YogaTube: A Video Benchmark for Yoga Action Recognition

Santosh Kumar Yadav
National University of Ireland
santosh.yadav@nuigalway.ie

Guntaas Singh
Birla Institute of Technology and Science
f20180269@pilani.bits-pilani.ac.in

Manisha Verma
Osaka University, Japan
mverma@ids.osaka-u.ac.jp

Kamlesh Tiwari
Birla Institute of Technology and Science
kamlesh.tiwari@pilani.bits-pilani.ac.in

Hari Mohan Pandey
Bournemouth University
profharimohanpandey@gmail.com

Shaik Ali Akbar
AcSIR, CSIR-CEERI, Pilani
saakbar158@gmail.com

Peter Corcoran
National University of Ireland
peter.corcoran@nuigalway.ie

***Abstract*—Yoga can be seen as a set of fitness exercises involving various body postures. Most of the available pose and action recognition datasets are comprised of easy-to-moderate body pose orientations and do not offer much challenge to the learning algorithms in terms of the complexity of pose. In order to observe action recognition from a different perspective, we introduce YogaTube, a new large-scale video benchmark dataset for yoga action recognition. YogaTube aims at covering a wide range of complex yoga postures, which consist of 5484 videos belonging to a taxonomy of 82 classes of yoga *asanas*. Also, a three-stream architecture has been designed for yoga *asanas* pose recognition using two modules, feature extraction, and classification. Feature extraction comprises three parallel components. First, pose is estimated using the part affinity fields model to extract meaningful cues from the practitioner. Second, optical flow is used to extract temporal features. Third, raw RGB videos are used for extracting the spatiotemporal features. Finally in the classification module, pose, optical flow, and RGB streams are fused to get the final results of the yoga *asanas*. To the best of our knowledge, this is the first attempt to establish a video benchmark yoga recognition dataset. The code and dataset will be released soon.**

***Index Terms*—Action recognition, Yoga, Multi-stream fusion, Deep Learning**

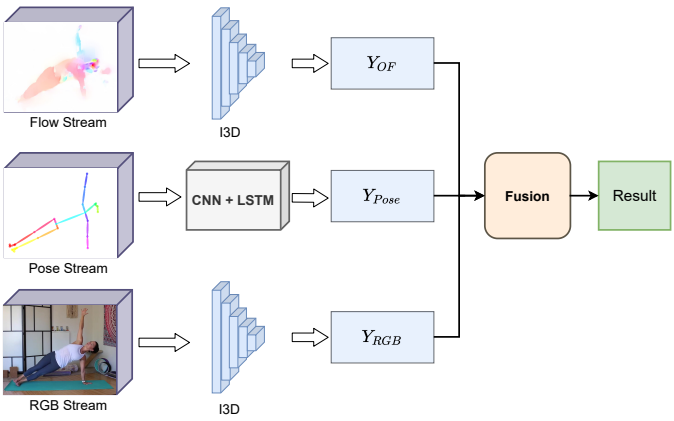

Fig. 1. The proposed architecture utilizes the RGB, pose, and optical flow cues to calculate the practitioner's asanas. To calculate high level features, raw RGB videos along with calculated optical flows and pose keypoints are inputted to the I3D and CNN+LSTM networks, respectively.

Introduction

Human action recognition is a well-motivated problem in the field of computer vision since the 1980s [1]. It is primarily a classification problem aiming towards determining what activity a human is performing in an image or video, which involves feature extraction from the image or video and classification in the most probable class. In the last few decades, the performance of computer vision algorithms has seen tremendous improvement in terms of complex human action recognition. This can be partly attributed to the introduction of more and more complex human action datasets on which these algorithms are evaluated. With the growth of online media, the amount of video and image databases are tremendously increasing on web platforms like YouTube, Bing, Flicker, *etc*. The computer vision community has made use of these and many works exploited the indefinite source of available data on the web to build large human activity datasets comprising a variety of actions, thus enhancing the application scope of recognition algorithms.

