ENHANCED ACTION RECOGNITION WITH VISUAL ATTRIBUTE-AUGMENTED 3D CONVOLUTIONAL NEURAL NETWORK

Yunfeng Wang[⋆], *Wengang Zhou*[⋆], *Qilin Zhang*[†], *Houqiang Li*[⋆]

[⋆]University of Science and Technology of China, Hefei, Anhui, China
[†]HERE Technologies, Chicago, Illinois, USA

ABSTRACT

Visual attributes in individual video frames, such as the presence of characteristic objects and scenes, offer substantial information for action recognition in videos. With individual 2D video frame as input, visual attributes extraction could be achieved effectively and efficiently with more sophisticated convolutional neural network than current 3D CNNs with spatio-temporal filters, thanks to fewer parameters in 2D CNNs. In this paper, the integration of visual attributes (including detection, encoding and classification) into multi-stream 3D CNN is proposed for action recognition in trimmed videos, with the proposed visual Attribute-augmented 3D CNN (A3D) framework. The visual attribute pipeline includes an object detection network, an attributes encoding network and a classification network. Our proposed A3D framework achieves state-of-the-art performance on both the HMDB51 and the UCF101 datasets.

Index Terms— Action Recognition, Visual Attributes, Detection, NetVLAD, Word2vec

1. INTRODUCTION

Action recognition has been extensively studied in past few years [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Among these methods, recognizing human actions in videos with convolutional neural networks (CNNs) has been a popular research topic [1, 12, 13, 7], thanks to the recent success of CNNs in various computer vision tasks [14, 15]. Typically, these methods incorporate 3D CNNs to capture spatio-temporal information, optionally with a separate optical flow stream to account for low-level motion. Due to the increased parameter size in 3D filters in 3D CNNs, 3D CNNs are typically much shallower than their 2D counterparts [2]. However, a new 3D CNN named "Two-Stream Inflated 3D ConvNet" (I3D) was recently proposed in [7], with much deeper but still computationally feasible network design.

This work was supported in part to Dr. Houqiang Li by 973 Program under contract No. 2015CB351803 and NSFC under contract No. 61390514, and in part to Dr. Wengang Zhou by NSFC under contract No. 61472378 and No. 61632019, the Fundamental Research Funds for the Central Universities, and Young Elite Scientists Sponsorship Program By CAST (2016QNRC001).

(a) Playing Dhol (b) Playing Tabla

Fig. 1: (a) "Playing Dhol", (b) "Playing Tabla" in UCF101, which confuse the I3D [7] algorithm. Visual attributes (dhol/tablas) could have eliminated the ambiguity.

One of the potential improvements to current 3D CNN based action recognition systems could be the explicit inclusion of visual attributes as auxiliary information [16, 17, 18] for classification. The visual attributes could be detected, encoded and classified with regular 2D CNNs effectively and efficiently. Visual attributes could play a vital role in some challenging action recognition cases, such as the one illustrated in Figure 1. The I3D [7] network struggles to distinguish "Playing Dhol" from "Playing Tabla". However, the visual attributes (e.g., marked by the red and blue boxes) could have helped to eliminate such confusions.

Based on this intuition, an enhanced action recognition framework is proposed, namely the visual Attribute-augmented 3D CNN (A3D), comprising both a 3D CNN pipeline and a visual attributes pipeline. The 3D CNN pipeline is a modified I3D network with temporally sub-sampled RGB/optical flow inputs, trained with filtered attributes. While the visual attributes pipeline consists of a YOLO9000 [19] visual attribute candidate's detector, a ResNet [20]/Word2Vec [21]+Mean-pool/NetVLAD [22] attributes encoders and different classifiers. Based on the output probabilities from both pipelines, the final prediction is constructed by global thresholding and weighted summation, as illustrated in Figure 2. In order to achieve fair comparison, we downloaded the network model pre-trained on dataset "Kinectics" and reimplemented the finetuning steps on both UCF101 and HMDB51 ourselves, and named this reimple-

