Analysis of Deep Learning Action Recognition for Basketball Shot Type Identification

Carlos Olea[1], Gus Omer[2], John Carter[2] and Jules White[1]
[1]*Dept. of Computer Science, Vanderbilt University, 1025 16th Ave. S, Nashville, U.S.A.*
[2]*NOAH Basketball, 26958 US Highway 72, Athens, U.S.A.*

Keywords: Sports Analytics, Action Recognition, Computer Vision, Deep Learning.

Abstract: Recent technologies have been developed to track basketball shooting and provide detailed data on shooting accuracy for players. Further segmenting this shot accuracy data based on shot types allows a more detailed analysis of player performance. Currently this segmentation must be performed manually. In this paper, we apply a state-of-the-art action recognition model to the problem of automated shot type classification from videos. The paper presents experiments performed to optimize shot type recognition, a unique taxonomy for the labeling of shot types, and discusses key results on the task of categorizing three different shot types. Additionally, we outline key challenges we uncovered applying current deep learning techniques to the task of shot type classification. The NOAH system enables the capture of basketball shooting data by recording every shot taken on a court along with its shooter and critical statistics. We utilized videos of 50,000 practice shots from various players captured through NOAH system to perform the task of classifying shot type. These three second video clips contain the shooting action along with movements immediately preceding and following the shot. On the problem of shot type classification, the Temporal Relational Network achieved an accuracy of 96.8% on 1500 novel shots.

1 INTRODUCTION

Today's game of basketball requires every player on the court to be able to shoot the ball well. However, the accuracy of a shot can vary wildly depending upon shot attributes such as the movements leading up to the shot (e.g. step back, side step, dribble pull up), court position, defender position, or if a pass was received immediately preceding the shot. For example, assuming equal shooting skill, a player's accuracy when shooting free throws is generally higher than when from the three point line while guarded. There are a multitude of factors that affect shot accuracy (Williams et al., 2020) (Zhen et al., 2016), one of which is shot ”type”. Accuracy can vary significantly between different types of shots (Erčulj and Štrumbelj, 2015), as such it is very useful to be able to analyze player shooting accuracy by shot type.

Teams often devote considerable resources to the analysis and drafting/trading of players in the NBA. This devotion of resources is justified as the median NBA yearly salary of 2.6 million USD (ESPN, 2020) makes drafting/trading a player a significant financial commitment. While many factors are considered when evaluating a player's value, one of the most important is their shooting ability. The most basic approach to defining shooting ability is a player's shooting percentage for free throws, 2 point shots, and 3 point shots. A next step could be a more granular approach to defining 2 point and 3 point shots (e.g. less than 10ft vs greater than 10ft or below the break three vs above the break three). A finer approach still, is using technology such as NOAH's system to define key shooting metrics such as entry arc, depth in the basket, and left/right control. Finally, segmenting the data by shot type provides the most detailed look at the overall ability of a shooter. The ability to automatically and efficiently analyze a player's accuracy on each individual shot type is indispensable when attempting to optimize the performance of a team as it allows coaches to address weaknesses via practice, lineup optimization, or player acquisition. Therefore, determining shot type proficiency in an unbiased way is of utmost importance. Doing so is nearly impossible without the careful identification (hereby referred to as labeling) and cataloguing of the type of each shot taken by a player. Prior work has covered adjacent topics, such as player position esti-

Olea, C., Omer, G., Carter, J. and White, J.
Analysis of Deep Learning Action Recognition for Basketball Shot Type Identification.
DOI: 10.5220/0010650200003059
In *Proceedings of the 9th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2021)*, pages 19-27
ISBN: 978-989-758-539-5; ISSN: 2184-3201

mation (Mortensen and Bornn, 2019), tracking player movements (Nistala Akhil, 2019) and analysis of shot biomechanics (Zhen et al., 2016), but automated classification of shots into semantic categories representing types such as free throw, catch and shoot and off the dribble has yet to be performed.

The manual labeling of shot type from video is arduous, time-consuming, and requires expertise to perform properly. The approach herein is able to assign shot type at a rate of 0.06 seconds per 3 second shot video on two RTX 2080ti, vastly outpacing even the most proficient human labeler. These shot videos have been provided by NOAH, specializing in the capture and analysis of shot data. This research was performed in collaboration with NOAH.

