RECOGNIZING OFFENSIVE TACTICS IN BROADCAST BASKETBALL VIDEOS VIA KEY PLAYER DETECTION

Tsung-Yu Tsai[⋆†] *Yen-Yu Lin*[⋆] *Hong-Yuan Mark Liao*[⋆] *Shyh-Kang Jeng*[†]

[⋆]Academia Sinica, Taipei, Taiwan [†]National Taiwan University, Taipei, Taiwan

ABSTRACT

We address offensive tactic recognition in broadcast basketball videos. As a crucial component towards basketball video content understanding, tactic recognition is quite challenging because it involves multiple independent players, each of which has respective spatial and temporal variations. Motivated by the observation that most intra-class variations are caused by non-key players, we present an approach that integrates *key player detection* into *tactic recognition*. To save the annotation cost, our approach can work on training data with only video-level tactic annotation, instead of key players labeling. Specifically, this task is formulated as an MIL (*multiple instance learning*) problem where a video is treated as a bag with its instances corresponding to subsets of the five players. We also propose a representation to encode the spatio-temporal interaction among multiple players. It turns out that our approach not only effectively recognizes the tactics but also precisely detects the key players.

Index Terms— group behavior analysis, offensive tactic recognition, key player detection, video understanding

1. INTRODUCTION

Basketball offensive tactic recognition is drawing attention, because it brings new insights into the games and has great impacts on the outcomes of the games. As an important topic of group behavior analysis, it helps review the performance of offense and defense executions, understand opposing team strategies, and even investigate certain players' habits. These tactics are often recognized and annotated by experts like assistant coaches. Due to the explosive growth of broadcast basketball videos, there has been a strong demand for an accurate tactic recognition system. However, tactic recognition is quite challenging because it involves multiple independent players with respective spatial and temporal variations.

We consider the players who execute a certain tactic *key players*, such as the two players who sequentially circle around the wing area of the court in tactic *wing-wheel*. Inspired by the observation that most unfavorable intra-class variations are caused by non-key players, we propose an approach where *key player detection* is integrated into *tactic recognition*. In this way, the performance degradation caused by large intra-class variations can be alleviated. Furthermore, it makes the recognition results more interpretable since we realize which players execute the predicted tactic.

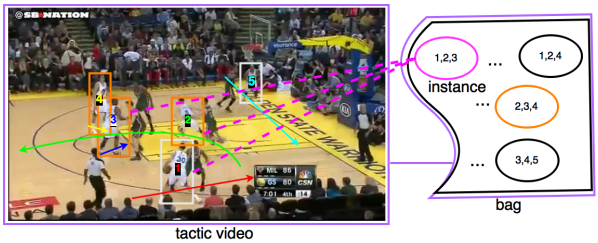


Fig. 1. We integrate key player detection into tactic recognition using multiple instance learning. Consider a tactic with three key players. The video is treated as a bag with C_3^5 instances, each of which corresponds to a particular group of three of the five offensive players. The positive instance is the one corresponding to the three key players.

Key player detection can be formulated as a supervised learning problem, but training data with player-level annotation are required. In this work, we save the annotation cost and assume training data with only video-level annotation, i.e. the ground truth of the offensive tactics, are available. Our approach carries out key player detection to facilitate tactic recognition, and deals with the absence of play-level annotation by using *multiple instance learning* (MIL) [1]. Specifically to recognize one particular tactic say wing-wheel, we treat a video clip as a *bag* with the player subsets as *instances*. A video is positive if tactic wing-wheel is performed in it. In a positive bag, the positive instance is the one that covers *exactly all* key players. An example is given in Fig. 1.

In our case, an instance corresponds to a subset of the five offensive players, and its representation is crucial to the performance. Therefore, we propose a novel representation, called *motion intensity map* (MIM), that accounts for multiple players simultaneously and is robust enough to characterize players of different temporal lengths and at arbitrary spatial locations. Specifically, MIM transforms a player's temporal positions and velocities into a distribution of motions over the quantized court regions. The activity among multiple players is encoded by summing up their distributions. With the two key components, MIM for player group representation and MIL for joint key player detection and tactic recognition, our approach achieves superior performance.

