WTM: Weighted Temporal Attention Module for Group Activity Recognition

Santosh Kumar Yadav
AcSIR, CSIR-CEERI, Pilani
santosh.yadav@pilani.bits-pilani.ac.in

Palaash Agrawal
Birla Institute of Technology and Science
f20180565@pilani.bits-pilani.ac.in

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

Ehsan Adeli
Stanford University
eadeli@stanford.edu

Hari Mohan Pandey
Bournemouth University
profharimohanpandey@gmail.com

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

***Abstract*—Group Activity Recognition requires spatio-temporal modeling of an exponential number of semantic and geometric relations among various individuals in a scene. Previous attempts model these relations by aggregating independently derived spatial and temporal features. This increases the modeling complexity and results in sparse information due to lack of feature correlation. In this paper, we propose Weighted Temporal Attention Mechanism (WTM), a representational mechanism that combines spatial and temporal features of a local subset of a visual sequence into a single 2D image representation, highlighting areas of a frame where actor motion is significant. Pairwise dense optical flow maps representing the temporal characteristic of individuals over a sequence are used as attention masks over raw RGB images through a multi-layer weighted aggregation. We demonstrate a strong correlation between spatial and temporal features, which helps localize actions effectively in a multi-person scenario. The simplicity of the input representation allows the model to be trained by 2D image classification architectures in a plug-and-play fashion, which outperforms its multi-stream and multi-dimensional counterparts. The proposed method achieves the lowest computational complexity in comparison to other works. We demonstrate the performance of WTM on two widely used public benchmark datasets, namely the Collective Activity Dataset (CAD) and the Volleyball Dataset. and achieve state-of-the-art accuracies of 95.1% and 94.6% respectively. We also discuss the application of this method to other datasets and general scenarios. The code is being made publicly available.**

***Index Terms*—Video Action Recognition, Human Activity Recognition, Transformers, Temporal Attention, Consensus, Convolutional Neural Networks**

Introduction

Group activity recognition aims to understand the collective behavior of a group of individuals, each involved in independent activities. It is an important subtask of general human action recognition [1], involving individual dynamics, relations between multiple people in a single sequence [2] as well as context [3]. The applications of group activity recognition include automatic video surveillance, sports video analysis, social behavior understanding, autonomous driving systems, *etc.*

Group activity recognition is challenging as it requires an understanding of higher-level relationships among individuals [4]. Group activity is majorly determined by activity occurring in a small fraction of the total area of the frame. Even within the more important areas, different individual activities contribute different amounts of information regarding the group activity. These important areas of the frame can be characterized by considering the temporal dynamics. Hence, the model search space in learning algorithms can be directed for activity recognition tasks, by developing methods that identify important features through proper temporal feature extraction. In this paper, temporal features have been used as an attention mechanism to highlight certain important areas of sequential frames.

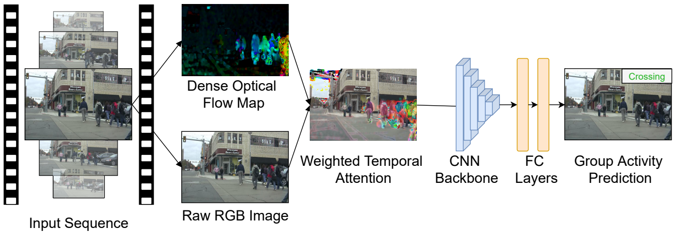

Fig. 1. Classification using weighted temporal attention mechanism (WTM).

This paper aims towards the prediction of activity through a simplified approach, by combining spatial information and temporal history into a single input representation, which can be classified through a replaceable image classification architecture. This provides the classification model with complete and comprehensive information in a single input image. The proposed method is highly efficient to extract scene-level understanding of the general activity. Moreover, the method is independent of the number of actors in the scene, and the exponential number of corresponding relations between various entities. To demonstrate this efficacy, the paper focuses on one of the harder subtasks of activity recognition - group activity recognition. Apart from the information derived through the proposed method, no explicit information (such as individual activity labels, pairwise relation annotations, bounding box annotations, *etc.*) is needed to identify group activities. This makes the model more feasible for practical real-world scenarios, where such labor-intensive annotations are infeasible.

We use dense optical flow [5] approach to derive trajectory maps corresponding to temporal movement. However, these maps contain noise in the form of erratic movement. A threshold filter has been implemented to suppress such noise. We demonstrate that the resulting optical flow maps are correlated with spatial representations. A layered weighted scheme is proposed to superimpose these optical flow maps onto raw images, acting as temporal attention over spatial distributions. The trajectorial maps include information about movements, directions, acceleration, *etc.* from immediate past as well as immediate future. We build local sub-sequences containing frames separated by a suitable frame stride. A brief description of the proposed approach is depicted in Fig. 1.

The major contributions of this paper are as follows.

