# COMPUTATIONAL SPORTS BROADCASTING: AUTOMATED DIRECTOR ASSISTANCE FOR LIVE SPORTS

*Christine Chen, †Oliver Wang, †Simon Heinzle, ‡Peter Carr, †Aljoscha Smolic, *†Markus Gross

*ETH Zurich, †Disney Research Zurich, ‡Disney Research Pittsburgh

## ABSTRACT

Live sports broadcast is seeing a large increase in the number of cameras used for filming. More cameras can provide better coverage of the field and a wider range of experiences for viewers. However, choosing optimal cameras for broadcast demands a high level of concentration, awareness and experience from sports broadcast directors. We present an automatic assistant to help select likely candidates from a large array of possible cameras. Sports directors can then choose the final broadcast camera from the reduced suggestion set. Our assistant uses both widely acknowledged cinematography guidelines for sports directing, as well as a data-driven approach that learns specific styles from directors.

***Index Terms*—** Computational Broadcast, Machine Learning, Shot Selection, Image Understanding

## 1. INTRODUCTION

Since its introduction in 1927, live TV sports broadcasting has grown into a major form of entertainment. In 2008, an estimated *4.7 billion* people turned on their TVs to follow the Beijing Olympics (roughly *two-thirds* the world's population). One crucial part of sports broadcasting is the role of the director, whose job is to select the best possible camera from all available cameras at all times: the story should be consistent and complete, important events should be emphasized, the cuts between cameras should be perceived naturally by the audience, and the overall video composition should be aesthetically pleasing. However, with a large number of cameras in stadiums, merely scrutinizing all camera signals at the same time can already be quite overwhelming (Fig. 1). As more cameras become introduced, this problem will only become more challenging.

In this work, we present an interactive, intelligent and intuitive system for sports broadcasting designed to assist directors in finding and recommending potential broadcast camera views from a large number of possible cameras. We propose two different approaches to determine camera ranking using machine learning techniques. Our first method is based on cinematographic rules supported by user data from a general audience, and attempts to learn an aesthetic quality for the different cameras. The second approach learns a director's specific directorial style based on their previously recorded footage. This method leverages a rich dataset of *already existing* training data in the form of expert-directed broadcast footage. Both methods show promising results, and show the potential of data-driven approaches for camera selection. In addition to TV broadcast, our framework is also applicable to virtual 3D environments that require virtual camera selection such as sports games.

**Fig. 1**. **Master control room for Beijing Olympics 2008.** A sports broadcast director is normally presented with more than 20 different camera views, making it difficult to choose the best one.

## 2. RELATED WORK

In this section, we will survey the following three areas: computational aesthetics, computer vision for sports, and recent works in computational sports broadcasting.

**Computational Aesthetics.** The goal of computational aesthetics is to accurately predict the perceived aesthetics of visual content. A promising approach is the use of machine learning techniques on human labeled data [1], where the influence of low-level features such as hue, saturation, rule of thirds, texture and depth of field is taken into account. Inspired by this work, researchers investigated the influence of additional low-level features [2, 3] as well as high-level fea-

tures such as human faces [4] and visually salient regions [5].

While much work exists for image aesthetics, video aesthetics is still in its infancy. One example is Moorthy et al [6], who presented an approach inspired by [1] that uses a combination of 97 low-level features to train the aesthetics predictor. While their approach is for general video, we show that for one specific genre of video we can achieve a better accuracy rate with a smaller number of more specialized features. In contrast to their work, we furthermore are able to personalize the perceived aesthetics to a certain directing style.

**Computer Vision in Sports.** Computer vision has been widely addressed in sports, and many methods are presented to identify and track players [7, 8, 9, 10]. The most important advances in this field are summarized in a recent survey of D'Orazio [11]. While our system uses vision analysis, we do not address it in this work and rather consider such data given as an input.