2. RELATED WORK

Group Behavior Analysis. Team sport tactic analysis is an important instance of group behavior analysis. Makris *et al*. [2] analyzed the group behavior of a shoal of fish and estimated the number of fishes around the world. They [3] further analyzed the group behavior in both the spatial and temporal domains, and more precisely predicted the number of fishes, the traveling route and so on. Solar *et al*. [4] suggested using proxemic theory, Granger causality, and *dynamic time warping* (DTW) to perform socially constrained structural learning and group detection. Tsai *et al*. [5] enhanced team sports player segmentation via image co-segmentation. Lin *et al*. [6] improved action recognition by leveraging auxiliary RGB-D visual clues. We are aware of two research trends in group behavior analysis. One is that scalability becomes critical in practice. The other is that finding coherent subgroups is crucial. The proposed approach does not rely on player-level annotation. Besides, it can identify key players, who are coherent across videos of the same tactic.

Sport Tactic Recognition. Players' trajectories are widely used for sport tactic recognition. For example, Intille and Bobick [7] recognized a football play by using Bayesian network to describe the interaction between players' trajectories. Siddiquie *et al*. [8] used a handcrafted spatiotemporal descriptor to classify elementary moves in American football. To bridge the semantic gap between low-level movements and high-level tactics, Perse *et al*. [9] designed a behavior detector from basketball players' trajectories, and identified tactic patterns as specific sequential combinations of behaviors. The performance of aforementioned approaches rely on high-quality players' trajectories. However, noisy trajectories are often present in practice. Bialkowski *et al*. [10] presented two representations, including *team occupancy* and *team centroid*, to alleviated the problems caused by noisy tracked sequences. Chen *et al*. [11] adopted DTW as the distance measure between trajectories, and further extended DTW to measure video-video similarity. Our approach detects key players and recognizes tactics simultaneously. More accurate tactic prediction can be achieved because the unfavorable intra-class variations from non-key players are excluded. To the best of our knowledge, the integration of key player detection and tactic recognition is novel in this field.

Multiple Instance Learning. MIL is a weakly supervised learning technology. It was firstly introduced by Dietterich *et al*. [1] for drug activity prediction. MIL has been used in various applications, such as image classification [12], object detection [13], text or document categorization [14], and semantic segmentation [15, 16]. Other than exploring the applications, many machine learning algorithms have been adopted to better solve MIL problems, such as *diversity density* [17], *support vector machines* [18], *artificial neural networks* [19], *decision trees* [20] and *AdaBoost* [21]. In this work, MIL is used to model the uncertainty caused by the absence of key player annotation in training data.

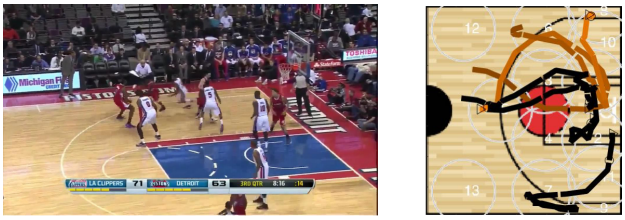


Fig. 2. A video frame of tactic *wing-wheel* and the five offensive players' trajectories with those of the two key players highlighted in orange. The centroid of each of the 13 tactic semantic sub-regions is shown in white.

3. OUR APPROACH

3.1. Notations and Problem Definition

We are given a set of N training videos for offensive tactic recognition, $D = \{(V_i, \ell_i)\}_{i=1}^{N}$, where V_i is the ith video with tactic $\ell_i \in \{1, ..., C\}$ label, and C is the number of tactic classes. Like our previous work [11], a player tracking algorithm is applied to extract players' trajectory, and each video is firstly represented by the five offense players' trajectories, i.e. $V_i = \{T_{i,p}\}_{p=1}^{5}$. Each trajectory is a time sequence of player position $T_{i,p} = \{T_{i,p_x}(t), T_{i,p_y}(t)\}_{t=1}^{F}$, where F is the frame number or the time interval of this video.

