The Action Recognition with Multi-Clue Sparse 3-Dimensional Convolutional Networks

1st Hongtao Man
Inspur Electronic Information Industry Co.
Beijing, China
manht@inspur.com

2nd Deqiang Zou
State Key Laboratory of High-End Server & Storage Technology
Beijing, China
zoudeqiang01@inspur.com

3rd Tuo Li
Inspur Electronic Information Industry Co.
State Key Laboratory of High-End Server & Storage Technology
Beijing, China
lituo@inspur.com

***Abstract*—The current approaches of deep neural network lack capacity on action recognition, yet ineffective structure for spatiotemporal feature learning, and heavy parameters of architecture are even harsh to optimize. In this paper, we propose multi-clue sparse 3-dimensional convolutional networks(MS3D ConvNet) based on landmark works that aims to make it far known in this field. The MS3D ConvNet enhances channels of input through low, medium, and fast pathways with disparate frames of video, frame sequences considering as temporal clue sampled in terms of variance rate of action. The frames of each pathway are convolved with commonly shared kernels forming cohorts as initial results, and pile up three-way cohorts to the 3D block transmitting to subsequent 3D network. The 3D network is modified by DenseNet architecture with randomly dropping out connections among the layers, and introduces 3D convolutional kernels to deal with temporal information, eventually generating 256-dimension codes of actions. Action recognition employs centers of vector clusters of actions instead of stubborn format, which provide a robust convergence environment. Spontaneously, the MS3D ConvNet fosters capacity of action information and hardly ever sustains heavy amounts of parameters which deserves double profits. This fancy structure also leaves an end-to-end fashion training and a possibility to pre-train parameters. Besides that, we devise tailor strategy based on yolo algorithm to endow disparate factors on pixels in compliance with bounding box frontier, regarded as the means of attention. The MS3D ConvNet has been applied to three public action datasets and a self-defined dataset, which performs advanced steps on accuracy, convergence, and robustness.**

***Keywords*—action recognition, multi-clue sparse 3-dimensional convolutional networks, variance rate of action, tailor strategy**

I. Introduction

Retrospecting the course of past decades, the Internet environment has changed a lot about information modality among message, audio, image, video, and even live streaming due to increased network bandwidth. This process just boosts vast transmission and ample storage of data, especially various kinds of video data produced with high speed. Video recognition is much requisite that conducts analyses and identifications of video contents. Action recognition, as part of video recognition, is a hotpot research among all walks of life that aims to identify object action of video which draws excellent attention on intelligent surveillance of manufacture, innovative healthcare, interactive traffic, physical action, etc [1]–[3]. Action recognition involves identifying actions from video clips consisting a sequence of frames, and thus it seems a natural extension of image recognition which much promoted by deep learning. While image recognition has made advanced progress, there is still no clear front running technique on video recognition which leaves a widely promoting room [4].

Video data is of great magnitude compared with image data which spontaneously deserves more extensive network architecture and parameters, and it requires much hardware supporting and training time, even models hardly ever converge [5]. Video frames generally exist redundancy phenomenon, though the current sampling strategy is a not bad scheme, and it somehow decreases the quality of data [6]. The over-proportion of the background of video frames, not like image data anymore, produces a series of adverse effects that disperses action clues of object and increase computation assignments, and variation of object area also disturbs action recognition [7]. The argument of input modality of optical flow and raw RGB is controversy [8]. Action recognition requires spatiotemporal information, and optical captures the variant of video frames which is not beneficial for subsequent works, though optical flow produces a good result. Original videos incorporate complete action information that fully applying them is advisable [9]–[12]. Action recognition of architecture is a lack of commonly admitted schemes.

Motivated by the above observations, we proposed Multi-Cue Sparse 3D ConvNet architecture accompanying with tailor strategy to carry forward action recognition. The MS3D structure aims to enhance the information density of video data with multi-pathways while relieving the information redundancy with interval sampling with action ratio distribution. Each pathway is independent at the first stage, and frames of each pathway are convolved to initial feature maps with share-kernel, just stack them into blocks of pathways, respectively. Following, blocks are convolved with corresponding 3D kernels to initially capture spatiotemporal information, and pile

them up to a 3D module passing to sparse Densenet3D. Sparse Densenet3D is introduced with 3D kernels and a dropping rate that randomly drops connection among layers, eventually generating a 256 dimensional vector. Action recognition would rely on centers of vector clusters of actions in the repository, which provides a robust convergence environment. Tailor strategy based on yolo algorithm, endows disparate work on pixels in compliance with bounding box frontier, which is regarded as a means of attention. We conduct experiments on HMDB51, UCF101, Kinetics, and self-defined dataset compared with current state-of-the-art models, MS3D ConvNet performs advanced effects.

