CSE 190a Project Report:
Golf Club Head Tracking

Ravi Chugh
rchugh@cs.ucsd.edu

Krystle de Mesa
kdemesa@cs.ucsd.edu

Abstract

Computer vision and graphics technologies have been used extensively in developing golf instruction and entertainment systems. In this project, we work towards constructing the trajectory of a golf swing from high quality video, since this information can enable many of these types of applications. We build on previous work that estimates a 2D spatio-temporal trajectory model from video, and we explore different ways of improving the estimation procedure. We also identify the importance of incorporating multiple camera angles to construct full 3D trajectory models.

1. Introduction

Innovations in computer graphics and computer vision have been applied in various golf-related applications – motion capture systems to provide detailed swing analysis, video systems to supplement golf instruction, high-precision launch monitors to track club and ball physics, and video games that continue to look more and more realistic. The ability to digitally analyze a golfer's swing can be used in both of these major applications: analyzing a golf swing and producing a 3D animation of a golf swing.

In this project, we work towards reconstructing the trajectory of a golf swing from video. Motion capture can be used to create very precise 3D models of a golfer's swing, but few have access to such studios. Most golfers have access to digital cameras, however – and many golfers record their swings to analyze and share with others – so work towards extracting swing information from videos may be useful to many golfers.

Most affordable cameras record video at a maximum of 30 frames per second. Because the golf club often travels over 100 miles per hour, there can be severe motion blur in videos recorded at this frame rate. We would like to eventually analyze videos of this quality, but for the scope of this project we will work with higher quality data. Although we had access to high quality cameras that record up to 200 frames per second, we use the following alternate approach that simulates video from a high frame rate camera: we capture a golf swing at a much slower speed than normal. Figure 1 shows a golfer making three similar swings with the same club that take two seconds (normal speed), four seconds, and eight seconds. The latter two serve as a simulation for the the quality of data from cameras that record approximately 60 and 120 frames per second. We will work exclusively with swings made at quarter speed (about 120 frames per second).

Our approach is based heavily on methods outlined in [3], which tracks golf clubs in video data from the face-on angle. We will describe how our approach differs in certain aspects and how we are continuing to try to make our approach more robust. We also begin considering how to make use of additional camera angles to extract a swing trajectory in 3D space.

2. Related Work

Several papers use the golf swing motion as a test case for evaluating various tracking algorithms. Lepetit et al. [4] argue that approaches that recursively predict motion at time t (based only on the position at $t-1$) suffer from 1) sensitivity to a small number of bad predictions and 2) motion that is hard to represent. They propose an approach to tracking that considers an interval of frames before and after time t in making a local prediction for time t. They demonstrate their approach to track the motion of a bouncing tennis ball and the club head of a golf swing. They use four model parameters to represent the golf swing based on the

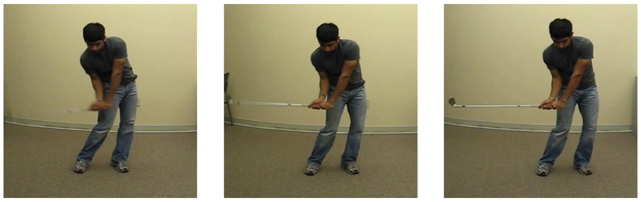

Figure 1. A golf swing that took about two seconds captured with a point-and-shoot camera that records 29 fps. Similar swings made with the golfer taking four and eight seconds, thus simulating the approximate quality of 60 and 120 fps cameras.

widely-taught "double pendulum" golf swing theory, and they are able to achieve online tracking with a slight delay (to account for interval size).

Gehrig et al. [3] model the trajectory of the club head through simple polar, polynomial approximations. Using empirical evidence, they found that a typical upswing can be accurately and stabally modeled by a 4th degree polar curve, and a typical downswing by a 6th degree one. After a series of processing steps to identify hypotheses for where the clubhead is in each frame, they use a RANSAC-like method to approximate these polar least-squares curves. They use a similar robust fitting to estimate the speed of the clubhead over time as well. Our work in this project is based largely on the approach described in this paper, and we will address how each step in our approach compares.