**Computational Sports Broadcasting.** Wang et al. [12] propose a sports broadcasting composition algorithm for automatic generation of replays and automatic camera selection. The camera selection algorithm is based on a hidden Markov model that has been trained with hand-labeled data, and only uses information of the camera motion parameters (pan, tilt, zoom, focus). As a result, their approach aims at selecting the view that is least blurred. Alternatively, Choi et al. [13] proposes the use of tracking information of the ball and to automatically select the views in which the ball is clearly visible. The approach is therefore biased toward wide-angle camera views and cannot cope with suitations in which no ball is visible. Compared with these two approaches, our method provides a more complete and robust solution. We use additional features, and provide both generalized aesthetic and personalized learning models.

## 3. OVERVIEW

The goal of our work is to select the best view from a list of possible cameras. We propose two different machine-learning approaches to solve this problem. The first approach is related to *computational aesthetics*, where sports-specific image features are used to learn the aesthetic quality of video sequences, based on training data acquired through a user study. The second approach is designed to learn *directorial style*, where low-level camera information and player location as features are used to learn how to predict shots, based on training from a qualified director's past broadcast recordings. In the following, we will first describe the data set used to evaluate our approaches. We will then describe both approaches in detail.

**Evaluation Dataset.** Both approaches will be evaluated on footage of a field hockey match, filmed with three calibrated, human-controlled cameras. In addition to the video streams, the data set contains a final broadcast video created by a professional director, and tracking data in the form of player locations in field coordinates. We break the video streams from each camera $i$ into short five-second clips $Clip_{(i,l)}$. Our goal then is to determine the best camera to broadcast for each time interval $l$.

## 4. COMPUTATIONAL AESTHETICS

Our first approach is based on *computational aesthetics*. More specifically, all video cameras are ranked based on automatically computed aesthetic features. As opposed to previous work, our features are targeted towards sports broadcasting and include the visibility of the ball, player distribution as well as the temporal motion of the players. The features are then used in a support vector machine (SVM) to compute an overall camera ranking. The influence of each feature in the SVM is trained from user-labeled preference data, which is determined by a user study.

### 4.1. Features

We define features over micro-shots, which consist of 15 consequential frames. We denote the beginning and end as $\text{MS}_{begin}$ and $\text{MS}_{end}$ respectively. We first define four features for each frame, $f_{\text{ball}}$, $f_{\text{players}}$, $f_{\text{thirds}}$, $f_{\text{size}}$, which we describe next.

**Ball Visibility.** The most common feature in live sports is often the ball, around which all the action is centered. In many cases, the position of the ball can be detected either using embedded sensors or image processing, or as in our case, human-labeled data. The feature descriptor $f_{\text{ball}}$ then describes whether the ball is visible in the current clip.

$$f_{\text{ball}} = \begin{cases} 1 & \text{if ball is visible} \\ 0 & \text{else} \end{cases} . \quad (1)$$

**Player distribution.** We use the position of the players as one of the main aesthetic rules. First, we assign an importance value $I_j$ to each player, where $\sum I_j = 1$ over all players. The importance value is determined based on the importance of each player: higher influence is assigned to strikers and superstars, which usually draw more attention from the audience. Then, the following scores are used as aesthetic features:

1. Most of the time the audience prefers camera views that cover a large *number of the players* on the field, as it provides a full picture and is less likely to miss out on major actions. Therefore, we define $f_{\text{players}}$ as the weighted sum of all visible players:

$$f_{\text{players}} = \sum_{j \in \text{visible}} \text{I}_j. \quad (2)$$

2. The *rule of thirds* is a popular aesthetic guideline in photography, which proposes that the object of interest should positioned along at the crossings of the third

lines of an image. More specifically, the four crossing points for an image of size $[w, h]$ are defined as

$$\mathbf{c}_{a,b} = \left[\frac{a}{3}w, \frac{b}{3}h\right] \quad \text{for} \quad a, b = 1, 2.$$

For each player $j$, we then quantify the rule of thirds as the minimum distance between the image position $\mathbf{x}_j$ of a players' head to the four crossings.

$$D_j = \min_{(a,b)} \|\mathbf{x}_j - \mathbf{c}_{a,b}\| \quad \text{for} \quad a, b = 1, 2. \tag{3}$$

The aesthetic score is then computed as weighted average of the rule of thirds for all visible players:

$$f_{\text{thirds}} = \frac{\sum_{j \in \text{visible}} I_j D_j}{\sum_{j \in \text{visible}} I_j}. \tag{4}$$

3. *Salient objects* should preferably have prominent sizes in the image. Such a score can be quantified using a weighted sum of all visible players:

$$f_{\text{size}} = \sum_{j \in \text{visible}} I_j A_j \,, \tag{5}$$

where $A_j$ denotes the image area occupied by player $j$.