Over the years, researchers have proposed various algorithms to analyze human actions. Despite several efforts of building large-scale datasets, not many datasets deal with complex human actions/poses. Heilbron et al. [2] proposed a video dataset comprising human activities of daily living. Their dataset introduces a large number of categories and a large number of samples per category, collected from YouTube. The manual annotations are handled through crowd-sourcing. Also, the Sports-1M [3] dataset has 487 sports-related categories, annotated by an automatic tagging algorithm. Furthermore, the automatic collection process introduces an undisclosed amount of label noise. Andriluka et al. proposed MPII dataset [4] that contains approximately 25,000 images of over 40,000 people. Each image is extracted from a YouTube video. The dataset

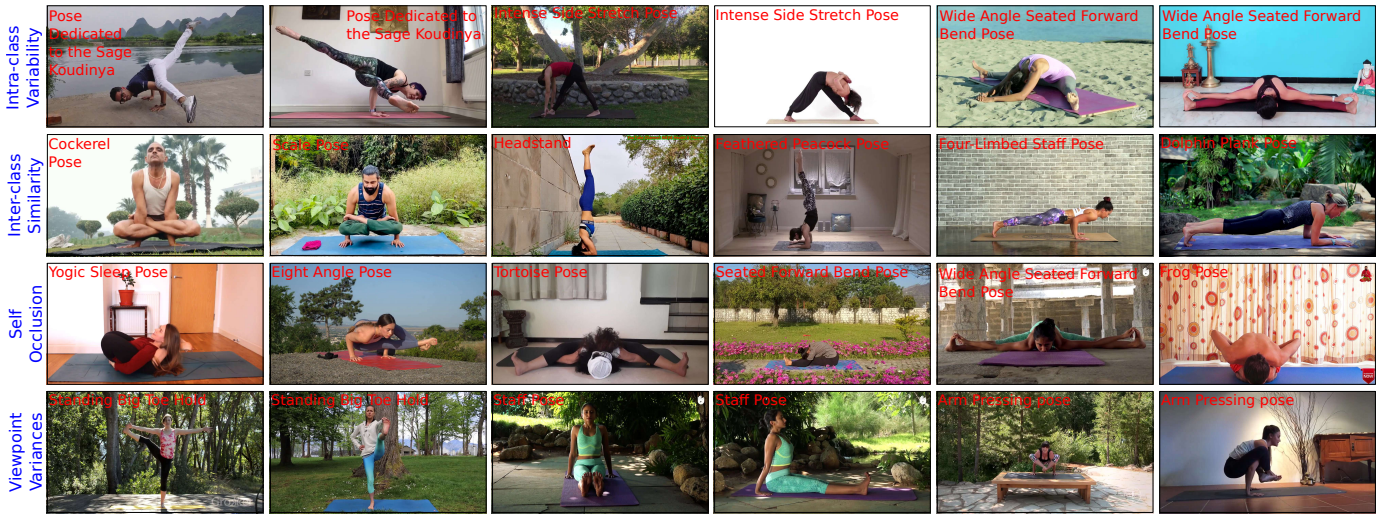

Fig. 2. Complexity of Yoga Postures.

Fig. 3. An example of cow pose as a sequence of atomic poses

covers 410 specific categories of human activity and 20 general categories. Also, UCF-101 [5] and HMDB-51 [6] datasets have been proposed for video action recognition. Though the aforementioned large-scale datasets have introduced activity diversity, they do not present enough challenges to the recognition algorithms in terms of individual action complexity and thus limiting its application scope. In this work, we propose a novel dataset for complex yoga action recognition (Fig. 1).

Yoga actions are a complex set of motions at various scales packed into a single *asana*. They consist of multiple poses that vary non-uniformly across space and time. For example, the asana Surya namaskar has 12 poses with varying body movements [7]. Though [8] comprises complex body postures, their work focuses on static images only. A typical yoga exercise is a combination of atomic poses where the sequence is important and that's why the temporal information plays a major role. For example, the yoga asana named *cow pose* is composed of many sub-postures where the sequence of postures is very significant while performing it (as shown in Fig. 3). Such information can not be retrieved using only static images. Such activities are best described with temporal information taken into consideration. That indicates the need for a video dataset based on yoga because these challenges can not be resolved using only the image-based dataset.

As of now, there is no video dataset covering large pose variations in Yoga. Therefore, in this work, we attempt to build an extensive video dataset comprising yoga postures of varying pose complexities to enhance the application scope of the recognition algorithms. This new dataset attempts to fill the gap in the following aspects. 1) We attempt to fill the existing gap in complex human activity recognition by introducing complex pose orientations accompanied by pose diversity (a large number of classes and complex body orientations) in our dataset. 2) We also attempt to imitate real-world scenarios by introducing interclass similarity, occlusion, and multi-viewpoints per class. 3) Also, we try to include all the variations of each class (intraclass variability). For example, the same yoga exercise can be performed with slight variations as shown in Fig. 2. Our dataset consists of inter-asana similarity and intra-asana variability to ensure the complexity of the dataset. Some of the asanas are very complex to perform and having a lot of occlusion due to their own body twisting making the recognition challenging. The videos are downloaded from YouTube and have diverse backgrounds as shown in the Fig. 2. The videos are trimmed based on the start and end of the asanas. The resolution of the videos differs adding proper complexity in the dataset for the recognition.