In this paper, we:

1. Catalog experiments conducted and results from attempting to train a deep learning network to perform the task of basketball shot type classification.
2. Detail the challenges that make the task of shot type classification from video difficult.
3. Present a taxonomy for creating shot type labels, which provides rules for the non trivial task of which and how many tags an individual shot should receive.
4. Utilize the Temporal Relational Network activity recognition model to create a basketball shot classification system.
5. Present results from 1,500 novel shots taken of players on NBA and NCAA teams and show strong results on classifying the labels of free throw (96.4% accuracy), catch and shoot (94.8% accuracy) and off the dribble (99.6% accuracy)

The remainder of this paper is organized as follows: Section 2 discusses important concepts in activity recognition, presents some of the state of the art models and discusses adjacent work, Section 3 describes the model we selected for the task of shot type classification, Section 4 discusses experiments performed and empirical results, and Section 5 discusses key challenges found for the task of shot type classification.

2 PRIOR RESEARCH ON ACTIVITY RECOGNITION

Activity recognition is a sub-field of computer vision concerned with assigning semantic labels, such as running, jumping, talking or playing a game to (usually human) actors in a video. Subsets of activity recognition range from general activity recognition, such as labeling videos for dancing, running, playing baseball, or working out to more specific tasks such as labeling hand gestures. The problem of labeling basketball shots by type is considered a part of specialized action recognition.

Labeling basketball shots by type is action recognition where the action classes are different types of shots. Most state of the art action recognition is done using deep learning which necessitates the usage of a large dataset. Therefore, we also use a deep learning approach based on tracking features between frames of videos (Zhou et al., 2018).

Before discussing our results and the challenges specific to the problem of shot classification, it is important to describe the dataset that is required for this task. Datasets in action recognition vary both in the generality and number of their labels.

The generality and variability in activity recognition dataset labels are referred to as the ”scope” and the ”variance” of their inputs (videos), hereby referred to as domains. Variance of input videos is constituted by differing background, lighting, actors, actions, etc. A dataset with a high number of general labels would have hundreds or more labels of actions from dancing to skiing to sleeping. Alternatively, a dataset with a smaller amount of specific labels may have 40-50 labels of different dance moves. Using the example of the theoretical dance move dataset, a dataset with a (relatively) wide or unconstrained domain may have instances of those 40-50 different labels in many different contexts, from dancing on the beach, to dancing in a crowd, to dancing on the street from a top down view. An example of the dataset with a highly constrained domain would be those same 40-50 classes of dance moves, but all with videos filmed in front of a green screen from the same distance and angle.

Datasets, such as Moments in Time (Monfort et al., 2019) have both wide scope and general domain with Moments containing well over 300 labels of vastly different activities. Datasets, such as sports1M (Karpathy et al., 2014) and Charades (Sigurdsson et al., 2016) both have arguably more specific scopes. Although sports1M has over 400 labels, both the labels and the domain are restricted to sports (and real life video capture, as Moments contains even cartoon actions). The same can be said of Charades and indoor (usually domestic chore) activities. Lastly, datasets, such as Jester (Materzynska et al., 2019) and Something Something (Goyal et al., 2017) have specialized scopes and constrained domains, being restricted to hand gestures or interactions at a standardized camera distance and angle. It should be noted that the wider the scope and the less constrained the domain, the more knowledge must be encoded into a model to accurately classify inputs. This is mentioned because the dataset that will be utilized falls into a specialized and constrained category and influences our model selection.

Some of the top deep learning models used for action recognition now include the following:

- Channel Separated Network for Action recognition: Utilizes group convolution where filters only receive input from output within their group, as opposed to all output nodes.
- Temporal Shift Module Network: Utilizes shifts along the temporal axis of a video to leverage efficient 2D techniques for 3D tensors (videos).
- Two-Stream Inflated 3D ConvNet: Based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and their parameters.
- Temporal Relational Network: Utilizes temporal relations between sets of n frames to extract additional temporal features to be used in prediction.

A considerable amount of work has been done in related categories. Work such as (Johnson, 2020) and (Nistala Akhil, 2019) revolve around player movement during game time, with an emphasis on positioning in certain game contexts (Tien et al., 2007) and trying to understand player positioning and ball tracking to detect possessions. Beyond the differences in goal, (Tien et al., 2007) utilizes a single angle camera, akin to what is publicly available in broadcast games while (Johnson, 2020) uses the SportVu capture system to gather data for analysis. There is a critical difference both in context and goal for these works, namely that we seek to categorize by shot type, a mostly disjointed task from those mentioned, and our primary dataset comes from recording practice/training sessions.