II. Related Works

Among various action recognition models based on deep learning, 3D convolutional models and two-stream convolutional models are two widely used models. Si monyan et al. [13] proposed the two-stream architecture of model, cohorts of optical flow and raw RGB as inputs, exploiting video frames with two modalities which improve accuracy on benchmarks. This optical way has been a foundation of amounts of competitive results in the literature. Nevertheless, optical flow barely captures the temporal sequence of action but leaves a piece of static information with invariant of appearance doing action recognition [14], which exists some arguments. Moreover, optical flow extracting is kind of complicated, and the end-to-end fashion training is in plight here. Raw RGB somehow incorporates complete information of actions, more than optical flow, and it is much valuable in future works [15]. From very early on, the 3D convolutional model performs along both spatial and temporal dimensions which is a natural way to capture spatiotemporal information. The 3D convolutional network(C3D) [16] was proposed with extensive search for the best 3D convolutional kernel and architecture, and that 3D convolution brings temporal segment function. Carreira et al. [17] proposed the model based on Inception-v1-3D and two-stream network(I3D), achieving a good performance, though the optical flow was infiltrated in this model. Temporal 3D ConvNets(T3D) [18] proposed temporal layers that model variable temporal convolution kernel depths, extending the DenseNet architecture with this temporal layer in 3D ConvNet formats. Efficient convolutional network(ECO) [19] proposed a 3D ConvNet that firstly uses a single frame to make initial classification of action and captures distant frames along initial frames feeding the feature representation into the 3D network to learn contextual relationship. Slow-Fast network [20] proposed two-pathway Slow-Fast model in that the slow pathway is designed to capture semantic information using sparse frames, and the fast pathway is responsible for capturing rapidly changing motion with dense frames. By treating the video at different temporal rates, the two pathways are fused by lateral connections with their own expertise tending to gain much. Though amounts of variants on 3D ConvNet of considering landmarks have been introduced to action recognition, the advanced running architectures are controversial in terms of accuracy and robustness, seemingly no exact winner scheme.

III. Multi-Clue Sparse 3D ConvNet(MS3D)

We believe that 3D ConvNet possesses the potential to capture spatiotemporal information to conduct action recognition. We proposed Multi-Clue Sparse 3D ConvNet architecture accompanying with tailor strategy to carry forward action recognition, which goal is to achieve a better effect.

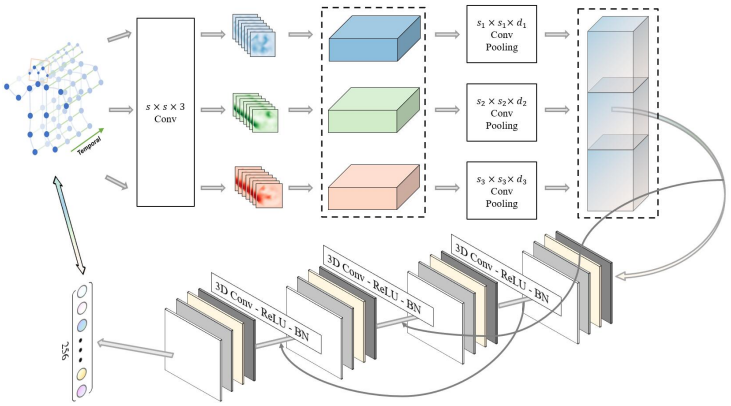


Fig. 1. Architecture of MS3D ConvNet.

A. Structure of MS3D ConvNet

MS3D ConvNet possesses an extra function that calculate the distribution of variance rate of action through divergence of adjacent frames,

$$p(z) = \frac{|F_{z+1}(x,y) - F_z(x,y)|}{\sum_{i=1}^{n-1} |F_{i+1}(x,y) - F_i(x,y)|}, \tag{1}$$

(x, y) is coordinate of pixel in frame F.

