Facilitating Photographic Documentation of Accessibility in Street Scenes

Marynel Vázquez
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
marynel@cmu.edu

Aaron Steinfeld
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
steinfeld@cmu.edu

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract
We present two interactive approaches for assisting users with visual impairments during photographic documentation of transit accessibility. We are working on an application for camera-enabled mobile devices that drives image composition towards highlighting visual information that is expected to be most relevant. In one interaction modality the user is guided trough small device motions that are expected to center the estimated region of interest in street photographs. In the other modality, the user captures the scene while pictures are processed, and the system alerts when enough data has been collected. The image that best aligns with our attention-getting composition model is then selected for documentation purposes. The specific design of these interactions is evolving to promote small motion behaviors by the user. Future work includes user studies.

Keywords
Transit, Photography, Visually Impaired, Accessibility.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User interfaces – Input devices and strategies, Interaction styles.

General Terms
Design, Experimentation, Human Factors

Introduction

Based on earlier results [11], we believe rich multimedia documentation of problems and solutions serves as persuasive evidence for promoting changes in transit accessibility. There is evidence that digital photographs are an attractive reporting method for transit riders [10]. From a citizen science and universal design perspective, we would like to support participation by transit riders who would normally have difficulty taking a good picture due to visual impairments. Therefore, we seek a method for assisting visually impaired users during photographic documentation of transit accessibility. Our sys- tem is intended to improve the composition of photographs by such users, for whom we expect a mobile phone implementation to be particularly useful. However, the limited interaction options for mobile phones and the lack of sup- porting infrastructure produce interaction challenges when providing spatial assistance. This paper describes our initial efforts towards appropriate solutions.

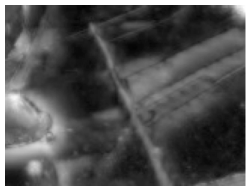
The problem of taking a "good" picture is dramatically simplified by the task characteristics. First, aesthetics are not an issue for problem documentation, thereby mitigating a significant challenge. Second, we do not need to know what the barrier is - we only need to know where it is. While being able to automatically annotate barriers might be useful for documentation purposes, it is not essential. This mitigates the need for object recognition. Third, we can assume the rider is able to localize the barrier in space and roughly aim a camera at the target. This means only small tilt, pan or roll camera motions are needed to improve photos for documentation purposes.

Consider Figure 1 as an example. From the initial view of the scene, we deduce with high probability that the accessibility barrier being documented is related to the stop sign. A better picture for documentation purposes would then have the sign centered in the image, since centering highlights evidence and increases the chance that the surrounding content will include relevant context. The second image in the Figure shows this view, as automatically proposed by our system in a simulation.

Figure 1. Initial view of a street scene (left) and automatically proposed view from a simulation test (right).

We propose two methods for facilitating the photographic documentation of transit accessibility. One approach is to guide the user so the key region of the picture is centered. The other is to let the user capture the scene using multiple images, and select the best for documentation purposes using a post-hoc approach. The former requires real-time computation, probably local to phone hardware. The latter can defer some computation, but an ideal system will alert the user when enough data has been collected, thereby also imposing real-time requirements. Efficient computer vision techniques, which can run on mobile phones, are necessary.

(a) Input image



(b) Saliency map

(c) Potential ROI

(d) Selected ROI (white)

(e) Proposed image center (rectangle)

Figure 2. ROI selection process.

The specific design of these interactions is evolving, but certain details are already known. First, humans cannot make precise and repeatable motions and, second, humans will not tolerate corrective aiming that takes a long time.

Image Composition Assessment

Our computer vision technique can be described as a method to avoid leaving out information that is expected to be most relevant. Our strategy was designed for the transit domain without explicit knowledge of object models. This leverages the fact that this domain is strongly composed by conspicuous elements.

Region of Interest

Our particular implementation constructs a model of visual attention in an image employing a modified version of saliency maps as defined by Itti and Koch [3]. These maps highlight visual stimuli that are intrinsically salient in their context, which tends to be the case for transit elements in street pictures.

Possible regions of interest (ROI) are generated by thresholding the saliency map of an image, and are later ranked based on their size and saliency. The region with highest score is selected as the ROI and its weighted center, using the saliency map for the weights, is the automatically proposed center of the image (Figure 2). Rationale, algorithm details, and evaluation of this approach can be found in [13].

Image quality

To measure the quality of a composition we propose to consider the distance from the selected center to the middle of the image, as well as the general orientation of the picture. Note the automatically selected center is biased towards the most salient point in the ROI, though it is possible that this is not the most salient point in the image. If we chose the most salient point in the image directly, then our proposed center is driven towards small salient regions that are not necessarily the center of attention. In the scene of Figure 2(a), the point of maximum saliency is very near the top-right corner of the image, where a small portion of grass is depicted near the sidewalk.

