# A Photo Booth That Finds Your Sports Player Lookalike

Mitsuru Nakazawa, Tomoyuki Mukasa, Björn Stenger
Rakuten Institute of Technology, Rakuten, Inc.
Rakuten Crimson House, 1-14-1 Tamagawa, Setagaya-ku, Tokyo, 158-0094
{mitsuru.nakazawa,tomoyuki.mukasa,bjorn.stenger}@rakuten.com

## Abstract

*This paper presents a digital signage system that encourages fan engagement at sports venues. The system detects users' faces and displays the team player who looks most similar to the user in a composite image that they can capture as a souvenir. We describe the design choices of this photo booth and evaluate the similarity search accuracy in several user studies. The system has been operating continuously at a baseball stadium with consistently high user engagement. So far over 12,000 people have taken their souvenir photo.*

## 1 Introduction

We describe a digital signage system with the goal to provide a fun experience to sports fans and enhance their engagement with the team that they support. As interactive digital signage systems are becoming more common, there is a risk that the novelty factor wears off, and people do not feel compelled to engage with the system. Several methods have been employed to engage potential users, such as generating a composite image or video of the user with a popular character, for example, *Hello Kitty* [2], which can be shared on social media. In this system, the same character is shown for all users, thus the resulting image is predictable. On the more interactive side of the spectrum, conversational systems aim to engage users by letting them talk with a virtual character, such as the virtual sales assistant systems *Millie* [9], or *Litchi* [3]. Motion-based approaches have also been employed in some digital signage systems for immediate user interaction [12]. The *MoveMirror* [1] provides a fun user experience by matching a user's movements with that of other people and showing those images in real-time. In this paper we propose a digital signage system acting as a photo booth that selects a lookalike player based on face similarity, providing a short, fun experience. As a side-effect, fans may learn about players that they are not yet familiar with. In the following we describe the design choices made to implement the system, and justify the approach in several user studies. We evaluate the accuracy of the similarity search compared to human perception. The photo booth has been in continuous operation for several months during the baseball season and has seen consistent user interest.

## 2 Photo booth system

The digital signage system consists of a large display, a laptop computer and an RGB-D sensor, allowing real-time image processing including foreground extraction, see Fig. 1a. To engage, one or two users stand in front of the display at about 1.5 meters distance. Once the system detects a user for a few seconds, it enters the *similar face search mode* as shown in Fig. 1b. In this mode, the three most similar baseball players are displayed on the sides of the display. After a few seconds delay the photo, biography and face similarity score of the top matching player are displayed, see Fig. 1c. Finally, the system switches into photo shooting mode, in which users can take a photo with their matching baseball player as shown in Fig. 1d. Throughout the experience, foreground regions are composed with a background stadium image for an immersive experience.

In our system, 24 popular players of a Japanese Baseball team are registered as retrieval targets for similar face search. For similar face search described in Subsection 2.1, we collect an image dataset from the internet for each player, containing at least 30 images for each. The images were taken under various lighting conditions and with various facial expressions, making the similarity search more robust.

### 2.1 Similar Face Search

In this section, we explain the the face similarity search procedure, where the face image of a photo booth user acts as a search query. First, we detect the face from both the input image and the image dataset of all retrieval targets using Multi-task Cascaded Convolutional Networks (MTCNN) [10]. We use FaceNet [8] to extract features from the detected faces. FaceNet embeds a face image into a space, in which matching faces of the same person are close to each other, which is achieved by minimizing a triplet loss during the training process. Let $\boldsymbol{v}_q$ be the feature vector of the query image, and $\boldsymbol{v}_{ij}$ be the feature vector


(a) Signage overview


(b) Similar face search mode


(c) Search result


(d) Photo shooting mode

Figure 1: **Overview of the proposed system.** *Users stand in front of the display (1) and after their face is detected, the most similar matches are displayed for up to two users (2). After a few seconds the closest match is shown (3) and users are prompted to pose for a composite photo shot (4).*

of the $j$-th image of the $i$-th retrieval target. The similarity score between the query and $i$-th retrieval target $s_i$ is calculated as follows:

$$s_i = \mathrm{med}_j^{N_i} \left( f\left(\boldsymbol{v}_q, \boldsymbol{v}_{ij}\right)\right), \tag{1}$$

where $f$ is the similarity function between two facial feature vectors and $N_i$ is the number of images of the $i$-th retrieval target. Because the feature space of FaceNet is an unit hypersphere [8], we use the cosine distance for $f$. Using the median of the distances for each player $i$ makes the score more robust to appearance variances and outliners.