This paper strives to recognize yoga asanas from videos. In this approach, we utilize the static and dynamic information of the action recognition. Static information is calculated using a pose estimation network followed by a spatiotemporal network consisting of convolutional neural networks (CNN) and long short-term memory (LSTM) networks. The dynamic information has been calculated using the RGB and optical flow streams, which are inputted to the I3D network to compute the spatiotemporal features. Finally, three streams are passed to the fully connected layers and fused to get the final classification score as shown in Fig. 4.

In particular, the major contributions of this paper are highlighted as follows. We propose a novel large-scale dataset for yoga action recognition in videos named YogaTube. The dataset contains 5,484 videos for 82 classes of yoga asanas with a total duration of 41 hours. The dataset comprises

complex actions of yoga asana which makes it challenging.We introduce a model for yoga asanas recognition utilizing three input modalities passing through a three-stream network, $i.e.$ pose, RGB video, and optical flow. Human joint keypoints and optical flow vectors are precomputed and all three inputs, $i.e.$, keypoints, RGB video, and optical flow vectors are passed to individual streams. The features obtained from three streams are fused using different fusion methods and results are presented.

The remainder of the paper is organized as follows. Section presents the literature review of the yoga asanas recognition. Section presents the architecture of the proposed network. Section describes the YogaTube dataset along with the data collection procedure, cleaning, and statistics. Section presents the experimental setup, implementation details, model evaluation, and discussion and comparison. Finally, Section presents the concluding remarks and future works.

BACKGROUND

This section presents recent action recognition datasets, methods, and yoga asanas recognition-related works.

Action recognition datasets: Spurred by the growth of online media, HMDB51 [6] and UCF101 [5] datasets contain 7,000 and 13,320 videos, respectively, for video action recognition. A large-scale video action recognition dataset Sports1M [3], consisting of one million YouTube videos spread into 487 sports classes, was proposed in 2014. Caba et al. proposed ActivityNet [2] in 2015, which consists of 27,801 videos for 203 classes. It comprises three tasks, $i.e.$ trimmed activity classification, untrimmed video activity classification, and human activity detection. YouTube8M [9], which was introduced in 2016, is the largest video dataset at present consisting of 8 million YouTube videos spread into 3,862 action classes. The Kinetics [10]–[12] datasets are one of the most widely used datasets for video action recognition. It consists of Kinetics400 [10], Kinetics600 [11], and Kinetics700 [12] datasets proposed in 2017, 2018, and 2019, respectively. AVA [13] is a large-scale spatiotemporal dataset consisting of 57,600 videos (385,446 samples) for 80 classes. 20BN-Something-Something V1 and V2 [14] datasets consist of 108,499 and 220,847 videos for 174 classes and were introduced in 2017. The videos consist of people performing daily actions and interacting with common objects. Following a similar trend, new datasets proposed for video action recognition include HACS Clips [15], HVU [16], AViD [17], FineGym [18], and HAA500 [19]. However, these datasets do not include complex postures such as yoga.

Action recognition using CNN architectures: With the availability of large-scale datasets, video action recognition has seen tremendous growth in recent years using deep learning. DeepVideo [3] is among the earliest architectures to apply deep CNN in video. There are two trends in video action recognition, $i.e.$ two-stream CNNs, and 3D-CNNs. The first trend utilizes two streams for spatial and temporal feature extraction. This trend started from the *Two Stream Networks* et al. [20], followed by many others like TDD [21], TSN [22], Fusion [23], TRN [24], $etc.$ TSN [22] samples video clips from evenly divided segments. To capture information along with temporal dimension, TRN [24] and TSM [25] utilize a shift module by replacing average pooling with an interpretable relational module. However, due to the utilization of 2D CNNs in two-stream networks, they do not capture the temporal dynamics. The second trend uses 3D convolutional kernels to jointly model the spatial and temporal semantics such as C3D [26], I3D [27], S3D [28], R3D [29], Non-Local [30], SlowFast [31], $etc.$

