Our framework consists of two key ingredients: 1) a hierarchical coarse-to-fine action classification framework, which divides the recognition processing into two stages: coarse-grained and fine-grained classification, so as to obtain a more precise feature representation for different granularity actions; 2) an stage-adaptive aggregation model which can select and aggregate different part patches at different classification stages, and thus better leveraging of the feature aggregation mechanism can be achieved. Experimental results show that our approach achieves the state-of-the-art or competitive performance.

6. ACKNOWLEDGEMENT

This work was partly funded by the National Key Research and Development Program (2017YFB1002401), NSFC (No.61 571297, 61521062, No.61420106008), and STCSM (18DZ227 0700).

7. REFERENCES

[1] Aaron F. Bobick and James W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[2] Paul Scovanner, Saad Ali, and Mubarak Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *International Conference on Multimedia 2007, Augsburg, Germany, September*, 2007, pp. 357–360.

[3] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[4] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 1, no. 4, pp. 568–576, 2014.

[5] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.

[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[7] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2014, pp. 4489–4497.

[8] Guilhem ChRon, Ivan Laptev, and Cordelia Schmid, "P-cnn: Pose-based cnn features for action recognition," in *IEEE International Conference on Computer Vision*, 2015, pp. 3218–3226.

[9] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black, "Towards understanding action recognition," in *IEEE International Conference on Computer Vision*, 2013, pp. 3192–3199.

[10] Jianlong Fu, Heliang Zheng, and Tao Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4476–4484.

[11] Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao, "A key volume mining deep framework for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1991–1999.

[12] Georgia Gkioxari and Jitendra Malik, "Finding action tubes," in *IEEE International Conference on Computer Vision*, 2014, pp. 759–768.

[13] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *IEEE International Conference on Computer Vision*, 2016, pp. 7445–7454.

[14] Du Tran, Jamie Ray, Zheng Shou, Shih Fu Chang, and Manohar Paluri, "Convnet architecture search for spatiotemporal feature learning," in *IEEE International Conference on Computer Vision*, 2017.

[15] Weiyao Lin, Yang Mi, Jianxin Wu, Ke Lu, and Hongkai Xiong, "Action recognition with coarse-to-fine deep feature integration and asynchronous fusion," in *AAAI Conference on Artificial Intelligence (AAAI 2018)*, 2017.

[16] Jialin Wu, Gu Wang, Wukui Yang, and Xiangyang Ji, "Action recognition with joint attention on multi-level deep features," in *In CoRR abs/1607*, 2016.

[17] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis Decoste, Wei Di, and Yizhou Yu, "Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2014, pp. 2740–2748.

[18] John S. Bridle, *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition*, Springer Berlin Heidelberg, 1990.

[19] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Computer Science*, 2014.

[20] Chih Chung Chang and Chih Jen Lin, *LIBSVM: A library for support vector machines*, ACM(Association for Computing Machinery), 2011.

[21] Bernt Schiele, "A database for fine grained activity detection of cooking activities," in *Computer Vision and Pattern Recognition*, 2012, pp. 1194–1201.