

# Let's Dance: Learning From Online Dance Videos

Daniel Castro  
Georgia Institute of Technology  
dcastro9@gatech.edu

Steven Hickson  
shickson@gatech.edu

Patsorn Sangkloy  
patsorn.sangkloy@gatech.edu

Bhavishya Mittal  
bmittal6@gatech.edu

Sean Dai  
sdai@gatech.edu

James Hays  
hays@gatech.edu

Irfan Essa  
irfan@gatech.edu

## Abstract

In recent years, deep neural network approaches have naturally extended to the video domain, in their simplest case by aggregating per-frame classifications as a baseline for action recognition. A majority of the work in this area extends from the imaging domain, leading to visual-feature heavy approaches on temporal data. To address this issue we introduce “Let’s Dance”, a 1000 video dataset (and growing) comprised of 10 visually overlapping dance categories that require motion for their classification. We stress the important of human motion as a key distinguisher in our work given that, as we show in this work, visual information is not sufficient to classify motion-heavy categories. We compare our datasets’ performance using imaging techniques with UCF-101 and demonstrate this inherent difficulty. We present a comparison of numerous state-of-the-art techniques on our dataset using three different representations (video, optical flow and multi-person pose data) in order to analyze these approaches. We discuss the motion parameterization of each of them and their value in learning to categorize online dance videos. Lastly, we release this dataset (and its three representations) for the research community to use.

## 1. Introduction

Video is a rich medium with dynamic information that can be used to determine, what is happening in a scene. In this work, we consider *highly dynamic video*, video that requires the parametrization of motion over extended sequences to identify the activity being performed. The main challenge with highly dynamic video is that a single frame cannot provide sufficient information to understand the action being performed. Therefore, multiple frames, leading to an extended sequence of frames, need to be analyzed for scene classification. One of the drawbacks of current action classification research is both a lack of approaches that can

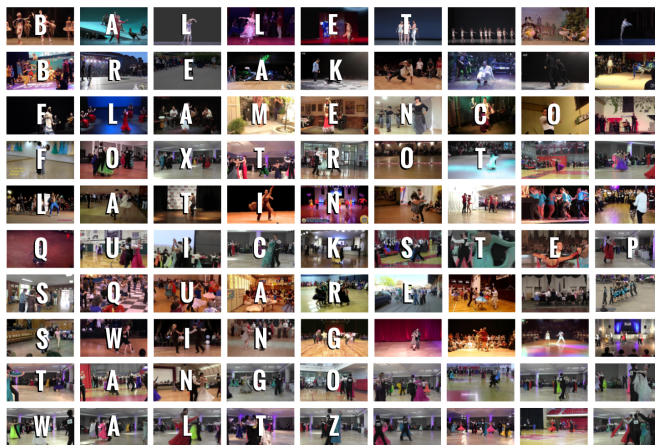


Figure 1. Each row contains frames from the class it represents. This figure is best viewed digitally.

be applied to extended/long sequences and datasets lacking in such highly dynamic videos. Our goal is to determine which methods best represent motion as opposed to methods that use a single (properly picked) frame [6] to identify the activity, as we feel such approaches devalue the necessity for video data. In this work we introduce a 1,000 video dataset and evaluate methods that focuses on highly dynamic videos requiring motion analysis for classification.

We choose the domain of dance videos as (a) there is large amount of dance videos available online and (b) the diversity of dynamics in these videos provides us with a challenging space of problems for highly dynamic video analysis. This enables us to conduct a focused study on the relevance of motion in dancing classification and the broader value of motion in improving video classification.

The core challenge of this task is attaining an adequate representation of human motion across a 10-second clip. In order to highlight the trajectory of this work, we will evaluate the current approaches and demonstrate the value of isolating motion for properly evaluating these approaches and this dataset.

Many video classification techniques exist, either utilizing single frames, late fusion architectures, temporal (3D) convolutional networks, or recurrent networks with long short-term memory (LSTM). Current classification problems can often be identified by a single frame. We present a more challenging problem wherein each class requires the use of multiple frames to adequately classify each category.

Specifically, we propose the use of optical flow and multi-agent pose estimation as motion representations which augment traditional video classification approaches. Comparing these approaches enables us to gain insights into the inherent encoding of motion in neural networks that is difficult to understand.