**Temporal smoothness.** Finally, we add a temporal smoothness term that computes the movement of players throughout a micro-shot. We define the $j^{\text{th}}$ player's position at frame $k$ as $\mathbf{x}_j[k]$, and define smoothness as:

$$f_{\text{smoothness}} = \sum_j \left\|\mathbf{x}_j[\text{MS}_{begin}] - \mathbf{x}_j[\text{MS}_{end}]\right\|. \tag{6}$$

**Total number of features.** For each feature, we compute six different measures, using the *mean*, *median*, *max*, *min*, *first quartile*, and *third quartile* of each feature over the micro-shot. Adding the additional temporal smoothness term $f_{\text{smoothness}}$, we then have 25 features per microshot.

### 4.2. Training

In order to train the support vector machine, we determined aesthetic labels for each micro-clip through a user study. In each task, users were explicitly asked to judge three randomly selected videos for quality scores based on content, composition, and smoothness by selecting which clips they preferred most and least (Figure 2). These three videos are selected from the *same* period of game shot by three different cameras, and the users are required to choose the camera that captures the game best during that event. In total, 30 tasks were presented to each user. A total of 35 participants, 26 male and 9 female, completed our user study, of this group, 85.7% of participants reported to watch sports on a regular basis.

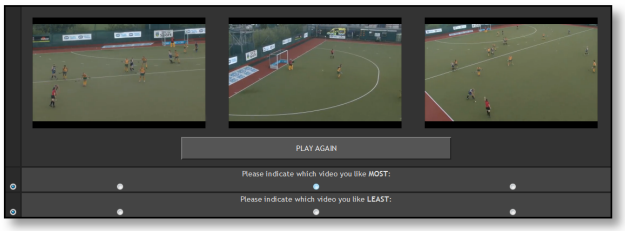


**Fig. 2**. **User Study.** Users are presented with a synchronized set of clips for each task. The order of the clips are randomly permuted. Users are asked to watch the video and select one which they like the most, and one which they like the least.

To learn how to predict the audience's preference on sports recordings, we use the labels *good* and *bad* to train a SVM model [14] with Gaussian radial basis functions (RBF), using the features presented in Section 4.1 as input. The labels *good* and *bad* are derived from the user study as follows. Let $V^+_{(i,l)}$ and $V^-_{(i,l)}$ be the number of votes 'most preferred' and 'least preferred' for clip $l$ from camera $i$. We compute the absolute score of a clip as $V_{(i,l)} = V^+_{(i,l)} - V^-_{(i,l)}$, and label the clips $V_{(i,l)} > 0$ with *good* and $V_{(i,l)} < 0$ as *bad*.

### 4.3. Results

We use a 2-class SVM to learn a *good/bad* predictor of clips. The model is trained on 740 labeled data points, using 10-fold cross validation, taking 6.961 seconds on a quad-core computer. Our validation data set includes 30 clips from 10 different time periods of a game. When applying the trained SVM classifier to the remaining validation dataset, we are able to achieve a validation result of $\mathbf{74.1}\%$ when comparing the SVM prediction to the user preferences.

To assist directors with camera choices, we need to rank the camera views for each frame. We use the margin of a prediction result in SVM as a ranking for views. To visualize the effectiveness of our method, we automatically generate a broadcast video with the test set which consists of 30 consecutive clips. We use our trained SVM model to predict the camera of choice for each clip. We show comparisons between the human directed broadcast video from the dataset and our automatically generated broadcast video using the same input videos in Figure 3. The video can be found in supplementary materials. As shown, the most recommended camera computed from our algorithm for each frame is consistently better than the two less recommended ones. Furthermore, our classifier is based on the audience's preference rather than the director's preference, and does not agree with the director's choice in roughly half the time. This suggests that directors often employ their own story-telling style, which gave rise to our next approach.

