EasySnap: Real-time Audio Feedback for Blind Photography

Samuel White, Hanjie Ji, and Jeffrey P. Bigham
University of Rochester Computer Science
Rochester, NY 14627 USA
samuel.white@rochester.edu, haji@ece.rochester.edu, jbigham@cs.rochester.edu

ABSTRACT
This demonstration presents EasySnap, an application that enables blind and low-vision users to take high-quality photos by providing real-time audio feedback as they point their existing camera phones. Users can readily follow the audio instructions to adjust their framing, zoom level and subject lighting appropriately. Real-time feedback is achieved on current hardware using computer vision in conjunction with use patterns drawn from current blind photographers.

Author Keywords: Photography, Non-Visual Interfaces, Blind Users

ACM Classification Keywords: H5.2 [Information interfaces and presentation]: User Interfaces. – Graphical user interfaces.

General terms: Design, Human Factors

INTRODUCTION
Blind people want to take photographs for the same reasons as others – to record important events, to share experiences, and as an outlet for artistic expression [1]. Both automatic computer vision technology and human-powered services can be used to give blind people feedback on their environment [3,4], but to work their best these systems need high-quality photos as input [4]. This demonstration introduces *EasySnap*, an iPhone application that enables blind people to take high-quality photographs in contexts important to them.

In a recent study, we showed that even framing or aligning a photo could be challenging for blind users unable to see the view screen [4]. Furthermore, without feedback, blind users have little opportunity to improve either their technique or the photos they take. EasySnap uses computer vision to provide appropriate audio feedback to blind photographers.

Mobile devices generally lack the resources to undertake complex computer vision quickly enough to provide real-time feedback. For example, the kNFB Reader device for mobile phones provides audio feedback to help blind people frame document for conversion to speech, but requires several seconds between each suggestion [3]. While waiting between suggestions, users must hold the device still and are not provided with immediate subject framing feedback. We have achieved real-time feedback on an existing phone (iPhone 3GS and 4) by adopting two principles. First, we use fast computer vision heuristics that are independently prone to errors but most often correct when averaged over several frames. Second, inspired by our conversations with current blind photographers (who use existing cameras without any feedback), we rely on users to set up initial conditions that make the computer vision problems easier.

EASYSNAP OVERVIEW
EasySnap assesses picture quality based on four criteria: subject or document framing, alignment (both text and object), exposure, and sharpness.

According to our user study results [4], poor subject or document framing is a common cause of low-quality photographs. In EasySnap, we recommend zooming out when the subject appears too close (by moving the camera away from the direction of the subject of the photo), and recommend moving the device in one of four directions when the subject overlaps the edge of the camera's view. Diagonal movements are indicated by multiple cardinal directions.

To determine the subject of a photo, EasySnap first asks users to choose one of the three modes: landscape, portrait/object, or document. *Landscape mode* is an unconstrained form giving users the freedom to take a picture of anything they want, and checks only exposure and sharpness. *Portrait/object mode* is designed for taking pictures of any object or human. In this mode, users first take a close-up picture of the subject matter at arm's length (which they can do by first touching the object and then moving the camera back). They then step back to frame and snap. EasySnap helps users frame the desired object by matching the original close-up image with each frame and providing feedback indicating whether the subject is entirely included and centered in the frame and how to center the subject if it is not. Finally, *document mode* is for taking pictures of books, newspaper, and banknotes, which assumes that user is 1 to 3 feet from the document. This mode includes feedback for alignment and rotation.

UIST'10, October 3–6, 2010, New York, New York, USA.
ACM 978-1-4503-0271-5/10/10.

After choosing the desired mode, the user begins the *framing process*, in which they are guided through centering and aligning the subject appropriate to their selected mode. Instructions include "Searching", "Zoom In", "Zoom Out", "Go Left", "Go Right", "Go Up", "Go Down", "Turn Left", "Turn Right", "Ready." The "Go *" instructions indicate which direction users should move the camera to center the subject of the photograph, while the "Turn *" instructions are given for alignment of documents and are not given until the document is centered.

