Beyond The Kiss-Cam: Measuring The Fan Using Computer Vision-based Analytics

George Williams Christoph Bregler Ian Spiro
Dept. of Computer Science, Courant Institute, New York University
george,chris,ian@movement.nyu.edu

Abstract

What are fans really doing at the game? When are they watching the action on the field/court, and when are they buried in their phone or tablet ? Which ads do they notice on the JumboTron, and which ones do they ignore ? Whos joining in the wave and whos not? Are they having fun ?

If you are a club trying to retain seat-buying fans, fill empty chairs, or effectively use an advertising budget, then these are important questions to answer. Video surviellance tools exist, but sifting through crowd footage does not scale.

In this paper, we describe a computer vision-based software method that goes beyond just capturing the fans on video. Through training, it can recognize distinct fan activity automatically. We combine both machine learning and inexpensive, crowd-sourcing techniques to achieve speed and scale. Moreover, we show how its possible to preserve the anonymity of the fans during the analysis.

We conducted experiments on a crowd at a college basketball game and describe our approach and results.

1. Introduction

In 2012, TicketMaster reported that 26% of available stadium seat tickets went unsold, representing an estimated loss of 100 millions dollars for the year [3]. Recent work in fanalytics seek to remedy this situation by analyzing mountains of information for nuggets of information about fans- from ticket purchase history to opinions and sentiment found on blogs and social media [4]. Little work has been done to learn precisely what fans enjoy or don't about the live game day experience. Presumably, such information would be useful for a club trying to attract and retain fans.

Customer interviews, surveys, and questionnaires might be useful but they are limited due to their ad-hoc nature and limited sampling. Moreover, it may not be clear what to ask fans about their experience. They may not remember why exactly they weren't watching a particular ad or why they weren't engaged in the game action at a specific point in time. Clearly, the direct observation of fans would be an invaluable tool. Yet, the detailed account of fan behavior during a game is so far an under-served dimension of fan-based analytics. Capturing fans on video is a start, but how can such observation scale to large stadium crowds ? How do you ensure fan privacy ?

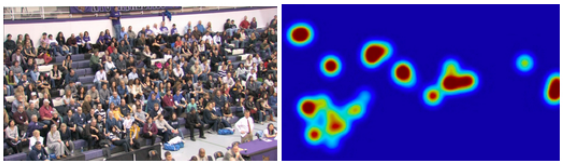

Figure 1. Heat map showing clapping fans in a frame of video.

We explore a software approach to overcome these obstacles, using a method we call 'site-specific learning.' The idea is to record complete video of the crowd at a game. Then, we use a portion of the crowd video footage to train a convolutional neural network. When trained properly, this neural network has the ability to recognize and detect chosen types of fan activity, in all of the footage.

Figure 1 demonstrates a result on a crowd watching a basketball game. We show one frame of video. The heatmap, on the right, shows locations of the fans, on the left, performing the same activity, in this case, clapping.

2 Related Work

"Fanalytics" covers a broad range of topics in fan-based analytics as it pertains to driving revenue, including ticketing analysis and dynamic pricing models [4] [3] [5] [6] [7].

From a computer vision standpoint, our work most closely fits into the categories of "crowd image analysis" or "surveillance tracking", for which there are dedicated workshops and conferences [8] [9]. In particular, we were inspired by the following papers [10] [11] in this domain.

Since our method detects individual fans and isolates their activity, we also draw upon a body of research in activity recognition that spans several decades and many conferences [12] [13] [14] [15] [16] [17] [18] [19] [20] [21].

This paper does does not address how to interpret fan activity, once its isolated in video. For example, an assessment of fan engagement may involve more than detecting obvious forms of expressive positive sentiment like clapping or cheering. Interpreting human gesture is an active area of research and involves a multidisciplinary approach from areas such as human anatomy and psychology.

3 Approach

The basic steps involved in our approach are summarized as follows:

- 1. Collect lots of video footage of the crowd at a game and venue.
- 2. Perform face detection on the video.
- 3. Isolate the body around each face detected and extrapolate short video clips of the fan (we call these "fidgets").
- 4. Randomly select a subset of fidgets and identifiy the fans behavior in each (we call this process "labeling").
- 5. From the set of labels, choose an activity (like, clapping).
- 6. Create a training database comprised of the labeled fidgets.
- 7. Train a neural network using the training database, and monitor its progress using a validation dataset (a portion of the original training set aside for this purpose).
- 8. Use the trained neural network to detect the activity of interest in all of the fidgets for all of the videos.
- 9. Repeat steps 4-8 for another activity of interest.