Using data from a video capture sequence to generate animated 3D models has also been explored in graphics and vision. [1] presents a data-driven method that uses a pose deformation model and a separate model of variation (based on body shape) to construct a 3D surface model with realistic muscle deformation for different people in different poses. Balan et. al [2] apply the previous method to generate 3D human body models from image data. Utilizing this SCAPE model [1], the parameters of the 3D body mesh is estimated directly through a cost function that measures the model's accuracy given image observations. A human tracking algorithm is used to initialize a stochastic search that helps optimize fitting the body shape and pose from the data.

Probabilistic approaches have also been applied to recovering 3D models from 2D data. Sigal and Black [5] estimate 3D human poses from monocular images by utilizing a hierarchical Bayesian inference framework that processes body part detections and computes a probability distribution over the different body poses comprised by these parts. The 2D poses are then probabilistically mapped onto a 3D pose through belief propagation, which infers 2D limb poses that are consistent with human body models.

3. Hypothesis Generation

The first major component of our project is the identification of the golf club in each frame. The pipeline for this component is summarized in Figure 2.

3.1. Motion Detection

We use a similar approach to the one described in [3] to identify the changing pixels in each frame, which assumes that the background is constant. We first scan each pair of consecutive frames to identify the pixels that differ, according to a color distance threshold. The difference between each pair of frames f_i and f_j is stored as a binary image d_{ij}. For a sequence of n frames, $n-1$ such "diff" images

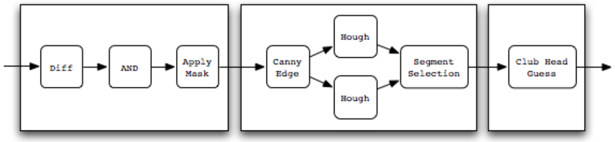

Figure 2. The three major steps in guessing where the club head is in each frame. First, the moving pixels in each image are identified. Then, line segments in these areas are detected, and the best guess for the club shaft is selected. Finally, the endpoint of each segment that is most likely the club head is identified.

are computed. We then consider each pair of consecutive diff images d_{ij} and d_{jk} to identify those pixels that correspond to motion at time j by pixel-wise ANDing d_{ij} and d_{jk}. The resulting binary image m_j is a mask that identifies the moving pixels in f_j, which will contain the moving parts we are interested in: the golfer and the club. Note that there are $n-2$ binary mask images. Figure 3 shows the construction of a mask from a particular sequence of three frames.

These binary masks suffer from two problems. The first is that even though we are working with synthetically high frame rate data, the club travels so fast that motion blur sometimes prevents the club from being detected in the diff images (and therefore also the masks). The second problem is that even though we are using a fixed camera, there are sometimes slight variations in background pixels, and these are evident in the masks. To address these problems, we

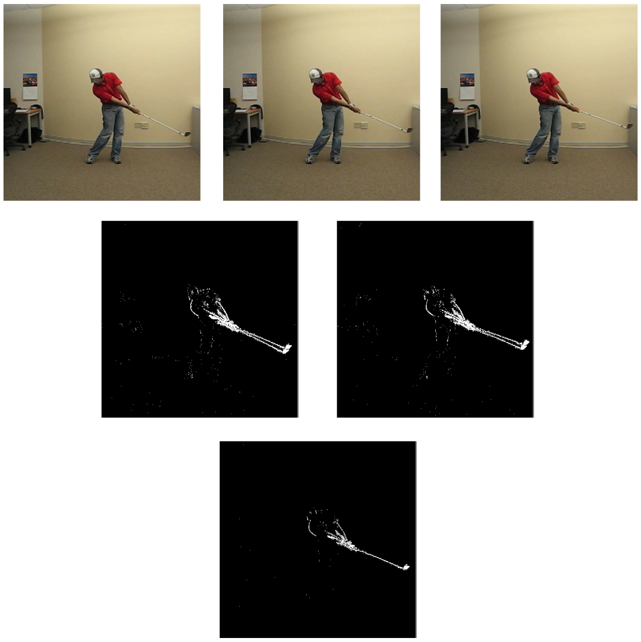

Figure 3. For each triple of consecutive frames, we compute a binary mask of the moving pixels by differencing the two pairs of consecutive frames and then ANDing these differences.

follow the method in [3] of applying a morphological closing on the masks to eliminate some of the noise and smooth out possibly-disconnected signal. Figure 4 demonstrates the kind of improvement that a closing can provide.