Our main contributions are: (1) An analysis of baseline and state-of-the-art approaches in video classification, (2) a general method for concurrently learning from multiple motion parameterizations in video, and (3) A 1000 video dataset of highly dynamic dance videos, contrasted with existing video datasets, to motivate further investigation and understanding of motion parameterization in video classification.

## 2. Related Work

In order to determine which competing state-of-the-art approaches to examine, we first present a literature review on video classification. While deep networks have been shown to be very effective at classifying, localizing, and segmenting images, it is still unclear how to properly extend these methods to the video domain. There are a few common approaches, some of which are: (1) Applying proven image classification deep network architectures to individual frames of a video; (2) Extending 2D convolutional operators to 3D convolutions acting on the time domain; and (3) Preprocessing the video into images that encode motion, such as optical flow, and running current image architectures on the processed frames.

A simple way to extend image-based neural networks to video classification is to extract features from each individual frame of a video [31]. While this technique does lead to some success if the network learns temporally-invariant features, it is commonly only used as a baseline approach to compare against networks that incorporate temporal data [9, 11]. One common variant is a two-stream late fusion architecture with a still frame-based “spatial” network stream running in parallel alongside a “temporal” network performing classifications based on optical flow calculations [3, 6, 8, 20]. This network architecture significantly outperforms approaches based solely on individual frame classification, suggesting that incorporating a temporal component is necessary. In our work we leverage the benefit of a temporal network by incorporating it into the design of our network architecture.

Karpathy et al. explore more direct methods of incorpo-

rating temporal data with each video frame by extending the convolution kernels from size  $m \times m \times 3$  to  $m \times m \times 3 \times T$ , where  $T$  represents a temporal extent[11]. They also point out one of the major challenges of using deep learning for video classification – there are no large-scale video datasets comparable to the quality and size of image recognition datasets. Similarly, 3D convolutional kernels that incorporate the spatial domain have been shown to be successful for action classification in both security camera and depth data recordings [9, 27]. Wang et al. use a similar two-stream late fusion approach [28], but they note that without incorporating the learned features into an ensemble method with handcrafted features, these deep-learned approaches still fail to outperform handcrafted approaches. We combine these methods in our work by incorporating preprocessed features (optical flow and multi-agent pose detection) with 3D convolutional kernels in order to integrate the representation of motion into the network architecture.

Another common approach is to leverage the sequential nature of a Long Short-Term Memory (LSTM) network—a specific type of recurrent neural network with additional gates to control the flow of information. LSTMs can process information over long term temporal sequences and have been applied in video for various tasks such as caption generation [26] and learning video representations [23]. Similarly, the long-term recurrent convolutional networks (LRCNs) proposed by Donahue et. al. introduce another variation of an LSTM for this task. Despite their temporal nature, these approaches have been less successful at encoding motion in comparison to two-stream networks [4] which encode the spatial and temporal domain in concurrent architectures.

The most effective method for classifying motion in video is still unclear. In the context of action recognition, many of these approaches are learning features based on the image’s context and not the inherent action. This is in part because commonly used video datasets such as UCF-Sports and more traditionally UCF-101 can generally be identified to moderately decent accuracy using single-frame approaches which do not encode motion parameters[11].

A specific method for encoding motion that has recently gained traction in action recognition is the use of pose detection over the temporal domain with neural networks [3][14][24][30]. Detecting pose over this domain provides us with the intrinsic motion of the subjects in the scene. As highlighted earlier, an initial breakthrough was achieved by Toshev et. al. [24] with state-of-the-art results in estimating the pose of a single individual from an image. The importance of pose was further demonstrated in [3], incorporating pose features from a CNN into action recognition. This work was extended over the next two years to attain joint-specific networks that work well with partial and occluded poses [30]. It was then most recently implemented to detect



Figure 2. Each of these examples represents a different class in our dataset (they are types of ballroom dancing). Top Left: Waltz, Top Right: Quickstep, Bottom Left: Foxtrot, Bottom Right: Tango.

multiple people within a single frame [2]. In our work we will leverage our own implementation of multi-agent pose detection to demonstrate the need for motion parameterization when classifying highly dynamic video.

## 2.1. Existing Datasets

There are a handful of relevant datasets that exist in the research domain. We highlight some of the more relevant video datasets that are appropriate to our work. All of these datasets demonstrate the growing need for understanding what type of motion features are relevant in classifying highly dynamic actions, which we explore in this work.