- A novel representational approach towards combining and correlating spatial and temporal features has been proposed for activity recognition. This reduces the problem to a single image, single-label classification problem, which can be modeled using a suitable 2D image classification architecture, replaceable in a plug-and-play fashion.
- We propose a temporal attention module that uses dense optical flow to derive trajectory-based information from the immediate frames, which is masked over raw RGB images and pooled through a layered weighting scheme to create a comprehensive representation of image sequences.
- The proposed approach results in reduced computational complexity in terms of both space and time. This highlights the effectiveness and importance of cognitively simpler approaches to information representation and learning.
- The performance of the proposed method has been evaluated on two benchmark datasets, namely the Collective Activity Dataset and the Volleyball dataset. The proposed method outperformed the state-of-the-art and achieved 95.1 % and 94.6% testing accuracies on the two datasets, respectively.

Related Work

Many works of literature [6]–[8] present a general approach of recognizing singular person activities and then building on this prior knowledge to recognize group activities. [6] proposed a spatiotemporal contextual relationship-based model, learning features hierarchically. In [8], a real-time inference framework is proposed for multi-person tracking and activity recognition at multiple hierarchical activity levels. [9]–[11] propose multi-stream convolutional networks as frameworks for group activity recognition in which predictions from CNN streams trained on different modalities are fused at the end. [7] proposed a semi-supervised multi-level sequential GAN architecture for group activity recognition.

Some works rely on extracting spatio-temporal information through recurrent neural networks. [12] presented a 2-stage deep temporal model based on LSTM to capture group-level dynamics for group activity recognition. Similarly, [13] proposed a semantic scheme with an LSTM based two-stage captioning and prediction model. [14] proposed a single forward pass architecture that performs localization and classification using LSTMs.

[15] focused on building end-to-end learnable relation graphs through matrix operations to simultaneously capture the appearance and position relation between actors in a multi-person scene. [16] proposed a CNN-based spatial relational scheme for group activity recognition. In a multi-person scenario, sometimes focussing on only a few key features suffices. This can be addressed by filtering irrelevant features. [17] proposed a reinforcement learning-based method to distill the low-level and high-level relations of group activities. Their approach involved constructing a relational graph for explicitly modeling the relations among persons. [12], [18] exploited RNN to detect only relevant events in videos by suppressing irrelevant information. Their multi-stage RNN model learns to detect events in videos while automatically attending to the key actors responsible for an event.

Some researchers have proposed attention mechanisms for group activity recognition. [18] proposed a soft attention-based model for detecting key actors and high-level activity. Tang et al. [19] assembled attention mechanisms to achieve compact representations by assigning varying pooling weights to a different person–group interactions. [20], [21] proposed relative spatiotemporal attention which is determined by a key actor. However, these methods treat and process attention in the spatial and temporal domains separately. This creates an information gap due to the independence of these features. The relation between spatial and temporal features can also be exploited through stronger visual representations to understand the latent structure of the sequential input. This often helps in better spatio-temporal correlation and results in lower computational complexity. [22] proposed a temporal pooling function to rank features of an entire video sequence in a 2D space, which can be classified through classifiers such as SVMs. Similarly, [23] proposes a ranking method, which creates alternative feature vectors corresponding to the entire sequence. However, such an approach results in the loss of the key original spatial distribution of images, including the background information.

Proposed Method

Group Activity Recognition is ultimately a problem of image classification. The approach used in this paper involves the use of the classification of images with temporal attention, which is fed into an image classification model for group activity recognition.

Elementary Approach

Initially, we train a convolutional neural network without any attention mechanism to understand the contribution of superficial spatial features in the final results. This is done by directly feeding single labeled images as input to the CNN. We feed all the labeled images and augment them using