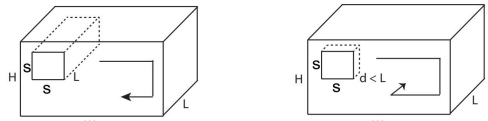
We put three pathways for Slow-Mid-Fast video clips of various frames which randomly sampled in accordance with $p(z)$, n_1 frames in slow pathway , n_2 frames in medium pathway, n_3 frames in fast pathway, $n_1 \leq n_2 \leq n_3$. Following, we employ share-kernel $s \times s \times 3$ to transform frames into initial feature maps, and then stack them up to three blocks. This part is fixed with no training process in contexts. We employ 3D convolution kernels to convolve features of each block, $s_1 \times s_1 \times d_1$ kernel size on slow pathway, $s_2 \times s_2 \times d_2$ kernel size on medium pathway, $s_3 \times s_3 \times d_3$ kernel size on fast pathway, $s_1 \leq s_2 \leq s_3$, and we pile them up to a 3D module, passing to sparse Densenet3D. We modify DenseNet structure with 3D kernels and introduce a dropping rate α which randomly drops connection among layers, eventually outputting 256 dimensional vectors. This sparsely adjusts the overall architecture of 3D network, and the l^{th} layer randomly receives outputs of previous layers, then

$$A_l = G_l(\{A_i\}_{i=1}^{(l-2)\alpha}), \tag{2}$$

G_l is conceptual function of the l^{th} layer, and the upper-neighbor layer is required. In the end, just concatenate 3D module to sparse Densenet3D and get the Multi-Cue Sparse 3D ConvNet Φ. The 3D convolution operation of position (x, y, z) at the i_{th} layer and the j_{th} feature map is performed,

$$a_{i,j}^{xyz} = g(b_{i,j} + \sum_{u=1}^{k} \sum_{p=0}^{s-1} \sum_{q=0}^{s-1} \sum_{r=0}^{d-1} w_{i,j,u}^{p,q,r} a_{(i-2),u}^{(x+p),(y+p),(z+r)}), \tag{3}$$

$k = (l-2)\alpha$ is number of the i^{th} layer randomly receiving.



(a) 2D convolution (b) 3D convolution

Fig. 2. 2D and 3D convolution operations.

Spontaneously, MS3D ConvNet fosters the capacity of action information and hardly ever sustains a heavy amount of parameters which deserves double profits in the meanwhile. This fancy structure also leaves an end-to-end fashion training and a possibility to pre-train parameters. The 3D ConvNet weight matrix $w^{p,q,r}$ with temporal dimension d initialized by Eq.(4),

$$w^{p,q,r} = \alpha_d w^{p,q}, \quad \alpha_d = \begin{cases} \frac{2d-1}{d}, & if d = 1, \\ -\frac{1}{d}, & otherwise. \end{cases} \tag{4}$$

B. Loss and Training

Loss design is a necessary part for convergence. We abandon the former ways that directly output action recognition results with Cross Entropy which require a tense convergence process. Instead, we construct an encoding network of action with N-pair loss. Assume **M** are positive type of actions, and **N** are negative type of actions, then cost loss of Φ

$$\theta : \arg\min_{\theta} \sum_{i=1}^{T} log(1 + \sum_{j=1}^{T-1} exp(\Phi(M_{i1})^T \Phi(N_j) - \Phi(M_{i1})^T \Phi(M_{i2}))), \tag{5}$$

After define MS3D ConvNet Φ and its cost loss, we employ adam optimization to train parameters of network with required executing components.

C. Full Repository

Assuming complete action dataset V contains T actions, $V = \{M_i\}_{i=1}^{T}$, each action possesses a few video fragments. Of each action, randomly choose w video fragments. Each fragment independently walks among described work flows for $p, p \geq 1$ times, and for i_{th} type of action, we encode it as

$$\mathbf{Z_i} = \begin{pmatrix} \Phi(M_{i11}) & \Phi(M_{i11}) & \cdots & \Phi(M_{i1p}) \\ \Phi(M_{i21}) & \Phi(M_{i22}) & \cdots & \Phi(M_{i2p}) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi(M_{iw1}) & \Phi(M_{iw2}) & \cdots & \Phi(M_{iwp}) \end{pmatrix},$$

then $\mathbb{Z} = \{\mathbf{Z_i}, y_i\}_{i=1}^{T}$, $T \times w \times p$, overall. This full repository considers clusters of actions that centers are

$$\dot{Z}_i = \frac{1}{w \times p} \sum_{m=1}^{w} \sum_{n=1}^{p} \Phi(M_{imn}), \tag{6}$$

exactly $\mathbb{Z} = \{\dot{Z}_i, y_i\}_{i=1}^{T}$.

D. Action Judgement

While conducting action recognition on experimental video c, it would be processed with described work flows first, then $\Phi(c) = Z_c$. Work out a minimum distance of Z_c among cluster centers of full repository $\mathbb{Z}$. Define threshold γ, only if $d(Z_c, Z_i) \leq \gamma$ video c is judged as action y_i, or 0.