The videos could be of arbitrary lengths and are unsynchronized. Namely, the value of F could vary from video to video, and the player correspondences across videos are not clear. To address these issues, we follow [11] and apply the DTW algorithm to equalize and synchronized the trajectories. The DTW algorithm estimates a proper player alignment reference and compiles the optimal warping matrices. After warping, we further divide each trajectory into SR equal-size intervals (temporal stages), and the features of each trajectory will be a concatenation of the characteristics extracted from individual intervals. This helps to have a compact feature representation, so the training and testing time can be reduced accordingly. Besides, it also reduces the synchronization error via quantization. However, it may lose the information about the chronological order of frames in a stage. Thus, the value of SR controls the trade-off between robustness and distinctiveness. We empirically set it to 10 in the experiments.

For reducing manual labeling effort, our approach can work without the labeling of key players in a training video is available, but we assume the tactic label based on [22] and the number, n_c, of the key players for each tactic c are given. We set the numbers of the key players for different tactics, as reported in Table 1. In Fig. 2 shows a video frame of tactic *wing-wheel* and the five players' trajectories. This tactic features two players sequentially circle around the wing area of a basketball court. Thus, the number of key players is set to two. Our goal in this work is to recognize the basketball offensive tactics and detect the key players. In the following, we will describe how to accomplish the two tasks simultaneously via multiple instance learning (MIL).

3.2. Joint Key Player Detection and Tactic Recognition

We consider one tactic c, say *wing-wheel*, and convert the training set into a binary one. That is, videos of tactic *wing-wheel*, are considered positive training data, while the rest are negative. We leverage MIL for joint key player detection and tactic recognition. Specifically, these videos are treated as *bags* with $C_{n_c}^5$ *instances*, each of which corresponds to a particular group of n_c offensive players. n_c equals 2 in *wing-wheel*. The positive instance is the one covering the n_c key players. Thus, a positive bag contains one positive instance, while the instances in a negative bag are all negative.

Specifically for tactic c, we have the converted training set of binary classes, $\{(V_i, y_i \in \{1, -1\})\}_{i=1}^N$ where bag V_i contains instances $\{\mathbf{x}_{i,j} | j \in \{1, 2, ..., C_{n_c}^5\}\}$. The feature vector for instance $\mathbf{x}_{i,j}$ will be given later. MIL works with the labels attached to bags instead of to instances. The relationship of bag label y_i and the unknown instance labels $\{y_{i,j} | j \in \{1, 2, ..., C_{n_c}^5\}\}$ in MIL is formulated below:

$$\sum_j \frac{y_{i,j}+1}{2} \geq 1, \text{ if } y_i = 1 \text{ and } y_{i,j} = -1, \text{otherwise.} \quad (1)$$

That is, positive bags have positive instances and instances in negative bags are all negative.

According to Andrew *et al.*'s model [18], MIL can be solved by a generalized soft-margin SVM. We use the *instance-level MIL* (mi-SVM) learning:

$$\min_{\{y_{i,j}\}} \min_{\mathbf{w}, b, \{\xi_{i,j}\}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \quad (2)$$
$$\text{s.t.} \quad \forall i, j : y_{i,j}(\langle \mathbf{w}, z_{i,j} \rangle + b) \geq 1 - \xi_{i,j}, \xi_{i,j} \geq 0,$$
$$y_{i,j} \in \{-1, 1\}, \text{ and (1) holds,}$$

where $\mathbf{w}$ is SVM hyperplane normal vector, b is hyperplane offset. $z_{i,j}$ is transformed variable of instance $\mathbf{x}_{i,j}$ with kernel function $\phi(\cdot)$, i.e. $z_{i,j} = \phi(\mathbf{x}_{i,j})$. Our work uses *radial basis function* (RBF) as the kernel function. Operator $\langle \cdot \rangle$ stands for inner product, C is the loss cost, and $\xi_{i,j}$ is the slack variable.

In general, the number of positive instances in a positive bag is not constrained. In our case, we look for only the instance that cover exactly all the key players. Thus for a bag predicted as positive, we use the instance posterior probability of SVMs to calculate the posterior probability. Specifically, we adopted the sigmoid probability approximation proposed by Platt *et al*. [23]. The instance with the highest probability is selected as top instance $\mathbf{x}_{i,j^*}$. The probability of a bag is equal to that of its top instance probability. For tactic classification, if the bag probability exceeds a threshold, the video is predicted as this tactic. Multi-class tactic recognition can be carries out by simply using one-vs-all fashion. In the next section, we will describe how to design the features of an instance $\mathbf{x}_{i,j}$ that represents multi-players' spatiotemporal positions and velocities.