(Foster and Binns, 2019) and (Kaplan et al., 2019) constitute other works in adjacent areas, however these deal primarily in indirect performance analysis and economic decisions and impacts related to drafting and player performance, rather than analyzing and annotating data for game and player analysis and description.

3 ACTIVITY RECOGNITION WITH THE TEMPORAL RELATIONAL NETWORK

Model selection was performed by aggregating top performing single (non ensemble) models in the area of action recognition. The primary reason for this choice is to support online classification. Although several other models were considered, the decision was made to use the Temporal Relational Network detailed in (Zhou et al., 2018). Our reasoning lies in the explanations and definitions given in the opening statement of section 2. This category of action recognition is both small in scope and small in domain. The domain is quite small as the contexts in which the actions can appear are relatively few, with the main difference being court marking color schemes, occasional non-target actors and low angle variance. Similarly, the scope is restricted to only 3-5 classes depending on which classes were selected for classification. The performance of the Temporal Relational Network on datasets with a similarly small scope and domain such as the Jester and Something Something datasets presented strong performance on small domain action recognition tasks.

As mentioned before in Section 2, the temporal relational network utilizes a technique from (Wang et al., 2016) where multiple frames from a video are sampled, the number of which may vary, and their features (derived using a standard CNN backbone) are fused and fed through a multi-layer perceptron. This process is end to end differentiable, and can be trained alongside the CNN. The multiscale version (which we utilize in this paper) selects multiple *n* frame relations to be used in training and prediction.

4 EMPIRICAL RESULTS FROM SHOT TYPE IDENTIFICATION WITH THE DATASET

4.1 The Dataset

The characteristics of the dataset are an incredibly important facet of the task of identifying shot type. NOAH provided a dataset containing 50,000 3 second videos with accompanying shot type annotations. Shot instances are collected via a multi-camera system affixed above the backboard, along with a depth sensor, and back-end image processing.

The capture system itself varies slightly from court to court, though the angle of incidence (discussed in 5.2) remains similar across locations. There are 4-5 RGB cameras each assigned to a differing (but partially overlapping) section of the court. The backend consists of a computer vision model that is used to both spatially and temporally crop videos to consist mostly of the shooter, the ball, and any immediately adjacent objects or players. This results in a set of 3 second videos that are used for shot identification.

Shots are carefully labeled and audited by NOAH personnel. Shots are given an array of labels, with at least one base shot type label and any number of appropriate additional, non base shot type labels. This is addressed in detail in 4.2. Videos are captured mostly from practice sessions of NBA and NCAA teams. The distribution of the labels over 24,480 shots is shown in Figure 1.

4.2 Tagging Semantics and Distribution

An important task for experimentation was determining and procuring labels. It was not immediately clear what label(s) should be used for a given shot nor the structure of the label(s). Further, there are no publicly available datasets with basketball shot labels. Tagging semantics is critical to ensure there is no label overlap and while still covering the experimental space completely. For example, the dataset initially provided by NOAH contained the ”jumpshot” label. However, both a catch and shoot and a step back shot could be labeled jumpshots. The shot annotations were reviewed and relabeled to remove the ”jumpshot” label. While many shot types exist, a number are rare and difficult to find and label. One example of this would be the disparity between catch and shoots and jab step fakes. This holds true in the dataset we utilized in training our model and could pose a considerable challenge to future attempts to classify less common shot types.

To address the issue of tagging semantics and scope, we introduce a shot type tagging system. Tags are divided into two distinct groups: base shot tags and supplementary tags. Base shot tags are required tags that are mutually exclusive; any given shot can only have one. Examples of this tag would be catch and shoot, free throw, jab step fake and step back shot. Supplementary tags may be applicable to one or several base shots, but are not necessary for the task of classifying shots unless it is at a level of granularity that requires it’s integration. The final set of base shot types used were as follows: Free throw, Catch and Shoot, and Off the Dribble. While this is only a fraction of the shots that could fall under the category of base shots, we lacked a sufficient number of examples for additional shots to be reliably classified, namely jab step fake and step back shot. The disparity in class distribution is shown in Figure 1.