3.2. Club Shaft Detection

Next we try to identify the shaft of the golf club in each frame, following [3]. To do this, we look for line segments in the moving parts of each frame. We apply the mask m_i to the frame f_i to isolate the moving pixels in that frame. We convert this image to grayscale and use Canny edge detection to identify the boundaries of moving objects.

The Hough transform is then applied in order to identify line segments among the set of edges found by Canny. In the ideal case, there is exactly one line segment detected by Hough and it corresponds to the club shaft. In practice, however, the results from the Hough transform are a major source of difficulty. If we use a strict threshold, most frames do not have any Hough lines. The frames that do, however, are likely to have a small number that correspond to the club shaft. If we use a loose threshold, most frames have a lot Hough lines, the vast majority of which are unrelated to the club. Figure 5 provides an example of the Hough lines for a particular frame using two different thresholds.

Because neither threshold alone provides enough information, our approach is to employ the results from both. For each frame, we collect Hough lines found with a strict threshold (we refer to these lines as GOOD lines) and Hough lines found with a looser threshold (BAD lines). Then, to hypothesize where the club shaft is in each frame, we run the following abstract procedure:

1. If there is a GOOD segment, choose it.
2. Otherwise, choose the best BAD segment, if any.

Step 1. We have observed that GOOD segments fall into several categories: 1) there is one segment that is co-linear with some part of the shaft; 2) there are two segments that are co-linear with either edge of the shaft; 3) there are three, four, or more segments that are co-linear with either edge of the shaft; and 4) there are several segments that do not necessarily correspond to the shaft.

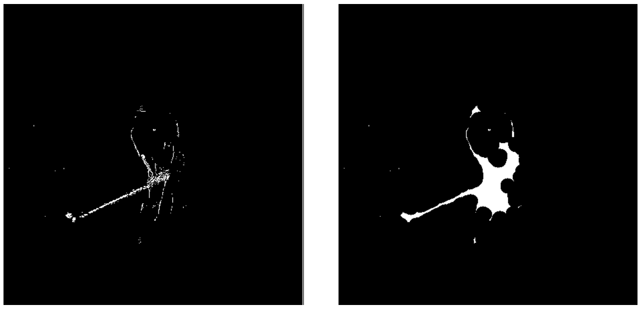
Figure 4. Before and after applying a morphological closing to a binary motion mask. In this case, it succeeds in smoothing out the region for the club shaft, but it also exposes more of the image. This is a tradeoff between creating the opportunity to identify the actual club shaft as well as potentially more erroneous segments.

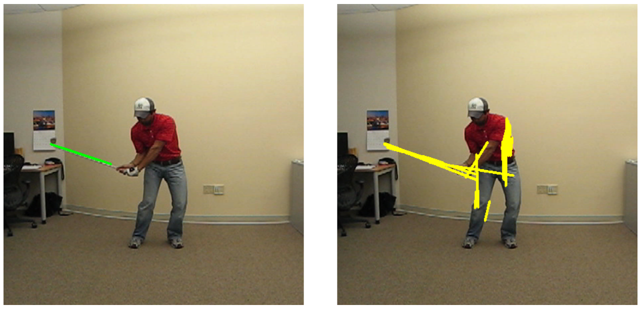
Figure 5. Hough transform lines using two different thresholds. The strict threshold does not identify as many segments, but when it does they usually correspond to the club shaft. The looser threshold on the other hand usually identifies many segments that are unrelated to the club.

We have empirically found that the first two cases are the most common with the particular strict threshold we use, so our simple approach to choosing GOOD segments caters to these cases. If there is exactly one segment, we choose it. If there are two segments, we merge them into one by averaging their endpoints and choose it. Otherwise, we conclude that there is no GOOD hypothesis for this frame. An improvement on this algorithm would be to consider the case where there are more than two segments and look for parallel and co-linear segments that could be merged and used as a hypothesis.