The following sections explore the stages of this pipeline in further detail. Mainly, we demonstrate how we created a clapping-fan detector using video footage of the crowd watching a college basketball game. The method generalizes to other types of fan activity.

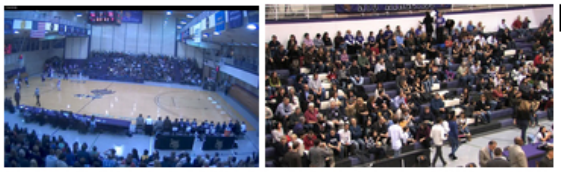

Figure 2. Two of the camera views of crowd footage.

3.1 Data Collection: Game Day

We set up 4 HD cameras at a college basketball game and positioned them above the audience at balcony level. We aimed 3 cameras at the most populated areas of the stands. The camera views did not overlap. We manually adjusted the zoom in order to capture as many fans as possible in each view, but not so far zoomed-out that each face spanned no less than around 20 pixels (the software-based face detector has difficulty detecting faces less than 20 pixels.) We aimed the final camera at the basketball game and adjusted the zoom to capture the action at both ends of the court, including the scoreboard which hovered above the center of the court. Figure 2 shows 2 of the camera views.

We recorded the entirety of two basketball games, a womens college game and a mens college game. We also captured the half-time show of each game as well.

3.2 Face Detection

The first stage in our data processing pipeline is software-based face detection. We used an off-the-shelf face detector from PittPatt.com, which was acquired by Google in 2012 [1]. Since we use face detection to detect fans, it was important to capture as many faces as possible. Highly accurate face detection can be time-consuming, so instead of each frame we performed face detection on every 50th frame. We felt we could justify this optimization because, over the course of a game, a fan is not likely to change their position much, especially over a 50 frame period.

3.3 Fidgets

For each face detected, we extracted a region of interest larger than the face itself. This was an attempt to capture the person's entire body, not just their face. We experimented with several parameters. We found that by increasing the height of the face detection bounding box by a factor of 4 and the length by a factor of 3, we could capture the fan not only sitting but also standing as well. We applied this region of interest to the following 39 frames that followed the face detection frame, resulting in a small 40 frame video clip of the fan that we call a "fidget."

Figure 3. The first five frames of a sample fidget.

The fidget may or may not contain anything interesting, and in general, most fidgets do not. Also, the length of 40 frames (a little over 1 second for our 30 FPS video) might seem short. But, as it turns out, a 40-frame fidget is enough to capture a myriad of distinct fan activities including clapping, talking, cheering, eating, using a cell phone, etc. It was important to keep the number of frames as small as possible as not to overload inputs of the neural network later on. Figure 3 shows the first five frames of an example fidget of a fan clapping:

3.4 Machine Learning

Our approach is based on machine learning, and specifically, a class of algorithms called convolutional neural networks. Convolutional Neural Networks (a corner stone of Deep Learning [2]) are not new, but recent advances in compute power make training and using these learning algorithms more practical than ever.

A treatment of the state-of-the-art in neural networks and machine learning is beyond the scope of this paper, but they all have one thing in common - they all require a learning phase in which the algorithm is trained on lots of examples.

3.5 Building A Training Database

To build a training database for our neural network, we used a small portion of the video we collected at the event. We randomly sampled a small subset of the fidgets extracted in a previous step and then identified the fan's activity, such as clapping, cheering, talking, etc. Labeling on this scale can be time consuming, so, to save on time, we crowd-sourced the job to hundreds of workers via Amazon Mechanical Turk. Figure 4 shows the task we presented to the workers.

To ensure label accuracy, we presented each fidget to three different Turkers. If a majority answered the same way, then we accepted the Turks label for this fidget. We paid two cents per task.

3.6 Training The Neural Network

This section summarizes how we trained the convolutional neural network, so some readers may want to skip it. For a tutorial on convolutional neural networks architecture, please see [2].