Table 1. Tactics considered in the experiments

tactic	abbr.	# video	# key players
2-3 Flex	F23	15	3
Elevator	EV	11	3
Hawk	HK	20	3
Pin-Down	PD	9	3
Princeton	PT	13	5
Back-Side Pick and Roll	RB	15	3
Side-Pick Slip and Pop	SP	15	2
Warrior Single	WS	13	3
Weave	WV	16	5
Wing-Wheel	WW	7	2

3.3. Group Feature Representation

Inspired by the *occupancy map* [10], we separate the basketball half court into 13 sub-regions, i.e. those in white in Fig. 2, by referring to [9, 24]. A drawback of using occupancy maps results from the hard assignment of player positions into sub-regions, which introduces the quantization error and is sensitive to noise in trajectories. Thus, soft assignment is used:

$$p_{ij} = \frac{f(\mathbf{l}_i | \mu_j, \tau)}{\sum_j f(\mathbf{l}_i | \mu_j, \tau)}, \text{ where } f(\mathbf{l}_i | \mu_j, \tau) = e^{-\frac{\|\mathbf{l}_i - \mu_j\|}{2\tau^2}} \quad (3)$$

$\mathbf{l}_i$ is the player position, μ_j is center of sub-region j and τ is the common deviation of Gaussian distributions. Thus, p_{ij} is the probability of player i in sub-region j. The other drawback of occupancy maps is that only static position information is recorded. We extend the maps by including dynamic velocity information via quantizing the orientations of velocity into $m = 8$ directions, i.e.

$$v_{ij} = \|V_j\| \frac{g(\theta_j | d_i, \kappa)}{\sum_{i=1}^m g(\theta_j | d_i, \kappa)}, \text{ where} \quad (4)$$
$$g(\theta_j | d_i, \kappa) = \frac{e^{\kappa \cos(\theta_j - d_i)}}{2\pi I_o(\kappa)}, d_i = \frac{(i-1)\pi}{4}, \quad (5)$$

$\|V_j\|$ is the velocity magnitude of player j, θ_j is the velocity direction, g is von mises distribution (normal distribution on unit circle), κ is the variation level, and $I_0(\kappa)$ is modified Bessel function of order 0.

After quantizing velocity magnitudes into m directions, we follow Eq. (3) and assign a player into the 13 subregions based on the location. We call this feature *motion intensity map* (MIM), which encodes both position and velocity information. Based on MIM, we will show in the experiments that simply summing over all the players in an instance suffices to describe multi-player interaction and the group behavior.

4. EXPERIMENTAL RESULTS

We adopt the dataset used in our previous work [11], which contains 134 videos of NBA 2013-2014 season. These videos

Table 2. Accuracy of five approaches for comparison.

method	average accuracy
unsupervised GMM [11]	0.8550
supervised GMM [11]	0.8867
team centroid [10]	0.7474
team occupancy [10]	0.8875
ours	**0.9467**

Table 3. The confusion matrix of our approach.

Accu.	F23	EV	HK	PD	PT	RB	SP	WS	WV	WW
F23	0.93	0	0	0	0.07	0	0	0	0	0
EV	0	0.90	0	0	0	0	0	0	0	0.10
HK	0	0	1.00	0	0	0	0	0	0	0
PD	0	0	0	1.00	0	0	0	0	0	0
PT	0	0	0.07	0	0.93	0	0	0	0	0
RB	0	0	0	0	0	1.00	0	0	0	0
SP	0	0	0	0	0	0	1.00	0	0	0
WS	0.07	0	0.07	0	0	0	0	0.76	0	0.10
WV	0	0	0	0	0.07	0	0	0	0.93	0
WW	0	0	0	0	0	0	0	0	0	1.00

are labeled by the experts with 10 different half-court offensive tactics summarized in Table 1. Each video clip comes with the trajectories of the five offensive players. Such trajectories are acquired automatically as indicated in [11].

