

Thesis: Final Report

Semantic Understanding of Table Tennis Video Clips

Zexuan Wang (998851773)

zexuan.wang@mail.utoronto.ca

Department of Engineering Science

University of Toronto

April 8th, 2016

Thesis supervisors:

Prof. Sanja Fidler and

Prof. Raquel Urtasun

Abstract

This thesis project demonstrates the possibility of using machine learning techniques to gain semantic understanding over table tennis games directly from its multimedia streamed video recordings, at the hit-by-hit level. In particular, it separates each point video clip into two temporal parts where the second part contains the last 10 image frames and the rest belongs to the first one. The first part is used to generate a sequence of player stroke annotations, alternating between the two players and the second part is used to classify on how the point ends and potentially who wins the point. For this project, a total of 178 table tennis point video clips over 8 different matches are collected and manually annotated for the training and testing datasets.

The machine learning task is performed using AlexNet [18] to select relevant features from the raw frame images and trains a CRF (Conditional Random Field) to model sequential relationship between each frame. Each input to the CRF is averaged over a sliding window of 10 consecutive frames images to minimize the label fluctuation over time. Using this methodology, we obtained a classification rate of 58.7% over 20 different classes for player stroke identification and 8 classes for the point-ending classification over our testing dataset, consisting of 35 table tennis point video clips.

Acknowledgement

I would like to thank Prof. Sanja Fidler and Prof. Raquel Urtasun for providing support and guidance throughout my thesis project as well as the Computer Science Department at University of Toronto for their computation servers.

Contents

1	Introduction	1
2	Background	2
2.1	Table Tennis Techniques	2
2.1.1	Forehand vs. Backhand	2
2.1.2	Stroke Classifications	3
2.1.3	End of a point	6
2.2	Sport Video Research	8
3	Training Data Preparation	10
3.1	Table Tennis Video Clip Collection	10
3.2	Data Label Creation Approach 1: Scoreboard Parsing	10
3.2.1	Scoreboard Parsing and Score Recognition	11
3.2.2	Training Set Creation	12
3.2.3	Problem with Above Approach	14
3.3	Data Label Creation Approach 2: Manual Labelling	15
3.3.1	Class Definition	15
3.3.2	Manual Labelling for Each Point	17
3.3.3	Training Data Preparation	17
3.4	Split between Training and Testing Datasets	19
4	Experimental Method and Results	20
4.1	Hit-level Classification	20

4.2 Point-ending Classification	22
4.3 Point-level Annotation Generation	22
4.4 Conditional Random Field (CRF)	24
4.5 Fine Tuning on AlexNet	25
5 Future Work	28
6 Conclusion	29
A Appendix	

1 Introduction

With the fast-paced development of Internet and multimedia service today, a large volume of high quality digital data become accessible to everyone where a large portion among them are sport game recordings. According to the study done in [1], in 2014, there are a total of 169 million sport fans in U.S who spend an average of 7.7 hours per week consuming sports content. On the other hand, in recent years, artificial intelligence and machine learning have been a major focus of active academic research. As a consequence, many computer science researches have been focusing on utilizing machine learning tools to create semantic understanding and / or automatic annotation of sport video clips.

Semantic understanding in sport video is important for a number of different reasons. For athletes, highlight videos serve as a good resume for them to impress coaches during recruit season. For coaches, tactic or strategy analysis allows them to discover the players' strengths and weaknesses and thus prepare for future training plans. For sport fans, highlight videos enable them to experience excitement from the game in a more efficient way and only watch the moments that they are interested in.

Table tennis is a racquet sport with a well-defined temporal structure. Each match is comprised of multiple games, usually best-3-out-of-5 or best-4-out-of-7. Each game is then separated into points, and the player who is able to get 11 points first wins the game. Within each point, two opposing players hit the ball alternatively. A rally is defined as the ball going back and forth once over the table. The player who is unable to return the ball on the opponent's side of the table loses the point. Moreover, table tennis also contains a clear definition of different player techniques with minimal ambiguity, ranging from backhand to forehand, from serve to loop. The temporal structure of table tennis and its players' clear stroke classification make table tennis an ideal environment for semantic understanding computer science researches. Furthermore, the methodology proposed in this table tennis research can be easily adapted to other racquet sports such as badminton and tennis.

The objective of this research is to apply different machine learning and computer vision techniques to create semantic understanding of table tennis videos. It investigates the possibility of using deep neural networks to differentiate different strokes of table tennis, such as serve, chop, loop, flick etc. It also explores the opportunity to learn the rule of table tennis and identify which player wins the point given a new point video clip, previously unseen to the model. In particular, this research will be conducted in the following high-level steps:

1. Collect table tennis video clips from YouTube, the popular multimedia service.
2. Split the table tennis videos into points and annotate them at the hit-by-hit level. The annotation can be either created manually and / or auto-generated using information from the video clip. An example source

of obtaining the useful information is the scoreboard at the bottom left corner of each frame image, OCR (Optical Character Recognition) can be performed to recognize the game score at any instant.

3. Split the annotated data above into training dataset and testing dataset. In particular, two different types of testing dataset is interesting for this project: 1) Testing data from the same game as the training data (i.e new points but players have been seen to the machine learning model); 2) Testing data from a new game where the machine learning model has never seen the players. Most practical applications should focus on the later type of testing data.
4. Using the labeled training and testing data prepared above, investigate the possibility of using different machine learning models to recognize each player stroke into different classes such as: serve, push, flip and loop, each then grouped into forehands and backhands as described in [2]. It will also explore the opportunity to utilize machine learning models to "learn" the rule of table tennis and having the model to recognize the winner given a new table tennis point video clip.

More details on exactly how each step described above is performed can be found in Section 4.

2 Background

2.1 Table Tennis Techniques

This section aims to provide some background information on different table tennis stroke types which can help understand the rest of this report. A player performs one particular stroke technique every time he / she hits the ball with the racket. In particular, this section explains the list of different table tennis strokes, classified similar to that of [2] with modifications based on the author's domain knowledge in table tennis.

2.1.1 Forehand vs. Backhand

Each table tennis hit can be classified into two broad classes: forehand and backhand. Forehand, by definition, is the stroke played with the palm of the hand facing in the direction of the stroke whereas backhand is the stroke where the back of the hand is facing in the direction of the stroke. In table tennis, particularly, they are usually differentiated by having rubbers of two different colors, one side having red and the other having black. A forehand shot is then differentiated from a backhand shot using the rubber color that has physical

contact with the ball. Figure 1 shows two sequences of frames from the video collected where the top one demonstrates an example of forehand shot and the bottom one demonstrates an example of backhand shot.



Figure 1: Demonstration of a forehand shot (top) vs. a backhand shot (bottom)

2.1.2 Stroke Classifications

Besides forehand and backhand, each table tennis stroke can be classified into one of the following five categories, as explained below:

1. Serve

Serve is the start of any point. Different from any other shots, for a legal service, one must ensure that the ball bounces on your side of the table first before landing on the opponent side. Figure 2 provides an example service represented as a sequence of image frames from one of the game videos collected.



Figure 2: Demonstration of a forehand serve

2. Loop

Loop, sometimes also referred as drive, is one of the most fundamental table tennis techniques where the player strokes from bottom upwards to lift the ball and put top-spin on that. Loop is usually referring to an offensive technique where the player either excercises spin and / or power on the ball. Figure 3 shows a sequence of images to demonstrate a forehand loop.



Figure 3: Demonstration of a forehand loop

3. Block

Block is a table tennis stroke that is similar to loop but a defensive strategy. It is usually used when your opponent performs an aggressive loop and you simply wants to get the ball back on the opponent's table. A block usually does not require the player to generate spin / power but utilizes power from your opponent instead. Figure 4 below illustrates a forehand block.



Figure 4: Demonstration of a forehand block

4. Flip

Flip, many times also referred as flick, is an offensive technique similar to a loop but the incoming ball is short or inside the table. The player usually has to move forward so that his / her body is on top of the table and flip generally requires the player to have a smaller action than a loop due to the spatial constraints created by the table. Figure 5 shows an example of a forehand flip performed by a left-handed player.



Figure 5: Demonstration of a forehand flip

5. Chop

Chop, often times referred as push, is different from all the techniques above where the player opens the racket such that the rubber is facing upwards and the stroke starts up high and goes downwards. It is intended to generate backspin on the ball causing the opponent harder to loop. Figure 6 shows the sequence of frames of a forehand chop.



Figure 6: Demonstration of a forehand chop

2.1.3 End of a point

As it is described above, a point ends when one player fails to return the ball to the opponent's side of the table. This can happen when one of the following scenarios takes place:

1. Player fails to hit the ball

In this scenario, the player fails to hit the ball with his / her racket. This is usually caused by the player being in a bad position and unable to reach the ball or the player made a wrong prediction on where the ball will land. Figure 7 shows a sequence of images where the player first predicts the ball coming to his backhand side and fails to hit the ball when the ball actually comes to his forehand side, thus losing the point.



Figure 7: Demonstration of a forehand miss hit

2. Player fails to return the ball over the net

In this case, the player reaches the ball but fails to hit the ball over the net, often times, caused by underestimation of the underspin on the ball. Figure 8 shows a sequence of images where the player attempts to perform a backhand loop against a backspin ball but unfortunately hits under the net and thus loses the point.



Figure 8: Demonstration of a player fails to return the ball over the net

3. Players hits the ball but it flies out of the table

This is the scenario where the player hits the ball but the ball does not land on the opponent's table and flies out of it instead. Figure 9 shows an example where the top player attempts to perform a backhand loop, but the ball unfortunately flies over the opponent table and the top player loses the point.



Figure 9: Demonstration of a player hits a ball but it does not land on the opponent's table

2.2 Sport Video Research

Prior to this research, a number of studies have been conducted in the sport video semantic understanding area. Generally speaking, most of previous researches can be classified into one of the following categories:

1. Event Detection

Studies in this category include detecting key events various sports given the sport video and / or audio recording. In soccer games, for example, people have been trying to identify special events such as goals [4], fouls, injuries etc [3]. In racquet sport like tennis, effort has been put into identifying particular events like end of a serve [5] or ball hit [6][7].

2. Sport video highlight

Research in this area includes extracting key excitement or important moment from a given sport video clip. For example, [8] is a previous research that also chooses table tennis as the target sport. It uses hidden Markov Model (HMM) on audio track and un-supervised shot clustering to extract individual points from a full table tennis match video. They claim to achieve 90% point-extraction accuracy based on the 7 table tennis matches they collected, all taken from television streaming. [9] applies a similar approach with specific baseball features and they claim that the highlight video created by their model overlaps with the ones created by human 70% of the time.

3. Player location / action recognition

Research in this area ranges from player stroke detection in tennis like [11] [12] or player location identification in a basketball game like [13]. [14] proposes an interesting method to extract the location of the player and the ball in a table tennis game. They propose to take the difference of consecutive frames images in a video to extract approximate positions of the players (since everything else in the video is stationary, including the table and background). In order to extract the ball trajectory, the approximate table location is first identified by an rectangle region filled with blue colour, and the ball is then extracted as a white circle within that region.

4. Game strategy / tactic analysis

This group of study analyzes what tactic the player is trying to perform and potentially evaluate their effectiveness. Research in this area usually involves collecting sport video under a highly monitored environment and / or manually extract features from a sport video and apply study directly on those features instead of raw sport video clip. For example, [10] records a table tennis game from several different camera angles to collect data including speed, direction and player. It then uses those information to predict whether a hit will go on the table or not. [2] takes the step even further, it manually extracts the stroke type out from a table tennis video and apply neural network on the sequences of strokes to classify

the tactic that the player is trying to perform and the likelihood that this particular hit will gain the player a point.