### 2.1.1 UCF-101

The UCF-101 dataset [22] contains approximately 13,000 clips and 101 action classes, totaling 27 hours of data. The clip length varies largely from 1 second to 71 seconds depending on the activity at a resolution of 320x240. This was one of the first datasets to tackle human actions in video. However, as we will demonstrate in this work, most per-frame (image-based) approaches still perform moderately well on the dataset, illustrating the main question which we seek to answer in this work – that being the representation of motion as a classification feature.

### 2.1.2 Kinetics

The Kinetics dataset [12] contains 300,000 clips and 400 action classes, with a minimum of 400 videos per class. The action classes are also loosely grouped in 32 parent classes which further break down the dataset. This dataset was collected semi-automatically with curation through image classifiers and use of Amazon Mechanical Turk to determine the action classes and if the video snippet was appropriate to that class.

### 2.1.3 Atomic Visual Actions (AVA)

The AVA dataset [15] contains 80 atomic visual actions in 57,400 movie clips which are localized within the frame. This work goes beyond simply understanding a simple action in a video clip to understanding the interaction, both between humans and with humans and objects. Although this is somewhat less relevant to our work, it demonstrates the need for understanding motion features in human interaction – specifically by localizing the action and its relevance in a scene that may contain multiple subjects / objects.

## 3. Let’s Dance Dataset

Our main challenge in this work was determining a reliable way of testing how well a specific method can parameterize motion. We realized that available video datasets such as UCF-101 [22] and UCF-Sports [18] tackled a known classification problem, one that could be evaluated using extensions of available image classification architectures.

With that in mind, we developed a new dataset that prioritizes motion as the key characteristic of the classification. We assembled a 1,000 video dataset containing 10 dynamic and visually overlapping dances. We chose the parent category of dancing because it has a variety of measurable features (*i.e.* rhythm, limb movement), and it is not represented in the Sports-1M and UCF-101 datasets [11, 22]. The categories included for this dataset are:

- Ballet
- Flamenco
- Latin
- Square
- Tango
- Break Dancing
- Foxtrot
- Quickstep
- Swing
- Waltz

The dataset contains 100 videos for each class. Each video is 10-seconds long at 30 frames per second. The videos themselves were taken from YouTube at a quality of 720p, and includes both dancing performances and plain-clothes practicing. Examples of each class can be seen in Figure 1.

We highlight that the dataset contains four different types of ballroom dancing (quickstep, foxtrot, waltz, and tango) as seen in Figure 2. The motivation behind picking these dances is that their parent category is specifically the setting in which the dance occurs (a ballroom). This satisfies our main challenge of selecting classes that exemplify highly dynamic video. On this note, we extract two different motion representations from our input data for use by the community; optical flow[5] and multi-person pose detection.

When attempting to detect pose, we found numerous methods that focused on single-person pose detection. We adapted these methods to multiple individuals (given that dancing is generally a group activity, see Figure 3) through

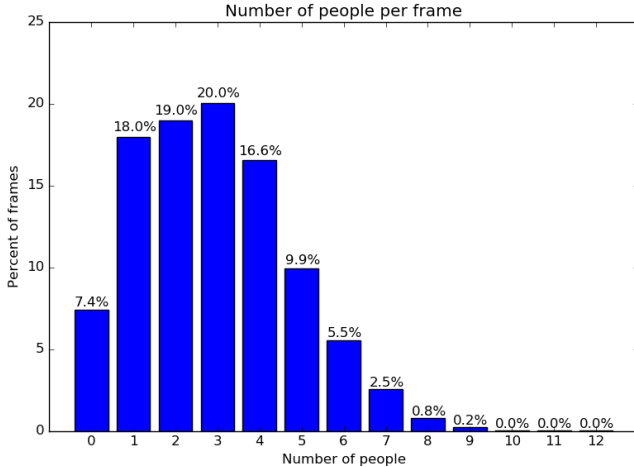


Figure 3. Distribution of number of people per frame using [17]. 75% of frames had at least two people detected in the dataset. 56% of the dataset has more than two people in the shot, which further illustrates the added complexity of this task.

the use of a recent real-time person detector[17]. Similar approaches can be seen in [7][16][2].