Yoga asanas recognition: Yoga is an ancient Indian exercise comprising body postures of various complexities. In the field of human activity recognition, yoga pose recognition is an emerging task for applications like self- and virtual-training systems [7], [32]. Works like [33], [34] proposed camera-based yoga recognition systems. The dataset of [33] had videos for 6 asanas, recorded with an RGB webcam, whereas [34] used Kinect to capture depth maps for the recognition of 12 yoga poses. However, their datasets contain relatively simple human poses and do not offer much challenge to the learning algorithms. Gochoo et al. [35] proposed an IoT-based privacy-preserving yoga posture recognition system by employing low-resolution infrared-sensors-based network for 26 yoga postures. In terms of body orientation, their dataset had only 4 complex poses (standing separate leg head to knee, camel, forward fold, and bow), while others were relatively simple. Hence, in this work, we aim to establish a large-scale video dataset for complex yoga postures that can be challenging enough to be used as a test-bed for human pose recognition algorithms.

YOGATUBE ARCHITECTURE

Inputs to the network

We rely on three input cues: raw RGB frames, optical flow, and human joint keypoints where each of them is provided as Spatio-temporal inputs.

Body pose keypoints: Estimation of human pose plays a vital role in obtaining a detailed understanding of people in videos. Pose estimation encodes the orientation of a person in space as a sequence of localized anatomical keypoints. It provides explicit information modeling spatial and temporal dynamics in a much lower-dimensional space as compared to raw videos. This can be useful for higher-level tasks like human action recognition [36].

We use body joint keypoints as one of the inputs in order to recognize the pose. We use a bottom-up pose estimation module based on Part Affinity Fields (PAFs) - OpenPose for extracting pose information from videos [37]. In this module, a multi-stage convolutional neural network is used for predicting and then, refining 2D confidence maps localizing various body parts. In addition, 2D vector fields are used for each limb to encode the degree of association for each part pair. These are subsequently combined via greedy bipartite matching to assemble complete poses.

Due to the complicated orientation of the human body in many yoga asanas, reliable pose estimation can be difficult

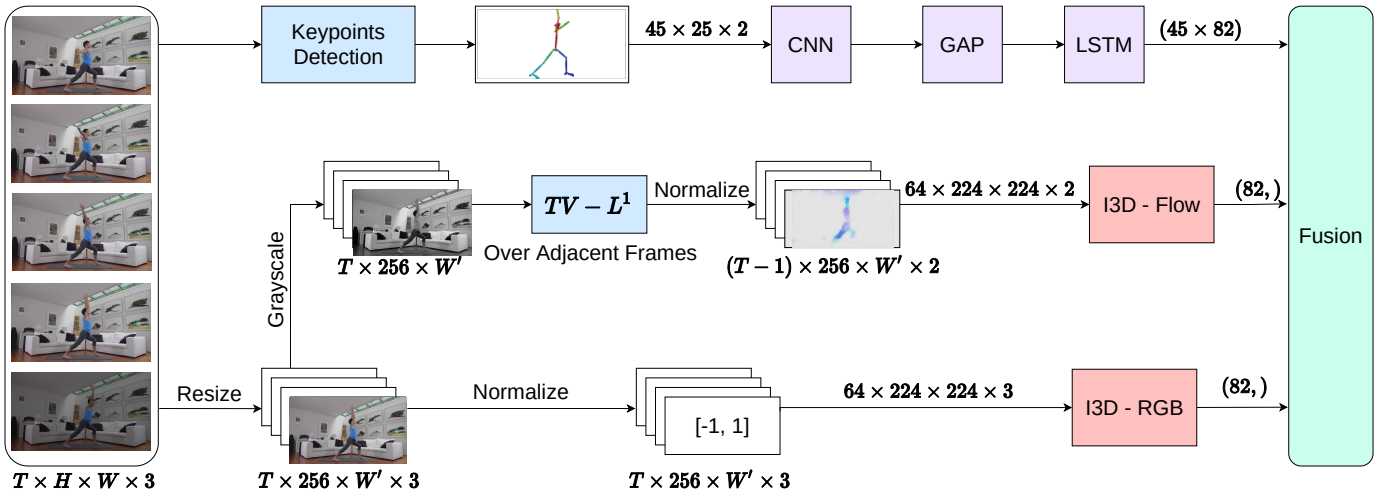

Fig. 4. Broad outline of the YogaTube architecture. We collect keypoint estimates from OpenPose for a sequence of frames and pass them to a LSTM-based model. Optical flow computed over a sequence of grayscale frames is normalized and passed to an inflated 3D CNN. Similarly, raw RGB frames are passed as input to another finetuned inflated 3D CNN. Finally, the output for all the streams are fused to obtain the final prediction.