Different from the above researches mentioned, this thesis project aims to gain semantic understanding of a table tennis game at the hit-by-hit level similar to other researches in the Game strategy / tactic analysis category but directly from recorded table tennis sport video clips without any manual preprocessing or feature extraction. In particular, it has two clearly defined objectives: 1) It attempts to classify each player hit into a pre-defined set of stroke classes such as serve, chop, loop etc. and automatically create a sequence of such annotations for each point; 2) It analyzes the ending of each point and classifies it into a list of point-ending classes, which teaches the model to understand the rule of table tennis and thus infer the winner of the point. To the author's best knowledge, it is the first study that creates annotation at such a detailed hit-by-hit level directly from streamed raw table tennis video clips.

The potential engineering application from this research is pretty straightforward. For players and coaches, auto-annotation at the hit-by-hit level allows automatic game statistic generation. For example, statistics such as "Percentage of point losses due to backhand blocks" or "Percentage of serves that the opponent is unable to return" are definitely useful data to help prepare players and coaches better understand the player's strengths and weaknesses, thus prepare for further trainings. Those game statistics will also be interesting for both the commentators and the audience just like those in basketball games. Another application of this research for audience would be selective game replay. It allows table tennis fans to watch only points that are interesting to them where they have the choice to watch points such as "Show me points with more than 20 rallies" or "Show me points where Player X serves and Player Y could not return". Furthermore, if the point-ending classifier is able to achieve a resonable accuracy, auto-genereation of the scoreboard that keeps track the player score would then be possible. Finally, the most useful application of this research would be an "auto-referee" robot that is able to keep track the score of a table tennis match on its own. Although referees play an important role in big international tournaments, it should also be noted that for small local tournaments or preliminary rounds of some tournaments, the organizer often do not have enough budget to hire dedicated referees. Combine this research with the point extraction research previously [8], a "referee robot" composed of a single computer with a camera will then become possible.

3 Training Data Preparation

3.1 Table Tennis Video Clip Collection

In order to create the machine learning model which understands the rule of table tennis, a training set of table tennis video clips need to be collected and properly labelled. In the following two sections, we will describe the exact steps on how such training set was created.

YouTube is a popular multimedia service that hosts the resource for many table tennis video clips. For the purpose of this research, 20 videos from the recent 2015 Swedish Open [15] are collected and they all share the following properties:

1. The videos are recorded from a recent tournament (in November 2015) with good quality. Each frame has a resolution of 1280 * 720 sampled at 24 frames per second. This high quality ensures the maximum amount of information captured in our input data.
2. Every video has a scoreboard overlayed at the bottom left corner, which we can parse out and extract additional information which may help us create the training set.
3. Every video is a highlighted version of a table tennis match such that the points are concatenated together. The time where players are picking up the balls and taking time-out have already been cut off. This not only helps us to save our storage space storing those high-resolution videos.
4. Videos from this tournament have minimal amount of shooting angle changes. This makes our learning easier as we do not have to accommodate with the change in camera position.
5. Videos from this tournament contain good audio track from the game, with minimal interference from the commentator and audience. This allows us to detect the sound when the racket hits the ball as well as the sound when the ball hits the table. Those audio clues may become useful if we ever decide to include them as features into our learning task.

The exact script (crawler.py) that was used to parse out the videos can be found from the GitHub repository:
<https://github.com/tabletennisr/thesis>

3.2 Data Label Creation Approach 1: Scoreboard Parsing

This section describes our initial approach to create the data labels for our training dataset, namely, by performing an OCR (Optical Character Recognition) on the scoreboard for every frame in the table tennis game

video. It also explains the problem with this approach and why it was not successful.

3.2.1 Scoreboard Parsing and Score Recognition

After the .mp4 videos are downloaded locally, a program called ffmpeg [16] is used to sample the videos into frames. For this particular research, the videos are sampled at 4 frames per second so that the beginning / end of points detection have an uncertainty of 0.25 seconds.

Along with each input video, a coordinate location at the top left corner of the scoreboard as well as the size of each score (in units of pixels) are also included. Figure 10 shows a sample frame image in the video as well as the four score images (player 1 game score, player 1 set score, player 2 game score and player 2 set score) cropped out. In order to better facilitate OCR (Optical Character Recognition) on the score images in the next step, every pixel is transformed to grey-scale, normalized and scaled down to 20*20 to match the resolution required for OCR.



Figure 10: Example of a frame image and the scores parsed out from the scoreboard

In order to convert the score image from a picture to an integer so that it can be understood by our Python

program. Google’s Tesseract OCR [17] is first attempted to recognize each score image, with settings set to recognize digits only and assuming a single uniform block of text. Although tesseract OCR has a good performance on recognizing hand-written digits, it actually has lots of trouble recognizing single digit and is really sensitive to small fluctuations in the image, like the brightness, location of the digit etc. Based on a manually created testing set using images parsed out from the table tennis game video (https://github.com/tabletennisr/thesis/tree/master/testing_set), it is tested that Tesseract OCR achieves approximately 80% of accuracy.

This percentage of accuracy leaves us a lot of noise if we were to create our table tennis video clip training set based on it, whose error will further propagate when we train our machine learning model to learn which player is winning given a table tennis point video clip. As a consequence, we decide to train our own single-digit based OCR to achieve better result. The training set for our score OCR is created by using a single image for each class (class includes digit 1, 2, 3 … 15, as it is unlikely for a table tennis game score to go beyond 15), shift and scale it into 1758 different locations. Figure 11 shows some of the images in the score recognition OCR training set for the digit 4. This method of digit OCR achieves 100% accuracy on the manually created testing set above. (https://github.com/tabletennisr/thesis/tree/master/testing_set) Given any new table tennis video clip as well as the approximate location of the scoreboard, a simple 3-NN (3-Nearest Neighbour) is then performed to find the closest class of the scoreboard cropped out, and the image can be converted into the corresponding integer represented by that particular class.



Figure 11: The same image of digit 4 is used to create training examples in our OCR training set after some shift and scale

3.2.2 Training Set Creation

Now that we have a reliable way to parse out the number on the scoreboard of any table tennis video clip, we can then create our training set according to the following simplified high-level algorithm:

Algorithm 1 Algorithm for creating the training set

```
procedure CREATE-TRAINING-SET(frames)
    prevScoreP1 ← 0
    prevScoreP2 ← 0
    startFrameNum ← 0
    P1WinningFrames ← {}
    P2WinningFrames ← {}

    for i = 0 to frames.length do
        curScoreP1 ← 3-NN OCR on P1 score region
        curScoreP2 ← 3-NN OCR on P2 score region
        if (curScoreP1.confidence < 1 or curScoreP2.confidence < 1) then
            startFrameNum ← i + 1
        else if (curScoreP1 != prevScoreP1 or curScoreP2 != prevScoreP2) then
            if (curScoreP1 == prevScoreP1 and curScoreP2 > prevScoreP2) then
                P2WinningFrames.append(Frames[pointStartFrameNum:i])
            else if (curScoreP1 > prevScoreP1 and curScoreP2 == prevScoreP2) then
                P1WinningFrames.append(Frames[pointStartFrameNum:i])
            end if
            startFrameNum ← i + 1
            prevScoreP1 ← curScoreP1
            prevScoreP2 ← curScoreP2
        end if
    end for
    return P1WinningFrames, P2WinningFrames
end procedure
```

Apply the above algorithm on the 20 videos from the recent 2015 Swedish Open [15], we have created a training set of 224 points where the bottom player is winning and a training set of 256 points where the top player is winning.



Figure 12: Left: Original image frame from the video;
Right: Cropped image prepared for our training set

Since the purpose of this research is to utilize machine learning tools to learn the rule of table tennis, any irrelevant information from the video should be cropped out, including the scoreboard and the umpire. Every image in the final training set has a resolution of 600*600 and is cropped out from the top middle part of the original 1280*720 frame. Figure 12 shows an example frame image before and after this preprocess. The left image shows a scenario where the umpire (on the left) puts up his left hand to indicate that the top player has won a point and the right picture shows the image prepared for our training to exclude such information.

3.2.3 Problem with Above Approach

The biggest problem with the approach above comes from noise in the training set created. Since the input videos are manually prepared by someone from ITTF (International Table Tennis Federation), the video creator sometimes leaves a few extra frames for some points comparing to others, and this creates noise or misleading information in data set. For example, left image in figure 13 is the 5th last image of a point in our training data set. It provides useful information to our machine learning model as it can be seen from the image, the top player fails to hit the ball over the net and the bottom player wins the point. On the other hand, right image of figure 4 is the 5th last image taken from another point in our training data set. It is obvious that this image will provide no information to our machine learning model since it is taken long after the point has really finished and the players have already walked away from the table to pick up the balls. Therefore, an immediate next step that should be taken is to get rid of frame images like the right image in Figure 13 from our data set. In order to achieve this, an algorithm that can automatically detect the exact end of point is required.



Figure 13: Left: 5th last image of a point that can provide useful information;
Right: 5th last image of a point that creates noise in our training data

This task of detecting the true ending time of a point is a very difficult machine learning problem itself. As it was discussed above in the background section, [8] uses hidden Markov Model (HMM) on audio track and un-supervised shot clustering to extract individual points from a full table tennis match video which claim to have 90% accuracy. However, a separate training dataset is required for that, which would take lots of extra effort to generate. Moreover, 10% inaccuracy is not ideal in the training dataset and would create un-necessary noise to our main research task: create semantic understanding of table tennis point video clips. Therefore, we decide to perform manual labeling to create the training dataset, as described below.

3.3 Data Label Creation Approach 2: Manual Labelling

This section describes the procedure on how we annotate our training dataset through manual labelling.

3.3.1 Class Definition

1. Hit-level Classes

As it was described in Section 2.1.2, each player stroke can be assigned to one of the 5 high-level types of techniques: serve, loop, block, flip and chop, where each technique can then be separated into forehands and backhands, leaving us 10 total classes for each player. Since the input to our machine learning model is defined to be directly taken from streaming table tennis videos with no pre-processing, both players will be present in any given frame, similar to the right image in Figure 12. This provides us with the following 20 different hit-level classes:

C1_TOP_PLAYER_FOREHAND_SERVE (See Figure 2)

C2_TOP_PLAYER_BACKHAND_SERVE
C3_BOTTOM_PLAYER_FOREHAND_SERVE
C4_BOTTOM_PLAYER_BACKHAND_SERVE
C5_TOP_PLAYER_FOREHAND_LOOP (See Figure 3)
C6_TOP_PLAYER_BACKHAND_LOOP (See bottom image of Figure 1)
C7_BOTTOM_PLAYER_FOREHAND_LOOP
C8_BOTTOM_PLAYER_BACKHAND_LOOP
C9_TOP_PLAYER_FOREHAND_BLOCK (See Figure 4)
C10_TOP_PLAYER_BACKHAND_BLOCK
C11_BOTTOM_PLAYER_FOREHAND_BLOCK
C12_BOTTOM_PLAYER_BACKHAND_BLOCK
C13_TOP_PLAYER_FOREHAND_FLIP (See Figure 5)
C14_TOP_PLAYER_BACKHAND_FLIP
C15_BOTTOM_PLAYER_FOREHAND_FLIP
C16_BOTTOM_PLAYER_BACKHAND_FLIP
C17_TOP_PLAYER_FOREHAND_CHOP (See Figure 6)
C18_TOP_PLAYER_BACKHAND_CHOP
C19_BOTTOM_PLAYER_FOREHAND_CHOP
C20_BOTTOM_PLAYER_BACKHAND_CHOP

2. Point-ending Classes

As it was discussed in Section 2.1.3, a table tennis point terminates either when a player fails to hit the ball or he / she fails to return the ball over the net or he / she hits the ball flying out of the table. Similar to the player stroke classes, all of these scenarios can apply to both the top player and the bottom player and we thus have the following classes for the end of a point:

C21_TOP_PLAYER_UNDER_NET (See Figure 8)
C22_TOP_PLAYER_HIT_OUT (See Figure 9)
C23_TOP_PLAYER_FOREHAND_MISS_HIT (See Figure 7)
C24_TOP_PLAYER_BACKHAND_MISS_HIT
C25_BOTTOM_PLAYER_UNDER_NET
C26_BOTTOM_PLAYER_HIT_OUT
C27_BOTTOM_PLAYER_FOREHAND_MISS_HIT
C28_BOTTOM_PLAYER_BACKHAND_MISS_HIT

3.3.2 Manual Labelling for Each Point

Given the player stroke classes and the point-ending classes defined above, we can now label our input training data accordingly. To avoid the problem described in Section 3.2.3, the starting time of each point is strictly defined to be the frame where the player starts to toss the ball and the ending time of each point is strictly defined to be 3-5 frames after the winner can be clearly identified by the person who performs the labelling. Each frame in this point duration will then be assigned to one of the player stroke classes listed in Section 3.3.1.1 with no gap in between except for the last 10 frames (correspond to the last 0.4 seconds of the point given a frame rate of 24fps) where one of the point-ending classes in Section 3.3.1.2 will be assigned upon. Figure 14 shows a visualization of an example labelled point where the x-axis is the frame number and the y-axis is the corresponding class ID. Figure 15 shows the original frame image at varies frame number throughout the point.