## 3 Experiments

### 3.1 Preference survey of similar face search

To evaluate the performance of our face similarity search, we conducted a preference survey. We compare the similarity based on FaceNet features with different methods:

**Random selection**: As a baseline, we compare against random selection of a player among the 24 team members.

**Dlib**: The method for face detection and feature extraction was replaced with dlib's CNN-based detection [5] and ResNet-based descriptor [6]. The similarity

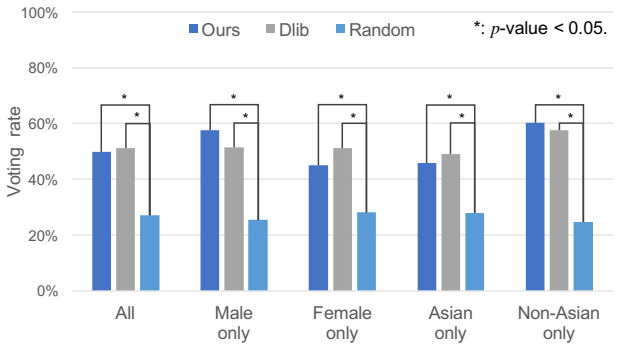


Figure 2: **User preference survey result.** *On average users shows slightly higher preference for the Dlib method, but the difference between FaceNet and Dlib is not statistically significant. Both methods were clearly preferred over random selection among the players.*

distance between two facial features was calculated by L2-distance as in [6]. The face detection and feature extraction parameters were set to default values.

As query images for the similarity search, we randomly selected 80 face images from the public *MegaAge* dataset [11]. The query images include 31 male and 49 female subjects, of which 58 are of Asian ethnic-

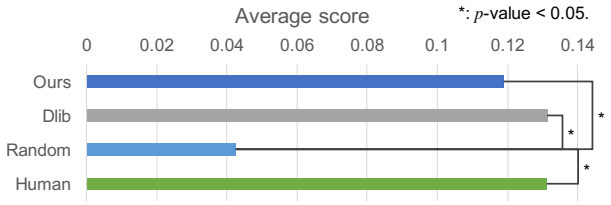


Figure 3: **Comparison with human search.** *To compare our system with human performance, we asked 10 subjects to select the most similar player to the same query images. The results show that our system has comparable accuracy to human selection.*

ity and 22 of other ethnicities. For each query image, the search results for each of the three methods were displayed in randomized order. In a user study with 107 subjects we asked the question: *Which of the three search results looks most similar to the query image?*.

Figure 2 shows the result of the preference survey. To measure the statistical significance, the $p$ values are calculated by the Steel-Dwass test. In Fig. 2, both FaceNet and Dlib clearly outperformed random selection with statistical significance. Between FaceNet and Dlib, the difference is statistically not significant ($p > 0.05$, see Fig. 2). The same trend can also be confirmed in the results of sub categories in Fig. 2.

### 3.2 Comparison with human similarity search

In a second user study we compare the system's similarity search to human ground truth. Face similarity is a subjective measure in this case, and different people may give different similarity rankings. Here we obtain ground truth by asking 10 volunteers to select the most similar player out of 24 for each query image. The query images were same as those of the preference survey. Using the manual search results, we calculated the accuracy of our method. In this evaluation, we define the accuracy as the average appearance frequency of the searched player by our method in 10 manual search results. In addition, the accuracy of manual search itself was also calculated by the same way as leave-one-out cross validation.

Figure 3 shows the performance comparison with manual search. The accuracy of the Dlib and random search are also shown in this figure. As with the preference survey, the $p$ values are calculated to measure the statistical significance. In Fig. 3, the accuracy of manual search is slightly better than ours. However, the difference was statistically not significant. The results show that the methods based on CNN-based features have comparable accuracy to human similarity search.

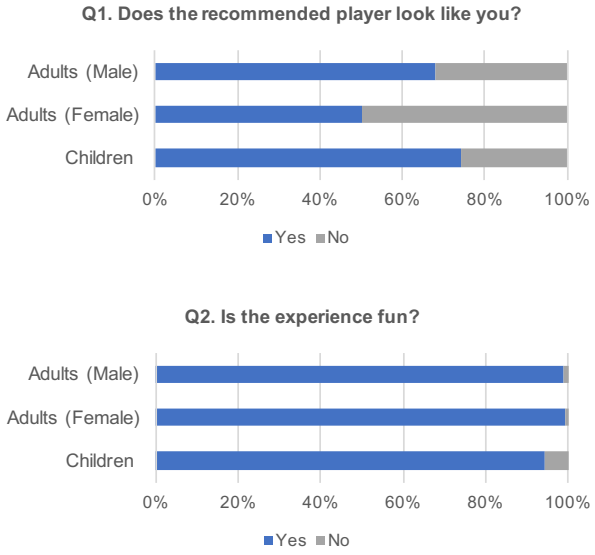


Figure 4: **Questionnaire results of the photo booth experience.** *Visitors mostly agreed that the displayed top matching players looked similar to them, and nearly all visitors enjoyed the experience.*

### 3.3 Run-time comparison

One of the most critical aspects for digital signage systems is response time in order to provide a smooth user experience. We therefore compared the run-times of face detection of ours and Dlib on a GPU server. Our average run-time was 70 milliseconds faster than the one of Dlib. Given that the similarity retrieval accuracy is not statistically significant, our method is selected in the final implementation in order to minimize the time spent for face similarity search.