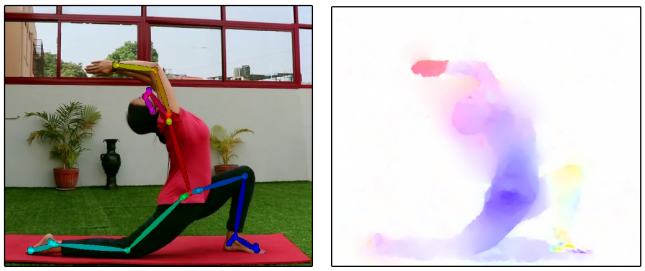

Fig. 5. Visualizations of skeleton pose annotations and optical flow representations.

to obtain. With multiple viewpoints, limbs can get occluded in a large number of ways that can not be modeled in a robust fashion using 2D pose estimation. These limitations hint towards the need for incorporating other modalities for yoga pose recognition.

RGB videos: The problems related to explicitly encoding information about the subject suggest that using an implicit representation in the form of RGB videos may help. They can provide rich appearance information and fine contextual cues to the system.

While fine appearance information about textures and other visual cues may be useful for tasks like image recognition, these features do not yield useful information for discerning between different yoga asanas. For example, the clothes worn by the subject or the appearance of the background are largely irrelevant for recognizing the yoga asana being performed and can act as a source of large intra-class variability. This can make action recognition challenging.

Optical flow: Optical flow is defined as the distribution of apparent velocities of movement of brightness pattern in an image [38]. It provides a representation that accurately encodes motion information and is independent of finer appearance features which can be useful for action recognition. Along with appearance features which can be decoded using RGB videos, optical flow can enhance temporal changes in the video. Fig. 5 highlights how optical flow can be used for effectively dealing with the problem of background clutter by eliminating static visual features so that the system can focus on the spatio-temporal dynamics relevant to the task.

For YogaTube, we estimate dense TV-L^1 optical flow using a method given by Zach et al. in [39]. It builds on the variational formulation first proposed by Horn and Schunck in [38]. Total variation regularization and an L^1 data penalty term are used to improve robustness with respect to changes in illumination, occlusion, and noise.

Feature extraction

As mentioned above, we use three types of inputs in this work. We use a three-stream network to utilize these inputs for an accurate yoga pose classification paradigm.

Pose network: For the Pose stream, we use a simple neural network that takes in an ordered set of 25 skeleton pose keypoints corresponding to a single person for a sequence of 45 frames. The network consists of three successive blocks of 1D convolutional layers and max pooling. This is followed by global average pooling before an LSTM with 100 units. Finally, a series of dense layers are applied before making predictions.

Inflated 3D CNN: In [27], Carreira et al. present inflated 3D (I3D) CNN for video-based action recognition that aims to build upon existing approaches in order to redefine their capability.The key idea behind the proposed approach is that 2D CNNs used for image classification can be converted into 3D networks by inflating their filters and introducing a new temporal dimension. It allows utilization of popular 2D-CNNs which work well in large-scale image datasets, in the area of video classification, and other tasks related to videos using 3D-CNN. The resulting 3D convolutional networks are very deep,

but have an exceedingly large number of parameters, making them difficult to train. The idea of the "boring-video" fixed point can be used to bootstrap weights for the 3D network from its pre-trained ImageNet counterpart, thus providing a valuable parameter initialization before training.

For YogaTube, we start with inflated 3D CNNs based on Inception-v1. We use two different networks for optical flow and RGB videos with the same network architecture. Pre-trained weights based on the ImageNet and Kinetics datasets are used to initialize the parameters for these models. Transfer learning is utilized to train RGB and Flow models for yoga action recognition. Since pre-trained models on action recognition are used, we only train the last Inception block while keeping other layers frozen in order to prevent over-fitting. Out of 12,378,594 parameters, we train 2,867,522, while the weights for the bulk of the layers are kept frozen.