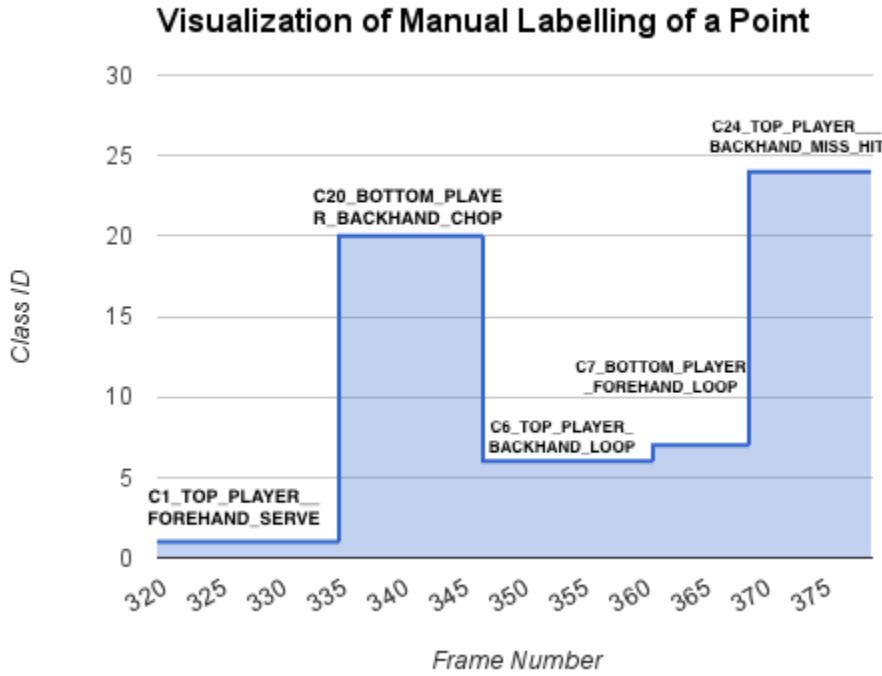


Figure 14: Point labelling visualization

3.3.3 Training Data Preparation

Given all table tennis point video clip data have been collected and properly labelled, two classifiers, one for the stroke class identification and the other for the point-ending classes are needed. Since each input



Figure 15: Original frame images for the point above
From left to right:

frame 327 (C1_TOP_PLAYER_FOREHAND_SERVE),
frame 340 (C20_BOTTOM_PLAYER_BACKHAND_CHOP),
frame 353 (C6_TOP_PLAYER_BACKHAND_LOOP),
frame 364 (C7_BOTTOM_PLAYER_FOREHAND_LOOP) and
frame 373 (C24_TOP_PLAYER_BACKHAND_MISS_HIT)

frame image has a very high dimension ($1280 * 720$), some feature extraction algorithm is needed. The feature extraction algorithm should not only select interesting features relevant to our training task but also reduce the input dimensionality so that the training can be performed on data with manageable size. For this project, we use the well-known AlexNet [18] to extract features from raw frame image. In particular, the normalized 4096-size vector from the second-last fully connected layer (see Figure 16) is taken as the input feature to our model as it empirically generates the best performance for our classification task.

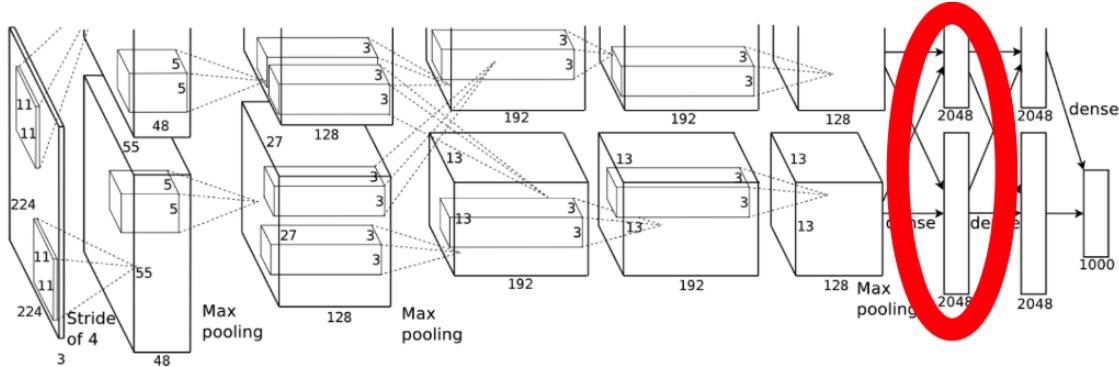


Figure 16: Famous Alex net structure, normalized vector from the second-last fully connected layer (circled in red) is used as feature vector

The pre-trained model of AlexNet is available as part of caffe [19], the open-source machine learning framework. The exact steps on how the training data is created is described in details below:

1. All sequences of frames of the same class are grouped together.
2. Choose 10 consecutive frame images with a running window of 5 frames (For example, if a particular C1_TOP_PLAYER_FOREHAND_SERVE contains 20 frames, then three sets of consecutive frame images will be used: the first 10 frames, frame No. 5 to frame No. 15 and the last 10 frames) This will correspond

to three different training data points in our classifiers.

3. Each image is cropped such that only the middle top $600 * 600$ pixels is remained. This is later reduced to $224 * 224$ and fed as input to the AlexNet above, the normalized vector from the second last fully connected layer of length 4096 is then averaged over 10 consecutive frames and considered as one properly prepared input data point.
4. The 4096 dimensional vector is considered the "selected feature" for the chosen sequence of 10 frame images. Simple classifiers such as SVM (Support Vector Machine) or Logistic Regression is trained directly on the 4096 dimensional vector with output being either the 20 player stroke classes or the 8 point-ending classes.

The experimental result for both the player stroke recognition as well as the point-ending classifications using the above procedure are explained in details in Section 4.

3.4 Split between Training and Testing Datasets

Using the manual labelling method described above, a total of 183 points from 8 matches in the 2015 Sweden Open have been collected, which corresponds to 1138 data points for player stroke identification and 227 data points for point-ending classification. The data collected is split into the following three different groups:

1. Training data constituted of 148 points from the first 7 matches, all data are collected from the first two games of each match.
2. Testing data constituted of 15 points from match No.7 where data are collected from the third and fourth game of that match.
3. Testing data constituted of 20 points from match No. 8. It should be noted that players in this testing data are new to our model.

Practically speaking, from the application's point of view, testing data from an entire new table tennis match should be desired as we cannot possibly train our model on every player. But it is interesting to compare results between test data of the same match but different points (i.e group No. 2 above) against test data of new table tennis matches (i.e group No. 3 above) to see how well our machine learning model generalizes.

4 Experimental Method and Results

In this section, we will describe how we utilize the dataset prepared above to perform our targeted machine learning tasks: Given a raw table tennis point video clip from streamed multimedia service, automatically generate annotation to not only identify the sequence of player strokes but also analyzes how a point ends and thus recognize the winner.

4.1 Hit-level Classification

Given the data preparation steps from 3.3.3, we now have a dataset where each input is a normalized continuous 4096 dimensional vector and each output is either one of the 20 player stoke classes labels or one of the 8 point-ending classes depending on the classification task. This establishes a very standard classification problem and a number of classifiers have been attempted and the raw experimental result is summarized below:

Classifier	Cross Validation Result	Testset Classification Rate
KNN (k=7)	63.3%	50.4%
Gaussian Naive Bayes	57.4%	38.8%
Logistic Regression	76.6%	57.6%
SVM (linear kernel with C=0.01)	78.2%	59.2%

Note that the hyper-parameters for the classifiers are selected based on the result from 4-fold cross validation, the original cross validation result can be found in Appendix A. Moreover, for all classifiers, the cross validation performance is way above the classification rate on the test dataset. This may attribute to the fact that for cross validation, every data point from the validation set has some corresponding data points in the training set coming from the same match (and thus same players) whereas for the actual test dataset, the input data points come from a new game and our machine learning model has never seen those players during the training phase. This is an indication that our model does not scale very well to new matches and new players. This may be improved by having a larger dataset and / or fine tuning AlexNet using our dataset, which will be discussed in the future work section.

Figure 17 shows the confusion matrix of the best classifier (i.e SVM with linear kernal and penalty parameter $C = 0.01$), each row corresponds the correct label where each column represents the player stroke class predicted by our model. So, the number at i-th row and j-th column represents the number of observations known to be in class i but our model predicts it to be in group j. The mapping from matrix indices to the actual player stroke label is also shown in Figure 17. Some intersting observations can be made from the confusion matrix:

```

[[ 1  0  0  0  0  0  0  2  0  1  0  0  0  2  2  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0]
 [ 0  0  1  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0]
 [ 5  0  0  0  0  0  0  6  1  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  0  2  0  0  0  0  0  0  0  0]
 [ 2  0  0  0  2  1  0  7  4  2  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  0  0  3  0  0  0  5  0  0  0  0  0  4  0  0]
 [ 0  0  0  0  0  0  0  0  0  5  39 0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  1  1  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0  59 0  0  4  0  0]
 [ 0  0  0  0  0  0  0  4  0  4  0  0  0  14 0  2  0  4]
 [ 1  0  0  0  1  0  0  0  4  3  0  0  0  7  2  1  1  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  1  0  2  0  14 1  0]
 [ 0  0  0  0  0  0  0  0  0  3  0  0  0  2  0  2  4  0]
 [ 0  0  0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  0]]]

0:'C10_TOP_PLAYER_BACKHAND_BLOCK',
1:'C11_BOTTOM_PLAYER_FOREHAND_BLOCK',
2:'C12_BOTTOM_PLAYER_BACKHAND_BLOCK',
3:'C13_TOP_PLAYER_FOREHAND_FLIP',
4:'C14_TOP_PLAYER_BACKHAND_FLIP',
5:'C15_BOTTOM_PLAYER_FOREHAND_FLIP',
6:'C16_BOTTOM_PLAYER_BACKHAND_FLIP',
7:'C17_TOP_PLAYER_FOREHAND_CHOP',
8:'C18_TOP_PLAYER_BACKHAND_CHOP',
9:'C19_BOTTOM_PLAYER_FOREHAND_CHOP',
10:'C1_TOP_PLAYER_FOREHAND_SERVE',
11:'C20_BOTTOM_PLAYER_BACKHAND_CHOP',
12:'C3_BOTTOM_PLAYER_FOREHAND_SERVE',
13:'C5_TOP_PLAYER_FOREHAND_LOOP',
14:'C6_TOP_PLAYER_BACKHAND_LOOP',
15:'C7_BOTTOM_PLAYER_FOREHAND_LOOP',
16:'C8_BOTTOM_PLAYER_BACKHAND_LOOP',
17:'C9_TOP_PLAYER_FOREHAND_BLOCK',

```

Figure 17: Top: Confusion matrix for the player stroke classification
Bottom: Corresponding player stroke class label for each index

1. Our model is very confident to identify player SERVE actions. Out of the 59 identified C3.BOTTOM_PLAYER_FOREHAND_SERVE data points, all of them are correct. Out of the 40 identified C1.TOP_PLAYER_FOREHAND_SERVE data points, only one is in-correctly labelled.
2. Our model is weak in identifying BLOCK actions. For example, out of the 8 C10.TOP_PLAYER_BACKHAND_BLOCK data points, our model only gets one of them correct, for the other three data points, it recognizes the bottom player's action instead. This may attribute to the fact that a player's BLOCK action may not be as intensive as other actions such as LOOP and the model seems to focus more on the other player instead.