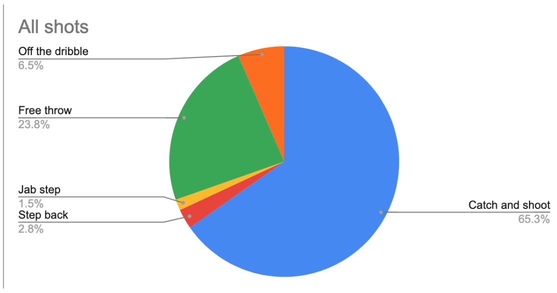

Figure 1: Image depicting the distribution of classes in the dataset.

4.3 Empirical Results on the Dataset

Several experiments were performed to determine the accuracy of our Temporal Relational Network tuned for shot type identification. They will be detailed here in chronological along with relevant results found in each.

4.3.1 Free Throw vs Non Free Throw

The initial task was to assign the labels ”free throw” and ”not free throw”. In this task we achieved an accuracy of 95.5% given equal amounts of both classes with no alterations to the base TRN network, using only RGB input. We believe this task came easily in large part because the location of free throws (and the camera used to capture them) are fixed and as such, locational features such as court markings, as well as the uniformity of the shots within the class ”free throw” could be used to easily identify them. Shot tags in the ”not free throw” class were catch and shoot (from any direction), jumpshot, step back shot and jab step fake.

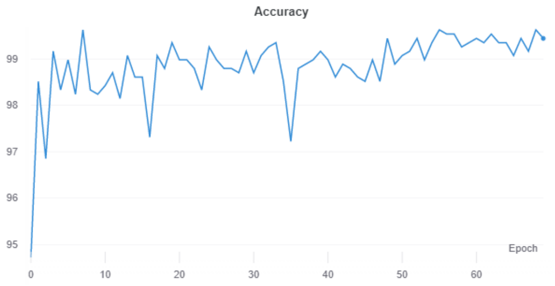

Figure 2: Validation accuracy (Free Throw vs Non Free Throw).

4.3.2 Free Throw vs Catch and Shoot vs Jumpshot

The ternary task of assigning the labels free throw vs catch and shoot vs jumpshot provided the first major challenges, as this dropped class accuracy to roughly 95%, 70% and 60% respectively. This was initially thought to be because of the variable position of ”catch and shoot” and ”jumpshot” shots on the court causing too much difference between shots of the same class. This was incorrect, however it was the cause of the next experiment.

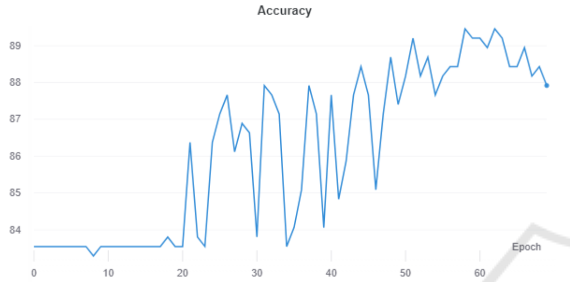

Figure 3: Validation accuracy (Free Throw vs Catch and Shoot vs Jumpshot).

4.3.3 Free Throw vs Catch and Shoot vs Jumpshot, Partitioned by Shot Location

To ameliorate the ”problem” of radically differing court position between shots, we implemented a system to partition shots based on their court position, and trained three sets of weights for the TRN model. One set for the portion of the court left of the key (respective to the backboard), one for the area behind the key (including the free throw line) and one for the area of the court to the right of the key. Before results for this change alone were found, it was discovered that the ”Jumpshot” tag violated the rules of our tagging semantics. It was treated as a ”base shot type” as opposed to a modifier. For this to be the case, it must be mutually exclusive of other base shot types, which it was not. This was heavily corrupting the training process, as catch and shoot shots were labeled as either catch and shoot or jumpshot. As seen in Figure 3, this was the cause of significant thrashing.

4.3.4 Free Throw vs Catch and Shoot vs Off the Dribble vs Jab Step Fake vs Step Back Shot

To address this, the jumpshot tag was removed and a new tag ”off the dribble” was introduced.The task then shifted to a quinary labeling task between the labels free throw, catch and shoot, off the dribble, step back shot and jab step fake. On this task, testing accuracies of 98.3%, 97.1%, 95.5%, 75.5%, and 60.8% respectively were achieved. During these tests, the shots were still being partitioned based on court position.