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

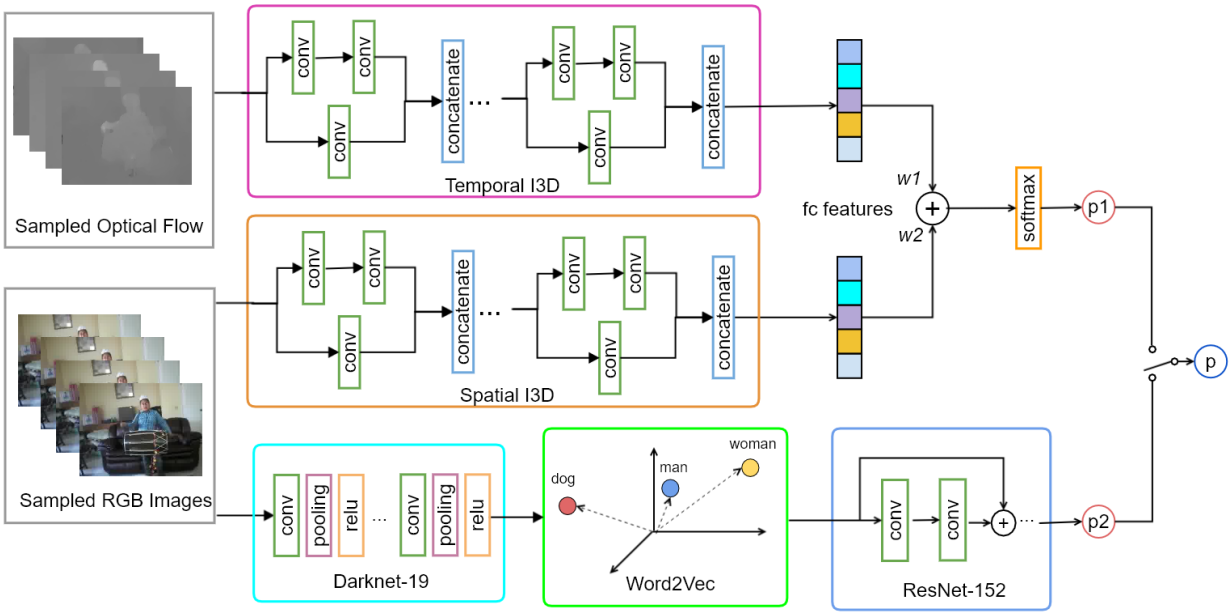

Fig. 2: A3D framework overview. Optical flow and RGB inputs are fed to the temporal stream and spatial stream of the I3D ConvNet, respectively, followed by early fusion which merges fc features and a softmax to obtain the 3D CNN pipeline prediction p_1. The visual attribute pipeline (at the bottom) samples RGB frames and detects visual attribute candidates (with e.g., Darknet-19 from YOLO9000), constructs video attribute representations (via e.g., Word2vec) followed by a classification network (finetuned ResNet-152). The prediction p_2 from this attribute pipeline is ultimately combined with p_1 by Eq. (1).

mented version as I3D*[1], which is later used for comparison in Table 4.

2. PROPOSED A3D FRAMEWORK

As in Figure 2, two pipelines are present in the A3D framework. The 3D CNN pipeline is illustrated on the top, with a temporal I3D stream and a spatial I3D stream, processing sampled optical flow images and regular RGB images, respectively. The fc feature outputs from both streams are merged before the softmax scoring layer (which is different from the original I3D implementation in [7]) to obtain pipeline prediction p_1. On contrary, the visual attribute pipeline is illustrated at the bottom of Figure 2, with only regular RGB images as inputs. This attribute pipeline contains a generic object (i.e., visual attribute candidate) detector (Darknet-19 from YOLO9000), a video attribute representation constructor (mean pooling, NetVLAD [22] or Word2vec [21]) and finally a CNN based classifier, which produces the pipeline prediction p_2. As an auxiliary information source, the attribute pipeline output p_2 is only activated when p_1 falls below a predefined global threshold. Details of major components of A3D are provided as follows.

Revised Two-stream Fusion in 3D CNN Pipeline. Specifically, the inflated Inception-V1 module proposed in [7] is used in both streams of the 3D CNN pipeline. After finishing network training, the outputs of the last fully connected layer from both streams are weighted (with $w_1 = 0.6$ and $w_2 = 0.4$, respectively. Values of the weights are empirically determined.), element-wise summed, and fed to a softmax scoring layer to obtain prediction p_1. Unlike the original fusion scheme in [7], where fusion is carried out after the softmax scoring layer, the revised fusion strategy in the 3D CNN pipeline of A3D leads to performance enhancements, as later detailed in Table 1.

Visual Attribute Candidate Detection. To exploit relevant visual attributes to action recognition tasks, Darknet-19 (from YOLO9000) object detector is applied on randomly selected video frames. Thanks to the very large number of object categories (approximately 9000), this object detector could serve as an off-the-shelf generic attribute mining tool. In order to remove obvious outliers with minimal useful information, attributes candidates with detected bounding boxes smaller than 20-pixels are removed. For an attribute a (there are approximately 9000 distinctive choices of a) selected from video of class t (t denotes the label of the video), we assign t as the label of a. Attribute candidates are cropped from original images and saved for subsequent steps.