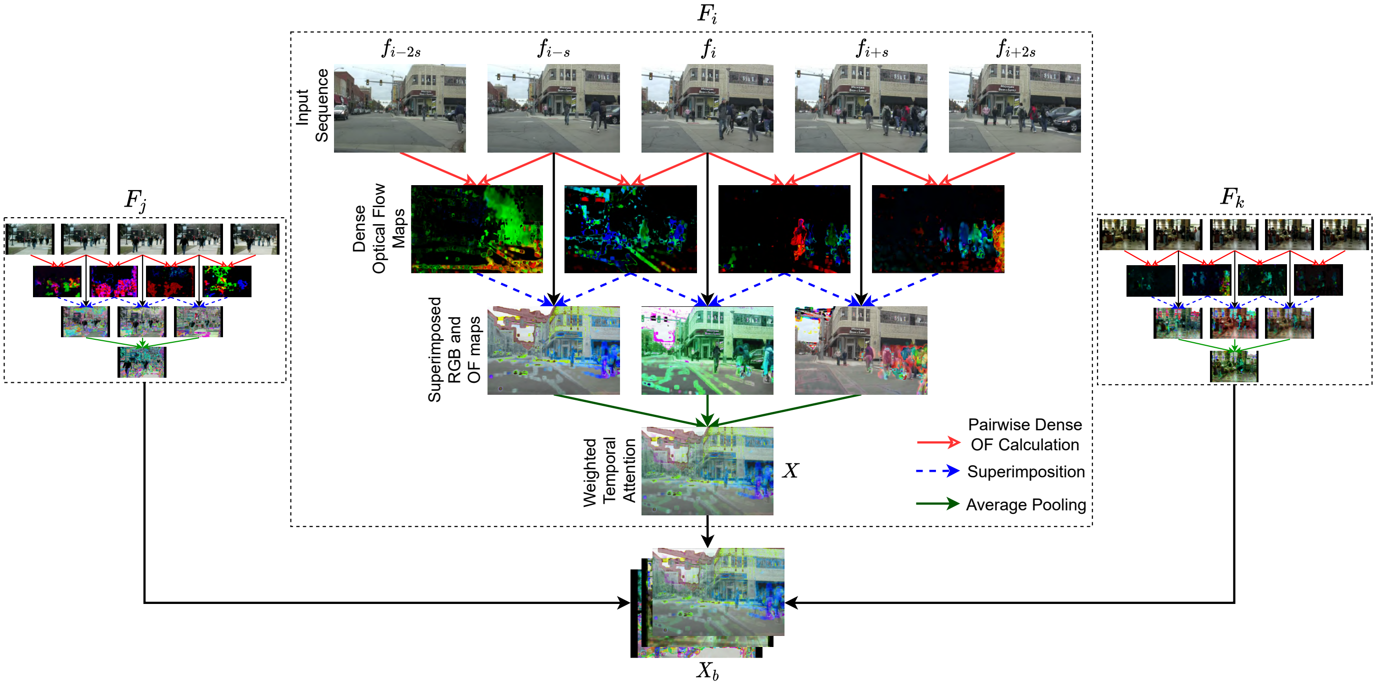

Fig. 2. Proposed method for weighted temporal attention mechanism. A frame window F_i consisting of N temporally continuous frames separated by a stride of s frames each. A batch X_b may contain image representations from various sequences since each sequence can be reduced to a single image with a single group activity label.

standard image augmentation techniques such as cropping, padding, rotation, and flipping. Along with this, we also add temporally adjacent frames to the training batch corresponding to a frame, as an equivalent of transformed images with the same label. Corresponding to one image, we create an augmented collection of $2N+1$ temporally adjacent frames of images, each pair separated by a stride of s frames.

$$frame\ f_i \xrightarrow{augment} frames\ \{f_{i+ns}\}$$

$$\forall n \in \{-N, -N+1, ..., N-1, N\}$$

$$s.t.\ dim(f_{i+\alpha s}) = dim(f_{i+\beta s})\ \forall\ \alpha, \beta\ \in n,\ \alpha \neq \beta$$

where N is the number of additional successive (or predecessive) frames appended to a batch. Any frame with non-uniform dimensions can be adjusted using an appropriate cropping technique. A non-unity stride s is preferred since two adjacent frames are more or less similar in spatial formation. Thus no significant temporal shift would be observed in that case. It is hypothesized that the addition of temporal information of each individual person would be required and sufficient to include the missing information needed to comprehensively derive information from the data. The drawbacks of traditional CNNs, in the context of group activity recognition, are (1) Any single frame cannot factor in temporal information, and its related features, such as trajectory, acceleration, speed, and the direction of movements of different people. Temporal information can be derived from a series of frames; and (2) The features extracted by CNNs are abstract, which means, it is difficult to guide the network to learn more meaningful features. For example, it is difficult to guide the network in a manner that gives more feature-based importance to the people themselves, compared to relatively non-essential features such as the setting of the background.

Weighted Temporal Attention Module

To include temporal information, we use Gunnar-Farneback [5] based estimate for optical flow (also known as dense optical flow) to find out movement-based details. Dense optical flow highlights the pixels of an image where the change in intensity of pixels is significant. In order to derive proper temporal information, an aggregate of optical flows is derived from a time window, including both immediate past frames and immediate future frames.

We calculate optical flow in a sliding window for a span of time t_i, spanning from frame f_{i-ns} to frame f_{i+ns}, separated by a stride of s frames since consecutive frames are likely to be very similar in pixels. Thus, optical flow is applied over a set of $2N+1$ images per window. A total of $2N$ pairwise optical flow maps are obtained corresponding to each frame-pair f_{i+ns} and $f_{i+(n+1)s}$ $\forall\ n \in [-N, N-1]$. We represent the optical flow map corresponding to the shift from frame f_{i+ns} and $f_{i+(n+1)s}$ as $O_{f_{i+ns} \to f_{i+(n+1)s}}$.

