Motivation & Objectives

- Recognition of human action in videos has generated substantial attention from the deep learning community in recent years [1-4].
- In this project, we aim to apply the video classification problem to detect tennis shots. We aim to build a deep neural network which takes in an RGB video and classifies it into one of 6 tennis shot classes.
- This work is also aimed at making progress toward a larger project that we are working on, where we envision a need to identify and correct player postures while playing various tennis shots.

Dataset

- The THETIS dataset [5] consists of 12 Tennis shots performed by 31 amateurs and 24 experienced players about 3 - 5 times/shot. 1980 R&B videos in total in .avi format, with about 80 frames per video.
- RGB video dataset split 0.8:0.1:0.1 into training (1584), validation (204) and test sets (192) with equal proportions of each shot.

Baseline and Bayes Error

- **Baseline LRCNN**: ‘LRCN’ model by Chow and Dibua [6]. To make their shot prediction, they extract features from video frames using the Inception V3 network pre-trained on ImageNet. Then, they feed the features into a many-to-many LSTM network, which is trained to output one of the 6 tennis shots as its prediction. Their results are shown in the first row of Results (baseline).

- **Bayes Error**: We asked 5 people with a good understanding of tennis to classify 24 videos into the 6 different shot categories. By using a voting method to combine their results, they obtained an 87.5% accuracy. Our Bayes error on these videos is thus 12.5%.

Models and Methods

- **Loss Function**: Categorical cross-entropy
- **Improved LRCNN Model**: We generated sequences for the THETIS RGB videos for 16 frames from Inception V3 network and used a Bi-directional LSTM, an extra hidden layer with 128 units, and an increased dropout rate of 50% (baseline dropout was 30%), the train and test accuracies and the F1 score are better than the baseline LRCNN. Results shown in Table in the next column.

Discussion and Error Analysis

- **Ensembling**: For the best results, we ensembled the outputs of Model 0 on the deep network and Model 1 on the shallow network. Results shown in the Table below.

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRCNN-Baseline</td>
<td>98.7%</td>
<td>82.3%</td>
<td>0.82</td>
</tr>
<tr>
<td>LRCNN-Improved</td>
<td>100.0%</td>
<td>84.4%</td>
<td>0.84</td>
</tr>
<tr>
<td>EvaNet+Transfer</td>
<td>99.8%</td>
<td>84.4%</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table: Results from our baseline, improved LRCNN and ensembled EvaNet+Transfer

Conclusion and Future Work

- On improved LRCNN and EvaNet+Transfer learning, we are able to identify 6 classes of shots with 84.4% accuracy.
- The cause for low accuracy is the shot quality. The large variance in the model performance is due to lack of sufficient data to train the model.
- Generating more quality data and training the model on it will help in removing the variance and improving accuracy, which is needed to further advance in predicting correctness of posture.

References