Although I can't See, I can Hear: Speech-Assisted Self Portraits on Mobile Devices

Nathan Sakunkoo, Patty Mink Sakunkoo
Stanford University
CA 94305, USA
{sakunkoo, psak}@stanford.edu

ABSTRACT
Self-portraiture is an important activity in photography. Yet, it remains difficult for mobile users who often capture self-portraits spontaneously. This demo presents a prototype that combines computer vision with audio interface to complement the conventional visual interaction, thus assisting users to capture better-framed self-portraits.

ACM Classification: H5.2. Information interfaces and presentation: User Interfaces

Keywords: Self-portrait; camera phone; audio interface

INTRODUCTION
Self-portraiture has long held a fascination for human beings. From the very earliest of times, people gazed at their own reflection in a pool of water, many a time with the desire to know and to capture their own likeness. That many Renaissance artists depicted themselves as the main subjects of their paintings further echoed the intensified popularity of self-portraiture [4]. Today, self-expression is not only an artistic pursuit but is also of high significance in casual photography. Especially in some cultures, self-shot is a dominant mode of camera phone use [3].

Despite the importance of self-portraits and advancement in camera technology, there remains a fundamental problem for users who capture images on casual impulses: since a camera phone user often takes pictures spontaneously [5], it is usually impossible for him to always carry a tripod to capture a nice self portrait with the desired background. In absence of a handy helper, he may use ad hoc means such as putting the mobile camera on a table, leaning it against the wall, or hanging it (Figure 1). Such ad hoc means often limit the user's access to the camera's viewfinder, thus making it difficult for him to compose the desired frame.

Figure 1: Existing solutions *vs.* proposed speech- and computer vision-based solution.

Copyright is held by the author/owner(s).
UIST'08, October 19-22, 2008, Monterey, California, USA
ACM 978-1-59593-975-3/08/10.

In this demo, we propose the use of conversational speech as a complement to conventional user-camera visual interaction. Instead of relying on extraneous hardware or letting the users perform the task by trial and error, *FindFrame* offers a simple computer-vision-based interface to verbally assist users to capture better-framed self-portraits. The rationale is that even though, in some circumstances, a user finds it impossible to see the viewfinder when he takes the picture, he can usually capture the approximate scene (without himself in) as close as possible to his desired scene (with himself in the foreground). Thus, *FindFrame* first lets the photographer take an approximate picture of the desired scene and keeps the scene as a reference (see Figure 2). Then, the photographer puts his camera on an ad hoc placeholder. Based on the reference frame and the current viewfinder frame, *FindFrame* gives audio guidance ("left, right, up, down") so that the photographer can adjust the camera's angle accordingly even though he could not see the viewfinder. When the camera says "OK," the user can stop adjusting the camera and is given ten seconds to situate himself while the camera utters, "will take picture in ten seconds." Finally, the camera takes the picture.

Figure 2: *Left:* The system says, "Please choose a (reference) frame." *Center:* The user takes a reference frame (top right). *FindFrame* then guides, "Left, Left, Up, OK, Will take picture in 10 seconds." *Right: FindFrame* counts down from three and takes the self/group portrait. The result is shown (bottom). A video demo can be found at: http://www.cs.stanford.edu/~nathans/projects/findframe/ .

IMPLEMENTATIONS

Image Feature Detection
FindFrame first scales down each input frame to a 115x86 pixel image, which is then converted from RGB to luminance and blurred with a Gaussian kernel. Next, our system extracts the corner features of objects within the image by implementing the standard Harris corner detector [2]. This operator detects points in which the image's horizontal and vertical gradients, I_x and I_y, change significantly along both directions. Because we only want distinctive interest points, *FindFrame* then performs non-maximal suppression to eliminate superfluous features.

Point Correspondence Matching and Displacement

A measure of similarity is computed from the maximal correlation score, which allows the system to match corresponding features from the viewfinder frame with those of the reference frame, despite their differences in perspective. Each of the frames' horizontal and vertical differences is categorized into a negative, equal to, or positive bin. For each viewfinder frame, the bin with the highest frequency is deemed the most significant distortion that needs to be corrected by camera adjustment. Accordingly, our output classification is: {"Left", "Right", "Up", "Down", "OK"}. The output is then uttered via the device's TTS. Since speech is inherently a slow medium [6], we attempted to make *FindFrame*'s dialog brief, relevant, and orderly [1]. We also eliminated extraneous words as far as possible.