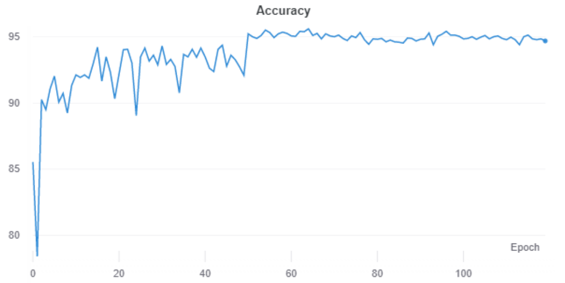

Figure 4: Validation accuracy (Free Throw vs Catch and Shoot vs Off the Dribble vs Jab Step Fake vs Step Back shot).

4.3.5 Free Throw vs Catch and Shoot vs Off the Dribble vs Other

To facilitate better accuracy on the minority shot types (e.g. step back shot and jab step fake) and ease of use in an environment with many shot types with few examples, we attempted to reduce the problem to a quaternary problem: Free throw vs Catch and shoot vs Off the dribble vs Other. This however resulted in inferior results, with the ”other” tag being completely unused by the model and the Free throw, Catch and shoot and Off the dribble tags being assigned with accuracies of 93.2%, 94% and 52.7% respectively. The poor accuracy of the off the dribble is due to a high rate of false positives from the jab step fake, step back shot and miscellaneous other categories.

We believe there are two possible causes for this. The initial being that the difference between off the dribble and both catch and shoot and free throw is much larger than the difference between off the dribble and any other label that without sufficient training videos in the ”other” category, the model will be trained to group them together, or that the off the dribble label has overlap with some tags in the other category, similar to the jumpshot tag and is perhaps insufficient as a base shot type. Given results from the previous experiment the former seems more likely to the latter, as if the off the dribble label were unsuitable as a base shot type, it would have likely hindered the accuracy of off the dribble in the previous task.

Simultaneously during the previous experiment, the partitioning of shots into 3 sets based on court position was undone, and the model was trained using one set of weights for all court positions. This was found to have no meaningful positive or negative impact on the accuracy of the model, debunking the idea that the position on the court (barring those very close to the hoop) or the angle of the camera relative to the backboard increase the number of contexts to an extent that requires significantly more knowledge to be encoded (thereby requiring multiple weight sets). This phenomenon is caused instead by the angle of incidence, discussed further in section 5.2.

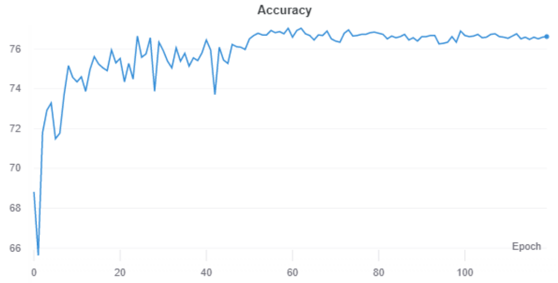

Figure 5: Validation accuracy (Free Throw vs Catch and Shoot vs Off the Dribble vs Other).

4.3.6 Free Throw vs Catch and Shoot vs Off the Dribble

Our final model, on a set of 1500 shots with 500 of each free throw, catch and shoot and off the dribble performs with a testing accuracy of 96.4%, 94.8% and 99.6% respectively. A measure of overall validation accuracy by epoch is show in Figure 6.

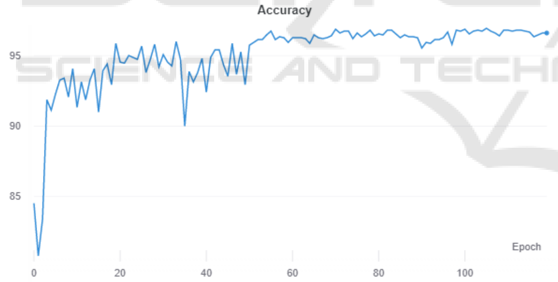

Figure 6: Validation accuracy (Free Throw vs Catch and Shoot vs Off the Dribble).

5 KEY CHALLENGES UNCOVERED FOR SHOT TYPE IDENTIFICATION

5.1 Semantically Cluttered Images

In action recognition, it important that key features to identify that action are visible. For example, to identify that an actor in a video is performing jumping jacks, the actor's arms and legs would ostensibly need to be visible. In this case, the arms and legs are high value features. A video that captures them entirely and without obstruction would contain these ”high value” features, while a video where they are obstructed or not fully and consistently in view would not or only have them in part. Conversely, in the same case the background or sky would be a low value feature, and would not be a priority to capture to create a video with high value features.