Step 2. For the looser threshold we use, the majority of the line segments are unrelated to the club, but we have observed that there are usually at least some segments that do correspond to the club. We therefore try to identify the most likely BAD segment that corresponds to the club by choosing the one that minimizes a simple error metric. We define this error to be a combination of the difference in slope and distance of endpoints as compared to the hypothesis for the previous frame. The BAD segment that minimizes this error is chosen as the hypothesis for the frame.

3.3. Club Head Detection

The final step in our hypothesis generation is to predict which end of each hypothesis corresponds to the club head and which end corresponds to the golfer's hands. We take advantage of the fact that the golfer's hands are above the clubhead at the beginning of the swing to define a reference point: in the first GOOD segment we find, we choose the endpoint with the smaller y coordinate (higher in the picture). Then, because we know the golfer's hands will (almost) always be closer to this initial hand position than

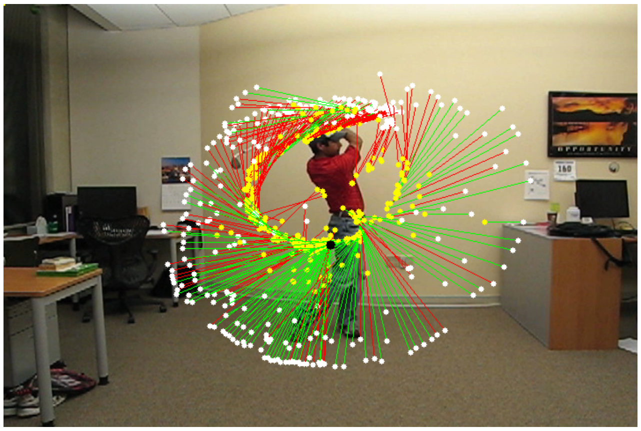

Figure 6. The green segments correspond to GOOD segments when there are any, and the red ones correspond to the best BAD segments. The white dots indicate the club head hypothesis in each frame. These hypotheses are used in the next phase to estimate polynomials that describe the swing trajectory in time and space.

the club, we guess that the club head in each frame is the endpoint of the segment furthest from this reference point.

Figure 6 shows results from hypothesis generation.

It is worth discussing the differences in our approach compared to that of [3]. Our approach to generating hypotheses diverges from theirs after we compute motion masks. Whereas their approach allows multiple hypotheses per frame, we try to choose the best segment as the single hypothesis for a given frame. In their paper, they discuss issues with inaccurate candidate hypotheses, and they try to improve them in a couple of ways. First, they eliminate segments that correspond to physically impossible club positions. Second, they try to improve some hypotheses that are too short by extending them in either direction by using color information.

In contrast, our approach to choosing the best BAD segment (based on slope and position) corresponds loosely to removing physically impossible hypotheses. Our approach employs no means for extending segments, however; we have observed that there are typically enough accurate hypotheses along the club head trajectory so as not to warrant it.

4. Trajectory Estimation

The second major component is taking the club head hypotheses and estimating a model for them in time and space. For this, we follow the approach taken in [3]. The pipeline for this component is summarized in Figure 7.

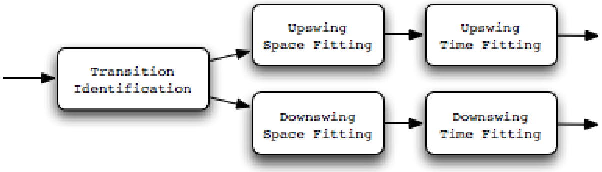

Figure 7. The upswing and downswing are processed separately, after obtaining the transition frame either automatically or from the user. Each part of the swing is fit with a polar curve using a RANSAC-like process. The inliers of each of the models are then used to estimate the trajectory as another polar function in time.

4.1. Transition Identification

We first need to identify which frames correspond to the upswing and which to the downswing. [3] evaluates the average y coordinate over time to conclude where the transition between upswing and downswing occurs. We use a slightly different approach. We consider the change in slope of hypotheses in consecutive frames, and we look for places where the derivative is zero, since the club must come to a stop when transitioning. Out of the candidate frames with zero (or close to it) derivative, we use an error metric to guess which is the transition. The error metric includes proximity to the middle of the video and use of GOOD segments over BAD. We have found our approach to identifying the transition to be fragile, however, so it might be worth trying the approach taken in [3]. As a temporary solution, we ask the user to provide the frame in which the transition from upswing to downswing occurs.