IV. Experiments

In this section, we demonstrate a detail study for the proposed MS3D ConvNet on action datasets. We first introduce datasets of public and private ones. Following, we display implementation details among preprocess, architecture optimization, training and evaluation. Afterwards, we show overall results of MS3D ConvNet in charts compared with current state-of-the-art models on action recognition.

A. Dataset

Commonly-used action datasets with well-trimmed videos have been brought to our experiments, also self-defined action dataset. The action datasets possess a few action classes, with more than one clip for each class. Generally, they do not require to localize action instance where interest lasts for nearly the entire duration.

HMDB51. $https://serre-lab.clps.brown.edu/$ [21];
UCF101. $https://www.crcv.ucf.edu/data/UCF101.php/$ [22];
Kinetics. $https://www.deepmind.com/open-source/kinetics$ [23];
Baseball Sport$_{Self-Defined}$.

B. Implementation Details

Preprocess. Over-proportion of static background is kind of distraction which leads to inferior performance on action recognition. To get rid of this, we shrink the background of videos close to action object with coordinate of bounding box of yolo, and promote pixels of action object with enhancement factor η,

$$F_{in}(x, y)' = F_{in}(x, y) \times \eta, \tag{7}$$

that is kind of attention on action object. To maintain coincident scale among frames, we properly pad the frames with edge decaying, that outward pixels of bounding box are divided into disparate intervals from inner to outer with 5 pixels units of bringing decaying factor ξ,

$$F_{out}(x, y)' = F_{out}(x, y) \times \xi^d, \tag{8}$$

d is number of interval.

Architecture optimization. The adjustment of MS3D structure is much flexible and insight is all you need here. After that, we sample frames to making Slow-Medium-Fast clips in temporal sequence with distribution of variance rate $p(z)$. We independently conduct this process with 7 frames in slow pathway, 17 frames in medium pathway, 25 frames in fast pathway. Then, we employ share-kernel size $3 \times 3 \times 3$ to transform frames into initial feature maps. Subsequent 3D convolution kernels are homogeneous with $3 \times 3 \times 3$ in slow pathway, $5 \times 5 \times 3$ in medium pathway, $7 \times 7 \times 3$ in fast pathway. When it comes to sparse Desnet3D, we leave the dropping rate $\alpha = 0.7$. Architecture optimization of MS3D ConvNet is attempted based on architecture search, and variants of MS3D ConvNet are existing just as landmark works.

Training. We generally train MS3D ConvNet from scratch on Kinetics and then take this set of parameters to deploy pre-trained formats of MS3D ConvNet on other video datasets. The expected frame size is 224×224. We adopt Adam with speeding up optimizers to train MS3D ConvNet with batch size 64, epochs 500, batch norm, etc. The other operating are considered as demands.

Evaluation. The evaluation action datasets are required walking through above work flows. The average results over action datasets are showed with standard evaluation protocols. Top-1 and Top-5 accuracies are evaluation criteria for action recognition. Top-1 accuracy is general ratio of predictions, and Top-5 accuracy is the ratio of that top 5 predictions contain target label.

C. Performance

After attempting architecture search on the fine configuration of MS3D ConvNet, we make comparisons of MS3D and current landmark models on Kinetics, training from scratch, and finetuning on HMDB-51, UCF-101 and Baseball Sport$_{Self-Defined}$.

Table I shows the performance of models which train from scratch on kinetics. C3D is average score based on single channel among Resnet3D and Densenet3D. T3D and ECO achieve better scores, but common. I3D's score is even better, and we hardly count the time consuming of which work is not end to end. Slow Fast achieves the best score on Top-5 93.2% which is out of our thinking, and it costs much computation. MS3D though doesn't achieve optimal results, it shows good robust effect and Top-1 82.9% is deserved.

TABLE I
PERFORMANCE OF MODELS ON KINETICS.

Model	Top-1	Top-5	Velocity(vps)
C3D	69.7	84.3	< 0.5
I3D	71.6	90.0	-
T3D	71.5	88.7	7.4
ECO	70.0	89.4	20.8
Slow-Fast	77.9	93.2	0.9
MS3D	82.9	92.1	11.9

Then, we finetune parameters of models training on HMDB-51 and UCF-101 in Table II and III, and leave scores training from scratch as well. Overall, the effect of models tend to make good performance on UFC-101, not HDMB-51 which may caused by inefficient data. Action data is kind of variance with image data, and convergence is more deserved, or unexpected scores. We notice that I3D shows the best Top-5 scores 95.6% on UCF-101, and optical flow is making sense here, even other scores are not bad which can be seen as a good model. MS3D though partially shows great scores, they are not far beyond the current state-of-the-art models, only improve $1\% \sim 3\%$. Robust effect of MS3D is somehow considerable which benefits from its cluster-based concept, even on HMDB-51. Generally, C3D as initial stage model, sole channel is losing influence. Other landmark models shows strong power as well which are famous for their architectures, but not fully exerting here.