After detecting the bounding boxes for each person in the scene we computed the pose for each individual using [30]. Positive and negative examples of this methodology can be seen in Figure 7.

#### 4. Baseline Methods

In order to better understand the need for motion parametrization in video, we have identified two commonly-used architectures to establish as our baseline. These are architectures which are commonly applied to video architectures but only take a single-frame as input (per architecture).

These approaches are extensions of very successful image classification techniques.

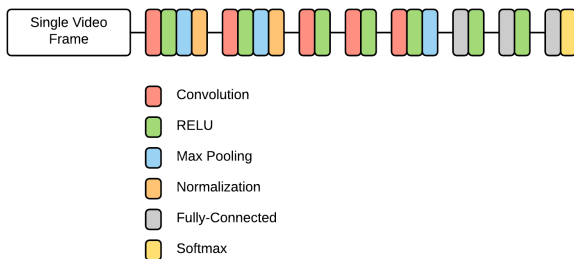


Figure 4. Frame-by-Frame Architecture: This is a traditional CNN, commonly used in image recognition.

#### 4.1. Frame-by-Frame

Using the architecture of a state-of-the-art convolutional neural network for image classification, such as VGG [21], a classification for the video can be achieved based on key image frames from a video. A sample architecture based on CaffeNet, a variation of AlexNet [13], is shown in Figure 4. This approach does not explicitly encode motion in determining the video’s classification but rather categorizes each frame and naively selects the majority label.

We do note that although there are numerous approaches for aggregating a single class from multiple per-frame classifications, the network itself does not encode the temporal domain.

#### 4.2. Two-Stream Late Fusion

A common way of adding a temporal component to deep networks is by separately performing a classification based on spatial data (a single frame) and temporal data (i.e. optical flow). Merging these results produces an overall classification for the video, as shown in Figure 5.

This approach computes optical flow from two frames (at time  $n$  and  $n - k$  where  $k$  is not necessarily 1) over the period of the entire video. Each frame in this case can be considered a single instance of motion that occurred in the video. For dancing we envision this as a specific move in a dance.

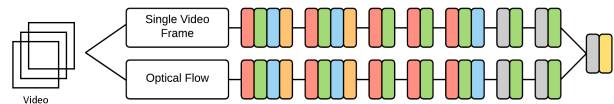


Figure 5. Two-Stream Late Fusion Architecture (color key in 4). This method incorporates motion (optical flow) into a traditional CNN pipeline.

### 5. Proposed Approaches

In order to address the challenge of categorizing highly dynamic videos we implement a number of methods which explicitly encode motion. At the core of these approaches is the notion of 3-dimensional kernels which process a series of video frames for classification. This enables us to pass in very short video clips (16 frames or approx. 1/2 second) for the network to learn. The overall objective was to incorporate motion in the learning pipeline of standard approaches and assess their performance. After testing these approaches it was evident that combining numerous motion parameterizations in a concurrent deep network architecture would best represent the input video.

#### 5.1. Temporal 3D CNN (RGB)

As stated, traditional convolutional neural networks can be extended to video by using 3-dimensional kernels that

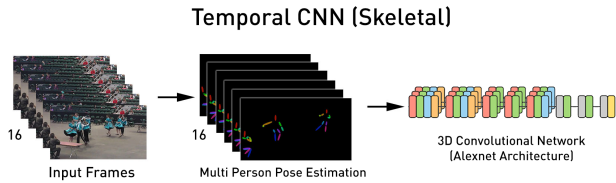


Figure 6. This pipeline displays a skeletal temporal CNN (3D Convolution) which processes the initial frames to obtain a multi person pose estimation from the input frames obtained by performing a bounding box person detection from [17] which is then processed by [30] for detecting the dancers’ pose.



Figure 7. Demonstration of outputs from our pose detection pipeline. Top: Latin dancing positively classified. Bottom: Break dancing being erroneously classified. The dancers’ left leg is accurate but his remaining limbs fail.

convolve in the temporal domain. We focus on testing this slow-fusion approach discussed in [9], which embeds the high-level spatial and temporal information at the initial convolutional layers by propagating the information through the network. One of the main setbacks of this proposed approach is the computational time it currently takes to compute these methodologies. We discuss this further in Section 7.