The Gunnar Flow [5] method returns a Hue Saturation Value (HSV) image, which is converted into an RGB image for better handling and interpretation. A threshold-based filter is used to remove noise from these optical flow maps. Raw OF maps contain high amounts of noise due to factors such as constant

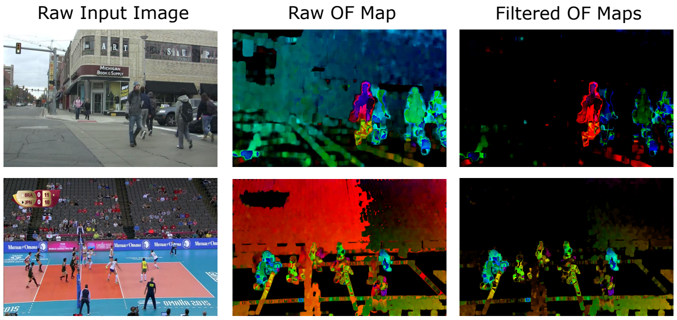

Fig. 3. Effect of filtering optical flow maps using the thresholding mechanism.

camera movement and blurriness. Pixel values across all channels are clipped within a range of values $\in [a_{min}, a_{max}]$, and scaled appropriately. This results in clean representations of temporal movement in the sequence of frames. This filter is represented as

$\phi(O_{f_i,f_{i+j}})$, and is defined as

$$\phi(O_{f_{i+ns}\to f_{i+(n+1)s}}) = u_{a_{min}}(x) - u_{a_{max}}(x)$$
$$\forall \text{ pixels } x \in O_{f_{i+ns}\to f_{i+(n+1)s}} \tag{1}$$

where, $u_a(x)$ is a step function defined as

$$u_a(x) = \begin{cases} 1 & x \geq a \\ 0 & x < a \end{cases} \tag{2}$$

and a_{min} and a_{max} are the lower and upper thresholds of the optical flow maps. This helps the representation in two ways. (1) Noise appearing due to erratic visual movement, sudden camera movement or blurriness, is suppressed. (2) The optical flow maps contain minor pixel value shifts. This can occur because of vibrating camera movement, or minute movements in the background. Clipping these activations to a zero value gives a clean representation of optical flow corresponding to the local temporal shift. The result of filtering can be seen in Fig. 3.

Out of the N frames in a sequence, we exclude the extreme ends and superimpose the optical flow maps on the remaining $2N-1$ frames, to obtain the superimposed image representation $S_{f_{t+i}}$, corresponding to frame f_{t+i}, using the following scheme.

$$S_{f_{i+ns}} =$$
$$f_{i+ns} + \frac{1}{k}\left[\phi(O_{f_{i+(n-1)s}\to f_{i+ns}}) + \phi(O_{f_{i+ns}\to f_{i+(n+1)s}})\right]$$
$$for\ n \in \{-N+1, -N+2, ..., N-2, N-1\} \tag{3}$$

where k is a scaling factor for the optical flow maps. The superimposed images are weight pooled to obtain the final representation corresponding to the middlemost frame f_i. This is given as:

$$x_{f_i} = \sum_{n=-N+1}^{n=+N-1} w_n S_{f_{i+ns}} \tag{4}$$

where w_n is a weight factor assigned to superimposed images. The middlemost frame is given the highest weightage, followed by the immediate frames in decreasing order. Successive and predecessive superimposed frames at equal distances from the middlemost frames have equal weightage. We define the weight factor w_n as $\frac{1}{2^{|n|+1}}$. The weighting factor was chosen in the form of $\frac{1}{2^n}$, because of the convergence of the corresponding infinite geometric series to unity. We assign equal weights to temporally equidistant frames. However, this is not strict and can be adapted depending on the dataset used, since the Neural Network is capable of fitting arbitrary, but reasonable data distribution. An equally valid weight factor formulation would be a normalized inverse exponential function $\frac{1}{k.e^{a(|n|+1)}}$ or $\frac{1}{k.x^{a(|n|+1)}}$ for some real number x, where k is a suitable normalizing factor, that not only ensures mathematical convergence of the corresponding infinite series superimposition but also ensures that pixel values stay within computational limits.

The parameters of WTM such as a_{min}, a_{max}, and w_n have been chosen to be compatible universally with any sequential image data. These do not have a strict value and were simply chosen through empirical logic. Therefore, experimentation with these values is encouraged.

Finally, we normalize the weighted image within the computation limit of the tensors. Additionally, we normalize the inputs to ImageNet statistics across individual channels. ImageNet Statistics enable us to use ImageNet pretrained 2D models. However, this normalization can b for training. This can be omitted if one decides to train the network from scratch.

$$x = N(x_{f_t}/max(x_{f_t}), \mu_{ImageNet}, \sigma_{ImageNet}) \tag{5}$$

This can be directly fed into an image classification model with the target class y equal to the corresponding label of frame f_i. A single batch x_b can thus contain data from multiple sequences, with different labels. A schematic of this process is illustrated in Fig. 2. A stronger hue is seen over portions where there is a significant change in intensity, which may occur by sudden movement, or shaking of the camera. The thresholding mechanism suppresses the noise to a great extent, as is seen in Fig. 3. We feed a batch of images with temporal attention to an image classification network with single labels and thus reduce the problem to a traditional 2-dimensional image classification problem. Thus we have eliminated the need for any multi-stream or augmented dimensional (such as 3-dimensional) architectures.