4.2 Point-ending Classification

Similarly to the hit-level classification task, the exact same procedure can be applied on the point-ending datasets created. The raw empirical result is shown below:

Classifier	Cross Validation Result	Testset Classification Rate
KNN (k=3)	25.7%	13.0%
Gaussian Naive Bayes	39.6%	13.0%
Logistic Regression	47.6%	8.7%
SVM (linear kernel with C=1.0)	61.3%	8.7%

Note that the classification rate for the point-ending classifier is a lot lower than that from player stroke classification, especially for the testset classification rate. In fact, the testset classification is merely around random guessing among the 8 different point-ending classes, indicating that our point-ending classifier is not scalable at all. This may attribute to the following reasons:

1. The training and testing data set for point-ending classifier is 5-6 times smaller than that of player stroke classification data set. This is caused by the nature of table tennis where for every table tennis point labelled, there is only one point-ending data point created whereas 5-6 player stroke classification data points are generated depending on how many times the ball gets hit back and forth.
2. Unlike player stroke recognition, all point-ending classes are very much dependent on ball instead of player's action. Since the ball is an object that is a lot smaller than the player themselves. Without special training, it is difficult for AlexNet to pick up that feature.
3. Unlike player strokes where each stroke lasts a duration of time ranging between 10 to 20 frames, some point-ending classes can happen during a very short period of time, within a frame or two. For example, whether a ball lands on the table or not should be determined between only two frames and all other frames in the dataset serve as noise or useless info. As a result, a model where an event is emitted only at certain frames and no event for most of the duration might be a better representation for this problem.

4.3 Point-level Annotation Generation

Combine the two classifiers above, we are now able to automatically generate annotations for any given table tennis point video clip using the following sliding window algorithm:

1. Given any table tennis point video clip, decompose it into image frames using ffmpeg [16], pass each

image through AlexNet to obtain the 4096 dimensional normalized vector from the second-last fully connected layer (sometimes also referred as 'fc6').

2. For any feature vector at a given time frame, average it with four frames before and five images after. The average of those 10 vectors is used as the input vector for this time instant. If the vector is at the beginning of a point and does not have four frames in front of it, then an average of as many as frame vectors in front of it is used to create the input data. Same situation applies if the vector is at the end of a point and does not have five frames after it.
3. If the frame vector has more than 10 frames after it until the end of the point, then player stroke classifier from Section 4.1 is used to create a label annotation at this frame instant. Otherwise, the point-ending classifier from section 4.2 is used.

The performance on the quality of the point-level annotation generated above is benchmarked using the accuracy of the labels it creates, defined as:

$$\text{Annotation Accuracy} = \frac{\text{Number of frames whose auto-generated label matches the manual annotation}}{\text{Total number of frame images in the video}} \quad (1)$$

It should be noted, however, that this method of benchmarking does not penalize wrong labelling of short player stroke as much as the long ones. For example, if a point consists of a sequence of player strokes starting with 'Top Player Serve' which lasts half of the time of the point duration and the rest 4 strokes share the second half. Let's say our classifier is really good at recognizing player's serves and get the 'Top Player Serve' event correct and everything else wrong, then the our annotation accuracy will be 50% although it may be arguable that 20% classification rate makes more sense since the classifier only gets one correct label out of five. The alternative benchmarking technique can be implemented as

$$\text{Error rate} = \frac{\text{Minimum edit distance between auto-generated label sequence and manual annotation}}{\text{Total number of strokes in the manual annotation}} \quad (2)$$

The former accuracy-based frame-level benchmarking technique is used for this project.

Result from the sliding window approach above is summarized in the table below:

Testing Dataset	Annotation Accuracy
15 new points from an old match	47.8%
20 new points from a new match	46.1%
Weighted average of the above	46.8%

This is a reasonably good result considering there are 20 different stroke recognition classes and 8 different point-ending types. Moreover, as it can be seen from the result, the result is pretty consistent between points from a seen match against that of a new match. This indicates that our model is relatively scalable and adapt to new players quite well. Figure 18 shows the annotation accuracy for each table tennis video clip and as it can be seen, it ranges from 23% to 75% depending on the exact video content.

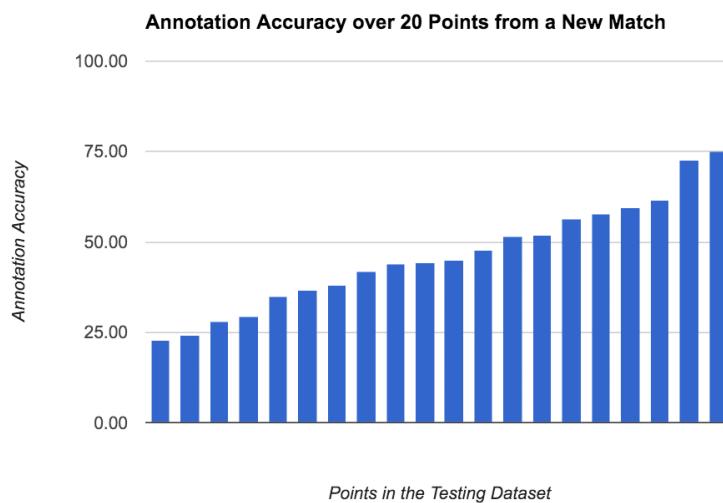


Figure 18: Annotation Accuracy among 20 different table tennis points

One problem with this approach is that it does not capture the sequential dependencies of different types of labels. For example the model has no knowledge of any historical nor any future labels and makes the prediction solely based on the frame image at that point of time. For example, even long after a point starts, the model may still predict that a player is "Serving" simply because his / her action looks like so and it has not knowledge that you cannot "Serve" in a table tennis game after your opponent loops the ball. Therefore, pairwise potential in a linear-chain conditional random field is introduced to capture such information.

4.4 Conditional Random Field (CRF)

In order to capture sequential relationship between different labels, a pairwise potential is introduced along with the unary discriminative classifier above. Instead of maximizing the product of probability like what we did in Section 4.3 where we assumed independence between data points across the time domain, we now add a pairwise potential term to maximize the likelihood of a linear combination of the unary potential and the pairwise potential whose weight is learned using gradient descent. In another word, we are trying to find a

sequence of labels that maximizes the following probability:

$$P(y_1, y_2, \dots, y_T | x) = \frac{1}{Z} \exp\left(\sum_{i=1}^T w_u \phi(y_i, x_i) + \sum_{j=2}^T w_p \phi(y_j, y_{j-1})\right) \quad (3)$$

where each x_i is our input 4096 dimensional input frame vector, averaged over 10 consecutive frames and each y_i is one of the 20 player stroke classes (or one of the 8 point-ending classes if it's within the last 10 frames of the point). The unary potential $\phi(y_i, x_i)$ is simply the discriminative classifier's softmax value trained above and the pairwise potential $\phi(y_i, y_{i-1})$ is obtained by counting the number of directed transitions from one class to another. w_u is the weight for the unary potential, w_p is the weight for the binary potential and Z is the normalization factor or the partition factor. Both w_u and w_p can be learned using gradient descent on the training dataset. Once the parameters w_u and w_p have been properly learned, inference on the most probable sequence y_1, y_2, \dots, y_T can then be calculated to be the sequence which maximizes the likelihood. This can be done efficiently using dynamic programming similar to the forward-backward algorithm.

The actual implementation of the CRF is done using the open-source project PyStruct [20] where it actually maximizes the SVM margin instead of the likelihood. The raw experimental result on the testing datasets are shown below (the source code that generates the result can be found in Appendix B):

Testing Dataset	Annotation Accuracy
15 new points from an old match	68.1%
20 new points from a new match	45.0%
Weighted average of the above	54.9%

The annotation accuracy for the 15 points coming from the old match improves significantly whereas that of the new match barely changes. This may attribute to the fact that the new match (8th game in the dataset) has a few un-common sequential label combinations. For example, it contains the combination of 'C1-TOP-PLAYER-FOREHAND-SERVE' followed by 'C28-BOTTOM-PLAYER-BACKHAND-MISS-HIT' which was never appears in the training dataset. In cases as such, the sequential dependency information offers no help to improve our annotation accuracy yet a larger training set that covers more state transitions between our training labels may help to alleviate this problem.

4.5 Fine Tuning on AlexNet

As it was mentioned in the previous section, the AlexNet [18] used above is trained based on the public ImageNet data set and has no knowledge about images from our training data set. In fact, it is very unlikely that the original ImageNet has any images on table tennis similar to the ones used in this project.

Customization on AlexNet is possible by fine tuning its weight parameters using frame images from our own training set via the back propagation algorithm and hopefully the last (fc7) or second-to-last (fc6) layer is able to capture features that are more relevant to our purpose. The neural net structure used is exactly identical to that from AlexNet, except for a output layer of 20 player stroke classes instead of the original 1000 different images classes and weights from the first 6 layers are copied over as a starting point. The set of training parameters (i.e solver.prototxt and train_val.prototxt) are attached in Appendix C. It should be noted that in addition to the structure shown in Figure 16, a droplet layer is introduced after the two fully-connected layers to prevent neural net overfitting our limited training data. Figure 19 shows the training loss over the 2000 iterations for both the neural net fine tuned using pre-trained AlexNet parameters (in blue) as well as a neural net with the exact same structure but trained from scratch (in green):

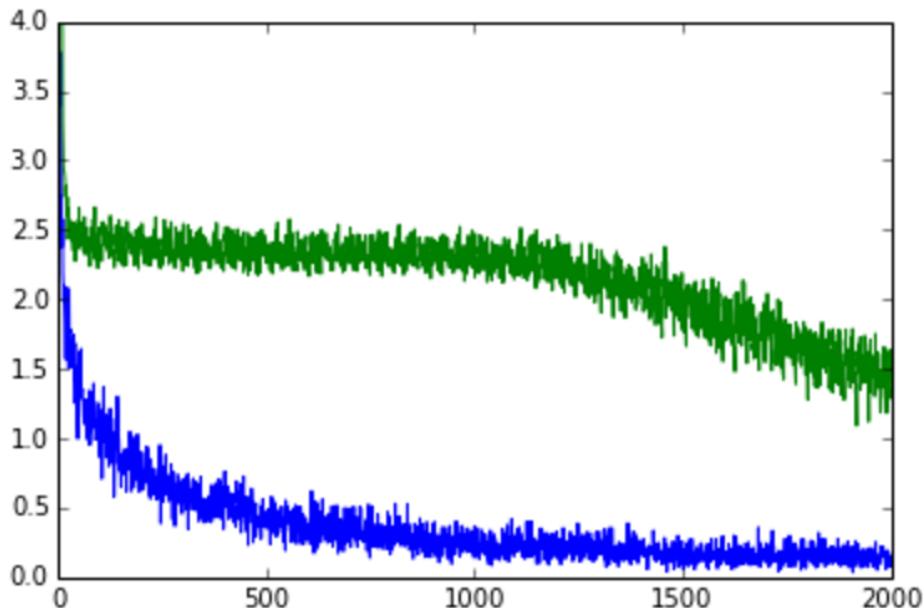


Figure 19: Training loss over 2000 iterations
 (Blue: Fine-tuned based on AlexNet;
 Green: Neural net with same structure but trained from scratch)

As it can be seen, fine-tuning based on a pre-trained AlexNet allows our model to find a relatively better local minimum on the error surface comparing to training it completely from scratch as it achieves a better training loss (cross entropy in this case). Apply the neural net trained after 2000 iterations, it was experimentally found that the fine-tuned net achieves a testset classification rate of 69% whereas the net that was trained from scratch achieves a classification rate of 48% (The original AlexNet with a SVM on the second last fully-connected layer achieves 59.2%).