We emphasize centering the selected center in the ROI before correcting orientation. Framing the ROI in pictures is a priority during problem documentation since it is important to retain relevant content.

Interaction Models

The problem of recentering a picture is not new from a computer vision perspective. There have been efforts for automating image cropping [6, 12], where the generated thumbnail is expected to be the most relevant portion of an image. There is work that focuses on image adaptation for small displays [2] or image and video retargeting [8, 9], where the same principle is applied. These methods, however, are designed for image post–processing and neither of them consider real–time user interaction. These approaches also tend to rely in composition heuristics that may not apply to photographers with visual impairments. For example, on–center compositions, where a dominant subject is geometrically centered in the image, are taken for granted in consumer photography and unlikely for users who are blind.

Our main contribution is integrating user interaction during the image capturing process, such that users

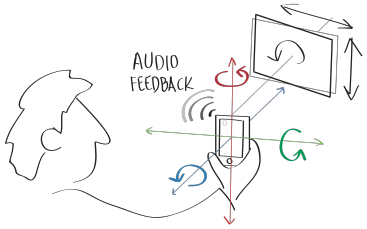


Figure 3. User interaction with our application.

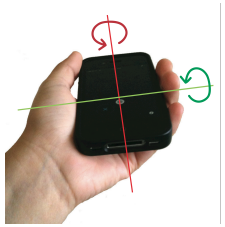


Figure 4. Prototype application for testing guidance through tones. As the user approaches a desired inclination, the pitch and tempo of feedback tones increase.

can take better pictures in real time. Given that our work is oriented towards assisting visually impaired users, audio feedback is a natural choice. We provide redundant onscreen information since assistance is also valuable to other users.

There has been work that focuses on leveraging proprioception to make mobile devices more accessible [4] and gesture based interactions for visually impaired people [14]. The former considers large-scale motion of a mobile device, which is held by the user with an arm extended. The rotation of the arm with respect to the body and its inclination are measured from the device, and this information is used to access virtual application shortcuts. The latter is also oriented towards facilitating the use of traditional mobile applications.

Our application is different from these approaches since we consider small camera (and mobile device) motions for recentering a photograph. The impact of small angle changes can be pronounced in our application, so fine adjustments are critical. Figure 3 shows a cartoon of the interaction.

Note that the system requires consistent tracking of the ROI in the image as the user moves the device. Our current implementation generally runs at 9 fps using a vision algorithm that estimates the perceived motion of visual features for this task [1]. The problem with this approach is that if the mobile device moves rapidly, then tracking tends to fail.

We are now working on sensor fusion so that our system is more robust to different tracking scenarios. We expect that data from accelerometers and gyroscopes, common in third generation mobile devices, will allow better real-time performance.

Guided Mode

One method to facilitate photographic documentation of transit accessibility is guiding the user in real-time. This method is based on estimating a relative motion of the camera that would result in a "better" view of the scene. Quality assessment is performed according to how well the initial view of the scene fits our attention–getting composition model. If the weighted center of the estimated ROI is far from the middle of the image, then there is evidence that suggests the picture can be improved. In this case a change in camera view is proposed.

Image improvement is achieved through small camera motions intended to center the estimated ROI. We propose to use non-verbal audio as feedback for guiding this motion. In particular, we favor tones over speech because they require less cognitive load [5].

We designed an application for the iPhone to test a simplified version of this interaction. The application was designed to guide the user towards a specific inclination of the device (Fig. 4). Note this is a simplification of the type of motion we consider for assisting photographers, since the application ignores displacements and reacts to only two out of three rotation types.

The application requires the user to hold the device in a position close to horizontal, such that tilt and roll motions generate significant readings for the accelerometer. Two marks are shown in the screen while the user plays with the application. One mark

renders the current inclination of the device, while another demarks a target inclination. As the user tilts the device, the mark that renders the current inclination moves along the screen. When this mark is close to the target, given a margin for error, the application indicates to the user that the desired inclination has been reached.

We asked for feedback from potential blind users and experts on auditory displays. The most concrete suggestion, which we implemented, focused on simple tone and tempo changes, and was based on development of applications for blind users by others [7]. The pitch and tempo of the tones generated by our application change based on the distance on the screen from the current position to the target. In both cases, higher values correspond to closer proximity to the desired orientation.