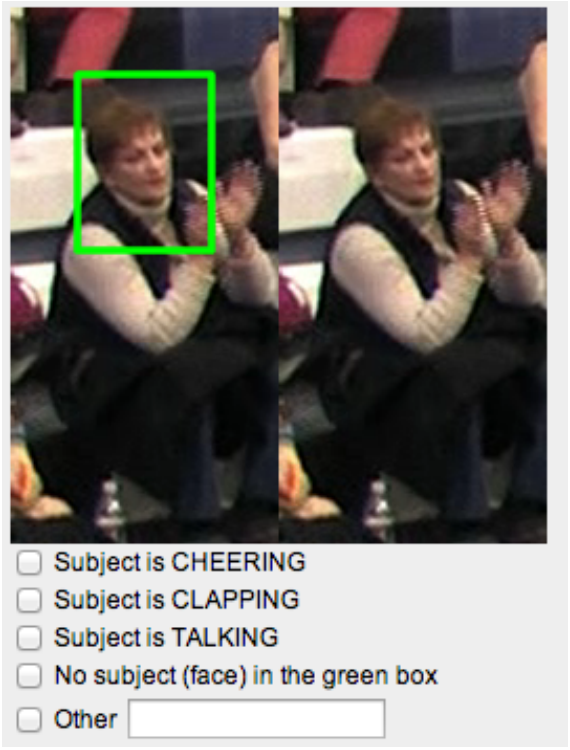

Figure 4. The labeling task we presented the Turkers.

The training database we presented to this neural network was comprised of a total of about 1840 Turk-labeled fidgets, 940 were "positive" training samples of fans clapping and 900 were "negative" training samples, comprised of fans that were either cheering or talking but not clapping. 20% of each set were set aside as a validation set, which was used during the training phase to guage its progress.

Before training, the neural network weights were randomized. During each training iteration, or epoch, a new batch of 8 fidgets were randomly sampled from the training database, 4 from the positive set, and 4 from the negative set.

The processing performed in each epoch is summarized as follows:

- each fidget in the batch is rescaled to the same dimensions 128w x 256h (40 frames is maintained in the time dimension)
- each fidget in the batch is then contrast normalized, using a 3X3 kernel of coefficient 1.591
- the first layer is a fully connected, containing 8 feature maps, each of size 9x9x5
- next is a downsampling stage of size 4x4x3

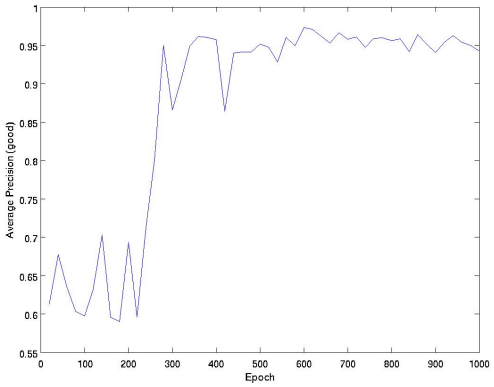

Figure 5. Training performance of the convolutional neural network over 1000 epochs.

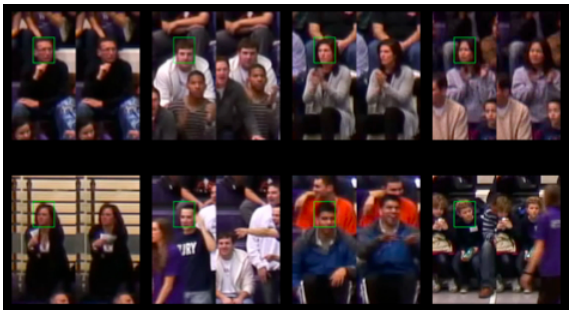

Figure 6. On the top row, our clapping-fan detector predicts accurately (true positives). On the bottom row, false positives are shown. Left-to-right: a fan fanning themselves, a fan moves quickly out of view, a fan disagrees with a call, another fan walks in front of fan.

- the next layer in the network containing 8 feature maps, each of size 9x9x5, not fully connected with instead 4 connections
- next is a downsampling stage of size 2x2x2
- finally, the output is a set of label predictions for each fidget in the batch

We evaluated the performance of the neural network at every 20th epoch using the Turk-labeled validation set. The results are shown in Figure 5, which appear to demonstrate a classic machine learning effect at around epoch 300.

4 Discussion

4.1 Performance

Its important to monitor the training of the neural network, because its very possible that the training database is insufficent. Likely this is due to a lack of positive training samples. For example, a boring game might not have a lot of "cheering". Thus there are too few "cheering" samples with which to train the network. Conversely, even in cases where there is demonstrable learning shown on the validation set, the detector may produce "false positives." This is likely due to a negative training set that is not complete. Figure 6 hows some false positives produced by our clapping-fan detector.

With respect to processing time, the site-specific learning method we demonstrate can initiate right after game capture, and results could be ready a day after if not sooner. Actually, much of the video processing like face detection and fidget generation can occur the moment cameras dump footage. Surprisingly, the labeling stage is not the most time consuming. It took less than an hour for Amazon Mechanical Turk's tens of anonymous workers to label the fan activity in 3000 fidgets. The most time- and compute-intensive stage is machine learning. Still this is not bad. On gpu-hardware, it takes only several hours to train a convolutional neural network for 300 epochs using the architecture we described.