Based on the fine-tuned AlexNet using our training data and apply the same CRF procedure as described in Section 4.4, we obtained the following result:

Testing Dataset	Annotation Accuracy
15 new points from an old match	66.8%
20 new points from a new match	52.1%
Weighted average of the above	58.7%

The comparison of labels (predicted by our model vs. manual annotation) for all 35 points in the testset can be found in Appendix D.

5 Future Work

Although some preliminary result on auto-generation of table tennis point video clips is presented above, there is a lot of room for future improvements. A short list can be found below:

1. Better Model Representation on Point-ending Classification

As it was seen in Section 4.2, the classifier on the point-ending classification task has terrible classification rate, merely around random guessing among the 8 classes. Besides the fact that the table tennis ball may not be picked up as a feature by AlexNet like discussed above, a better machine learning model is also desired. This is because many point-ending event occurs at an instant, instead of a duration in the player stroke classification tasks. An example would be when a player hits the ball out of the table, only two frames should be enough to provide the most relevant information: First frame where the ball is still on top of the table and a second frame where the ball drops below the table. All other images in that 10-frame sliding window simply adds noise to the data and confuses the classifier. As a result, a better model to represent this problem may be having the model trained only on those two images and emit a "Bottom Player Hits Out" event when its confidence exists some threshold value.

2. Larger Dataset

Our current dataset only contains a total of 187 table tennis video points, which is a reasonable size for the player stroke classification task as each point can generate 5-6 hit-level data points. However, it is not sufficient for the point-ending task as each table tennis point only generates one valid input data point for our classifier, split that between the training and testing datasets, leaving us only 35 data points for testing (an average of 4 data points per class) and outliers may influence our classification result.

3. Explore Other Machine Learning Models

Because of time constraints, the author did not get a chance to explore other deep learning models that may be more suitable for this sequential-based supervised learning task. Other popular approaches in recent years includes the LSTM (Long-short Term Memory) RNN (Recurrent Neural Network) http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf which can be combined with BRNN (Bidirectional Recurrent Neural Network) for the suitable offline learning applications.

4. Utilize Audio Information to Facilitate the Classification Tasks

Many times, as human beings, it is difficult to tell whether the ball lands on the table or not by just looking at the frame images. [8] proposes a way to combine visual and audio information in a table tennis video and may be an interesting direction to investigate further once other improvements above improvements have been made.

6 Conclusion

In conclusion, this project investigated the possibility of using machine learning techniques to create semantic understanding of table tennis game directly from multimedia streamed point video clips. It uses AlexNet [18] to select relevant features from the raw frame images and trains a CRF (Conditional Random Field) where each input is averaged over a sliding window of 10 consecutive frames images. We have obtained a classification rate of 58.7% over 20 different classes for player stroke identification and 8 classes for the point-ending classification over our testing dataset, the 35 table tennis point video clips. All source code with documentation on how the above result can be reproduced can be found online: <https://github.com/tabletennisr/thesis>

There are many potential applications that can derive from this project given our hit-by-hit level annotation. For coaches and athletes, auto-generated annotation allows them to analyze player's strength and weaknesses. For sport spectators, annotated sport video clips allows them to only watch the parts that they are interested in. A far reach of this project would be to create an "Auto-referee" that keeps track of the score for a table tennis game in realtime, yet it's not quite feasible with current point-ending classification rate.

References

- [1] Chris Harper, Frank Dunne, Charlotte Pratten, Ian Westover, Ben Speight and Simon Banoub *The Global Sports Media Consumption Report 2014 - US Overview*. SportBusiness Group, 2014.
- [2] D. Sun. *Neural Network-Based Table Tennis Competition Technique and Tactics Diagnosis and Evaluation*. The Open Cybernetics & Systemics Journal, 2015, Vol 9, Pp. 1722-1727
- [3] C. Xu, J. Wang, K. Wan, Y. Li and L. Duan.
Live Sports Event Detection Based on Broadcast Video and Web-casting Text 14th annual ACM international conference on Multimedia (MM '06), 2006
- [4] K. Wickramaratna, M. Chen, S. Chen and M. Shyu.
Neural Network Based Framework For Goal Event Detection In Soccer Videos IEEE International Symposium on Multimedia (ISM), 2005
- [5] H. Harb, L. Chen
Highlights Detection in Sports Videos Based on Audio Analysis IEEE Content-Based Multimedia Indexing (CBMI), 2003
- [6] B. Zhang, W. Dou and L. Chen
Ball Hit Detection in Table Tennis Games Based on Audio Analysis IEEE International Conference on Pattern Recognition (ICPR), 2006
- [7] D. Connaghan, P. Kelly and N. E. O'Conn
Game, Shot and Match: Event-based Indexing of Tennis IEEE Content-Based Multimedia Indexing (CBMI), 2011
- [8] B. Zhang, W. Chen, W. Dou, J. Yu, L. Chen
Content-based Table Tennis Games Highlight Detection Utilizing Audiovisual Clues Image and Graphics (IGIG), 2007, pp. 833-838.
- [9] Y. Rui, A. Gupta, and A. Acero
Content-based Table Tennis Games Highlight Detection Utilizing Audiovisual Clues ACM Multimedia Conference (ACMMM) 2000, pp. 833-838. Automatically Extracting Highlights for TV Baseball Programs
- [10] L. Draschkowitz, C. Draschkowitz, H. Hlavacs
Predicting Shot Success for Table Tennis Using Video Analysis and Machine Learning Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2014, Vol 136, pp. 12-21

- [11] M. Petkovic, W. Jonker and Z. Zivkovic
Recognizing Strokes in Tennis Videos Using Hidden Markov Models Visualization, Imaging and Image Processing, 2011
- [12] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing
Player Action Recognition in Broadcast Tennis Video with Applications to Semantic Analysis of Sports Game ACM Multimedia, 2006, pp. 431-440
- [13] W. Lu, J. Ting, J. Little, and K. Murphy
Learning to Track and Identify Players from Broadcast Sports Videos IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2013
- [14] W. Chen, Y. Zhang
Tracking Ball and Players with Applications to Highlight Ranking of Broadcasting Table Tennis Video Int. Conf. Computational Engineering in Systems Applications, 2006, vol. 2, pp.1896–1903.
- [15] International Table Tennis Federation (ITTF)
https://www.youtube.com/watch?v=NC-uMLzI8g&list=PLl_mJ3zK7B-LRKA4RL9o9GR95nugytDD
- [16] A complete, cross-platform solution to record, convert and stream audio and video.
<https://wwwffmpeg.org/>
- [17] Tesseract Open Source OCR Engine.
<https://github.com/tesseract-ocr/tesseract>
- [18] Krizhevsky, A., Sutskever, I., and Hinton, G.
ImageNet classification with deep convolutional neural networks Advances in Neural Information Processing Systems (NIPS), 2012.
- [19] Jia, Y.
An open source convolutional architecture for fast feature embedding
<http://caffe.berkeleyvision.org/>, 2013.
- [20] Andreas C. Mueller, Sven Behnke.
PyStruct - Structured prediction in Python
<https://pystruct.github.io/>, 2014.
- [21] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A
Going deeper with convolutions Large Scale Visual Recognition Challenge (ILSVRC), 2014

A Appendix

all_class_classification_base

April 8, 2016

```
In [1]: %matplotlib inline
import matplotlib
import matplotlib.pyplot as plt
matplotlib.use('Agg')
from sklearn import svm, cross_validation
import pylab as pl
from PIL import Image
import numpy as np
import os

/usr/lib/pymodules/python2.7/matplotlib/__init__.py:1173: UserWarning: This call to matplotlib.use() has
because the backend has already been chosen;
matplotlib.use() must be called *before* pylab, matplotlib.pyplot,
or matplotlib.backends is imported for the first time.

warnings.warn(_use_error_msg)

In [2]: # Make sure that caffe is on the python path:
caffe_root = '/u/zexuan/caffe/caffe/' # this file is expected to be in {caffe_root}/examples
caffe_real_root = '/pkgs/caffe/'
thesis_root = '/ais/gobi2/pingpong/thesis/'
#!ls /pkgs/caffe
import sys
sys.path.insert(0, caffe_real_root + 'python')
import caffe

plt.rcParams['figure.figsize'] = (10, 10)
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

if not os.path.isfile('~/caffe/caffe/models/bvlc_reference_caffenet/bvlc_reference_caffenet.caffemodel'):
    print("Downloading pre-trained CaffeNet model...")
    !~/caffe/caffe/scripts/download_model_binary.py ~/caffe/caffe/models/bvlc_reference_caffenet.caffemodel

Downloading pre-trained CaffeNet model...
Model already exists.

In [3]: caffe.set_device(0)
        caffe.set_mode_gpu()

In [4]: # MODEL_FILE = caffe_root +'models/bvlc_reference_caffenet/deploy.prototxt'
# PRETRAINED = caffe_root +'models/bvlc_reference_caffenet/bvlc_reference_caffenet.caffemodel'
# caffe.set_mode_cpu()
# net = caffe.Classifier(MODEL_FILE, PRETRAINED,
```

```

#           mean=np.load(caffe_root + 'python/caffe/imagenet/ilsvrc_2012_mean.npy'),
#           channel_swap=(2,1,0),
#           raw_scale=255,
#           image_dims=(600, 600))
# print net.blobs['data'].data.shape
# [(k, v.data.shape) for k, v in net.blobs.items()]

# caffe.set_mode_cpu()
print caffe_root + 'models/bvlc_reference_caffenet/deploy.prototxt'
#!ls caffe_root+'models/bvlc_reference_caffenet/deploy.prototxt'
net = caffe.Net(caffe_root + 'models/bvlc_reference_caffenet/deploy.prototxt',
                 caffe_root + 'models/bvlc_reference_caffenet/bvlc_reference_caffenet.caffemodel',
                 caffe.TEST)

# input preprocessing: 'data' is the name of the input blob == net.inputs[0]
transformer = caffe.io.Transformer({'data': net.blobs['data'].data.shape})
transformer.set_transpose('data', (2,0,1))
transformer.set_mean('data', np.load(caffe_root + 'python/caffe/imagenet/ilsvrc_2012_mean.npy'))
transformer.set_raw_scale('data', 255) # the reference model operates on images in [0,255] ran.
transformer.set_channel_swap('data', (2,1,0)) # the reference model has channels in BGR order
print net.blobs['data'].data.shape
[(k, v.data.shape) for k, v in net.blobs.items()]

```

/u/zexuan/caffe/caffe/models/bvlc_reference_caffenet/deploy.prototxt
(10, 3, 227, 227)

Out[4]: [('data', (10, 3, 227, 227)),
 ('conv1', (10, 96, 55, 55)),
 ('pool1', (10, 96, 27, 27)),
 ('norm1', (10, 96, 27, 27)),
 ('conv2', (10, 256, 27, 27)),
 ('pool2', (10, 256, 13, 13)),
 ('norm2', (10, 256, 13, 13)),
 ('conv3', (10, 384, 13, 13)),
 ('conv4', (10, 384, 13, 13)),
 ('conv5', (10, 256, 13, 13)),
 ('pool5', (10, 256, 6, 6)),
 ('fc6', (10, 4096)),
 ('fc7', (10, 4096)),
 ('fc8', (10, 1000)),
 ('prob', (10, 1000))]