People on the project team informally tried the application with their eyes closed, which led us to identify specific weaknesses of this approach and reveal how our application differs from others. First, the interaction is influenced by the amount of smoothing applied to accelerometer readings and the type of tone used. Smoothing was needed to overcome sensor and pose noise, though too much smoothing delays feedback from the application. Desired configurations for these parameters tend to change according to personal preferences. Second, this sort of interaction seems to encourage large motions by the user, which is clearly a concern for our assisted photography approach. We believe one potential reason is the tendency to move the device in ways not measured by the application. For example, when the device is held horizontally (as in Figure 4) and it is rotated along an axis perpendicular to its screen, then the system does not alter the feedback tone. This scenario generates confusion in the user, leading to larger motions in order to trigger an audible response. The situation is even aggravated if the application is set to generate long tones, because their duration impose an extra constraint in the reaction time of the application. We suspect that incorporating gyroscope readings will improve the interaction significantly, since gyroscope data allow better device motion estimates. Short tones seem to be better suited for the application with respect to long ones.

We believe the fundamental idea of guidance is sound. However, we need to refine the design to promote small motion behaviors by the user.

Paparazzi Mode

An alternative approach is to let the user capture a scene continuously and select the best still image using image quality metrics calculated by our computer vision technique. In other words, the system selects the image that brings the selected ROI of the initial view closest to the image center. If two images are equivalent in terms of recentering, then we favor the one with the more horizontal or vertical camera orientation. We assume that horizontal or vertical orientations are optimal in the Manhattan world of street photographs.

Two different pictures can be equivalently good in the system by having the same orientation and distance from their ROI to the center of the image. The problem of defining a generic and robust metric for image composition comparison is still unsolved and, clearly, a somewhat subjective matter.

An almost optimal picture is captured if the ROI is nearly centered and camera orientation is close to vertical or horizontal. In this case, the system informs the user that a good picture has been obtained, under the assumption that he or she roughly aimed at the target from the beginning of the capturing process.

Future Work

A complete evaluation of our framework for assisting visually impaired photographers requires user studies. From our experience, we believe it is important to provide consistent feedback to all types of camera motions generated by the user. We are currently working on incorporating both accelerometer and gyroscope readings into our application to improve the tracking of the regions of interest and its real-time performance. This is important for a consistent interaction, which we will evaluate with real users in the near future. Long-term goals include accounting for the added complexity of dynamic scenes and adding more semantic information.

Acknowledgments

The Rehabilitation Engineering Research Center on Accessible Public Transportation (RERC-APT) is funded by grant number H133E080019 from the United States Department of Education through the National Institute on Disability and Rehabilitation Research.

References

[1] Bouget, J.-Y. Pyramidal Implementation of the Lucas Kanade Feature Tracker: Description of the algorithm. Tech. rep., MRL, Intel Corporation (1999).

[2] Chen, L., Xie, X., Fan, X., Ma, W.-Y., Zhang, H.-J., and Zhou, H. A visual attention model for adapting images on small displays. Tech. Rep., MSR-TR-2002-125, Microsoft Research (2002).

[3] Itti, L., and Koch, C. Computational Modelling of Visual Attention. *Nature Reviews Neuroscience 2*, 3 (2001), 194–203.

[4] Li, F. C. Y., Dearman, D., and Truong, K. N. Leveraging Proprioception to Make Mobile Phones More Accessible to Users with Visual Impairments. In *Proc. ASSETS'10.*

[5] Loomis, J. M., Golledge, R. G., and Klatzky, R. L. Navigation System for the Blind: Auditory Display Modes and Guidance. *Presence: Teleoper. Virtual Environ. 7* (April 1998), 193–203.

[6] Marchesotti, L., Cifarelli, C., and Csurka, G. A framework for visual saliency detection with applications to image thumbnailing. In *Proc. ICCV'09.*

[7] Miele, J. A. Personal communication, 2010.

[8] Rubinstein, M., Shamir, A., and Avidan, S. Improved Seam Carving for Video Retargeting. *ACM Trans. Graph. 27*, 3 (2008), 1–9.

[9] Setlur, V., Takagi, S., Raskar, R., Gleicher, M., and Gooch, B. Automatic Image Retargeting. In *Proc. MUM'05*.

[10] Steinfeld, A., Aziz, R., Von Dehsen, L., Park, S. Y., Maisel, J., and Steinfeld, E. Modality Preference for Rider Reports on Transit Accessibility Problems. TRB 2010 Annual Meeting. Transportation Research Board, (2010).

[11] Steinfeld, A., Maisel, J., and Steinfeld, E. The Value of Citizen Science to Promote Transit Accessibility. In *First Intl. Symposium on Quality of Life Tech., (*2009).

[12] Suh, B., Ling, H., Bederson, B. B., and Jacobs, D. W. Automatic Thumbnail Cropping and its Effectiveness. In *Proc. UIST'03*.

[13] Vázquez, M., and Steinfeld, A. An Assisted Photography Method for Street Scenes. In *Proc. WACV'11*.

[14] Vidal, S., and Lefebvre, G. Gesture Based Interaction for Visually-Impaired People. In *Proc. NordiCHI'10.*