## 5.2. Temporal 3D CNN (Skeletal)

In this pipeline we compute a temporal CNN on multi-person pose information. We visualize the pipeline in Figure 6. This architecture demonstrates the importance of leveraging context for particular videos. Dance videos inherently benefit from this representation given that there are generally multiple people in the scene. Through the use of a visualization of pose we are able to attain comparable results to single-frame CNN approaches. It is key to note that this method does not use visual information from the original frame but solely visualized pose information as shown in Figure 7. Similar to our optical flow approach, it is likely that this method benefits heavily from encoding the number of people in the frame in addition to the motion over the 16

frames that are convolved in the temporal domain.

## 5.3. Three-Stream CNNs

We tested both single-frame and temporal approaches for a three-stream convolutional network in order to directly compare the potential importance of embedding multiple frames into the learning pipeline in addition to providing multiple representations of your original input. We highlight that these temporal convolutions are computing 2D convolutions over each of the input frames. Although this increases the complexity of our model it still remains significantly more tractable than computing 3D convolutions which require approximately twice the computational power.

### 5.3.1 Frame-by-Frame Architecture

For the frame-by-frame architecture, the first stack of our network processes the spatial representation of our input which is our RGB image. Our second stack processes the optical flow representation which was computed from frames  $n$  and  $n - 10$  in order to accentuate particular motions from a given dance. Our third stack processes our multi-person pose visualization explained in Figure 6. As discussed earlier, this stack is essentially encoding the number of participants detected for a given dance frame and their current pose.

### 5.3.2 Temporal Architecture

The temporal architecture utilizes the same three stacks but processes chunks of 16 frames at a time in order to incorporate a temporal component into the loss of the network. This enables us to learn motion parameters from the spatial, optical flow and multi-person pose representations. A visualization of our pipeline can be seen in Figure 8 whose convolutional and fully connected layers are based on the standard AlexNet architecture[13].

## 6. Baseline Experiments

We implement our proposed approaches with the goal of determining which approach is most effective at highly dynamic video classification. Implementation details for each approach are given below.

### 6.1. Dataset Splits

We extract individual frames from the Let’s Dance dataset (1000 10-second videos at 30fps, resulting in 300000 frames), which we then randomly split per video into 80% training, 10% testing and 10% validation (consistent across experiments). Optical flow and pose detection was split in the same manner in order to consistently test the approaches.

## Three-Stream Temporal CNN

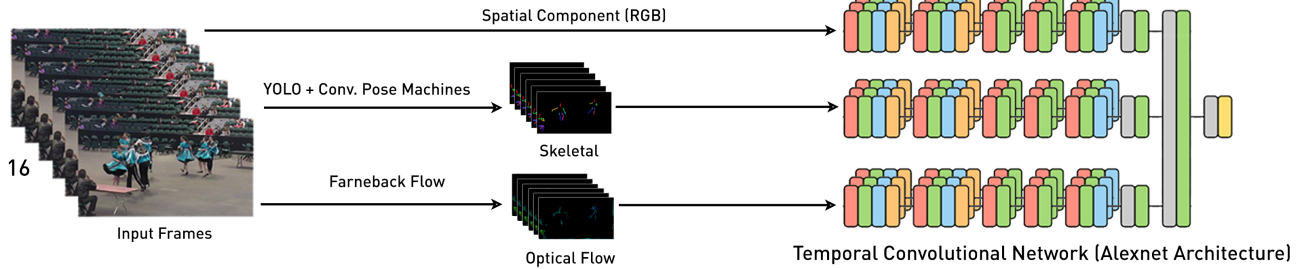


Figure 8. This visualizes the workflow for our three-stream temporal CNN which uses three convolutional stacks to process the spatial and respective motion components. It aggregates the fc7 layers into one and outputs the dance classification for a 16 frame input.

## 6.2. Frame-by-Frame

To perform a baseline video classification experiment, we implemented the architecture shown in Figure 4 in Tensorflow[1]. The weights for the network’s convolutional layers are initialized to values from a network pre-trained on the ILSVRC 2012 dataset [19]. Final video classification results can be determined by classifying each frame in a video and voting to determine the video’s overall class.

For an initial comparison, we also tested the network with optical flow imagery as the input.