4.2. Computing Space Models

Once we have the transition, we process the upswing and downswing frames independently. As empirically determined in [3], we seek to fit a 4th degree polar curve to the upswing hypotheses and a 6th degree polar curve to the downswing ones. For this we convert all hypotheses from Cartesian to polar coordinates, but we use the standard direction of the θ axis, unlike the one used in [3].

We use a RANSAC approach to estimating these curves by randomly sampling a subset of hypotheses. The sample size is two more than the parameters of the polynomial (five points for the upswing and seven points for the downswing) to overconstrain the curve to prevent unstable behavior. For each randomly selected subset of points, their least squares fit is computed. Once these parameters are computed, each hypothesis is checked to see if it is within some error threshold away from the value predicted by the curve. This process is repeated many times, and the model with the highest number of inliers is selected.

Figure 8 shows the trajectories estimated on both parts of a swing. The results vary on our set of sample videos. The

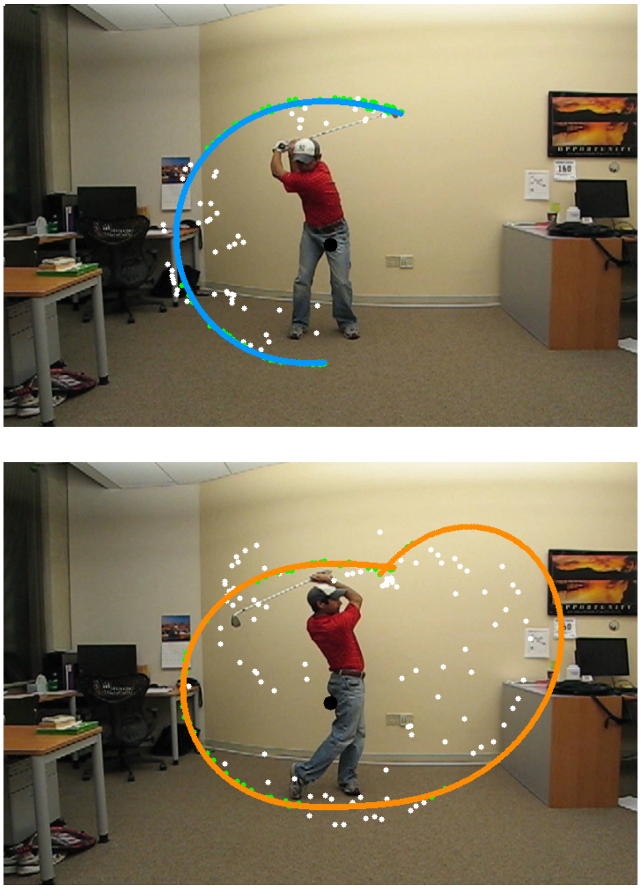

Figure 8. A fourth degree polar curve estimated for the upswing and a sixth degree polar curve estimated for the downswing using a RANSAC procedure.

upswing fitting is generally pretty good, but the downswing results are largely inaccurate.

4.3. Improving Downswing Fitting

There are two major problems with downswing fitting:

- There are typically few good club head hypotheses in the last part of the downswing, after the golfer's hands pass from the second quadrant into the first quadrant.
- When passing from the second quadrant to the first quadrant, the angles of our hypotheses need to be offset by 2π.

Our current procedure for addressing the second of these concerns is fragile and is susceptible to failure when we have bad hypotheses around the 2π portion of the swing. The first problem results from particularly bad segment identification in that portion of the downswing.

We have decided to eliminate these problems by simply considering the downswing only until the club head reaches 2π. Although we will eventually return to fitting the entire downswing, we have made this simplification to make further progress on at least a large portion of the downswing.