IMPLEMENTATION

We chose to develop EasySnap on the Apple iPhone because of its popularity among the blind community due to its free screen reading technology[1] and because of the processing capabilities of the devices' ARM processor.

In an effort to limit the number of use cases that we needed to support, and the computational complexity of our approach as a whole, we grounded three separate photography use cases in the experiences of blind photographers: landscape, portrait/object, and document. In the document case, the image is converted to a binary image using a Gaussian adaptive threshold and then rescaled to 320×480 for faster processing. Edges are detected using canny edge detection, and a Hough transform is applied to extract lines of text. A bounding box depicting the contour of the document is then generated from the computed lines to decide whether the document is in the center of the frame. This series of operations is repeated roughly 4-5 times a second in order for the device to discard statistical anomalies before generating appropriate feedback.

Angle of rotation is computed for each line of text and subsequently classified into one of three categories: left, right and aligned. The margin of error in terms of rotation is set as ± 5°. The angle of the entire document is then determined based on the category containing the most classifications. The user is given audio instruction to achieve alignment until they get to the "ready" condition or the instruction status need to be changed.

The portrait mode detects the subject area by matching it against a close-up image taken by the user at roughly arms length. We determined this distance to be ideal for both capturing an initial SURF feature set [2] and allowing blind users to physically orient themselves with respect to their subject. A bounding box is generated when the number of matched points passes a threshold determined experimentally, and is used to determine where the subject is in relation to the camera's view. The user is notified when they have properly framed the original subject matter while standing at their newly chosen distance.

Alignment and inclusion are difficult to judge for landscape photos because we do not know what the subject matter is and there may be multiple subjects. As a result, landscape mode simply uses device's gyroscope and accelerometer to help the user aim the device lower or higher than a set of boundary angles determined experimentally.

The main challenge of providing real-time feedback to users lies in the limited computational power of mobile devices. Our method of categorizing use cases allows us to focus our limited computational resources on a finite number of possible scenarios. In doing this we are able to compute a much larger dataset to use for determining appropriate audio feedback. Our initial experimentation has shown that this approach is more accurate than relying on a smaller yet more processed data set.

The final images captured in the above scenarios are each given a final check for proper exposure and sharpness. Exposure detection takes precedence and works by creating a grayscale histogram to check for any large concentrations of dark pixels indicating insufficient lighting. Sharpness is estimated by computing the mean and standard deviation of an image from its binary map and evaluating these values using a set of pre-built covariance matrices created from images known to be blurry or sharp [5]. At the user interface level, users are warned by an audio alert after successfully capturing a photo if it may be too blurry or too dark, and are given the option to either retake the picture or continue. It is important to note that neither darkness nor blur detection can work all the time, such as when the user takes a picture of a black sheet or an abstract painting with soft lines.

CONCLUSION

We have presented EasySnap, a picture-taking application with almost real-time audio feedback to help blind and visually impaired people take a picture. EasySnap not only guides the users through framing and aligning the main subject with real-time feedback but also checks the exposure and sharpness of the image ensuring the quality of final image output. As blind people are able to take better photographs, it will become easier to provide both automatic and human-powered services to help them better interpret and interact with the visual environment.

REFERENCES

1. Sensory Photography: Photography for blind and visually impaired people. Available at: http://www.photovoice.org/html/methodology4sp/why.htm.
2. Bay *et al*. "SURF: Speeded Up Robust Features", *CVIU 2008*, 110, 3, 346—359, 2008.
3. knfb Reader. knfb Reading Technology Inc., 2008. http://www.knfbreader.com/.
4. Bigham *et al*. Vizwiz: Nearly Real-time Answers to Visual Questions. *UIST 2010*. To appear
5. Ko, J. and Kim, C. Low Cost Blur Image Detection and Estimation for Mobile Devices. *ICACT 2009*, 1605–1610, 2009.
6. Google Goggles, 2010. http://www.google.com/mobile/goggles.

[1] http://www.apple.com/accessiblity/voiceover