Experimental Results

We evaluate the results and trends of our model with respect to various experiments and hypotheses. We also provide a critical analysis of the approaches used in this paper compared to the approaches of other papers. However, before discussing the performance of the model, we briefly analyze the datasets upon which the model has been built.

Datasets

We train and compare our results over two widely used benchmark datasets - the Collective Activity Dataset and the Volleyball Dataset.

Collective Activity Dataset (CAD). The dataset contains 44 short sequences of five different collective activities, namely - walking, talking, crossing, waiting, and queueing, recorded through a handheld camera. Each short sequence contains various continuous frames, that are sampled at a rate of 25 frames per second (fps). Every tenth frame is labeled. The dataset contains a fairly balanced distribution of different activities, background settings, and camera angles.

Volleyball Dataset This dataset is curated from publicly available Volleyball matches found on YouTube. It contains 4830 annotated frames from 55 different videos [12], containing 8 group activities specific to this sport, and 9 individual activity labels. Labeled items are separated by 40 unlabeled frames, exclusive of the labeled items, meaning for every labeled frame, the authors provide 20 continuous succeeding unlabeled frames and 20 continuous preceding unlabeled frames, which are intended to be used along with the labeled frame to factor in temporal motion within the sub-sequence. The quality of the imagery is fairly good by modern standards.

Performance Evaluation

We present and compare the results over both the datasets used. The proposed model achieves State-of-the-Art testing accuracy on both datasets. The individual results of the Group Activity Classification are 95.1% and 94.6% for CAD and Volleyball datasets, respectively. The image classification backbone is replaceable with any suitable backbone. As mentioned before, we demonstrate the results on two backbones - *ResNet* [24] and *EfficientNet* [25]. The two architectures demonstrate similar performances. However, the ResNet variant slightly outperforms the latter and also achieves state-of-the-art results.

Fig. 4 demonstrates the performance of the model on the most difficult cases. We illustrate through various examples how the model has been able to identify the characteristics of various activities despite underlying ambiguity in actions and labels.

Ablation Study

Base Architecture Selection We primarily evaluate our method on various architectural variations of the ResNet and EfficientNet backbones. We perform experiments to analyze the contribution of spatial information in group activity datasets. We initially train the CNN backbones on images without optical flow-based temporal information. All other data transformations and augmentation methods remain the same. The model is trained using the same training method as described in earlier sections. The only variation occurs in hyperparameter selection [26]. The comparison of various ResNet architectures is shown in TABLE I. We use an ImageNet pre-trained ResNet architecture as the backbone for training the image classification model. The table shows a comparison of three ResNet variants, each tested on two size variants of the input - 128×128 pixels and 256×256 pixels sequentially. All values reported are testing accuracies as pre-determined by the authors of the datasets.

TABLE I
COMPARISON OF BASE ARCHITECTURE WITHOUT OPTICAL FLOW-BASED ATTENTION MECHANISM BASED ON TESTING ACCURACY

Dataset	**ResNet18**		**ResNet34**		**ResNet51**	
Size	128	256	128	256	128	256
CAD	88.3	88.3	85.0	**89.0**	82.6	84.1
Volleyball	50.20	72.2	67.7	75.5	73.3	**75.7**

We observe in TABLE I that a ResNet34 architecture outperforms its related architectures for the Collective Activity Dataset by a significant margin, while the ResNet51 architecture outperforms its related architectures for the Volleyball dataset. However, the ResNet34 counterpart for the Volleyball Dataset only underperforms the ResNet51 architecture by 0.2%. Hence we select the ResNet34 architecture as the baseline for our model training. The table also shows the final size of the pre-sized input image, over which the model is trained. As mentioned before, we use the method of progressive resizing, meaning the image size is increased in stages, starting from size 128×128, followed by size 256×256 pixels. This not only helps train the network faster but helps train essential features with a smoother contour, hence giving better overall accuracy at the end of the training.

Architecture performance with WTM Following this, we add temporal attention to input images through WTM. This involves optical flow-based attention maps overlayed on top of the raw input RGB images. This completes the information representation of our proposed method. We train the model over multiple parametric variants of the WTM module to analyze the effect of all contributing temporal attention representations within the input image. For symmetry, N assumes a positive odd integer value not equal to 1 (since 1 frame per subsequence would not account for any temporal information), while s can be any positive integer. However, we set an upper limit for experiments on N, such that the corresponding weight factor w_n does not exceed a value of 1000 since beyond that, the contribution of temporal representation becomes insignificant. In other words, N takes a maximum value of 9. Similarly, we set an upper limit on s equal to 5, since, beyond that, the noise in the optical flow maps becomes significant, and affects the attention mechanism, even after careful thresholding. A detailed analysis of experiments on the weighted temporal attention module is shown in TABLE III.

The results can be seen in TABLE IV and TABLE V. The proposed weighted temporal attention mechanism beats the state-of-the-art results on both datasets. These results have been compared with previous works in the following subsection.