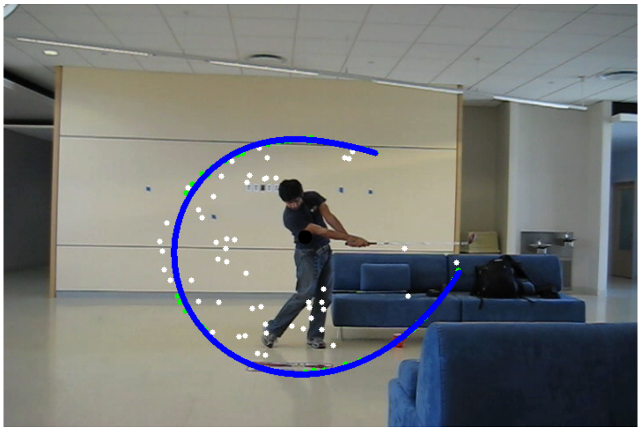

Figure 9. After obtaining the frame in which the downswing crosses over 2π, we fit this portion of the downswing with a fourth degree polar curve.

We show the user our club shaft and head hypotheses for each frame and the reference point we use as the origin, and we ask the user to supply the frame in which the club crosses 2π. We then fit the upswing and the downswing up until this frame. Because the portion of the downswing we are considering is similar in shape to the upswing, we use a 4th degree polar curve to model it.

Figure 9 shows an example where this process works fairly well for the downswing. But in general, this process still fails to provide consistently accurate models.

We have noticed that on all of our videos and for both upswings and downswings, the extreme hypothesis in every direction (up, down, left, right) over the course of the swing portion is in fact an actual location of the club head. These hypotheses are unfortunately often not well-fit by the model that RANSAC chooses. Put another way, we have never observed any outliers that are beyond the reach of the swing trajectory; they are all inside it.

This observation suggests that we want to try and fit the outermost hypotheses more than the inner ones. We tried a simple weighting scheme in the counting of inliers during the RANSAC procedure. We allowed the error threshold to increase as the distance from the origin increases, but this approach does not produce accurate results. More points end up agreeing with the final model, but usually because of the increase in error threshold and not because of a good fit.

Our next approach was to bias the samples that RANSAC selects to compute models. Because we want the extreme points to heavily influence the resulting model, we require that RANSAC choose all four of them as part of its sample. It then randomly selects two more points and then computes a model as usual. It is important to note that

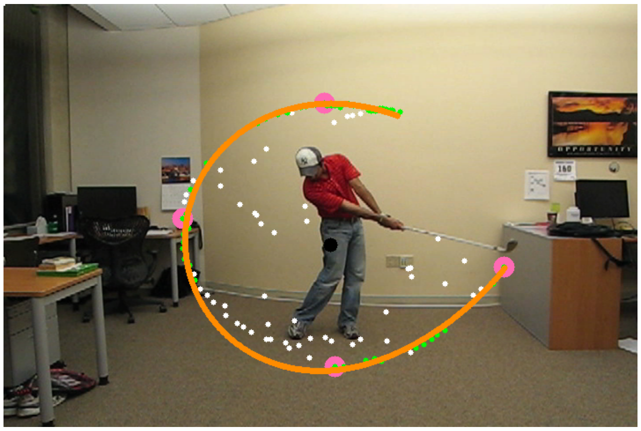

Figure 10. The best model return by guiding RANSAC with the four superdelegates – the extreme hypotheses in every direction, whose votes count for more than the other points.

although each sample contains the four extreme points, the best model does not necessarily fit each of these points well; it is simply more likely that it will. This approach does in fact improve the quality of the best model on most videos, but there are still some with undesirable results.

The last way in which we guide RANSAC towards a better solution builds upon the previous one. Not only do we require RANSAC to use the extreme points in computing models, we also give more weight to them when they agree with a model. We consider the extreme points to be "superdelegates" because when they agree with a model, they have more influence over whether the model will be kept than all other points that agree with the model.

The danger in biasing RANSAC in these two ways is that the popular vote solution computed by random consensus may be overturned in favor of a model that is possibly overfit to these four extreme points. However, on the set of sample videos we have used, this guided RANSAC approach produces consistently acceptable results. Figure 10 shows one such result.