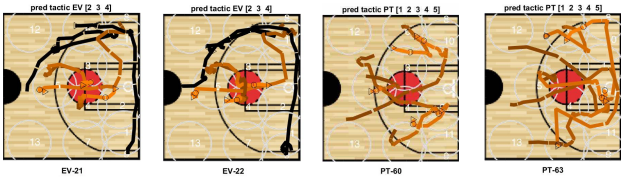
We compare our method with four different methods. The first one is our prior work [11] which uses unsupervised *Gaussian mixture model* (GMM) to build tactic spatio-temporal templates. For comparison, we further generalize it to the supervised setting by providing tactic labels to learn GMM. The other two competing methods are based on [10] with *team centroid* and *team occupancy* as the group features, respectively. Our approach adopts the proposed MIM feature representation. Five-fold cross validation is used for tuning the hyper-parameters in MIL.

As Table 2 shows, the four competing approaches reach the accuracy rates of $0.75 \sim 0.89$. Our approach achieves the accuracy of 0.95. The performance gain of $6 \sim 20\%$ is quite significant. Except the first one, all the competing methods are supervised. The poor performance of using team centroid as features results form that the centers of all tactics lie in the center line of the court in the half court offense, so the centroid actually reduces the discrimination. We consider our performance gain result from two factors. First, our approach leverages the additional information, the numbers of key players, and integrates key player detection into tactic recognition. Second, the proposed MIM includes both the information position and velocity for better feature representation.

To gain insight into the average accuracy, the confusion matrix of our approach is given in Table 3, where the predicted tactics are given in columns. Diagonal entries represent correct classification. The confusion matrix indicates that except tactic WS, all the tactics have accuracies higher than 90%. The lower performance of tactic WS is due to mis-classifying it as other tactics like F23 (7%), HK (7%), and WW (10%). The mis-classification is caused by the less

Table 4. Performance of different feature representations

feature	# dim	tactic acc.	key player acc.
`MIM P`	20	0.89	0.6813
`MIM V`	20	0.8133	0.5857
`MIM HA`	130	0.8967	0.632
MIM	1040	0.9467	0.7853



(a) (b)

Fig. 3. Visualization of the detected key players.

accurate key player detection, because it is more difficult to distinguish the key players in tactic WS from the rest.

To further examine the performance of the proposed MIM, we compare MIM with its degenerated versions, including 1) `MIM P`: only position information is included in the representation; 2) `MIM V`: only velocity information is encoded in the representation; 3) `MIM HA`: MIM uses hard assignment in quantization. Our approach with the MIL formulation is applied to MIM and the three variants. The recognition rates are reported in Table 4. The results point out the importance of joint consideration of position and velocity information in tactic recognition as well as the critical issue of quantization errors in spatio-temporal analysis.

Fig. 3 visualizes the detected key players. The two videos of tactic EV. in Fig. 3(a) were originally mis-classified until the integration of key player detection, which excludes unfavorable intra-class variations caused by non-key players. In Fig. 3(b), the detected key players in tactic PT have long-distance running. The yielded variations can be well handled by using merely either position or velocity information. It supports the use of MIM which joint consider position and velocity information for group behavior description.

5. CONCLUSIONS

Automatic group behavior analysis is always in demand due to the explosive growth in broadcast team sports videos where valuable information is included. In this work, we have presented an approach that integrates key player detection into basketball offensive tactic recognition. It works with training data with only video-level annotation, but can carry out key player detection, player-level classification, via formulating the task as a multiple instance learning problem. Besides, a new feature presentation MIM is proposed to better encode both spatial and temporal information. Both the quantitative and visualization results confirm that our approach achieves effective and remarkably superior performance.

Acknowledgement. This work was supported in part by grants MOST 104-2628-E-001-001-MY2, MOST 105-2221-E-001-030-MY2, and MOST 105-2221-E-001-018-MY3.

6. REFERENCES

[1] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[2] Nicholas C Makris, Purnima Ratilal, Deanelle T Symonds, Srinivasan Jagannathan, Sunwoong Lee, and Redwood W Nero, "Fish population and behavior revealed by instantaneous continental shelf-scale imaging," *Science*, vol. 311, no. 5761, pp. 660–663, 2006.