Overall, we observed significant amounts of overfitting in the original training accuracy which hints at the network learning too much about the appearance of the specific videos in the training set for each class. As we hypothesized, using image frames alone results in the network learning features that do not generalize well to the dancing categories, since it has no way to observe the motion inherent in the video. Testing accuracy peaks at 56.4% over 10,000 iterations of fine-tuning the network. We compare these results to a similar framework introduced by [29] which tested the frame-by-frame baseline on UCF-101, attaining an accuracy of 72.8%. This directly demonstrates the possibility of solving the classification problem by carefully selecting the right frame versus understanding the underlying motion of the video.

We also ran the identical setup using optical flow estimation. Before training we pre-compute optical flow for the entire dataset. We used Farneback’s method for calculating dense optical flow [5] to obtain a per-pixel estimate of the horizontal and vertical components of motion and then incorporate this into the same network architecture.

In this case we saw slightly worse performance at approximately 45% for testing. We do note that the overfitting for optical flow images is subdued given that the per-frame images no longer contain background information. Given that a number of our dances occurred in similar or identical settings, background information was a strong confounding factor for the original images. The overall result for optical flow performs worse than training on RGB images given

that it is merely embedding the motion between two frames. We will later demonstrate that larger frame chunks provide significant improvements to this approach.

## 6.3. Two-Stream Late Fusion

We implemented the two-stream late fusion architecture shown in Figure 5 in Caffe[10]. The two-stream approach follows intuitively from the previous subsection in which we discuss the effects of both a frame-by-frame method on images and on optical flow. Each individual stream uses the CaffeNet architecture, with weights initialized to a network pre-trained on the ILSVRC 2012 dataset [19]. We then fine-tune the network by training only the fully-connected layers at the end of each stream, which are then concatenated and passed through a final fully-connected layer which outputs the respective classifications.

We note that each architecture in the two-stream method is still using a single frame as input, and as such the network is trained on a frame-by-frame basis. We chose to use the CaffeNet architecture for each frame, initialized with the ILSVRC 2012 weights, to be consistent with the baseline frame-by-frame experiment described in the previous section. This allows us to perform a direct comparison between the two-stream and frame-by-frame approaches, to determine the benefit of optical flow on this dataset. As with the frame-by-frame approach, final video classifications can be determined by classifying each individual frame and optical flow image pair, followed by voting to determine an overall class. It is interesting to note that the total per video classification accuracy of this method was 68.89% which is much higher than the single frame-by-frame accuracy of 56.40%. Although one may be compelled to argue that single-frame motion is key to this classification, we refer back to Figure 3. This figure demonstrates that the frames throughout the dataset also contain a tremendously varied number of participants. As we can further see in Figure 9, optical flow tends to visually separate the dancers from the background, which also explains the significant increase in the algorithm’s performance. In addition to the motion in a single frame pair,

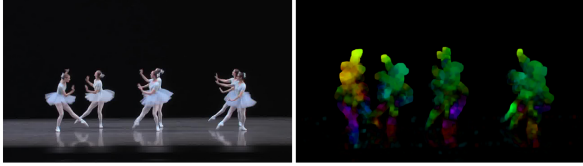


Figure 9. An image of dancers performing ballet and their optical flow estimation. As we can see, optical flow does a good job of segmenting the subjects in the scene in addition to encoding their motion.

the foreground’s shape and representation is playing a key role in the classification of the network.

The results demonstrate an improvement over each independent approach, with a classification accuracy of 64.69% per-frame. This is a significant increase of 10% above the imaging method and 20% over the optical flow method. This increase was attained by combining the same architecture as the previous two methods, with the addition of a single concatenation node to fuse the data at the end of the network. It demonstrates that directly incorporating temporal data into a network can be immediately beneficial towards classifying video.

Leveraging the network to perform full video classification (rather than only per-frame classification), we tested the trained network on our test set of videos, taking the class with the largest number of per-frame votes as the final video label. This resulted in a per-video classification accuracy of 66.14%. After further experimentation with the network architecture, we saw a significant improvement from computing a unique mean image to subtract from the optical flow, which increased our accuracy to a final per-video classification rate of 68.89%.

The network performs well at classifying Ballet, Waltz, Tango, Flamenco, and Foxtrot, with poor classification accuracy on Break and Swing dancing. Of particular interest is the network’s performance on Waltz, Tango and Foxtrot which occur in similar settings. As such, the network shows that it’s capable of performing fine-grained classification within the Let’s Dance dataset.