4.4. Computing Time Models

Once we have estimated the magnitude of the club head as functions of angle, we then estimate the angle of the club head as functions of time. As in [3], we use the inliers for each of these curves to compute the least squares fits for the desired time models. Examples of complete spatio-temporal models can be found on the project website.

5. Additional Camera Angles

The approach we have described produces a 2D estimation of the golf swing. What we would like to produce, however, is an estimation of the swing in 3D. Having a complete representation of the swing in space would be useful for applications to golf instruction and to animating a golfer's swing. To produce a 3D estimate, additional camera angles are needed. In addition to the face-on angle we have been dealing with, the down-line (directly behind the golfer on the target line) and up-line (directly in front of golfer on the target line) are the other two popular camera angles for golf swing instruction and analysis.

To demonstrate the additional information provided by the down-line angle, for instance, in Figure 11 we show three different positions at the top of the swing from the face-on and down-line angles.

We have run the early stages of our pipeline – motion detection and line segment detection – on videos from these two perspectives, and the quality of the line segments we have observed are encouraging. We leave it to future work to identify which end of the segments correspond to the club head and how to estimate the motion of these club head hypotheses as functions of time and space.

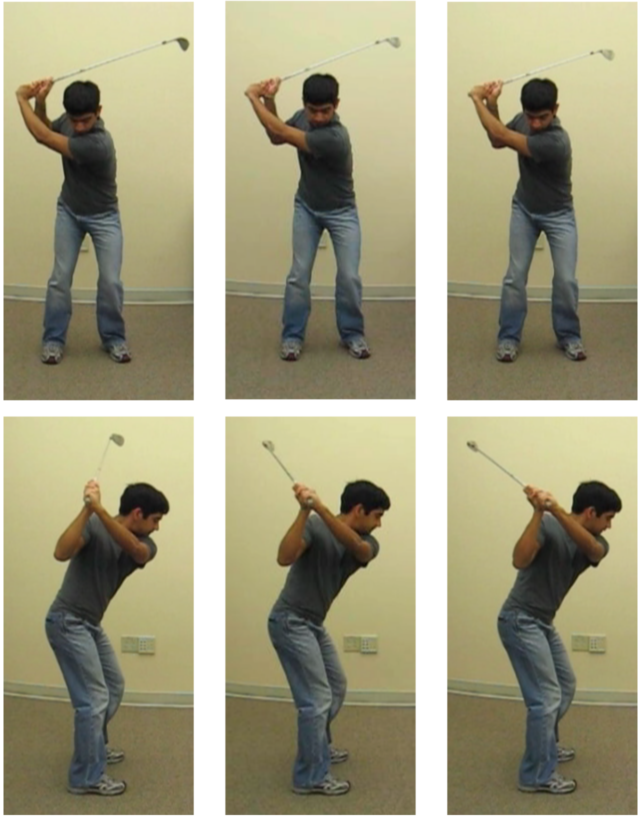

Figure 11. Three different positions at the top of the backswing. Although they look similar from the face-on view, the down-line view reveals significant differences. Thus, analyzing additional perspectives is crucial for constructing 3D trajectory models. Note: the images for the face-on and down-line view were taken separately, while the golfer re-enacted the same positions.

6. Conclusion

We have made progress towards our goal of reconstructing a golfer's 3D swing trajectory out of videos from multiple camera angles. Our approach to tracking the club from the face-on view is based heavily on previous work in [3], but our algorithm for identifying and selecting hypotheses for the club head is fundamentally difficult. Our results are preliminary but promising, and improving the robustness of several steps in our pipeline is likely to produce very good results in the face-on view.

An aspect that requires further consideration is the end of the downswing, the portion that occurs in the first quadrant. We have excluded this section for simplicity, but separately fitting points in this region sometimes produces very good results, as in Figure 12. Future work may choose to continue fitting the downswing in these two separate parts or return to the approach that tries to fit the entire downswing at once.

Whereas [3] spent more effort in improving the quality of hypotheses, we focused more effort on overcoming the presence of many bad hypotheses in the fitting process. The guided RANSAC approach we employ demonstrates encouraging results, but as we have mentioned, second-guessing RANSAC can lead to models that are overfit to the superdelegates.