TABLE II
PERFORMACE OF MODELS ON HMDB-51.

Model	HMDB-51			
	Pre-Train		From-Scratch	
	Top1	Top5	Top1	Top5
C3D	56.8	63.3	34.3	55.7
I3D	69.1	74.8	66.5	70.0
T3D	61.1	71.3	56.6	59.2
ECO	68.5	72.4	66.9	68.1
Slow-Fast	65.2	73.7	68.5	71.2
MS3D	72.1	74.3	70.1	72.2

TABLE III
PERFORMACE OF MODELS ON UCF101.

Model	UCF-101			
	Pre-Train		From-Scratch	
	Top1	Top5	Top1	Top5
C3D	82.3	85.9	53.2	64.6
I3D	93.5	95.6	70.7	89.8
T3D	91.7	94.7	69.7	90.4
ECO	90.2	92.8	68.4	84.5
Slow-Fast	93.7	92.0	76.4	89.1
MS3D	92.4	94.5	80.4	91.2

Performance of MS3D on Baseball$_{Self-Defined}$ is admittable, but we take multi-sensors to record action information that each record of actions transmits to pathways of MS3D [24]. It seems to get advanced performance that Top-1 score 94.5% and top-5 score 97.7% training from scratch, which is conformed with statements of multi-clue models. MS3D is flexible which allows properly adjust for improvements in compliance with applying states.

V. CONCLUSION

We propose a Multi-Clue Sparse 3D ConvNet accompanying with tailor strategy to carry forward action recognition. The MS3D ConvNet aims to enhance information density of video data with multi-pathways while relieving information redundancy with interval sampling with variance ratio distribution. Action recognition employs centers of vector clusters of actions instead of stubborn format, which provide a robust

convergence environment. Tailor strategy is a means of attention that endows disparate factors on pixels with bounding box frontier of yolo to improve action clue. The MS3D ConvNet achieves advanced steps on accuracy and convergence.

VI. Acknowledgment

The work is supported by Inspur Electronic Information Industry Co. and State Key Laboratory of High-End Server & Storage Technology. We would like to thank for insightful suggestions from cooperation partners.

References

[1] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, B. Sierra, I. Rodriguez, and E. Jauregi, "Video activity recognition: State-of-the-art," *Sensors*, vol. 19, no. 14, 2019.

[2] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2678–2687.

[3] Y. Kong, Z. Wei, and S. Huang, "Automatic analysis of complex athlete techniques in broadcast taekwondo video," *Multimedia Tools and Applications*, vol. 77, no. 11, pp. 13 643–13 660, 2018.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] Z. Liu, D. Luo, Y. Wang, L. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and T. Lu, "Teinet: Towards an efficient architecture for video recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 669–11 676.

[6] G. Peng, B. Pang, and C. Lu, "Efficient 3d video engine using frame redundancy," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3791–3801.

[7] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.

[8] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. Tianyi Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Information Sciences*, vol. 480, pp. 287–304, 2019.

[9] D. T. Concha, H. D. A. Maia, H. Pedrini, H. Tacon, A. D. S. Brito, H. D. L. Chaves, and M. B. Vieira, "Multi-stream convolutional neural networks for action recognition in video sequences based on adaptive visual rhythms," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 473–480.

[10] S. Yun, S. J. Oh, B. Heo, D. Han, and J. Kim, "Videomix: Rethinking data augmentation for video classification," *arXiv preprint arXiv:2012.03457*, 2020.

[11] S. Vyas, Y. S. Rawat, and M. Shah, "Multi-view action recognition using cross-view video prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 427–444.

[12] S. Bai, Q. Wang, and X. Li, "Mfi: Multi-range feature interchange for video action recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6664–6671.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[14] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in *German conference on pattern recognition*. Springer, 2018, pp. 281–297.

[15] M. B. Shaikh and D. Chai, "Rgb-d data-based action recognition: A review," *Sensors*, vol. 21, no. 12, 2021.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[17] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," 2017.

[18] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," 2017.

[19] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 695–712.

[20] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[21] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[22] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[23] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017.

[24] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, and V. H. C. de Albuquerque, "A comprehensive survey of multi-view video summarization," *Pattern Recognition*, vol. 109, p. 107567, 2021.