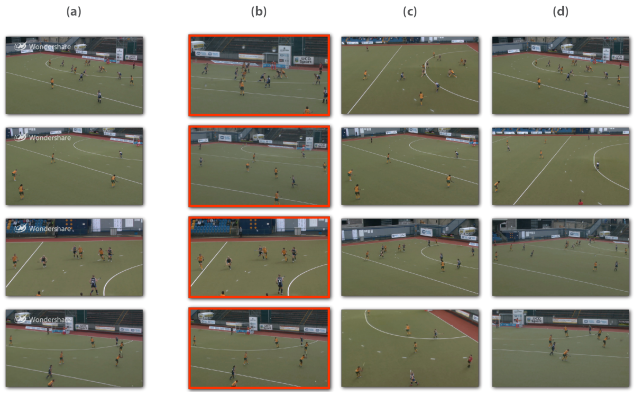


**Fig. 3. Comparison between the automatic suggestions with the computational aesthetics approach and the director's choice.** We show side-by-side comparisons of 4 frames drawn with equal spacing (15 seconds apart). (a) The director's choice for the 4 frames as shown in the broadcast video. (b) The most recommended camera generated with our algorithm. (c) The second most recommended camera. (d) The least recommended camera.

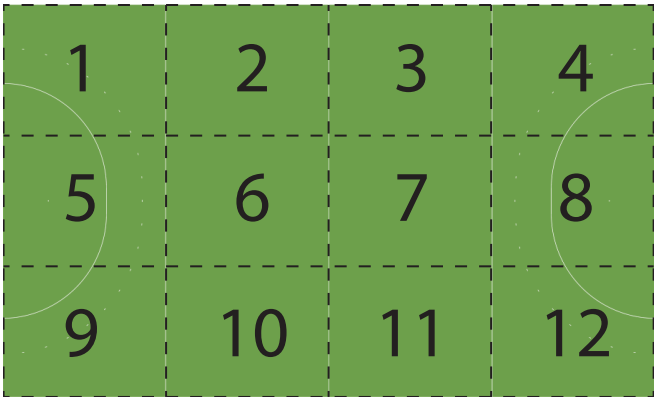


**Fig. 4. Labeled Regions of a Hockey Pitch.** We divide the field into 12 regions of equal size. We count the number of players in each region to generate the player location features.

## 5. LEARNING DIRECTOR STYLES

While our first approach allows to please a general audience, many directors may want to preserve their distinct style for the automatic camera selection. To acknowledge these individual artistic styles, we propose a second prediction method, where we learn a predictor based on previous footage of a given director.

### 5.1. Features

In order to evaluate the smoothness of a video, we use the same notation of a 15-frame micro-shot as defined in Section 4.1. We consider a different set of features, as the high level concepts we are trying to quantify are fundamentally different. In the previous section, we employed features that were associated with sports aesthetics to please a general audience. In this section, we use features that reflect director's choices based on higher-level game information.

1. **Player distribution**: Our first class of features describes the distribution of the players on the field. We divide the pitch into 12 equal areas, and perform player counting in these areas only (Figure 4). The first 12 features are then computed as the number of players in each area averaged over one micro-shot.

2. **Game flow**: As second class of features we describe the flow of the game. Denote the number of players for the region $d$ in frame $F_k$ as $\text{PlayerCount}_{(k,d)}$, then

$$f'_{\text{flow}}(d) = \text{PlayerCount}_{(k_0,d)} - \text{PlayerCount}_{(k_0-14,d)}$$

is defined as the temporal flow of players for one area over one micro-shot. Note that one such feature is computed for each of the 12 areas.

3. **Camera Movement**: As last set of features, we include the temporal motion of the cameras. For each camera $i$ at frame $k$, we use the parameters $\text{Pan}_{(i,k)}$ and $\text{Tilt}_{(i,k)}$ to construct rotation features. Further, we include $\text{Zoom}_{(i,k)}$ to construct magnification features. We compute the first five camera features for the $i$-th camera as the average of $\sin(\text{Pan}_{(i,k)})$, $\sin(\text{Tilt}_{(i,k)})$, and $\text{Zoom}_{(i,k)}$. In addition, we include the magnitude and direction of the camera movement over one micro-shot, and construct a second set of five features as the difference between the value at the start and end of the micro-shot for each of the above three measures.