Classification module

We get a classification score from each of the three streams individually and use four different techniques for fusion. **Average fusion** combines the logits of all the streams using mean pooling. **Max fusion** combines the logits of all the streams using max pooling. **Weighted Average fusion** allows us to control how the logits from each stream are weighed. In **fusion with FC layer**, we concatenate the logits from all the streams and train a small neural network to derive the final prediction using this vector as input. The network learns per-class weights for logits from each stream, allowing us to weigh the predictions from each stream in a fine-grained manner - separately for each class.

BUILDING YOGATUBE

Yoga asanas consist of many complex and diverse postures that a human body can perform. It is very complex to capture these asanas from a single point of view. The complexity further increases with changes in the video resolution and occlusion. A dataset dedicated to yoga poses can be utilized as a baseline in the area of yoga pose monitoring, tutoring, $etc.$ which is an inspiration behind creating this dataset.

Video collection: We begin the data collection by defining our tasks. These must satisfy four criteria: they must entail one person performing an asana correctly; the practitioner is properly visible while doing the asana; the duration of the video must not be more than one hour, and the camera is not moving abruptly so that we can assure the continuous tracking of the person. Using search queries in both English and Sanskrit, we downloaded videos for all 82 classes in the best quality available from YouTube.

Complexity: Many of the asanas consist of such a complex posture that the body parts are not even properly visible due to self-occlusion. This makes it more challenging for the recognition algorithms as the full-body can not be properly tracked. The dataset contains all the possible viewpoints such as back and forth, left and right, and from different angles of practitioners from the camera.

TABLE I
RESULTS ON THE YOGATUBE DATASET USING DIFFERENT STREAMS.

Sr.	Model Name	Precision	Recall	F1-Score	Accuracy
1	Pose Stream	83.99%	81.12%	81.02%	83.45%
2	RGB Stream	83.98%	81.37%	81.65%	83.45%
3	Optical Flow Stream	82.22%	77.36%	77.29%	80.03%

Cleaning the dataset: After downloading the long compilation videos from YouTube, we delineated them to separate unrelated actions. For this, each video was validated by at least two human observers to ensure consistency. This ensures the asanas being practiced correctly in a proper sequence. However, while annotating the data, there is no distinction made between the asanas being performed indoor or outdoor. The dataset involves videos recorded in various complex real-world environments such as gardens, yoga centers, beaches, $etc.$ We deliberately consider these types of videos to enhance the generalization-ability of the system. In this process, a video is trimmed from the point where asana is being started till the end of the asana. Each asana contains a specific sequence, which starts with the practitioner being at the neutral position followed by performing the specific asana and then ending it by coming to the neutral position again. The transition of the asanas is also considered to ensure the practicality of the system and the sequential nature of the asanas are maintained in the dataset.

Dataset statistics: The proposed dataset consists of 82 distinct yoga asanas. The dataset contains a varying number of videos in each class, which is ranging from 50 (minimum) to 130 (maximum). On average, there are 75 videos per class. The total number of videos is 5,484. The dataset comprises a total of 41 hours of trimmed videos. To the best of our knowledge, to date, this is the largest and perhaps most diverse dataset for yoga action recognition using videos. We plan to release the dataset, along with the pre-computed optical flow and pose features. We believe that this dataset will facilitate the development and evaluation of models to recognize yoga actions.

Fig. 6 shows the sample distributions among all the classes. We split the YogaTube dataset into 4,000 and 1,484 video samples for training and testing, respectively. The average length of the videos is 27 seconds. The total number of frames in the dataset is 4,352,916.

EXPERIMENTAL RESULTS

In this section, we present the implementation details of our proposed approach and evaluate its performance on the YogaTube, after giving a concise description of our experimental setup.

Experimental setup

All experiments were performed on an HP Z420 Workstation having a single NVIDIA TITAN Xp GPU along with 26 GB RAM and an Intel Xeon E5-1620 CPU clocked at 3.60 GHz. All models were implemented and trained using TensorFlow [40]. A consistent 80:10:10 split of the YogaTube