Video Attribute Representation and Classification. The number of attributes detected from different video is variant, thus it is necessary to encode them in a format (i.e., video attribute representation) with consistent size. Three strategies

[1]The reimplmented I3D* slightly underperforms the official one.

of attribute representation construction are tested. We find the strategy *Word2Vec+ResNet-152 finetune* achieves the best accuracy. Due to limited space, please refer to the extended version of this paper [23] for detailed experiment settings and results.

Joint Inference of Two Pipelines. In the testing phase, both pipelines jointly perform classification inference. Given a test video, p_1 and p_2 are obtained by the 3D CNN pipeline and visual attribute pipeline, respectively. Final prediction p is obtained by,

$$p = \begin{cases} p_1 & \text{if } p_1 > T \\ p_2 & \text{if } p_1 \leq T \end{cases}, \quad (1)$$

where T is a global threshold of prediction confidence, empirically set to 0.1.

3. EXPERIMENTS

The proposed A3D framework is evaluated on both the HMDB51 and UCF101 datasets. the standard evaluation protocol is used and the average accuracies over three splits are reported (unless otherwise specified). In the following experiments, the pretrained (on Kinetics dataset) inflated Inception-V1 models are finetuned on UCF101 and HMDB51 dataset, separately. The temporal I3D stream and spatial I3D stream are also separately trained on optical flow images and RGB images, respectively. For each video, 64 regular RGB frames and 64 optical flow frames are randomly selected as the input to finetune the 3D CNN pipeline. Regular stochastic gradient descent (SGD) and back propagation (BP) are used to optimize training loss, with initial learning rate of 0.001, a learning rate decay of $0.8\times$ every 10 epochs and a total of 50 epochs. In the original I3D implementation [7], fusion is carried out after the softmax scoring layer. On contrary, the revised fusion in the proposed 3D CNN pipeline happens immediately after obtaining the features from fully connected layers. FC layer features are weighted by $w_1 = 0.4$ and $w_2 = 0.6$ (weight values are determined empirically) and summed element-wise. As shown in Table 1 (action classification based purely on p_1), the revised fusion outperforms original fusion on the "split1" of both datasets. Therefore, revised fusion is used throughout the remainder of the paper.

Thanks to the relatively simple appearance of trimmed videos in UCF101/HMDB51, one randomly selected frame is empirically sufficient for visual attribute extraction. We exploit YOLO9000 [19] by setting a low threshold of 0.02

Table 1: 3D CNN Pipeline: 2 fusion strategies.

Method	UCF101(split1)	HMDB51(split1)
Original Fusion	95.67%	79.00%
Revised Fusion	**97.09%**	**79.22%**

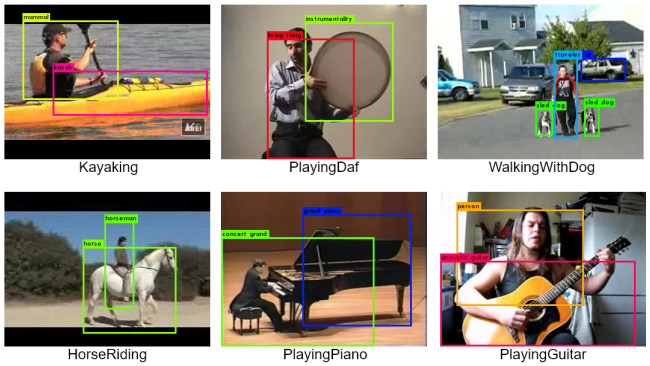

Fig. 3: Sample results of visual attributes detection

Table 2: Comparison of 3D CNN pipeline p_1 based classifier, video attribute pipeline p_2 based classifier and p based joint classifier on split1 of UCF101 and HMDB51.

Classifier based on	UCF101(s1)	HMDB51(s1)
3D CNN pipeline p_1	97.09%	79.22%
Attribute pipeline p_2	37.10%	27.05%
Joint A3D framework p	**97.44%**	**79.35%**

and generate abundant visual attribute candidates (samples are shown in Figure 3).

In addition, an ablation study is carried out and the results are summarized in Table 2. Although attribute pipeline p_2 based classifier substantially underperforms the counterpart based on p_1 (possibly due to excessive noises incurred by irrelevant objects), it stills helps to incorporate p_2 via a reasonable combination function (e.g., Eq. (1)). The classifier based jointly on both pipelines p achieves the highest accuracy on split1 of both datasets.

Computational complexity wise, the proposed A3D framework is only marginally heavier than the I3D algorithm. A comparison of running time[2] (including both the training phase and testing phase with split1 of both datasets) is summarized in Table 3. The bottom row in Table 3 contains the percentile overhead running time of A3D over I3D, which are all under 10%. In Table 4, the end-to-end action recognition accuracies are compared over all 3 splits of both datasets. The proposed A3D framework achieves the highest overall accuracies on both datasets.