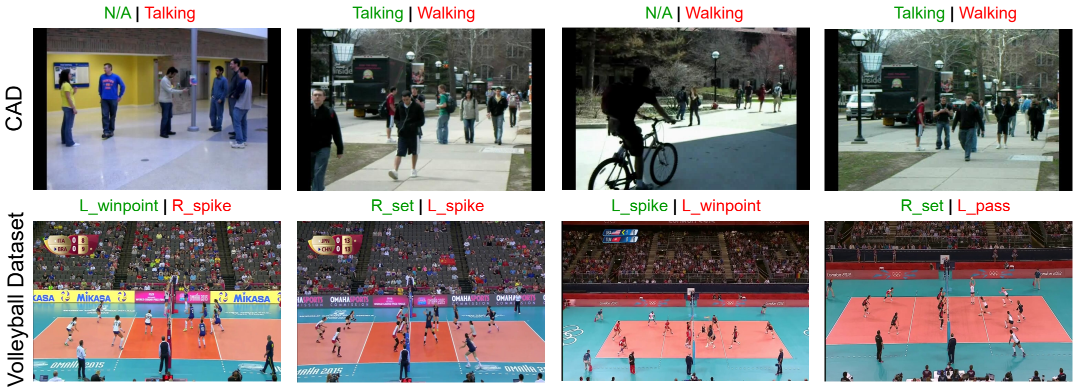

Fig. 4. Model prediction on difficult cases. **Green**: Actual Label in the dataset. **Red**: Model Prediction. The figure illustrates wrong predictions on the two datasets. **Collective Activity Dataset**: The figure illustrates that the model has correctly learned features of activities such as talking, crossing, queuing, *etc.*, despite ambiguity and error in labeling logic. **Volleyball Dataset**: Most errors occur in cases where two activities have similar actions. For example, the majority of misclassification pairs constitute win points vs spikes. Both actions have similar physical characteristics.

TABLE II
EXPERIMENTS ON THE PROPOSED WTM ATTENTION METHOD. WE VARY THE NUMBER OF FRAMES PER SUBSEQUENCE N AND THE STRIDE BETWEEN TWO CONSECUTIVE FRAMES s. THE ACCURACIES ON BOTH CAD AND VOLLEYBALL DATASETS HAVE BEEN TABULATED.

s	N			
	3	5	7	9
Accuracy on Collective Activity Dataset (%)				
1	89.3	92.3	91.9	89.0
2	91.4	92.4	92.8	91.9
3	93.1	**95.1**	94.9	92.5
4	91.0	94.3	93.7	89.1
5	87.6	89.4	89.9	88.8
Accuracy on Volleyball Dataset (%)				
1	75.7	81.9	80.6	73.0
2	87.5	87.8	89.3	83.4
3	93.9	91.3	87.2	77.2
4	**94.6**	94.2	84.3	82.7
5	81.3	87.4	89.0	74.6

TABLE III
EXPERIMENTS ON THE PARAMETERS OF WTM.

	Testing Accuracy on CAD (%)				Testing Accuracy on Volleyball Dataset (%)			
s	N				N			
	3	5	7	9	3	5	7	9
1	89.3	92.3	91.9	89.0	75.7	81.9	80.6	73.0
2	91.4	92.4	92.8	91.9	87.5	87.8	89.3	83.4
3	93.1	**95.1**	94.9	92.5	93.9	91.3	87.2	77.2
4	91.0	94.3	93.7	89.1	**94.6**	94.2	84.3	82.7
5	87.6	89.4	89.9	88.8	81.3	87.4	89.0	74.6

Discussion and Comparison

The proposed method achieves the state-of-the-art results of 95.1% and 94.6% on the collective activity dataset and volleyball dataset, respectively. The results are tabulated in the TABLE IV and TABLE V, along with a comparison with various past works. All reported values are testing accuracies as pre-determined by the authors of the datasets.

Space-Time Complexity Analysis : For a proper generalization and comparison with various models, we define computational complexities in terms of some general variables, as follows.

- d: dimensional (pixel) size (height or the width) of the input. This affects the number of convolutions in a convolutional neural network through a squared relation.
- l: average number of layers across all streams.
- m: number of streams used in the model. Thus the total

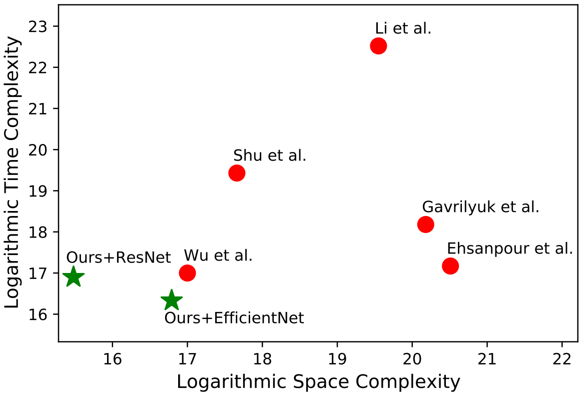

Fig. 5. Space Time Complexity Analysis for various models.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART ON THE COLLECTIVE ACTIVITY DATASET. RED DENOTES THE BEST AND BLUE DENOTES THE SECOND BEST RESULTS.