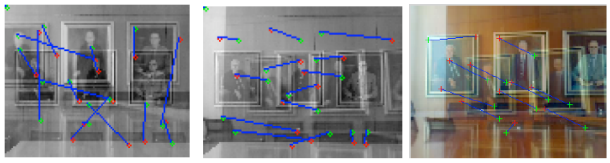

Figure 3: Examples of *FindFrame*'s features and displacement detection. The features detected in the desired scene are shown in red while those of the viewfinder frame are in green. *Left, Center, Right* – Vertical, Horizontal, and Diagonal displacements

Hardware

Our implementation platform is the Nokia N95, which has an ARM11 330 MHz processor with 8M memory. The system can be easily ported to other mobile platforms that allow *FindFrame* to take pictures without user interference.

EVALUATION

We evaluated our prototype by 1) measuring the performance on a set of scenes, and 2) soliciting informal user feedback, which informed our iterative design.

Performance Study: Simulated Tests

To systematically analyze performance, we captured video streams from the camera phone and ran the algorithm on a PC. Six test sets were captured from: 1) Outdoor (with landmark), 2) Outdoor (vast field), 3) Indoor (good lighting), 4) Indoor (low light, assorted objects) 5) Indoor (white wall, small object), and 6) Indoor (white wall, no object). Twenty five frames were tested for each test set.

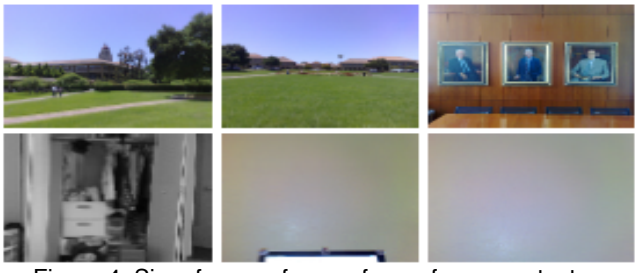

Figure 4: Six reference frames for performance tests.

The system performed well for the first four scenes, with over 80 percent guidance accuracy. The performance in the fifth scene was good except when the camera was panned upward too much. Intuitively, the system was disoriented in the sixth scene, which did not have an identifiable object. The study also shows that the degree of displacement does not hinder the system's capability to provide accurate guidance as long as an identifiable object is still in the scene.

Informal User Feedback

We sought informal feedback from four regular camera-phone users. We demonstrated how to use the system and then asked them to try out both in the same settings and freely. With preliminary feedback, we solved major usability problems. Then, further feedback was sought, the interface was further refined, and the users were again asked to validate changes. We observed their behaviors closely.

Through informal discussions, several problems were uncovered and solved. For example, two users initially mentioned that *FindFrame* gave confusing guidance. After investigation, we found that this is because asking the users to adjust the camera along both horizontal and vertical axes at the same time (i.e. diagonal axis such as "up-right") was impractical. Therefore, we simplified the dimensions to only horizontal and vertical axes, beyond which the alignment task becomes too confusing for the users. We chose to provide horizontal guidance first, then vertical, and so on because we observed that horizontal adjustment was easier. Furthermore, there was a considerable lag time between users' hearing and freezing their hands. We therefore split "OK" into two steps, "Freeze" and "OK", so as to give ample time for final adjustment. Another challenge was the difficulty of adjusting the vertical tilt of the N95 because it is slightly slanted down while most users preferred flat or slightly upward angles. Perhaps, future cameras could include an adjustable peg for solving this problem. Lastly, the photographers' heads were occasionally out of frame. A face detector could be promisingly incorporated in future versions to solve this problem.

Although we only tested the prototype on four users, the number too low to show statistical significance, the performance tests and qualitative feedbacks are encouraging. To follow up from this demo, we plan to conduct a formal, controlled user study, as well as incorporate a face detector algorithm to further aid alignment of foreground elements.

REFERENCES

1. Grice, H. P. *"Logic and Conversation,"* Syntax and Semantics: Speech Acts, Cole & Morgan, editors, V. 3, Academic Press, 1975.
2. Harris, C., and Stephens M. A combined corner and edge detector, *Proc. Fourth Alvey Vision Conference*, 1988, pp. 147-151.
3. Hjorth, L. Snapshots of Almost Contact: The Rise of Camera Phone Practices and a Case Study in Seoul, Korea. *Continuum 21.2* (2007), pp. 227-238.
4. Kelly, S. *The Self-Portrait: A Modern View*. London: Sarema Press, 1987.
5. Kindberg, T., Spasojevic, M., Fleck, R., Sellen, A. I saw this and thought of you: some social uses of camera phones, *Proc. CHI 2005*.
6. Lai, J., Yankelovich, N. Conversational Speech Interfaces. *The Human-Computer Interaction Handbook*. Lawrence Erlbaum Associates. 2003, pp. 698-713.