### 3.4 User feedback on the proposed system

We conducted a quality test based on a questionnaire to obtain feedback from the booth visitors during an actual baseball game. We received 394 replies during one day as shown in Fig. 4. The questionnaire consists of two questions:

1. Was the top matching player similar to you?
2. Was the booth fun?

As the number of baseball players registered in the system is limited, similar search results are not always good matches, especially for female users (see top of Fig. 4). However, as can be seen from the responses to Question 2 (see bottom of Fig. 4), almost all visitors enjoyed the booth. As a result of the positive visitor feedback, we installed two photo booths at a baseball

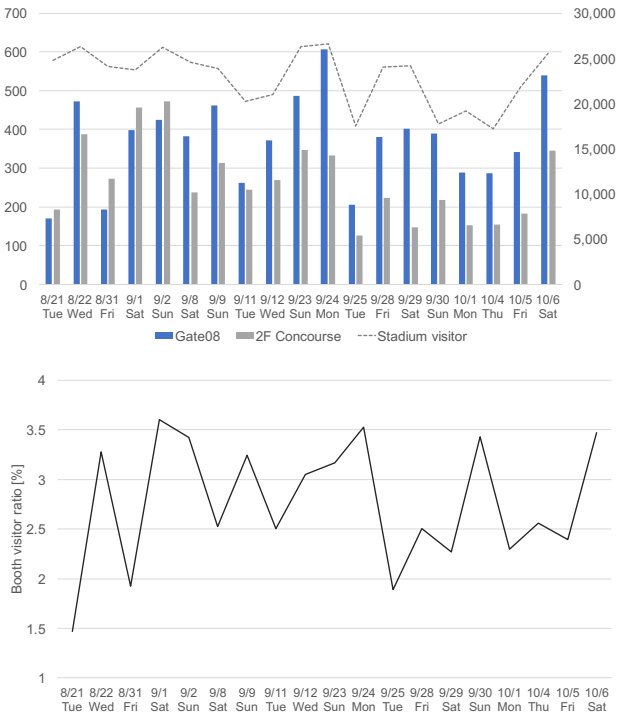


Figure 5: **Visitor statistics.** *The number of visitors on different game days (Top) and the ratio of booth visitors to total stadium visitors (Bottom) show that the engagement with the digital signage system remains high throughout the baseball season. The number of stadium visitors represented as the dot line in Top is quoted from [4].*

stadium in different locations (one at a gate, another one at a second floor concourse, indicated as "Gate08" and "2F concourse", respectively in Fig. 5). So far the booths had over 12,000 visitors during 19 league games. Figure 5 shows the number of visitors to the booths. In general, there are many fans who regularly attend the games to support the team. Therefore we may assume that there are several people who try the booth several times. In Fig. 5, we can see the both booth visitor numbers and the ratio compared to the total stadium visitors did not decrease much over the course of the season. This indicates that our system succeeded in keeping people's attention over time.

## 4 Conclusion

In this work, we proposed a photo booth style digital signage system that draws peoples' attention by showing a lookalike person to each visitor. To calculate the face similarity, we confirmed that an approach using Multi-Task Cascaded Networks for detection and FaceNet features for similarity search shows high retrieval performance at perceptual quality comparable to humans. In the future it will be interesting to use features designed specifically for perceptual face similarity instead of recognition performance as in the recent work by Sadoynik *et. al.* [7]. We validated the novelty of our system by demonstrating it to more than 12,000 people in a public space with largely positive feedback in our user feedback.

## Acknowledgments

The authors would like to thank Rakuten Baseball, Inc. and the Brand & Marketing Strategy Department, Rakuten, Inc. for kindly supporting this work.

## References

[1] Google Creative Lab. Move Mirror. `https://experiments.withgoogle.com/move-mirror`, accessed on Mar. 17, 2019.

[2] Pony Canyon, Inc. HELLO KITTY the digital signage vending machine. `https://news.ponycanyon.co.jp/2018/09/26727/`, accessed on Dec. 11, 2018.

[3] Alibaba. `https://bit.ly/2Cd6hmb`, accessed on Dec. 14, 2018.

[4] Baseball freak. `https://baseball-freak.com/audience/18/eagles.html`, accessed on Mar. 19, 2019.

[5] Dlib. Easily create high quality object detectors with deep learning. `http://blog.dlib.net/2016/10/easily-create-high-quality-object.html`, accessed on Dec. 11, 2018.

[6] Dlib. High quality face recognition with deep metric learning. `http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html`, accessed on Dec. 11, 2018.

[7] A. Sadovnik, W. Gharbi, T. Vu, and A. Gallagher. Finding your lookalike: Measuring face similarity rather than face identity. In *Proceedings of the 1st CVPR Workshop on Visual Understanding of Subjective Attributes of Data (V-USAD)*, 2018.

[8] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)*, 2015.

[9] TwentyBN. `https://20bn.com/`, accessed on Dec. 14, 2018.

[10] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[11] Y. Zhang, L. Liu, C. Li, , and C. C. Loy. Quantifying facial age by posterior of age comparisons. In *Proceedings of the 28th British Machine Vision Conference (BMVC)*, 2017.

[12] ZiiCON. HILLINGER. `https://www.youtube.com/watch?v=DT_-1wsDQkE&t=2s`, accessed on Mar. 22, 2019.