4. CONCLUSIONS

In this paper, the A3D action recognition framework is proposed based on the explicit incorporation of visual attributes to a two-stream 3D CNN pipeline with a new fusion strategy. Different designs of multiple components (video attribute representation constructor, classifier) in the new video

[2]Based on our hardware setup, GPU: $1\times$ Nvidia GTX 1080Ti, CPU: $2\times$ Intel(R) Xeon(R) E5-2640 v4 2.40GHz and 512 GB memory.

Table 3: Running time comparison on split1 of 2 datasets

Method	UCF101(s1)		HMDB51(s1)	
	Training	Testing	Training	Testing
I3D	41.67h	0.53h	16.70h	0.22h
A3D	45.32h	0.61h	17.82h	0.26h
Overhead	8.76%	1.56%	6.71%	1.82%

Table 4: Comparison of A3D with competing methods

Method	UCF101	HMDB51
iDT [4]	85.9%	57.2%
Two-Stream [1]	88.0%	59.4%
C3D [2]	85.2%	-
TDD + iDT [24]	91.5%	65.9%
TSN [13]	94.2%	69.4%
P3D ResNet + iDT [25]	93.7%	-
ST-ResNet + iDT [26]	94.6%	70.3%
I3D*	97.1%	79.4%
Proposed A3D	**97.4%**	**80.5%**

attribute pipeline are tested and empirically determined. An ablation study confirms the value of the additional video attribute pipeline.

5. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[3] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.

[4] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.

[5] Xingyang Cai, Wengang Zhou, Lei Wu, Jiebo Luo, and Houqiang Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 141–154, 2016.

[6] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li, "Video-based sign language recognition without temporal segmentation," in *AAAI*, 2018.

[7] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[8] Yunfeng Wang, Wengang Zhou, Qilin Zhang, Xiaotian Zhu, and Houqiang Li, "Low-latency human action recognition with weighted multi-region convolutional neural network," *arXiv preprint arXiv:1805.02877*, 2018.

[9] Xin Lv, Le Wang, Qilin Zhang, Zhenzing Niu, Nanning Zheng, and Gang Hua, "Video object co-segmentation from noisy videos by a multi-level hypergraph model," in *ICIP, Athens, Greece*, 2018.

[10] Xuhuan Duan, Le Wang, Changbo Zhai, Qilin Zhang, Zhenzing Niu, Nanning Zheng, and Gang Hua, "Joint spatio-temporal action localization in untrimmed videos with per-frame segmentation," in *ICIP, Athens, Greece*, 2018.

[11] Jinliang Zang, Le Wang, Ziyi Liu, Qilin Zhang, Zhenxing Niu, Gang Hua, and Nanning Zheng, "Attention-based temporal weighted convolutional neural network for action recognition," in *AIAI, Rhodes, Greece*, 2018.

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016.

[13] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: towards good practices for deep action recognition," in *ECCV*, 2016.

[14] Lingyan Ran, Yanning Zhang, Wei Wei, and Qilin Zhang, "A hyperspectral image classification framework with spatial pixel pair features," *Sensors*, vol. 17, no. 10, pp. 2421, 2017.

[15] Lingyan Ran, Yanning Zhang, Qilin Zhang, and Tao Yang, "Convolutional neural network-based robot navigation using uncalibrated spherical images," *Sensors*, vol. 17, no. 6, pp. 1341, 2017.

[16] Qilin Zhang, Gang Hua, Wei Liu, Zicheng Liu, and Zhengyou Zhang, "Can visual recognition benefit from auxiliary information in training?," in *ACCV, Singapore*, 1-5 November, 2014, pp. 65–80.

[17] Qilin Zhang, Gang Hua, Wei Liu, Zicheng Liu, and Zhengyou Zhang, "Auxiliary training information assisted visual recognition," *IPSJ Trans. Comput. Vis. and Appl.*, vol. 7, pp. 138–150, 2015.

[18] Qilin Zhang and Gang Hua, "Multi-view visual recognition of imperfect testing data," in *ACM Multimedia*, 2015, pp. 561–570.

[19] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016.

[23] Yunfeng Wang, Wengang Zhou, Qilin Zhang, and Houqiang Li, "Visual attribute-augmented three-dimensional convolutional neural network for enhanced human action recognition," *arXiv preprint arXiv:1805.02860*, 2018.

[24] Limin Wang, Yu Qiao, and Xiaoou Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *CVPR*, 2015.

[25] Zhaofan Qiu, Ting Yao, and Tao Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *ICCV*, 2017.

[26] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes, "Spatiotemporal multiplier networks for video action recognition," in *CVPR*, 2017.