A notable challenge of basketball shot classification is that most available basketball videos are wide angle frames (Gu et al., 2020) (Tien et al., 2007). While these frames are useful for other tasks such as player tracking, in the task of shot classification, only one actor, which is the player shooting the ball, is of primary interest for determining the type of shot being taken. An example of semantic clutter is shown in Figure 7, which shows a commonly available frame for many recorded competitive basketball games at the professional and college level. As detailed in the image, high value features are the player skeleton and the ball, as well as their respective movements as the video progresses. Lower, but still relevant features include court markings for deriving player position, non-target players (players that are not shooting the ball) and their movements, as many maneuvers require an opponent or teammate. Lowest/negligible semantic value features include most things outside this category, including players on the sideline, crowd members, hoop base and pole, referees, etc. In this specific case, the area of highest interest composes less than 2% of the overall image (less than 4500 pixels). A low percentage of high semantic value area is inherent to most if not all video capture systems that utilize a wide angle, such as rafter-mounted or similar systems, and presents marked difficulty in the extraction of useful features for the task of shot classification as they constitute less of the available pixels for the image and are thereby less detailed and discernible.

However, the frame shown in Figure 7 is the ideal case for videos procured through the wide frame video capture systems. In reality, many useful features are visually occluded by, most commonly, other players. While this may be useful for the identification of some classes, in most it obscures player movement and prevents access to valuable features. This occlusion is demonstrated in Figure 8.

5.2 Angle of Incidence

Variable angle created the challenge of addressing a less constrained domain. As mentioned earlier an unconstrained domain can make for a more difficult classification problem as more contexts for each class often requires the model to encode additional knowl-

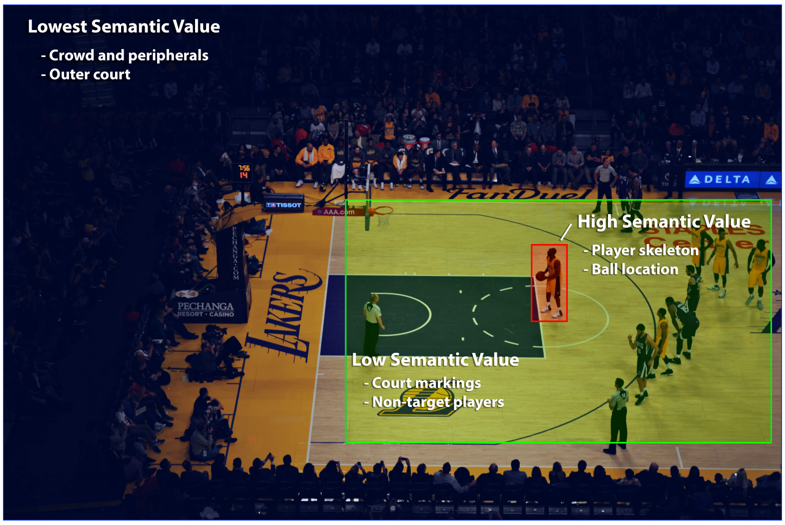

Figure 7: Detailed image of areas of semantic value and their percentage of the overall photo. Ramiro Pianarosa, *The Last Throw*, 4928 x 3264 px. Courtesy of Unsplash, accessed May 4, 2021, https://unsplash.com/photos/8hW2ZB4OHZ0.

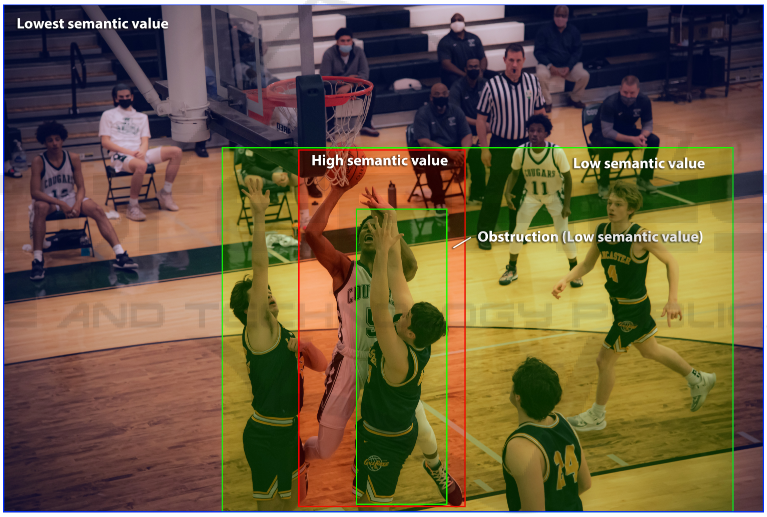

Figure 8: Image detailing the effects of player obstruction on feature extraction. Matthias Cooper, *Untitled*, 6016 x 4016 px. Courtesy of Unsplash, accessed May 4, 2021, https://unsplash.com/photos/8TYh5icTQVc.

edge to label each class accurately. One way to increase or reduce the number of contexts is not only by viewing it in a different physical location, but also from a different angle or viewpoint.