In [5]: import skimage
import skimage.io
import time

start_time = time.time()
training_set = {}
training_set['target'] = []
container_path = 'labeled_img_selected_all_classes/'

folders = [f for f in sorted(os.listdir(container_path)) if os.path.isdir(os.path.join(container_path, f))]
print 'folders: ', folders

```

result = []
for folder in folders:
    folder_path = os.path.join(container_path, folder)
    documents = [os.path.join(folder_path, d) for d in sorted(os.listdir(folder_path))]
    input_10_imgs = []
    for pic in documents:
        if not pic.endswith('png'):
            continue
        img = skimage.img_as_float(skimage.io.imread(pic)).astype(np.float32)
        input_10_imgs.append(img)
        if len(input_10_imgs) >= 10:
            #net.blobs['data'].data[...] = transformer.preprocess('data', training_set['data'][...])
            #net.blobs['data'].data[...] = map(lambda x: transformer.preprocess('data', x), tra
            net.blobs['data'].data[...] = map(lambda x: transformer.preprocess('data', x), input_
            out = net.forward()
            result.append(np.mean(net.blobs['fc6'].data, axis=0))
            input_10_imgs = []
    training_set['target'].append(folder)
#training_set['data']=data
#training_set['target']=target
print 'Total time loading training data: ', time.time()-start_time, ' seconds'
print 'number of images: ', len(training_set['target'])
#print 'dimensionality of each image: ', input_10_imgs[0].shape

folders: ['C10_TOP_PLAYER_BACKHAND_BLOCK_selected', 'C11_BOTTOM_PLAYER_FOREHAND_BLOCK_selected', 'C12_BOT'
Total time loading training data: 458.498327971 seconds
number of images: 11680

In [6]: start_time = time.time()
testing_set = {}
testing_set['target'] = []
container_path = 'labeled_img_selected_all_classes_testset/'

folders = [f for f in sorted(os.listdir(container_path)) if os.path.isdir(os.path.join(container_
print 'folders: ', folders

for folder in folders:
    folder_path = os.path.join(container_path, folder)
    documents = [os.path.join(folder_path, d) for d in sorted(os.listdir(folder_path))]
    input_10_imgs = []
    for pic in documents:
        if not pic.endswith('png'):
            continue
        img = skimage.img_as_float(skimage.io.imread(pic)).astype(np.float32)
        input_10_imgs.append(img)
        if len(input_10_imgs) >= 10:
            #net.blobs['data'].data[...] = transformer.preprocess('data', training_set['data'][...])
            #net.blobs['data'].data[...] = map(lambda x: transformer.preprocess('data', x), tra
            net.blobs['data'].data[...] = map(lambda x: transformer.preprocess('data', x), input_
            out = net.forward()
            result.append(np.mean(net.blobs['fc6'].data, axis=0))
            input_10_imgs = []
    testing_set['target'].append(folder)
#training_set['data']=data
#training_set['target']=target

```

```

print 'Total time loading training data: ', time.time()-start_time, ' seconds'
print 'number of images: ', len(testing_set['target'])

folders: ['C10_TOP_PLAYER_BACKHAND_BLOCK_selected', 'C11_BOTTOM_PLAYER_FOREHAND_BLOCK_selected', 'C12_BOT'
Total time loading training data: 100.246762991 seconds
number of images: 2500

In [7]: # print caffe.io.load_image(thesis_root + 'labeled_image_selected/bottom_player_winning_selecte
# print caffe.io.load_image(thesis_root + 'labeled_image_selected/bottom_player_winning_selecte
# print training_set['data'][0][0][0]

    ,
result = []
#for i in xrange(2200, 2300, 10):
for i in xrange(0, traning_set_size, 10):
    #net.blobs['data'].data[...] = transformer.preprocess('data', training_set['data'][0])
    net.blobs['data'].data[...] = map(lambda x: transformer.preprocess('data', x), training_set
out = net.forward()
#print net.blobs['fc7'].data.shape
result.append(np.mean(net.blobs['fc6'].data, axis=0))
,
print 'Number of input to SVM: ', len(result), 'Size of each input', result[0].shape

print transformer.deprocess('data', net.blobs['data'].data[0]).shape
print 'Sample image from the training set'
#plt.imshow(transformer.deprocess('data', net.blobs['data'].data[4]))

Number of input to SVM: 1418 Size of each input (4096,)
(227, 227, 3)
Sample image from the training set

In [8]: import pickle

pickle.dump(result, open("nn_result_fc6_unnormalized_clf_7videos.pickle", "wb" ) )

In [9]: #from sklearn.preprocessing import normalize
print len(result)
print np.max(result[0])
print result[0].shape
normalized_result = []
#normalized_result = normalize(result)
for r in result:
    #normalized_result.append(r/np.max(r))
    normalized_result.append((r-np.mean(r))/np.std(r))

1418
45.1747
(4096,)

In [10]: import pickle

pickle.dump(normalized_result, open("nn_result_fc6_normalized_clf_7videos.pickle", "wb" ) )

In [11]: print len(normalized_result)
print np.max(normalized_result[0])
print normalized_result[0].shape
print result[0].shape

```

```

1418
13.611
(4096,)
(4096,)

In [15]: from sklearn import decomposition

    traning_set_size = len(training_set['target'])
    testing_set_size = len(testing_set['target'])
    print 'traning_set_size: ', traning_set_size
    print 'normalized_result size: ', len(normalized_result)
    #X_train, X_test, y_train, y_test = cross_validation.train_test_split(result, training_set['ta
    #X_train, X_test, y_train, y_test = cross_validation.train_test_split(normalized_result, train
    X_train = normalized_result[:traning_set_size/10]
    X_test = normalized_result[traning_set_size/10:]
    y_train = training_set['target'][0:traning_set_size:10]
    y_test = testing_set['target'][0:traning_set_size:10]

    print 'X_train: %d; X_test: %d; y_train: %d; y_test: %d'%(len(X_train), len(X_test), len(y_trai

traning_set_size: 11680
normalized_result size: 1418
X_train: 1168; X_test: 250; y_train: 1168; y_test: 250

In [30]: for c in [0.001, 0.005, 0.01, 0.05, 0.1, 1, 10, 500]:
            svc = svm.SVC(kernel='linear', C=c)
            kfold = cross_validation.KFold(len(X_train), n_folds=4, shuffle=True)
            cv_result = cross_validation.cross_val_score(svc, X_train, y_train, cv=kfold, n_jobs=1)
            print 'C = %g, Cross validation: %s'%(c, cv_result)
            print 'C = %g, Mean value: %g'%(c, np.mean(cv_result))

C = 0.001, Cross validation: [ 0.71232877  0.69520548  0.69178082  0.76027397]
C = 0.001, Mean value: 0.714897
C = 0.005, Cross validation: [ 0.80136986  0.78424658  0.72945205  0.78082192]
C = 0.005, Mean value: 0.773973
C = 0.01, Cross validation: [ 0.78767123  0.7739726   0.77054795  0.79452055]
C = 0.01, Mean value: 0.781678
C = 0.05, Cross validation: [ 0.81506849  0.75342466  0.78082192  0.71575342]
C = 0.05, Mean value: 0.766267
C = 0.1, Cross validation: [ 0.7260274   0.76027397  0.74315068  0.70890411]
C = 0.1, Mean value: 0.734589
C = 1, Cross validation: [ 0.77054795  0.68835616  0.74315068  0.7260274 ]
C = 1, Mean value: 0.732021
C = 10, Cross validation: [ 0.73287671  0.77739726  0.75           0.67123288]
C = 10, Mean value: 0.732877
C = 500, Cross validation: [ 0.71232877  0.70890411  0.73972603  0.78082192]
C = 500, Mean value: 0.735445

In [31]: for c in [0.01, 0.05, 0.1, 1, 10, 100, 1000, 1000000]:
            svc = svm.SVC(C=c, kernel='poly', degree=1)
            kfold = cross_validation.KFold(len(X_train), n_folds=4, shuffle=True)
            cv_result = cross_validation.cross_val_score(svc, X_train, y_train, cv=kfold, n_jobs=1)
            print 'C = %g, Cross validation: %s'%(c, cv_result)
            print 'C = %g, Mean value: %g'%(c, np.mean(cv_result))

```

```

C = 0.01, Cross validation: [ 0.21232877  0.21575342  0.2260274   0.15410959]
C = 0.01, Mean value: 0.202055
C = 0.05, Cross validation: [ 0.32534247  0.30821918  0.37328767  0.37328767]
C = 0.05, Mean value: 0.345034
C = 0.1, Cross validation: [ 0.44520548  0.45205479  0.43493151  0.46917808]
C = 0.1, Mean value: 0.450342
C = 1, Cross validation: [ 0.69520548  0.56164384  0.63356164  0.57876712]
C = 1, Mean value: 0.617295
C = 10, Cross validation: [ 0.77054795  0.77054795  0.73630137  0.72945205]
C = 10, Mean value: 0.751712
C = 100, Cross validation: [ 0.77054795  0.76712329  0.79109589  0.71917808]
C = 100, Mean value: 0.761986
C = 1000, Cross validation: [ 0.77054795  0.73630137  0.73287671  0.73972603]
C = 1000, Mean value: 0.744863
C = 1e+06, Cross validation: [ 0.69178082  0.72945205  0.75684932  0.72260274]
C = 1e+06, Mean value: 0.725171

In [32]: for c in [0.01, 0.05, 0.1, 1, 10, 100, 1000, 1000000]:
    svc = svm.SVC(C=c, kernel='poly', degree=2)
    kfold = cross_validation.KFold(len(X_train), n_folds=4, shuffle=True)
    cv_result = cross_validation.cross_val_score(svc, X_train, y_train, cv=kfold, n_jobs=1)
    print 'C = %g, Cross validation: %s' %(c, cv_result)
    print 'C = %g, Mean value: %g' %(c, np.mean(cv_result))

C = 0.01, Cross validation: [ 0.21575342  0.18835616  0.19178082  0.21232877]
C = 0.01, Mean value: 0.202055
C = 0.05, Cross validation: [ 0.37671233  0.34589041  0.38013699  0.42808219]
C = 0.05, Mean value: 0.382705
C = 0.1, Cross validation: [ 0.50684932  0.44863014  0.46917808  0.48630137]
C = 0.1, Mean value: 0.47774
C = 1, Cross validation: [ 0.65410959  0.63356164  0.65410959  0.6609589 ]
C = 1, Mean value: 0.650685
C = 10, Cross validation: [ 0.79109589  0.78767123  0.75684932  0.75342466]
C = 10, Mean value: 0.77226
C = 100, Cross validation: [ 0.76369863  0.73287671  0.74657534  0.73972603]
C = 100, Mean value: 0.745719
C = 1000, Cross validation: [ 0.76712329  0.75342466  0.76369863  0.76712329]
C = 1000, Mean value: 0.762842
C = 1e+06, Cross validation: [ 0.70890411  0.71917808  0.73972603  0.76369863]
C = 1e+06, Mean value: 0.732877

In [33]: for c in [0.01, 0.05, 0.1, 1, 10, 100, 1000, 1000000]:
    svc = svm.SVC(C=c, kernel='poly', degree=3)
    kfold = cross_validation.KFold(len(X_train), n_folds=4, shuffle=True)
    cv_result = cross_validation.cross_val_score(svc, X_train, y_train, cv=kfold, n_jobs=1)
    print 'C = %g, Cross validation: %s' %(c, cv_result)
    print 'C = %g, Mean value: %g' %(c, np.mean(cv_result))

C = 0.01, Cross validation: [ 0.21575342  0.22945205  0.15753425  0.20547945]
C = 0.01, Mean value: 0.202055
C = 0.05, Cross validation: [ 0.39041096  0.42808219  0.45205479  0.40410959]
C = 0.05, Mean value: 0.418664
C = 0.1, Cross validation: [ 0.48287671  0.47260274  0.4760274   0.5239726 ]
C = 0.1, Mean value: 0.48887
C = 1, Cross validation: [ 0.67465753  0.71917808  0.62328767  0.64383562]