4.2 Cost

4.3 Privacy

Legally speaking, there is no "expectation of privacy" at a public performance such as a sports event. And, in general, we are all gradually coming to terms with the fact that we have little to no privacy from video surveillance in many areas of their lives. Still, game fans may not be too wild having their every move monitored and recorded at a game, especially where there is alcohol in play. To remedy this, venues can instead enable viewing of the data via only the resulting heatmap 1. The heat-map stores the salient information anyway-indicating the presence and the location of the activity of interest. The heat-map essentially decouples the measurement of the fan's behavior from their identity. The venue could go further, and, once processing an event has completed, store only the result of the computations (the label predictions of the neural network) and delete the original crowd footage. Technicians can pull up the information later and visualize them in the heat-map form.

5 Future Work

6 Conclusion

This paper described a machine learning-based method that can be trained to detect specific kinds of fan activity in crowd footage. We believe that this technique and others like it are a key component in a

broader framework to understand audience engagement at sports events and other public performances. Clubs and event venues can use these analytics to improve customer experience, retain fans, and ultimately, increase ticket revenue.

7 Acknowledgements

We would like to thank Will Freeman, Peggy Hackney, Chris Shell for useful discussions and pointers, and the Office of Naval Research (ONR N000140910076, N000140910789).

References

[1] PittPatt.com, Techcrunch. http://techcrunch.com/2011/07/22/google-acquires-facial-recognition-software-company-pittpatt/. news story, Techcrunch, 2011.

[2] DeepLearning.net. http://deeplearning.net/. general information, 2013.

[3] J. Forese Ticketmaster LiveAnalytics: Big Data and Sports Ticketing Talks and Presentations, Sloan Sports Conference 2012.

[4] T. Brosnan, M. Cuban, N. Hubbard, J. Kraft, J. Walsh, B. Simmons Fanalytics Panels, Sloan Sports Conference 2012.

[5] J. Forese Ticketmaster: Fan Analytics And Insights Talks and Presentations, Sloan Sports Conference 2012.

[6] S. Springer, W. Townshend, B. Lafemina, C. Gahagan, L. DePaoli Ticketing Analytics Panels, Sloan Sports Conference 2012.

[7] B. Lafemina, C. Gahagan, L. DePaoli, D. Maged, J. Gelman Ticketing Analytics, Presented By Stub-Hub Panels, Sloan Sports Conference 2012.

[8] Performance Evaluation of Tracking and Surveillance http://pets2012.net Conference, 2009-2012

[9] Workshop on Modeling, Simulation, and Visual Analysis of Large Crowds http://vision.eecs.ucf.edu/ICCVWorkshop/home.html Conference, 2011-2012

[10] M. Rodriguez,S. Ali,T. Kanade Tracking In Unstructured Crowd Scenes. In *ICCV*, 2009.

[11] G. Weina, R. Collins, B. Ruback Automatically Detecting Small Group Structure of a Crowd. *Applications of Computer Vision*, IEEE, 2009.

[12] J. O'Rourke, N. Badler. Model-based image analysis of human motion using constraint propogation. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2(6):522-536, 1980.

[13] D. Hogg. Model-based vision: a program to see a walking person. In *Image and Vision Computing*, 1(1):5-20, 1983.

[14] M. Yahamoto, K. Koshikawa. Human motion analysis based on a robot arm model. In *Computer Vision and Pattern Recognition*, 1991.

[15] D. Gavrila, L. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *Computer Vision and Pattern Recognition*, 1996.

[16] K. Rohr. Towards model-based recognition of human movement in image sequences. In *CVGIP-Image Understanding*, 59(1):94-115, 1994.

[17] J. Rehg, T. Kanade. Model-based tracking of self-occluding articulated objects. In *Computer Vision*, 1995.

[18] C. Bregler, J. Malik. Tracking people with twists and exponential maps. In *Computer Vision and Pattern Recognition*, 1998.

[19] L. Kakadiaris, D. Metaxas. Model-based estimation of 3d human motion. In *Pattern Analysis and Machine Intelligence*, 22(12):1453-1459, 2000.

[20] J. Deutscher, A. Blake, I. Reid. Articulated body motion capture by annealed particle filtering. In *Computer Vision and Pattern Recognition*, 2000.

[21] H. Sidenbladh, M. Black, D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Computer Vision ECCV*, pages 702-718, 2000.

[22] Laban Movement Analysis. http://en.wikipedia.org/wiki/Laban_Movement_Analysis general information, 2013.

[23] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome general information, 2013.