### 5.2. Training

In contrast to the aesthetics-based approach, meaningful labels can be easily derived from a director's previous footage. More specifically, we assign a camera identifier to the label $L_k$ that has been chosen as 'live' camera at frame $k$. As a result, we can use a higher-dimensionality feature set, and provide many more training examples. We therefore use a random forest model [15] to train the predictor, rather than an SVM, which provides faster performance for training and testing on larger datasets.

### 5.3. Results

We used a random forest model implemented by OpenCV to train the 3-class classifier. The number of trees is determined automatically by the algorithm. The training was performed on 2000 frames randomly selected from 35 minutes of data, and took 1.037 seconds on a quad-core computer. We chose the same number of data points from each class. Inside each

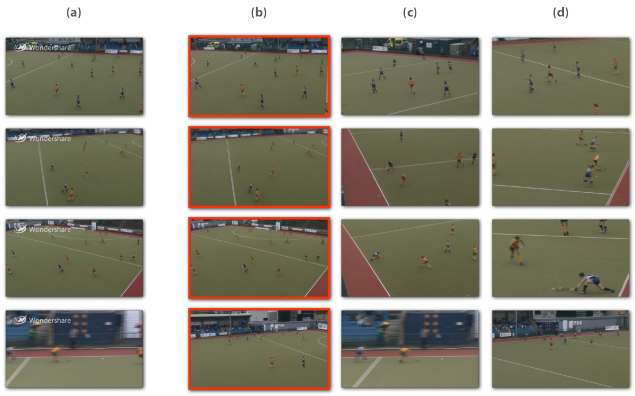


**Fig. 5**. **Comparison between the automatic suggestions trained with raw data and the director's choice.** We show side-by-side comparisons of 4 frames drawn with equal spacing (10 seconds). (a) The director's choice for the 4 frames as shown in the broadcast video. (b) The most recommended camera generated with our algorithm. (c) The second most recommended camera. (d) The least recommended camera.

class, we picked the training data randomly and made sure that the micro-shots for the selected data do not overlap.

For testing, we used 35 minutes (52500 frames) of data. We compared the first choice suggested by our algorithm with the director's choice and achieved an accuracy of $\mathbf{51.59}\%$. If we also consider the second choice, the director's choice in $\mathbf{93.79}\%$ of the frames were in either the first or second choices suggested by our algorithm. The results are promising and the accuracy is significantly higher than using a random choice ($33.33\%$).

To compare the automatic suggestion and the director's choice, we composed a video from the best shot suggested by our algorithm and compared it side by side with the director's choice for the testing data set. In general, our automatic camera ranking does reflect the director's preference very well. A few screen shots are shown in Figure 5, and an analysis is described in the next section. The full video can be found in the supplementary materials. It can be seen that our suggested choice echoes the director's choice in the majority of the frames.

## 6. COMPARISONS BETWEEN METHODS

We proposed two different data-driven approaches above for automatic shot suggestion. Despite computing features from the same dataset, the features and labels are constructed in different ways and with different intents. Features used to learn aesthetic models are based on heuristics, such as the rule of thirds, and player/ball visibility. They are designed predict how appealing a given camera is at any time, and do a good job of predicting user preference. However, they do not perform as well at predicting the high-level information that directors use to determine camera cuts, such as the flow and current state of the game.

The features used in the directorial style approach are designed to provide a more high level model of game semantics. One disadvantage of the directorial features is that the dimensionality is much higher compared to the aesthetic rules (25 vs 54 dimensions). In order to learn an accurate model of such high dimensionality a large amount of data would be needed, which is hard to acquire from user studies. The directorial style approach does not suffer from this limitation, as existing broadcast video can be used as training data.

## 7. CONCLUSION

We have presented two methods for automatic shot suggestion using computational approaches. The first method uses computational aesthetics and a visual preference-based user study to learn how to predict whether a clip of sports recording is visually appealing. The second method uses camera and tracking-based features to learn an individual director's style. Both methods show promising results, and are able to predict the preferred cameras faithfully.