[3] Nicholas C Makris, Purnima Ratilal, Srinivasan Jagannathan, Zheng Gong, Mark Andrews, Ioannis Bertsatos, Olav Rune Godø, Redwood W Nero, and J Michael Jech, "Critical population density triggers rapid formation of vast oceanic fish shoals," *Science*, vol. 323, no. 5922, pp. 1734–1737, 2009.

[4] Francesco Solera, Simone Calderara, and Rita Cucchiara, "Socially constrained structural learning for groups detection in crowd," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, , no. 1, pp. 1–1.

[5] Tsung-Yu Tsai, Yen-Yu Lin, Hong-Yuan Mark Liao, and Shyh-Kang Jengg, "Precise player segmentation in team sports videos using contrast-aware co-segmentation," in *Proc. Int'l Conf. Acoustics, Speech and Signal Processing*, 2016.

[6] Yen-Yu Lin, Ju-Hsuan Hua, Nick C. Tang, Min-Hung Chen, and Hong-Yuan Mark Liao, "Depth and skeleton associated action recognition without online accessible rgb-d cameras," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014.

[7] Stephen S Intille and Aaron F Bobick, "A framework for recognizing multi-agent action from visual evidence," *Proc. Nat'l Conf. Artificial Intelligence*, vol. 99, pp. 518–525, 1999.

[8] Behjat Siddiquie, Yaser Yacoob, and Larry S Davis, "Recognizing plays in american football videos," *University of Maryland, Tech. Rep*, vol. 111, 2009.

[9] Matej Perše, Matej Kristan, Stanislav Kovačič, Goran Vučkovič, and Janez Perš, "A trajectory-based analysis of coordinated team activity in a basketball game," *Computer Vision and Image Understanding*, vol. 113, no. 5, pp. 612–621, 2009.

[10] Alina Bialkowski, Patrick Lucey, Peter Carr, Simon Denman, Iain Matthews, and Sridha Sridharan, "Recognising team activities from noisy data," in *Proc. Conf. Computer Vision and Pattern Recognition Workshops*, 2013.

[11] Ching-Hang Chen, Tyng-Luh Liu, Yu-Shuen Wang, Hung-Kuo Chu, Nick C Tang, and Hong-Yuan Mark Liao, "Spatio-temporal learning of basketball offensive strategies," in *Proc. ACM Conf. Multimedia*, 2015.

[12] Oded Maron and Aparna Lakshmi Ratan, "Multiple-instance learning for natural scene classification.," in *Proc. Int'l Conf. Machine Learning*, 1998.

[13] Cheng Yang and Tomas Lozano-Perez, "Image database retrieval with multiple-instance learning techniques," in *Proc. Int'l Conf. Data Engineering*, 2000.

[14] Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth, "From group to individual labels using deep features," in *Proc. ACM Conf. Knowledge Discovery and Data Mining*, 2015.

[15] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang, "Augmented multiple instance regression for inferring object contours in bounding boxes," *IEEE Trans. on Image Processing*, vol. 23, no. 4, pp. 1722–1736, 2014.

[16] Feng-Ju Chang, Yen-Yu Lin, and Kuang-Jui Hsu, "Multiple structured-instance learning for semantic segmentation with uncertain training data," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014.

[17] Oded Maron and Tomás Lozano-Pérez, "A framework for multiple-instance learning," in *Conf. in Neural Information Processing Systems*, 1998.

[18] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann, "Support vector machines for multiple-instance learning," in *Conf. in Neural Information Processing Systems*, 2002.

[19] Zhi-Hua Zhou and Min-Ling Zhang, "Neural networks for multi-instance learning," in *Proc. Int'l Conf. Intelligent Information Technology*, 2002.

[20] Hendrik Blockeel, David Page, and Ashwin Srinivasan, "Multi-instance tree learning," in *Proc. Int'l Conf. Machine Learning*, 2005.

[21] Peter Auer and Ronald Ortner, "A boosting approach to multiple instance learning," in *Proc. Euro. Conf. Machine Learning and Principles and Practice of Knowledge Discovery*, 2004.

[22] Philip Jing, "Nba tactic breakdown in graphics," 2014, http://blog.xuite.net/philip741012/blog.

[23] John Platt et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[24] hoopTactic, "Basketball basic court area," 2016, http://hooptactics.com/Basketball_Basics_Court_Areas.