Fine Grained Action Recognition in Sports Videos

Amit Nagpal
(anagpal1@stanford.edu)

Predicting

Video, today, is searched and browsed primarily by its cover. To make "Stephen Curry's 3-pointers" accessible, publishers must create a separate video just for Stephen Curry's 3-pointers and label it as such. This is inefficient and has not scaled for obvious reasons. Moreover, an increasing population of sports fans do not wish to consume the entirety of every game. They are only interested in watching key games and "interesting parts" of all other games. Finer grained consumption of video content, through finer grained tagging such as action recognition, is key to the ecosystem of sports.

Unlike in everyday life, action recognition in sports is particularly hard due to the chaotic nature of game play. Occlusion is everywhere, players take on extreme body postures and subtle variations mean completely different moves.

We break our problem down into action proposal and action classification of short half a second sequences of sports videos. For this project we will focus on classifying a sequence of frames as containing a jump as a proposal of action and its location and player density to classify a sequence as a dunk or a three pointer.

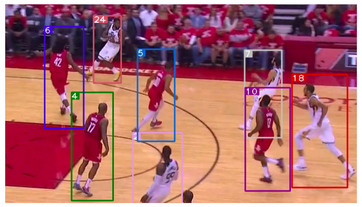
Data

240, five second videos that were weakly tagged to contain specific NBA moves. These video segments were extracted from 2 minute and 10 minute game highlight videos downloaded from nba.com. Videos titles were used to match them to Game schedules available at sports.yahoo.com. Rough 5 second video segments were then tagged to contain a move such as "dunk", "three pointer" and a "layup". This was done by matching game clock in the videos (using OCR) and textual play by play (using IR techniques). Overall the data was noisy.

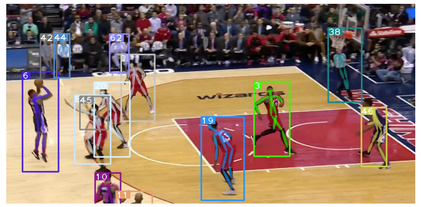
Features

Due to scarcity of training data and noise, extensive work was required to extract features that we could use to provide meaningful semantics to our classifier.

2D pose - We annotate every frame of the 5 second videos with player poses [1]



Tracking - We use deepsort's [2] people model to track players



Matching - We then assign poses to tracks based on maximum body containment,

Basket as origin - we use Amazon rekognition to locate banner ad on basket holder to estimate basket location. We then transform all location and distance vectors with basket location as the new origin.

Aggregate features - (A) Per Frame - **Player density** - We calculate density of players in different regions of the court
(B) Per Track - **Y travel, Arms** - We calculate min, max change in Y and count of raised arms over last 5 frames

Models - SVM

X,Y = mean distance of pose from basket
dY = change in Y w.r.t. last 5 frames of a track
A = # of time pose's arms were up (last 5 frames)

Model-A: 3 frame sequence: jump / no-jump

Feature vector F = [min(dY), max(dY), A]

$$\underset{w,b,\xi}{\text{argmin}} \quad \frac{1}{2}||w||^2 + C\left[C_1 \sum_{\{i|y_i=1\}} \xi_i + C_{-1} \sum_{\{i|y_i=-1\}} \xi_i\right]$$
$$\text{subject to} \quad y_i(w^T x + b) \geq 1 - \xi_i.$$

C-SVM to counter class imbalance and a degree 3 polynomial kernel.

Model-B: 3 frame seq: Dunk / 3 ptr / Nothing

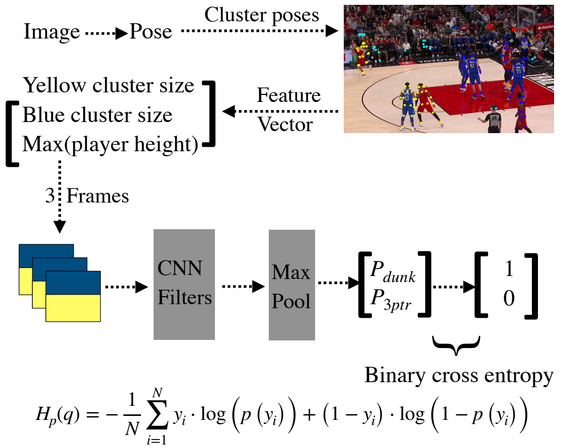
Feature vector F = [min(dY), max(dY), A, X, Y]

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\rho} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i$$
$$\text{subject to} \quad y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq \rho - \xi_i,$$
$$\xi_i \geq 0, i = 1, \ldots, l, \quad \rho \geq 0.$$

nu-SVM with degree 3 polynomical kernel and multi class classification to a dunk, 3 pointer or no move.

Models - CNN

Model-C: Change in player density: Dunk / 3 pointer



$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right) \cdot \log\left(1 - p\left(y_i\right)\right)$$

Results

Model	Train Accuracy	Test Accuracy
A. SVM - Jump / No jump	77.97% (308/395)	82.22% (74/90)
B. SVM - Dunk / 3 ptr / None	87.59 (346/395)	85.5556% (77/90)
C. CNN - Dunk / 3 ptr	<tbd - see paper>	<tbd - see paper>

Discussion

- We have taken an objective to break down moves into their semantic components, such as run, jump, dribble etc. And it works as we show it with SVMs. So, we don't need lots of training data for each individual move from the broadcast videos and can stitch moves with more generically learn sub-moves.
- Key frames are often occluded. Player poses are often missing key joints and limbs right when they are most needed. Right when the player hangs on the basket or jumps to make a dunk or three pointer.
- Tracking too has similar issues as pose when utilized for multiplayer sports and needs to be domain adapted.
- Audience introduces a lot of noise for indoor games such as NBA. It is often hard to distinguish a player in the backdrop of hundreds.

Future

Domain adapt Pose and Tracking models to work on multi player sports videos, so they don't miss the key frames.

Performance optimization. Currently each of 2D pose, Tracking, Matching, Basket location require long GPU and CPU cycles. Cut down on higher level inferences not required for our classification.

Utilize other weakly supervised data of basket ball coaching videos.

Borrow concepts from Homography to pin point exact location and elevation of a player in 3D space.

References

1. Zhe Cao et al. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". https://arxiv.org/abs/1611.08050
2. Nicolai Wojke et al. "Simple Online and Realtime Tracking with a Deep Association Metric". https://arxiv.org/abs/1703.07402
1. Victor Escorcia et.al. "Deep Action Proposals for Action Understanding". http://www.eccv2016.org/files/posters/P-2B-10.pdf
2. Huanyu Yu et. al. "Fine grained video captioning for sports Narrative". https://bit.ly/35LEndY