Our implementations have several limitations that can be addressed in future work. The evaluation data set included a limited set of tracking features only, and we believe that more features such as the position of the ball could improve the results significantly. Similarly, we only considered the number of players in different regions, but robust player tracking would allow us to consider individual player movement. Furthermore, while we applied our work to field-based sports, it would be interesting to extend our learning method to other sports. We believe that our method should be extendable to any ball-based sport.

Despite these limitations, we believe that computational assistance has the potential to become a very important part of live sports broadcast, and contains many areas for future work. Given camera parameters and player tracking data, our system aims at real-time performance in shot suggestion. Generating replays in sports broadcasts is also a very interesting but little-explored area. To do this, the camera ranking system would have to be extended to not only recommend the camera with the best shots matching the compositions, but also determine the boundary of the shot for the replay.

In addition, our algorithms might very well be applicable to sports computer games. In such virtual environments, all the data on player location and camera parameters is readily available, and it would be very easy to incorporate our best shot suggestion and scripting algorithms. This would allow to essentially learn a director's style, and apply it to an in game virtual camera. Users could then select from a range of director's styles to find one that best suits their viewing expectations.

## 8. REFERENCES


[1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Wang, "Studying aesthetics in photographic images using a computational approach," *Computer Vision–ECCV 2006*, pp. 288–301, 2006.

[2] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato, "Sensation-based photo cropping," in *Proceedings of the Seventeen ACM International Conference on Multimedia - MM '09*, New York, New York, USA, Oct. 2009, p. 669, ACM Press.

[3] Pere Obrador, Ludwig Schmidt-Hackenberg, and Nuria Oliver, "The role of image composition in image aesthetics," in *2010 IEEE International Conference on Image Processing*. Sept. 2010, pp. 3185–3188, IEEE.

[4] Congcong Li, Andrew Gallagher, Alexander C. Loui, and Tsuhan Chen, "Aesthetic quality assessment of consumer photos with faces," in *2010 IEEE International Conference on Image Processing*. Sept. 2010, pp. 3221–3224, IEEE.

[5] Lai-Kuan Wong and Kok-Lim Low, "Saliency-enhanced image aesthetics class prediction," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. Nov. 2009, pp. 997–1000, IEEE.

[6] Anush K. Moorthy, Pere Obrador, and Nuria Oliver, "Towards computational models of the visual aesthetic appeal of consumer videos," in *Proceedings of the 11th European Conference on Computer vision*, Sept. 2010.

[7] Chris J. Needham, Chris J Needham, and Roger D Boyle, "Tracking multiple sports players through occlusion, congestion and scale," *British Machine Vision Conference*, vol. 1, pp. 93 – 102, 2001.

[8] Sachiko Iwase and Hideo Saito, "Parallel tracking of all soccer players by integrating detected positions in multiple view images," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 4, pp. 751–754.

[9] Tiziana D'Orazio, Nicola Ancona, Grazia Cicirelli, and M Nitti, "A ball detection algorithm for real soccer image sequences," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. IEEE, 2002, vol. 1, pp. 210–213.

[10] Xinguo Yu, Changsheng Xu, Hon Wai Leong, Qi Tian, Qing Tang, and Kong Wah Wan, "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video," in *Proceedings of the eleventh ACM international conference on Multimedia - MULTIMEDIA '03*, New York, New York, USA, Nov. 2003, p. 11, ACM Press.

[11] T DOrazio and M Leo, "A review of vision-based systems for soccer video analysis," *Pattern Recognition*, vol. 43, no. 8, pp. 2911–2926, 2010.

[12] Jinjun Wang, Changsheng Xu, Engsiong Chng, Hanqing Lu, and Qi Tian, "Automatic composition of broadcast sports video," *Multimedia Systems*, vol. 14, no. 4, pp. 179–193, Mar. 2008.

[13] Kyuhyoung Choi, Sang Wook Lee, and Yongduek Seo, "A friendly location-aware system to facilitate the work of technical directors when broadcasting sport events," Tech. Rep., School of Media, Sogang University, Seoul, Korea, 2011.

[14] Christopher J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[15] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.