Method	Backbone	2D/3D	Single/ Multi-Stream	Group Activity
[27]	N/A	N/A	N/A	79.7%
[28]	N/A	N/A	N/A	80.4%
[2]	AlexNet	2D	Multi	81.2%
[12]	AlexNet	2D	Single	81.5%
[29]	N/A	N/A	N/A	83.4%
[16]	I3D	3D	Multi	85.8%
[13]	Inception-v3	2D	Multi	86.1%
[14]	VGG16	2D	Multi	87.2%
[21]	VGG16	2D	Multi	89.1%
[30]	I3D	3D	Single	89.4%
[15]	Inception-v3	2D	Multi	91.0%
[31]	I3D	3D	Multi	92.8%
Ours (WTM)	**EfficientNet-B4**	2D	Single	**95.0%**
Ours (WTM)	**ResNet-34**	2D	Single	**95.1%**

TABLE V
COMPARISON WITH STATE-OF-THE-ART ON THE VOLLEYBALL DATASET. RED DENOTES THE BEST AND BLUE DENOTES THE SECOND BEST RESULTS.

Method	Backbone	2D/3D	Single/ Multi-Stream	Group Activity
[12]	AlexNet	2D	Single	81.9%
[14]	VGG16	2D	Multi	83.3%
[21]	VGG16	2D	Multi	89.3%
[32]	VGG19	2D	Single	89.5%
[33]	Inception-v3	2D	Single	90.6%
[15]	Inception-v3	2D	Multi	92.5%
[16]	I3D	2D	Multi	93.0%
[30]	I3D	3D	Single	93.1%
[31]	HRNet + I3D	3D	Multi	**94.4%**
Ours (WTM)	**EfficientNet-B4**	2D	Single	**92.2%**
Ours (WTM)	**ResNet-34**	2D	Single	**94.6%**

TABLE VI
COMPARISON OF SPACE-TIME COMPLEXITY OF A GENERAL GROUP ACTIVITY RECOGNITION MODEL VS THE PROPOSED MODEL.

	General Model	**Proposed Model**
Time Complexity	$O(Nd^2k^2lm)$	$O(d^2k^2l+N)$
Space Complexity	$O((p+l)m+Nd^2)$	$O(p+l+Nd^2)$

number of layers in the model is lm, which increases the computational complexity linearly.

- N: The number of frames used to derive temporal information. The computational complexity increases linearly with the number of frames considered per sequence to model temporal representations.
- p: average number of parameters in the model accross all streams. Hence the total number of parameters in the model equals pm This constitutes the space complexity of the model directly.

Time Complexity: We initially analyse the time complexity of the group activity recognition model architecture. For a convolutional neural network based architecture, a forward pass constitutes of $O(d^2)$ convolutions per layer. Hence for lm layers in the model, there are $O(d^2lm)$ convolutions. If we consider a kernel size k, the time complexity of a single forward pass can be expressed as

$$O(d^2k^2lm)$$

Considering the effect of multi-dimensional convolutional network based approaches, a forward pass constitutes of $O(Nd^2)$ convolutions per layer. This expression also applied to 2 dimensional approaches, that convolute over multiple frames independently. Hence, in such cases, the time complexity of a single forward pass can be expressed as

$$O(d^2Nk^2lm) \tag{6}$$

Besides this, any extra component in the architectural pipeline (if exists), such as graphs, can be expressed through a separate term as follows.

$$O(d^2Nk^2lm+A) \tag{7}$$

The proposed method eliminates many factors from the above complexity expression. The classification architecture uses a single stream 2D convolutional neural network, the time complexity of which can be expressed as

$$O(d^2k^2l)$$

The addition of the proposed weighted temporal attention module introduces an additional time complexity of the order of $O(Nd^2)$. Hence the overall time complexity of the proposed method is expressed as

$$O(d^2(k^2l+N)) \tag{8}$$

Space Complexity: The general space complexity of a group activity recognition model is directly proportional to the total parameters in the model. There is also an $O(lm+1)$ space complexity corresponding to additional hyperparameters, which results in a total architectural space complexity of

$$O(pm+lm+1) \approx O((p+l)m)$$

Apart from this, the memory requirements of the input data are of the order of $O(Nd^2)$.Hence, the total space complexity of a general model can be expressed as

$$O((p+l)m+Nd^2) \tag{9}$$

The proposed architectural approach, by virtue of being single-stream in nature, reduces the architectural space complexity to

$$O(p+l)$$

The memory requirements corresponding to input data along proposed weighted temporal attention module is an additional $O(Nd^2)$ order memory per subsequence. Hence the proposed model has an overall space complexity of

$$O(p+l+Nd^2) \tag{10}$$

A side by side comparison of the expressions of space and time complexities of a general group activity recognition model versus the proposed method is presented in Fig. VI.