Dataset	Frame-by-Frame	Two-Stream
UCF-101 [22]	72.8%	88.0%
Let’s Dance	56.4%	68.89%

Table 1. Method Comparison of UCF-101 and Let’s Dance. UCF Frame-by-Frame results obtained from [29], Two-Stream results obtained from [20]

Lastly, we revisit our accuracy results with UCF-101, a well-established activity recognition dataset. Table 1 illustrates high levels of accuracy on UCF-101 using the standard extensions of image classification techniques which we discuss in this section. It is important to note that the

two-stream comparison is comparing a two-stream accuracy for UCF-101 that utilizes an SVM to combine its streams whereas we concatenate the final layers of both convolutional streams into the fully connected output. As stated earlier, this illustrates the core issue we encountered in looking for a highly dynamic dataset which further validates our motivation to introduce the “Lets Dance” dataset to the research community.

## 7. Results & Discussion

In order to assess our temporal architectures we compare with a number of state-of-the-art approaches that explicitly encode motion in order to determine their performance. Overall it has become clear to us that we need to transition from traditional per-frame CNN approaches when conducting video classification.

It is evident from Table 2 that methods which embed motion significantly outperform traditional methods and that metrics to evaluate these approaches are necessary in order to better understand what each network architecture is learning.

### 7.1. Temporal 3D CNN

In order to evaluate this approach we restructured our data into 16-frame chunks that were needed as the input for the 3D convolution. The network could be trained on the 3D features from 16-frame non-overlapping chunks of the video. We fine-tuned from the network trained on UCF101 by [25]. This method yielded a per-video accuracy of 70.11%. This result was particularly impressive because it demonstrated the inherent ability of a 3D convolution to extract motion features that are not explicitly computed. The major drawback of this approach is its complexity. A 3D convolution inherently takes significant computation for a single-stream.

We were unable to perform multi-stream approaches using 3D convolutions due to this complexity. In order to combat this we introduce more tractable approaches for state-of-the-art graphics cards (Our current systems utilizes Titan Z Pascal graphics cards) that achieve comparable performance by explicitly encoding motion into the network architecture. In addition to this we note that 3D convolutions are limited to the initial input-size which in our case was 16 frames. This makes it difficult to encode more complex motions that last more than 1/2 second without sub sampling frames which will invariably lead to a loss in detail. Most temporal methods will invariably suffer from this limitation given that variable inputs into a convolutional network has not been fully explored.

### 7.2. Skeletal Temporal 3D CNN

In order to embed human motion data, we incorporate skeletal images into a temporal CNN. We visualize each

Approach	Testing Accuracy
Frame-by Frame CNN	56.4%
Two-Stream CNN	68.89%
Temporal 3D CNN (RGB)	70.11%
Temporal 3D CNN (Skeletal)	57.14%
Three-Stream CNN	69.20%
Temporal Three-Stream CNN	<b>71.60%</b>

Table 2. Comparison of numerous approaches and their testing accuracies on our dataset

pose into a single image which represents the pose for that particular frame. We attained an accuracy of 57.14%. We note that this accuracy still performs marginally better than a frame-by-frame approach despite the fact that it does not utilize the spatial (RGB) representation. Due to the computational complexity of running concurrent 3D convolutional networks we propose a stacked 2D convolutional method which allows us to combine multiple streams in a single state-of-the-art graphics card.

### 7.3. Frame-by-Frame Three-Stream CNN

Our Three-Stream Frame-by-Frame architecture utilizes all three data modalities. We assess this as both a single-frame and as a stacked architecture in order to compare their benefits and drawbacks. As shown in Table 2, this approach attains an accuracy of 69.20%. This three-stream network performs comparably to the two-stream fusion approach we conducted as one of our baselines which indicates that there is not a significant amount of information added from the use of both skeletal and optical flow representations.

### 7.4. Temporal Three-Stream CNN

Logically, we extended our frame-by-frame approach into the temporal domain by stacking the image input layers to produce a 16-frame chunk. This approach utilizes the same input as the Temporal 3D CNN we implemented at a much lower complexity for three streams. We saw this method attain the best performance out of all of the methods we evaluated, at an accuracy of 71.60%.