Future work may explore "safer" ways of guiding RANSAC. For instance, we have noticed that the motion history image summarizing all moving pixels over the course of a video provides a very good approximation for the span of the swing. One approach may be to use the boundary of such a motion history to seed RANSAC with more hypotheses. Figure 13 shows a way to approximate this boundary using a Radon transform.

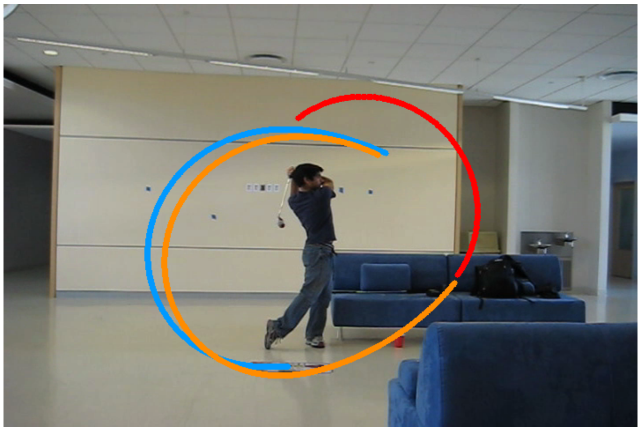

Figure 12. Results from guided RANSAC fits for an upswing and both parts of a downswing. In this example, fitting a second degree polynomial to the second part of the downswing produces a good result. More work is needed to reliably fit the second part of the downswing or to return to the approach that fits the entire downswing at once.

Finally, we have also identified some possibilities for club tracking from the down-line and up-line views, which can be used to reconstruct a complete 3D swing trajectory. Much of the future work for this project lies in this area. When successfully completed, 3D trajectory information will enable technology for use in swing analysis and in animation.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: Shape completion and animation of people. In *ACM Transactions on Graphics (SIGGRAPH)*, volume 24, 2005. 2

[2] A. Balan, L. Sigal, M. J. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, June 2007. 2

[3] N. Gehrig, V. Lepetit, and P. Fua. Golf club visual tracking for enhanced swing analysis tools. In *British Machine Vision Conference*, September 2003. 1, 2, 3, 4, 6, 7

[4] V. Lepetit, A. Shahrokni, and P. Fua. Robust data association for online applications. In *IEEE Computer Science Conference on Computer Vision and Pattern Recognition*, 2003. 1

[5] L. Sigal and M. J. Black. Predicting 3d people from 2d pictures. In *IV Conference on Articulated Motion and Deformable Objects, AMDO*, 2006. 2

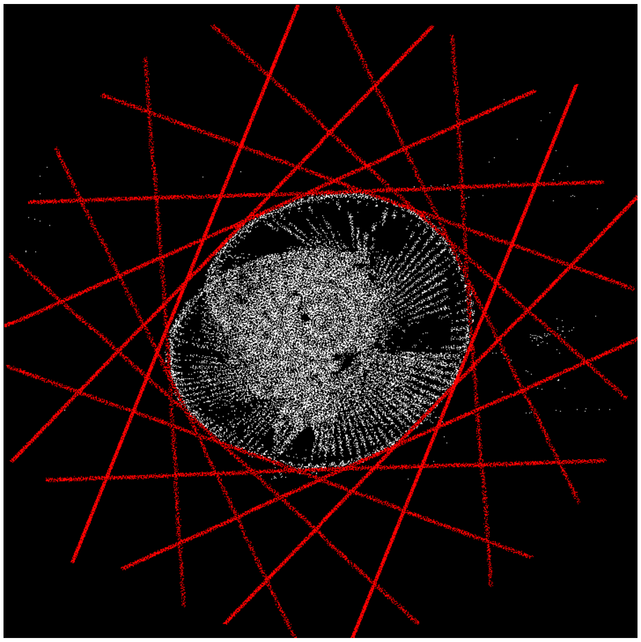

Figure 13. The boundary of a motion history image using a Radon transform. The resulting polygon is a good approximation for the trajectory of the swing, and it may be useful to guide RANSAC towards points along its boundary.