Following this scheme, we compare the space and time complexities of various works in Fig. 5. As can be seen, WTM based image classification achieves a lower space and time complexity in comparison to other previously proposed methods. For uniformity, it has been assumed that a single forward pass is carried out on an image of size $d = 128$ pixels over convolutions of kernel size $k = 3$.

The proposed model effectively reduces time and space complexity in comparison to multi-stream networks and multi-dimensional (3D) models. In comparison to general single stream, 2D networks, the proposed approach has a reduced time complexity and no additional space complexity. This proves the efficiency of the proposed method.

SUMMARY AND FUTURE WORK

In this paper, we propose WTM, a novel temporal attention-based method that combines spatial and temporal information for the task of group activity recognition. This method uses temporal flow as an attention mask, which can be overlayed on raw input images using a layered weighted scheme. The resulting input images can be fed into any suitable image classification model. The proposed method achieves state-of-the-art results on two benchmark datasets. We demonstrate how a simple and intuitive approach towards information representation is effective in extracting high-level features from data. Further work can explore the effect of temporal attention in feature derivation, and how the search space for effective features can be further reduced while retaining maximal information from the data. The method described in this paper provided flexibility in the selection of image classification model architecture, which can be replaced with any suitable architecture in a plug-and-play fashion. We also discussed the features and limitations of the two benchmark datasets that are widely used to evaluate group activity recognition models. There is a need to layout guidelines about identifying and prioritizing activities. The quality of datasets is another concern that needs to be addressed, especially in modern times, where image quality standards are improving every day.

REFERENCES

[1] C. Zalluhoglu and N. Ikizler-Cinbis, "Collective sports: A multi-task dataset for collective activity recognition," *Image and Vision Computing*, vol. 94, p. 103870, 2020.

[2] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4772–4781.

[3] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou, "Graph interaction networks for relation transfer in human activity videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2872–2886, 2020.

[4] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "Higcin: Hierarchical graph-based cross inference network for group activity recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[5] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[6] S. Vahora and N. Chauhan, "Deep neural network model for group activity recognition using contextual relationship," *Engineering Science and Technology, an International Journal*, vol. 22, no. 1, pp. 47–54, 2019.

[7] F. Ferland, R. Agrigoroaie, and A. Tapus, "Assistive humanoid robots for the elderly with mild cognitive impairment," 2017.

[8] W. Li, M.-C. Chang, and S. Lyu, "Who did what at where and when: simultaneous multi-person tracking and activity recognition," *arXiv preprint arXiv:1807.01253*, 2018.

[9] S. M. Azar, M. G. Atigh, and A. Nickabadi, "A multi-stream convolutional neural network framework for group activity recognition," *arXiv preprint arXiv:1812.10328*, 2018.

[10] J. Zang, L. Wang, Z. Liu, Q. Zhang, G. Hua, and N. Zheng, "Attention-based temporal weighted convolutional neural network for action recognition," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2018, pp. 97–108.

[11] L. Wang, J. Zang, Q. Zhang, Z. Niu, G. Hua, and N. Zheng, "Action recognition by an attention-aware temporal weighted convolutional neural network," *Sensors*, vol. 18, no. 7, p. 1979, 2018.

[12] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1971–1980.

[13] X. Li and M. Choo Chuah, "Sbgar: semantics based group activity recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2876–2885.

[14] T. Shu, S. Todorovic, and S.-C. Zhu, "Cern: confidence-energy recurrent network for group activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5523–5531.

[15] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9964–9974.

[16] S. M. Azar, M. G. Atigh, A. Nickabadi, and A. Alahi, "Convolutional relational machine for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7892–7901.

[17] G. Hu, B. Cui, Y. He, and S. Yu, "Progressive relation learning for group activity recognition," *arXiv preprint arXiv:1908.02948*, 2019.

[18] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3043–3053.

[19] Y. Tang, P. Zhang, J.-F. Hu, and W.-S. Zheng, "Latent embeddings for collective activity recognition," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6.

[20] L. Lu, H. Di, Y. Lu, L. Zhang, and S. Wang, "Spatio-temporal attention mechanisms based model for collective activity recognition," *Signal Processing: Image Communication*, vol. 74, pp. 162–174, 2019.

[21] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "stagnet: An attentive semantic rnn for group activity recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 101–117.

[22] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 773–787, 2016.

[23] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3034–3042.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[25] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.

[26] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay," 2018.

[27] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 8, pp. 1549–1562, 2011.

[28] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 215–230.

[29] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori, "Visual recognition by counting instances: A multi-instance cardinality potential kernel," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2596–2605.

[30] M. Ehsanpour, A. Abedin, F. Saleh, J. Shi, I. Reid, and H. Rezatofighi, "Joint learning of social groups, individuals action and sub-group activities in videos," *arXiv preprint arXiv:2007.02632*, 2020.

[31] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. Snoek, "Actor-transformers for group activity recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 839–848.

[32] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 721–736.

[33] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4315–4324.