Looking at our most successful approaches, three-stream methods and 3D convolution, we note that both achieve very similar performance in per-video classification. However, the two methods are not equivalent in terms of computational resources. Beyond the increased workload and restrictions inherent in appropriately formatting the data for the temporal CNN, 3D convolution is much more computationally-intensive at both training and testing time. We observe that even though the temporal CNN was our most successful approach, it may be sub-optimal when a much simpler three-stream stacked convolutional network approach is available.

## 8. Conclusion and Future Work

In this work we sought out to understand the effect of motion on classifying videos. Recent work in the are has demonstrated the relevance of these type of videos, most recently seen in [15] and [12]. The work we have conducted demonstrates that traditional CNN approaches do not properly or intentionally encode motion in their methodology. This fact is frequently overlooked by testing on videos that do not inherently require motion. That was the primary motivator of this work. As we can see in Table 2, 3D convolution methods outperform more traditional approaches by inherently encoding motion into their computation and prediction. Similarly, two-stream methods that incorporate optical flow can also leverage temporal features to significantly improve video classification.

This also opens up some potential for future work in incorporating optical flow and pose data. Hybrid approaches, such as a three-stream temporal CNN, have the potential to increase an algorithm’s understanding of the video. We have also developed a more focused dataset that we believe the research community will benefit from by intentionally selecting highly dynamic actions in one specific class. We tested a variety of traditional and more complex methods in order to begin to understand the composition of our dataset and its baseline performance. The Let’s Dance dataset will continue to help us to assess adequate motion parameterization and hopefully assist in improving how we learn from video data.

One of the biggest problems we ran into throughout this research endeavor was determining the best classes to select for our dataset. Initially we had some intuition for dancing and martial arts being adequate parent categories but we quickly saw that martial arts represented a multi-class problem. Although dancing exhibits similar overlaps the separation was much more evident when performing the data collection. We also had to alternate between different dances partly due to availability on YouTube and our own understanding of these dances.

One of the next steps we have considered in this work is modifying the input data in order to blur out regions of the video which are not in motion or considered background. This would further enforce motion parameterization and help us better understand how we can accomplish that to improve more general video classification algorithms. This could also be explored by independently classifying the pose estimation which is significantly more challenging.

We seek to extend this work by continuing to develop human-motion representations that intentionally target these highly dynamic actions.

For more information, please visit: <https://www.cc.gatech.edu/cpl/projects/dance/>



## References

- [1] M. Abadi and A. A. et. al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 6
- [2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. June 2017. 3, 4
- [3] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3218–3226, 2015. 2
- [4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015. 2
- [5] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image analysis*, pages 363–370. Springer, 2003. 3, 6
- [6] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. *arXiv preprint arXiv:1604.06573*, 2016. 1, 2
- [7] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3582–3589, 2014. 4
- [8] G. Gkioxari and J. Malik. Finding action tubes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 759–768, 2015. 2
- [9] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, Jan 2013. 2, 5
- [10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 6
- [11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2, 3
- [12] 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. 3, 8
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 4, 5
- [14] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016. 2
- [15] C. Pantofaru, C. Sun, C. Gu, C. Schmid, D. Ross, G. Toderici, J. Malik, R. Sukthankar, S. Vijayanarasimhan, S. Ricco, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. 2017. 3, 8
- [16] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3178–3185. IEEE, 2012. 4
- [17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2015. 4, 5
- [18] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 3
- [19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6
- [20] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014. 2, 7
- [21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4
- [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. 3, 7
- [23] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *CoRR*, abs/1502.04681, 2015. 2
- [24] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014. 2
- [25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *arXiv preprint arXiv:1412.0767*, 2014. 7
- [26] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4534–4542, 2015. 2
- [27] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo. 3d human activity recognition with reconfigurable convolutional neural networks. In *Proceedings of the 22Nd ACM International Conference on Multimedia, MM '14*, pages 97–106, New York, NY, USA, 2014. ACM. 2
- [28] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

- [29] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015. [6](#), [7](#)
- [30] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *arXiv preprint arXiv:1602.00134*, 2016. [2](#), [4](#), [5](#)
- [31] W. Zou, S. Zhu, K. Yu, and A. Y. Ng. Deep learning of invariant features via simulated fixations in video. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3203–3211. Curran Associates, Inc., 2012. [2](#)