```

```

C = 1, Mean value: 0.66524
C = 10, Cross validation: [ 0.73972603  0.80136986  0.80479452  0.7739726 ]
C = 10, Mean value: 0.779966
C = 100, Cross validation: [ 0.73972603  0.78082192  0.75342466  0.73630137]
C = 100, Mean value: 0.752568
C = 1000, Cross validation: [ 0.73630137  0.74315068  0.78082192  0.70205479]
C = 1000, Mean value: 0.740582
C = 1e+06, Cross validation: [ 0.70205479  0.75342466  0.77054795  0.72945205]
C = 1e+06, Mean value: 0.73887

In [34]: from sklearn import svm, cross_validation
X_train = np.asarray(X_train)
svc = svm.SVC(kernel='precomputed')
kernel_train = np.dot(X_train, X_train.T) # linear kernel
svc.fit(kernel_train, y_train)

# Testing
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
kernel_test = np.dot(X_test, X_train.T)
y_pred = svc.predict(kernel_test)
print accuracy_score(y_test, y_pred)

import pandas as pd
y_true = pd.Series(y_test)
y_pred = pd.Series(y_pred)

#print pd.crosstab(y_true, y_pred, rownames=['True'], colnames=['Predicted'], margins=True)

print confusion_matrix(y_test, y_pred).shape
print confusion_matrix(y_test, y_pred)
#print '\n'.join(sorted(list(set(y_test))))
print len(set(y_test).union(set(y_pred)))

```

0.592
(18, 18)

```

[[ 1  0  0  0  0  0  0  2  0  1  0  0  0  0  2  2  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0]
 [ 0  0  1  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0]
 [ 5  0  0  0  0  0  0  6  1  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  0  2  0  0  0  0  0  0  0  0  0]
 [ 2  0  0  0  2  1  0  7  4  2  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  0  0  3  0  0  0  5  0  0  0  0  0  0  4  0  0]
 [ 0  0  0  0  0  0  0  0  0  5  39 0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  1  1  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0  59 0  0  4  0  0  0]
 [ 0  0  0  0  0  0  0  4  0  4  0  0  0  14 0  2  0  4  0]
 [ 1  0  0  0  1  0  0  0  4  3  0  0  0  7  2  1  1  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0  2  0  14 1  0  0  0]
 [ 0  0  0  0  0  0  0  3  0  0  0  2  0  2  0  2  4  0  0]
 [ 0  0  0  0  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0]]

```

18

```

/u/zexuan/.local/lib/python2.7/site-packages/sklearn/utils/validation.py:386: DeprecationWarning: Passi
  DeprecationWarning)

C16_BOTTOM_PLAYER_BACKHAND_FLIP_selected C8_BOTTOM_PLAYER_BACKHAND_LOOP_selected
DIFFERENT: C7_BOTTOM_PLAYER_FOREHAND_LOOP_selected C8_BOTTOM_PLAYER_BACKHAND_LOOP_selected
DIFFERENT: C10_TOP_PLAYER_BACKHAND_BLOCK_selected C8_BOTTOM_PLAYER_BACKHAND_LOOP_selected
DIFFERENT: C5_TOP_PLAYER_FOREHAND_LOOP_selected C8_BOTTOM_PLAYER_BACKHAND_LOOP_selected
DIFFERENT: C5_TOP_PLAYER_FOREHAND_LOOP_selected C8_BOTTOM_PLAYER_BACKHAND_LOOP_selected
DIFFERENT: C18_TOP_PLAYER_BACKHAND_CHOP_selected C9_TOP_PLAYER_FOREHAND_BLOCK_selected
DIFFERENT: C5_TOP_PLAYER_FOREHAND_LOOP_selected C9_TOP_PLAYER_FOREHAND_BLOCK_selected

/u/zexuan/.local/lib/python2.7/site-packages/sklearn/utils/validation.py:386: DeprecationWarning: Passi
  DeprecationWarning)
/u/zexuan/.local/lib/python2.7/site-packages/sklearn/utils/validation.py:386: DeprecationWarning: Passi
  DeprecationWarning)
/u/zexuan/.local/lib/python2.7/site-packages/sklearn/utils/validation.py:386: DeprecationWarning: Passi
  DeprecationWarning)

In [59]: from sklearn.neighbors import KNeighborsClassifier
neigh = KNeighborsClassifier(n_neighbors=7)

kfold = cross_validation.KFold(len(X_train), n_folds=4, shuffle=True)
cv_result = cross_validation.cross_val_score(neigh, X_train, y_train, cv=kfold, n_jobs=1)

print 'Cross validation result', cv_result, '. Avg:', np.mean(cv_result)

clf = neigh.fit(X_train, y_train)
print 'Testing result', clf.score(X_test, y_test)

Cross validation result [ 0.62328767  0.59931507  0.64383562  0.66438356] . Avg: 0.632705479452
Testing result 0.504

In [61]: from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
cv_result = cross_validation.cross_val_score(gnb, X_train, y_train, cv=kfold, n_jobs=1)
print 'Cross validation result', cv_result, '. Avg:', np.mean(cv_result)

clf = gnb.fit(X_train, y_train)
print 'Testing result', clf.score(X_test, y_test)

Cross validation result [ 0.6130137   0.51712329   0.58561644   0.58219178] . Avg: 0.57448630137
Testing result 0.388

In [62]: from sklearn
          Logistic Regression & 63.3\% & 57.6\% \\ \hline
          \hline
          lr = LogisticRegression()

```

```
cv_result = cross_validation.cross_val_score(lr, X_train, y_train, cv=kfold, n_jobs=1)
print 'Cross validataion result', cv_result, '. Avg:', np.mean(cv_result)

clf = lr.fit(X_train, y_train)
print 'Testing result', clf.score(X_test, y_test)

Cross validataion result [ 0.76369863  0.74315068  0.78424658  0.7739726 ] . Avg: 0.766267123288
Testing result 0.576

In [26]: pickle.dump(clf, open("fc6_normalized_lr_clf_7videos_2.pickle", "wb" ) )

In [27]: print sorted(list(set(y_train)))
print len(sorted(list(set(y_train)))))

['C10_TOP_PLAYER_BACKHAND_BLOCK_selected', 'C11_BOTTOM_PLAYER_FOREHAND_BLOCK_selected', 'C12_BOTTOM_PLAYER_
18

In [ ]:
```

Appendices B

```
1 from pystruct.datasets import load_letters
2 import numpy as np
3 import fileinput
4 from glob import glob
5 import pickle
6 import sys
7 import time
8
9 label_to_index_pt_end = {
10     'C21.TOP.PLAYER.UNDER.NET': 0,
11     'C22.TOP.PLAYER.HIT.OUT': 1,
12     'C23.TOP.PLAYER.FOREHAND.MISS.HIT': 2,
13     'C24.TOP.PLAYER.BACKHAND.MISS.HIT': 3,
14     'C25.BOTTOM.PLAYER.UNDER.NET': 4,
15     'C26.BOTTOM.PLAYER.HIT.OUT': 5,
16     'C27.BOTTOM.PLAYER.FOREHAND.MISS.HIT': 6,
17     'C28.BOTTOM.PLAYER.BACKHAND.MISS.HIT': 7}
18
19 index_to_label_pt_end = {
20     0: 'C21.TOP.PLAYER.UNDER.NET',
21     1: 'C22.TOP.PLAYER.HIT.OUT',
22     2: 'C23.TOP.PLAYER.FOREHAND.MISS.HIT',
23     3: 'C24.TOP.PLAYER.BACKHAND.MISS.HIT',
24     4: 'C25.BOTTOM.PLAYER.UNDER.NET',
25     5: 'C26.BOTTOM.PLAYER.HIT.OUT',
26     6: 'C27.BOTTOM.PLAYER.FOREHAND.MISS.HIT',
27     7: 'C28.BOTTOM.PLAYER.BACKHAND.MISS.HIT'}
28
29 label_to_index = {
30     'C10.TOP.PLAYER.BACKHAND.BLOCK': 0,
31     'C11.BOTTOM.PLAYER.FOREHAND.BLOCK': 1,
32     'C12.BOTTOM.PLAYER.BACKHAND.BLOCK': 2,
33     'C13.TOP.PLAYER.FOREHAND.FLIP': 3,
34     'C14.TOP.PLAYER.BACKHAND.FLIP': 4,
35     'C15.BOTTOM.PLAYER.FOREHAND.FLIP': 5,
36     'C16.BOTTOM.PLAYER.BACKHAND.FLIP': 6,
37     'C17.TOP.PLAYER.FOREHAND.CHOP': 7,
38     'C18.TOP.PLAYER.BACKHAND.CHOP': 8,
39     'C19.BOTTOM.PLAYER.FOREHAND.CHOP': 9,
40     'C1.TOP.PLAYER.FOREHAND.SERVE': 10,
41     'C20.BOTTOM.PLAYER.BACKHAND.CHOP': 11,
42     '# C2.TOP.PLAYER.BACKHAND.SERVE': 11,
43     'C30.BOTTOM.PLAYER.BACKHAND.LOB': 12,
44     'C3.BOTTOM.PLAYER.FOREHAND.SERVE': 13,
45     '# C4.BOTTOM.PLAYER.BACKHAND.SERVE': 14,
46     'C5.TOP.PLAYER.FOREHAND.LOOP': 14,
47     'C6.TOP.PLAYER.BACKHAND.LOOP': 15,
48     'C7.BOTTOM.PLAYER.FOREHAND.LOOP': 16,
49     'C8.BOTTOM.PLAYER.BACKHAND.LOOP': 17,
50     'C9.TOP.PLAYER.FOREHAND.BLOCK': 18,
51     'C21.TOP.PLAYER.UNDER.NET': 19,
52     'C22.TOP.PLAYER.HIT.OUT': 20,
53     'C23.TOP.PLAYER.FOREHAND.MISS.HIT': 21,
54     'C24.TOP.PLAYER.BACKHAND.MISS.HIT': 22,
55     'C25.BOTTOM.PLAYER.UNDER.NET': 23,
```

```

56     'C26_BOTTOM_PLAYER_HIT_OUT':24 ,
57     'C27_BOTTOM_PLAYER_FOREHAND_MISS_HIT':25 ,
58     'C28_BOTTOM_PLAYER_BACKHAND_MISS_HIT':26 }

59 index_to_label={
60     0: 'C10_TOP_PLAYER_BACKHAND_BLOCK' ,
61     1: 'C11_BOTTOM_PLAYER_FOREHAND_BLOCK' ,
62     2: 'C12_BOTTOM_PLAYER_BACKHAND_BLOCK' ,
63     3: 'C13_TOP_PLAYER_FOREHAND_FLIP' ,
64     4: 'C14_TOP_PLAYER_BACKHAND_FLIP' ,
65     5: 'C15_BOTTOM_PLAYER_FOREHAND_FLIP' ,
66     6: 'C16_BOTTOM_PLAYER_BACKHAND_FLIP' ,
67     7: 'C17_TOP_PLAYER_FOREHAND_CHOP' ,
68     8: 'C18_TOP_PLAYER_BACKHAND_CHOP' ,
69     9: 'C19_BOTTOM_PLAYER_FOREHAND_CHOP' ,
70     10: 'C1_TOP_PLAYER_FOREHAND_SERVE' ,
71     11: 'C20_BOTTOM_PLAYER_BACKHAND_CHOP' ,
72     # 11: 'C2_TOP_PLAYER_BACKHAND_SERVE' ,
73     12: 'C30_BOTTOM_PLAYER_BACKHAND_LOB' ,
74     13: 'C3_BOTTOM_PLAYER_FOREHAND_SERVE' ,
75     # 14: 'C4_BOTTOM_PLAYER_BACKHAND_SERVE' ,
76     14: 'C5_TOP_PLAYER_FOREHAND_LOOP' ,
77     15: 'C6_TOP_PLAYER_BACKHAND_LOOP' ,
78     16: 'C7_BOTTOM_PLAYER_FOREHAND_LOOP' ,
79     17: 'C8_BOTTOM_PLAYER_BACKHAND_LOOP' ,
80     18: 'C9_TOP_PLAYER_FOREHAND_BLOCK' ,
81     19: 'C21_TOP_PLAYER_UNDER_NET' ,
82     20: 'C22_TOP_PLAYER_HIT_OUT' ,
83     21: 'C23_TOP_PLAYER_FOREHAND_MISS_HIT' ,
84     22: 'C24_TOP_PLAYER_BACKHAND_MISS_HIT' ,
85     23: 'C25_BOTTOM_PLAYER_UNDER_NET' ,
86     24: 'C26_BOTTOM_PLAYER_HIT_OUT' ,
87     25: 'C27_BOTTOM_PLAYER_FOREHAND_MISS_HIT' ,
88     26: 'C28_BOTTOM_PLAYER_BACKHAND_MISS_HIT' }

89
90
91 def load_data(files):
92     test_data = []
93     test_label = []
94     for dat_file in files:
95         test_data_pt = []
96         test_label_pt = []
97         prev_label = None
98         with open(dat_file) as f:
99             for line in f:
100                 line = line.strip().split()
101                 test_label_pt.append(label_to_index[line[-1]])
102                 data_val = map(float, line[:-1])
103                 test_data_pt.append(data_val)
104             test_data.append(np.asarray(test_data_pt, dtype=np.float32))
105             test_label.append(np.asarray(test_label_pt))
106     return np.array(test_data), np.array(test_label)

107
108
109 training_files = glob('./seq_data_fc6_normalized_new/point_0*.dat')
110 testing_files = glob('./seq_data_fc6_normalized_8videos_testset/point_0*.dat')
111
112 X_train_pre, y_train = load_data(training_files)