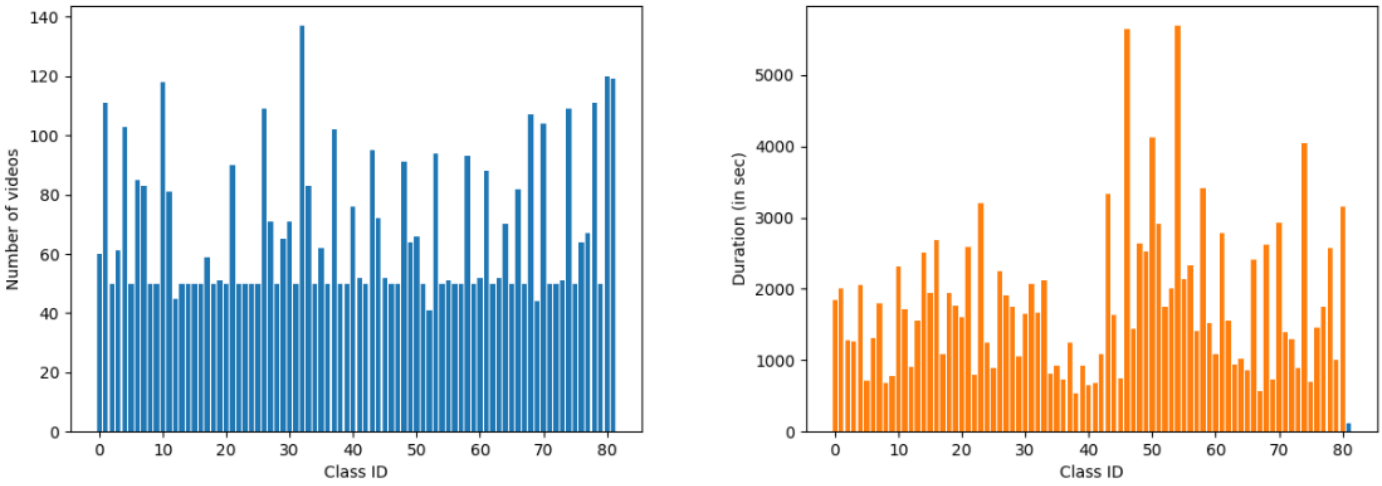

Fig. 6. Dataset statistics of the YogaTube dataset.

TABLE II
RESULTS ON FUSION OF STREAMS USING DIFFERENT METHODS.

Sr.	Fusion method	Precision	Recall	F1-Score	Accuracy
1	Average Fusion (RGB + Flow)	88.18%	85.41%	85.60%	87.20%
	Average Fusion (RGB + Pose)	86.70%	83.65%	83.79%	85.66%
	Average Fusion (Flow + Pose)	86.13%	82.82%	82.96%	84.81%
	Average Fusion (RGB + Flow + Pose)	86.95%	83.87%	84.03%	85.84%
2	Max Fusion (RGB + Flow)	83.99%	81.37%	81.65%	83.45%
	Max Fusion (RGB + Pose)	85.64%	82.13%	82.41%	84.30%
	Max Fusion (Flow + Pose)	85.71%	82.17%	82.42%	84.30%
	Max Fusion (RGB + Flow + Pose)	85.64%	82.13%	82.41%	84.30%
3	Weighted Average Fusion (RGB + Flow)	88.18%	85.42%	85.61%	87.20%
	Weighted Average Fusion (RGB + Pose)	86.70%	83.66%	83.80%	85.67%
	Weighted Average Fusion (Flow + Pose)	86.14%	82.82%	82.96%	84.81%
	Weighted Average Fusion (RGB + Flow + Pose)	86.95%	83.87%	84.03%	85.84%
4	Fusion with FC Layer (RGB + Flow + Pose)	91.85%	90.44%	91.14%	90.96%

dataset was used for training, validation, and evaluation across all three streams.

Implementation details

Pre-processing: For the RGB and Flow streams, since we use ImageNet and Kinetics pre-trained weights for initializing our models, the same pre-processing pipeline itecarreira2017quo is used in this work. First, the video is re-sampled to a uniform frame rate of 25 frames per second. The smaller dimension out of width and height is resized to 256 pixels, maintaining aspect ratio. For the Flow stream, $TV\text{-}L^1$ optical flow is computed between adjacent frames after conversion to grayscale, following which the flow values are truncated to the range $[-20, 20]$. For both - RGB and Flow streams, the pixel values are normalized to the range $[-1, 1]$. The frames are cropped spatially and temporally to obtain a uniformly shaped set of 64 224×224 frames, with 2 and 3 channels respectively, for Flow and RGB modalities.

Data augmentation: Considering the large parameter space for the I3D architecture and the modest size of our dataset, data augmentation is of critical importance to obtain a model with good generalization. To this end, we apply random spatial and temporal cropping and horizontal flipping on training videos. To maintain a stable evaluation measure across all experiments, these transformations are not applied to validation and test sets.

Model training: For the RGB and Flow streams, categorical cross-entropy was used as the loss function in conjunction with the Adam optimizer [41] with a learning rate scheduler.