The angle in question for the task of shot classification is the angle created by two intersecting lines in 3d space, the first being the line created from the camera lens to the player, the second from the player to the intended shot location. This angle will be referred to as the angle of incidence from this point onward. The angle of incidence of most available basketball video data is highly variable.

As suggested by progress in the field of image subject rotation (Tran et al., 2019), it is likely that the higher the delta between two images'(same subject) angle of incidence, the more disparate the features. Highly disparate features from instances of the same class results in a minimum amount of features that must be encoded for a computer vision model at least several times larger than if the angle can be standardized between and within classes. For shot type classification, variable angle affects most if not all high and low value features, including player skeleton movement, ball movement, court markings relative to the player and non-target player positioning and movement.

The vast difference in angles is shown in differences between Figures 9 and 10. Both figures show a

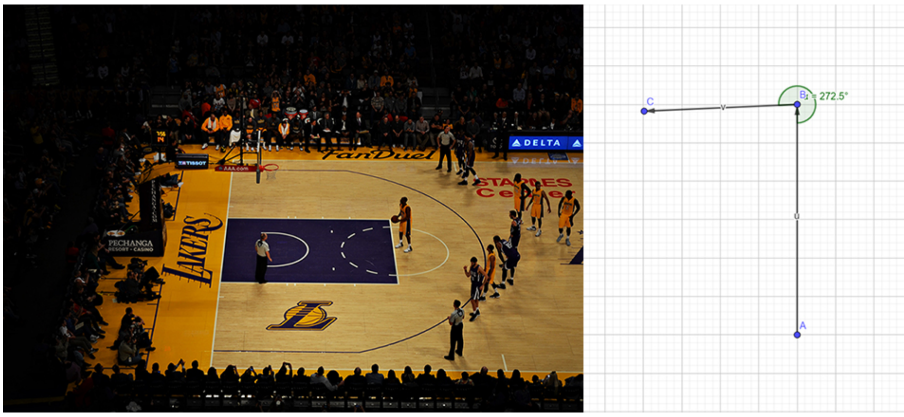

Figure 10: Image depicting the angle of incidence of a shot in from the free throw line. Ramiro Pianarosa, *The Last Throw*, 4928 x 3264 px. Courtesy of Unsplash, accessed May 4, 2021, https://unsplash.com/photos/8hW2ZB4OHZ0.

player taking a shot, with the shot's associated angle of incidence from a top-down perspective. Although technically the effect of angle of incidence occurs in 3d space, the shot is shown from a top down perspective for ease of depiction and understanding. Ignoring other differentiating features, the angle of incidence is widely different between the two. A model would be required to encode a drastically increased number of features and likely require far more training data to accurately classify a dataset with the desired classes of shot type labeled utilizing angles within a range of roughly 20° of the two shown in Figure 9 and Figure 10 than it would if it were to utilize a dataset utilized only one set of similar angles.

In describing the dataset provided by NOAH, we mention that despite using multiple cameras as a means of capturing videos of shots, and camera configurations occasionally varying slightly between locations, the angle of incidence remains the same. This property of the dataset contributes to the result in 4.3.6 where we find there is no need to group shots by location on court or which camera recorded them so long as the angle of incidence is similar.

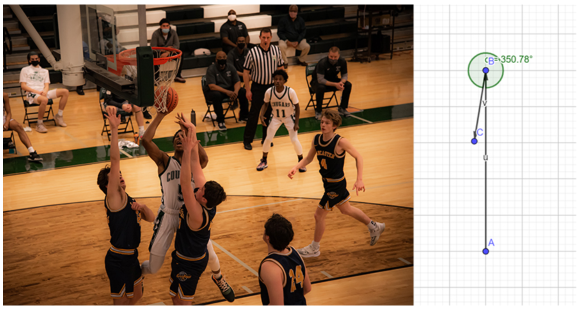

Figure 9: Image depicting the angle of incidence of a shot in the key. Matthias Cooper, *Untitled*, 6016 x 4016 px. Courtesy of Unsplash, accessed May 4, 2021, https://unsplash.com/photos/8TYh5icTQVc.