```

```

113 X_test_pre, y_test = load_data(testing_files)
114 X_train = np.copy(X_train_pre)
115 X_test = np.copy(X_test_pre)
116
117 for i, _ in enumerate(X_train_pre):
118     numCorrectPred = 0
119     for j, _ in enumerate(X_train_pre[i]):
120         min_ind = max(0, j-4)
121         max_ind = min(len(X_train_pre[i]), j+5)
122         sum_dat = np.array(X_train_pre[i][min_ind], dtype=np.float64)
123         for k in xrange(min_ind+1, max_ind):
124             sum_dat += np.array(X_train_pre[i][k], dtype=np.float64)
125         dat = sum_dat / 9
126         X_train[i][j] = dat
127
128 for i, _ in enumerate(X_test_pre):
129     numCorrectPred = 0
130     for j, _ in enumerate(X_test_pre[i]):
131         min_ind = max(0, j-4)
132         max_ind = min(len(X_test_pre[i]), j+5)
133         sum_dat = np.array(X_test_pre[i][min_ind], dtype=np.float64)
134         for k in xrange(min_ind+1, max_ind):
135             sum_dat += np.array(X_test_pre[i][k], dtype=np.float64)
136         dat = sum_dat / 9
137         X_test[i][j] = dat
138
139 from pystruct.models import ChainCRF
140 from pystruct.learners import FrankWolfeSSVM
141 from pystruct.learners import SubgradientSSVM
142 import cPickle as pickle
143
144 model = ChainCRF(directed=True)
145 ssvm = FrankWolfeSSVM(model=model, C=1000, max_iter=100, verbose=3,
146   show_loss_every=0)
147 ssvm = SubgradientSSVM(model=model, C=.1, max_iter=10)
148 ssvm.fit(X_train, y_train)
149 pickle.dump(ssvm, open("crf_ssVm_player_stroke.pickle", "wb"))
150
151 ssvm = pickle.load(open("crf_ssVm.pickle", "rb"))
152 y_pred = np.array(ssvm.predict(X_test))
153
154 accuracies = []
155 for i, _ in enumerate(y_test):
156     # print 'Point %d' % (i)
157     num_of_correct_labels = 0
158     for j, _ in enumerate(y_test[i]):
159         print 'Predict:%s; Actual:%s' % (index_to_label[y_pred[i][j]], index_to_label
160 [y_test[i][j]])
161         # print 'Predict:%s; Actual:%s' % (index_to_label_pt_end[y_pred[i][j]],
162         index_to_label_pt_end[y_test[i][j]])
163         if y_pred[i][j] == y_test[i][j]:
164             num_of_correct_labels += 1
165     accuracies.append(num_of_correct_labels * 100.0 / len(y_test[i]))
166     print "Accuracy: %.3f" % (accuracies[-1])
167 print "Overall Accuracy: %.3f" % (sum(accuracies) / len(accuracies))

```

Appendices C

```
net: "/u/zexuan/caffe/caffe/models/finetune_stroke_cls/train_val.prototxt"
test_iter: 100
test_interval: 1000
base_lr: 0.001
lr_policy: "step"
gamma: 0.1
stepsize: 20000
display: 20
max_iter: 100000
momentum: 0.9
weight_decay: 0.0005
snapshot: 10000
snapshot_prefix: "/u/zexuan/caffe/caffe/models/finetune_stroke_cls/finetune_stroke_cls"

name: "FlickrStyleCaffeNet"
layer {
    name: "data"
    type: "ImageData"
    top: "data"
    top: "label"
    include {
        phase: TRAIN
    }
    transform_param {
        mirror: true
        crop_size: 227
        mean_file: "/u/zexuan/caffe/caffe/data/ilsvrc12/imagenet_mean.binaryproto"
    }
    image_data_param {
        source: "/u/zexuan/caffe/caffe/data/stroke_cls/training_rand.txt"
        batch_size: 100
        new_height: 256
        new_width: 256
    }
}
layer {
    name: "data"
    type: "ImageData"
    top: "data"
    top: "label"
    include {
        phase: TEST
    }
    transform_param {
        mirror: false
        crop_size: 227
        mean_file: "/u/zexuan/caffe/caffe/data/ilsvrc12/imagenet_mean.binaryproto"
    }
    image_data_param {
        source: "/u/zexuan/caffe/caffe/data/stroke_cls/testing_rand.txt"
        batch_size: 100
        new_height: 256
        new_width: 256
    }
}
layer {
    name: "conv1"
    type: "Convolution"
    bottom: "data"
```

```
top: "conv1"
param {
    lr_mult: 1
    decay_mult: 1
}
param {
    lr_mult: 2
    decay_mult: 0
}
convolution_param {
    num_output: 96
    kernel_size: 11
    stride: 4
    weight_filler {
        type: "gaussian"
        std: 0.01
    }
    bias_filler {
        type: "constant"
        value: 0
    }
}
layer {
    name: "relu1"
    type: "ReLU"
    bottom: "conv1"
    top: "conv1"
}
layer {
    name: "pool1"
    type: "Pooling"
    bottom: "conv1"
    top: "pool1"
    pooling_param {
        pool: MAX
        kernel_size: 3
        stride: 2
    }
}
layer {
    name: "norm1"
    type: "LRN"
    bottom: "pool1"
    top: "norm1"
    lrn_param {
        local_size: 5
        alpha: 0.0001
        beta: 0.75
    }
}
layer {
    name: "conv2"
    type: "Convolution"
    bottom: "norm1"
    top: "conv2"
    param {
        lr_mult: 1
        decay_mult: 1
    }
    param {
        lr_mult: 2
        decay_mult: 0
    }
}
```

```
        }
    convolution_param {
        num_output: 256
        pad: 2
        kernel_size: 5
        group: 2
        weight_filler {
            type: "gaussian"
            std: 0.01
        }
        bias_filler {
            type: "constant"
            value: 1
        }
    }
}
layer {
    name: "relu2"
    type: "ReLU"
    bottom: "conv2"
    top: "conv2"
}
layer {
    name: "pool2"
    type: "Pooling"
    bottom: "conv2"
    top: "pool2"
    pooling_param {
        pool: MAX
        kernel_size: 3
        stride: 2
    }
}
layer {
    name: "norm2"
    type: "LRN"
    bottom: "pool2"
    top: "norm2"
    lrn_param {
        local_size: 5
        alpha: 0.0001
        beta: 0.75
    }
}
layer {
    name: "conv3"
    type: "Convolution"
    bottom: "norm2"
    top: "conv3"
    param {
        lr_mult: 1
        decay_mult: 1
    }
    param {
        lr_mult: 2
        decay_mult: 0
    }
    convolution_param {
        num_output: 384
        pad: 1
        kernel_size: 3
        weight_filler {
            type: "gaussian"
        }
    }
}
```

```
        std: 0.01
    }
    bias_filler {
        type: "constant"
        value: 0
    }
}
layer {
    name: "relu3"
    type: "ReLU"
    bottom: "conv3"
    top: "conv3"
}
layer {
    name: "conv4"
    type: "Convolution"
    bottom: "conv3"
    top: "conv4"
    param {
        lr_mult: 1
        decay_mult: 1
    }
    param {
        lr_mult: 2
        decay_mult: 0
    }
    convolution_param {
        num_output: 384
        pad: 1
        kernel_size: 3
        group: 2
        weight_filler {
            type: "gaussian"
            std: 0.01
        }
        bias_filler {
            type: "constant"
            value: 1
        }
    }
}
layer {
    name: "relu4"
    type: "ReLU"
    bottom: "conv4"
    top: "conv4"
}
layer {
    name: "conv5"
    type: "Convolution"
    bottom: "conv4"
    top: "conv5"
    param {
        lr_mult: 1
        decay_mult: 1
    }
    param {
        lr_mult: 2
        decay_mult: 0
    }
    convolution_param {
        num_output: 256
```

```

    pad: 1
    kernel_size: 3
    group: 2
    weight_filler {
        type: "gaussian"
        std: 0.01
    }
    bias_filler {
        type: "constant"
        value: 1
    }
}
layer {
    name: "relu5"
    type: "ReLU"
    bottom: "conv5"
    top: "conv5"
}
layer {
    name: "pool5"
    type: "Pooling"
    bottom: "conv5"
    top: "pool5"
    pooling_param {
        pool: MAX
        kernel_size: 3
        stride: 2
    }
}
layer {
    name: "fc6"
    type: "InnerProduct"
    bottom: "pool5"
    top: "fc6"
    param {
        lr_mult: 1
        decay_mult: 1
    }
    param {
        lr_mult: 2
        decay_mult: 0
    }
    inner_product_param {
        num_output: 4096
        weight_filler {
            type: "gaussian"
            std: 0.005
        }
        bias_filler {
            type: "constant"
            value: 1
        }
    }
}
layer {
    name: "relu6"
    type: "ReLU"
    bottom: "fc6"
    top: "fc6"
}
layer {
    name: "drop6"
}

```

```

type: "Dropout"
bottom: "fc6"
top: "fc6"
dropout_param {
    dropout_ratio: 0.5
}
}
layer {
    name: "fc7"
    type: "InnerProduct"
    bottom: "fc6"
    top: "fc7"
    # Note that lr_mult can be set to 0 to disable any fine-tuning of this, and any other, layer
    param {
        lr_mult: 1
        decay_mult: 1
    }
    param {
        lr_mult: 2
        decay_mult: 0
    }
    inner_product_param {
        num_output: 4096
        weight_filler {
            type: "gaussian"
            std: 0.005
        }
        bias_filler {
            type: "constant"
            value: 1
        }
    }
}
layer {
    name: "relu7"
    type: "ReLU"
    bottom: "fc7"
    top: "fc7"
}
layer {
    name: "drop7"
    type: "Dropout"
    bottom: "fc7"
    top: "fc7"
    dropout_param {
        dropout_ratio: 0.5
    }
}
layer {
    name: "fc8_flickr"
    type: "InnerProduct"
    bottom: "fc7"
    top: "fc8_flickr"
    # lr_mult is set to higher than for other layers, because this layer is starting from random while the others are pre-trained
    param {
        lr_mult: 10
        decay_mult: 1
    }
    param {
        lr_mult: 20
        decay_mult: 0
    }
    inner_product_param {
}

```

```
    num_output: 20
    weight_filler {
        type: "gaussian"
        std: 0.01
    }
    bias_filler {
        type: "constant"
        value: 0
    }
}
layer {
    name: "accuracy"
    type: "Accuracy"
    bottom: "fc8_flickr"
    bottom: "label"
    top: "accuracy"
    include {
        phase: TEST
    }
}
layer {
    name: "loss"
    type: "SoftmaxWithLoss"
    bottom: "fc8_flickr"
    bottom: "label"
    top: "loss"
}
```

Appendices D