Consecutive sequences of 64 adjacent frames are passed to the model. Videos having fewer than 64 frames are looped, while for longer videos, a sliding window is used to sequentially cover frames across the entire duration. The logits for each sequence obtained from a single video are then aggregated by average-pooling to derive the final prediction for the complete video.

Model evaluation

In order to develop a better understanding of the impact of different modules on the performance of our model and the effectiveness of different modalities, the performance of the Pose, RGB, and Flow streams on testing data is presented in TABLE I. The results of experiments with various fusion techniques are also provided in TABLE II.

While the Pose and RGB streams produce a similar accuracy of 83.45% each, the Flow stream produces a lower accuracy of 80.03%. This indicates that the elimination of visual cues in the optical flow representation might be detrimental to recognition. However, fusing the Flow stream with any combination of the other streams was found to lead to improved classification performance in all cases - indicating that the information that it captures, complements and aids activity recognition. In case of average fusion, adding the pose stream was found to have a detrimental impact on model performance. This may be attributed to noise introduced by erroneous pose estimation - which is particularly difficult for the complex yoga poses present in our dataset. Average fusion was found to outperform max fusion for all combinations. Finally, fusion with FC layer produced the best results, indicating that each modality captures complementary information and fine-grained weighing of predictions from each stream helps us to incorporate this information into our predictions in an effective manner.

Summary and Future Work

This paper provides a new benchmark dataset for yoga action recognition named as YogaTube. We propose a novel three-stream model for yoga action recognition along with different fusion techniques for classification. The code and YogaTube dataset will be made publicly available to support future research in this area. We hope this dataset will prove to be a testbed for developing novel video representation learning algorithms.

References

[1] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[2] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[4] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[5] 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.

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[7] T. K. K. Maddala, P. Kishore, K. K. Eepuri, and A. K. Dande, "Yoganet: 3-d yoga asana recognition using joint angular displacement maps with convnets," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2492–2503, 2019.

[8] M. Verma, S. Kumawat, Y. Nakashima, and S. Raman, "Yoga-82: a new dataset for fine-grained classification of human poses," *arXiv preprint arXiv:2004.10362*, 2020.

[9] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[10] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[11] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[12] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[13] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[14] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5842–5850.

[15] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "Hacs: Human action clips and segments dataset for recognition and temporal localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8668–8678.

[16] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen, and L. Van Gool, "Large scale holistic video understanding," in *European Conference on Computer Vision*. Springer, 2020, pp. 593–610.

[17] A. Piergiovanni and M. S. Ryoo, "Avid dataset: Anonymized videos from diverse countries," *arXiv preprint arXiv:2007.05515*, 2020.

[18] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2616–2625.

[19] J. Chung, C.-h. Wuu, H.-r. Yang, Y.-W. Tai, and C.-K. Tang, "Haa500: Human-centric atomic action dataset with curated videos," *arXiv preprint arXiv:2009.05224*, 2020.

[20] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *arXiv preprint arXiv:1406.2199*, 2014.

[21] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[22] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[23] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[24] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–818.

[25] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[26] 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.

[27] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[28] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 305–321.

[29] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[30] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

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

[32] Z. Wu, J. Zhang, K. Chen, and C. Fu, "Yoga posture recognition and quantitative evaluation with wearable sensors based on two-stage classifier and prior bayesian network," *Sensors*, vol. 19, no. 23, p. 5129, 2019.

[33] S. K. Yadav, A. Singh, A. Gupta, and J. L. Raheja, "Real-time yoga recognition using deep learning," *Neural Computing and Applications*, vol. 31, no. 12, pp. 9349–9361, 2019.

[34] H.-T. Chen, Y.-Z. He, and C.-C. Hsu, "Computer-assisted yoga training system," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 23 969–23 991, 2018.

[35] M. Gochoo, T.-H. Tan, S.-C. Huang, T. Batjargal, J.-W. Hsieh, F. S. Alnajjar, and Y.-F. Chen, "Novel iot-based privacy-preserving yoga posture recognition system using low-resolution infrared sensors and deep learning," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 7192–7200, 2019.

[36] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[37] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[38] B. K. Horn and B. G. Schunck, "Determining optical flow," in *Techniques and Applications of Image Understanding*, vol. 281. International Society for Optics and Photonics, 1981, pp. 319–331.

[39] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*. Springer, 2007, pp. 214–223.

[40] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.