6 CONCLUDING REMARKS AND LESSONS LEARNED

The problem of efficient basketball shot type labeling is one that can aid teams and coaches in making more well-informed decisions, as well as allow for more specialized and efficient training of players. In this paper, we (1) design a novel label assignment system for labeling basketball shot types, as well as propose 5 base shot type labels for use in this system, (2) train a Temporal Relational Network to identify shot types, (3) present empirical results detailing the model's performance under several variations of the shot type labeling task, and (4) identify key challenges and problems that must be addressed when performing the task of basketball shot type labeling. The following are key lessons from applying and analyzing the performance of the Temporal Relational Networks on the problem of basketball shot type labeling:

- **Classes Must Be Ensured to Be Mutually Exclusive.** Tags should, before labeling occurs be separated into base shot type tags and supplementary tags. This will avoid situations where training and results are corrupted by overlap between classes, such as in the case of ”jumpshot”.
- **Angle of Incidence Should Be Fixed or Otherwise Accommodated for.** Widely varying angles of incidence will likely increase the required learning/knowledge necessary to be encoded. Controlling the angle of incidence, or otherwise addressing the additional contexts that varying angle of incidence creates is advised in further applications.
- **Insufficient Number of Class Instances Can Lead to Class Assimilation.** As pointed out in

4.3.3, it is possible that shot type labels can go ignored and/or be assimilated by other labels if there are insufficient instances, or the model is unable to learn features to define it from a similar class.

REFERENCES

Erčulj, F. and Štrumbelj, E. (2015). Basketball shot types and shot success in different levels of competitive basketball. *Plos One*, 10(6).

ESPN (2020).

Foster, B. T. and Binns, M. D. (2019). Analytics for the front office: Valuing protections on nba draft picks. *MIT SLOAN Sports Analytics Conference*, 13.

Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fründ, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., and Memisevic, R. (2017). The ”something something” video database for learning and evaluating visual common sense. *CoRR*, abs/1706.04261.

Gu, X., Xue, X., and Wang, F. (2020). Fine-grained action recognition on a novel basketball dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2563–2567.

Johnson, N. (2020). Extracting player tracking data from video using non-stationary cameras and a combination of computer vision techniques. *MIT SLOAN Sports Analytics Conference*, 14.

Kaplan, S., Ramamoorthy, V., Gupte, C., Sagar, A., Premkumar, D., Wilbur, J., and Zilberman, D. (2019). The economic impact of nba superstars: Evidence from missed games using ticket microdata from a secondary marketplace. *MIT SLOAN Sports Analytics Conference*, 13.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

Materzynska, J., Berger, G., Bax, I., and Memisevic, R. (2019). The jester dataset: A large-scale video dataset of human gestures. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., et al. (2019). Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8.

Mortensen, J. and Bornn, L. (2019). From markov models to poisson point processes: Modeling movement in the nba. *MIT SLOAN Sports Analytics Conference*, 13.

Nistala Akhil, G. J. (2019). Using deep learning to understand patterns of player movement in the nba. *MIT SLOAN Sports Analytics Conference*, 13.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. *CoRR*, abs/1604.01753.

Tien, M.-C., Huang, C., Chen, Y.-W., Hsiao, M.-H., and Lee, S.-Y. (2007). Shot classification of basketball videos and its application in shooting position extraction. volume 1, pages I–1085.

Tran, L., Yin, X., and Liu, X. (2019). Representation learning by rotating your faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3007–3021.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Gool, L. V. (2016). Temporal segment networks: Towards good practices for deep action recognition. *CoRR*, abs/1608.00859.

Williams, C. Q., Webster, L., Spaniol, F., and Bonnette, R. (2020). The effect of foot placement on the jump shot accuracy of ncaa division i basketball players. *The Sport Journal*, 21.

Zhen, L., Wang, L., and Hao, Z. (2016). A biomechanical analysis of basketball shooting.

Zhou, B., Andonian, A., Oliva, A., and Torralba, A. (2018). Temporal relational reasoning in videos. *Computer Vision – ECCV 2018 Lecture Notes in Computer Science*, page 831–846.