Composition Context Photography

Daniel Vaquero*
Nokia Technologies
daniel.vaquero@nokia.com

Matthew Turk
Univ. of California, Santa Barbara
mturk@cs.ucsb.edu

Abstract

Cameras are becoming increasingly aware of the picture-taking context, collecting extra information around the act of photographing. This contextual information enables the computational generation of a wide range of enhanced photographic outputs, effectively expanding the imaging experience provided by consumer cameras. Computer vision and computational photography techniques can be applied to provide image composites, such as panoramas, high dynamic range images, and stroboscopic images, as well as automatically selecting individual alternative frames. Our technology can be integrated into point-and-shoot cameras, and it effectively expands the photographic possibilities for casual and amateur users, who often rely on automatic camera modes.

1. Introduction

Taking compelling pictures is a complex task. Effectively communicating the intended message through imagery requires selecting adequate camera parameters, as well as properly framing the shot using an interesting composition. This is typically accomplished by looking at the camera's viewfinder while pointing at the scene and adjusting parameters such as point of view, zoom, aperture, and focus.

Digital cameras have features for automatically estimating optimal parameters, such as autofocus and auto-exposure. These functionalities reduce the photographic process to pointing the camera at the scene of interest, adjusting the point of view to capture the desired composition, and shooting a photograph. While professional photographers still tend to favor a manual or semi-automatic selection of parameters, point-and-shoot cameras are very popular and have made photography accessible to the general public.

The adjustment of capture parameters and framing, together with the dynamism brought in by the photo subject, which can be constantly changing, show that the moment being experienced is broader than the single instant of time captured by a photograph. If additional information could be gathered by a camera before or after the photographer captures an image, this information could be explored to enhance the imaging experience, and the photograph then ceases to be a mere representation of a single time instant [7]. New image capture options could be provided, such as going back and forth in time to choose the desired moment (Figure 1), or compositing multiple images to create panoramas or photomontages.

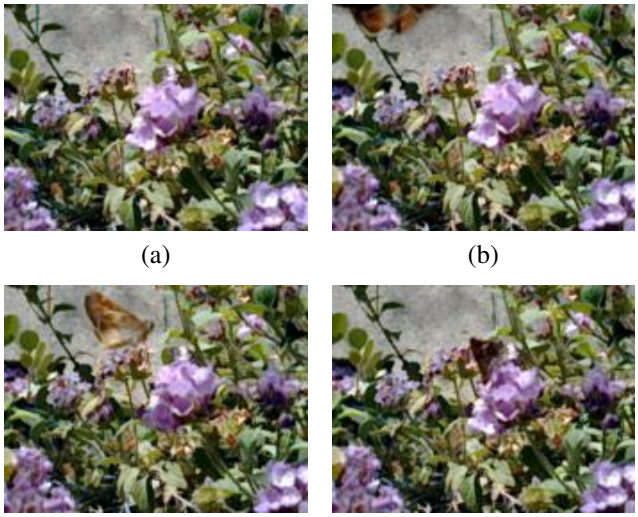


Figure 1. While framing a picture of a flower, a butterfly briefly lands on the flower, but immediately flies away. The photographer quickly triggers the shutter to capture the butterfly, but the final picture misses it (a). By considering frames captured before the shutter was triggered (b-d), a better view of the intended scene is found (c).

While current camera architectures still have limitations, such as memory and bandwidth constraints, that prevent developers from fully capturing and exploring visual information before and after a picture is taken, some information can now be gathered and used in interesting ways. In this paper, we summarize our experience developing a real

*Most of this work was conducted at the University of California, Santa Barbara, while Daniel was a Ph.D. student.

camera prototype and exploring the ideas of silently capturing visual information while the photographer is framing a picture, and using this information to extend the range of photographic possibilities. Our promising results motivate a research agenda that we call *composition context photography*.

2. Related Work

There has been increasing interest in using contextual information in photography, by collecting extra information before and after a picture is taken. Different types of information can be useful, such as audio; additional sensors such as accelerometers, gyroscopes, compasses, light sensors, and thermometers; location data, typically obtained from a Global Positioning System (GPS) sensor; image databases from the Internet; and viewfinder images. While not every camera is able to acquire such information, modern mobile devices often provide means to access several of these data sources.

A major motivation for using contextual information is to increase the chances of capturing the ideal moment in scenes with motion. Professional photographers often use the built-in burst functionality in current cameras and later select the best picture from the set [19]. When the goal is to minimize blur due to handshake, the "lucky imaging" application [1] uses inertial sensor data to pick the least blurry shot from a sequence of images. In portrait photography, machine learning has been used to select candid portraits in a sequence of frames [11].

A few commercial photography applications that include contextual information have been recently released. A general pattern is to capture a short sequence or burst of images and use its information to achieve interesting effects. Examples include taking pictures on smile, editing faces in a group picture, choosing the best moment from an image burst, creating stroboscopic images, adding artificial blur to motion areas, removing undesired moving subjects, creating cinemagraphs by "freezing" parts of the image while keeping other regions moving, capturing self-portraits once the user's face is framed, and automatically metering for high dynamic range. We provide a unified framework for incorporating contextual information in photography, and these applications deal with use cases that are encompassed by our framework.

There have also been other works based on sensors. Holleis et al. [14] presented a preliminary study suggesting that inertial sensor data obtained during framing could be applied to enhance the photographic experience, since it carries information on how the photographer handles the camera. Håkansson et al. [13] presented a camera that applies graphics effects, such as pixelation, shadows, and waves, to its photographs based on contextual audio and motion. More recently, Bourke et al. [4] and Yin et al. [25] used GPS, compass, auto-exposure parameters, and Internet image collections to recommend alternative photographs or propose scene compositions.

Finally, viewfinder images are a very important element of context, since they capture whatever the camera "sees" while the photographer is framing a picture. The automatic mode of several cameras relies on so-called 3A (auto-exposure, autofocus, and auto-whitebalance) algorithms that analyze viewfinder images to automatically find an optimal set of capture parameters that provide a properly exposed image. However, those frames are usually discarded after this process. In contrast, we utilize those frames to generate variations of the captured image that may be more compelling than the captured image itself.

3. Composition Context Photography

We explored the idea of using contextual information from a standard point-and-shoot camera, collected while the user is framing a picture, to provide additional image results. The contextual information includes the viewfinder frames; metadata consisting of capture parameters (exposure time, sensor gain, white balance, focus, zoom level) for every viewfinder frame; and camera motion data obtained from accelerometers and gyroscopes. We recorded this data for a period of time before the user triggers the shutter to capture an image[1]. We refer to this type of contextual data as the "composition context" of the picture, since the collected data typically corresponds to the period while the photographer is adjusting the photo composition and framing.

To generate the photo suggestions, we either pick individual frames from the composition context, targeting the problem of missing a desired moment (Figure 1), or we create image composites by exploring variations in capture parameters (Figure 2), such as field of view, exposure time, and focus, that naturally happen in the framing procedure either because of the user's actions or induced by the 3A algorithms. Some composites also explore the presence of moving objects. To generate the results, we employ computer vision and computational photography algorithms.

3.1. Exploratory Study

In order to obtain initial hands-on experience and better understand the composition context in photography, we developed a prototype camera and performed an exploratory study with users to gather composition context data in real-world scenarios. The goal was to acquire knowledge about the actions performed by users of point-and-shoot cameras in "automatic" mode in preparation for taking a photograph.

To enable silent background recording of viewfinder frames, their capture parameters, and inertial sensor data,

[1] It is also useful to gather data after the shutter is triggered; we focused on pre-shutter data due to constraints of our implementation platform.

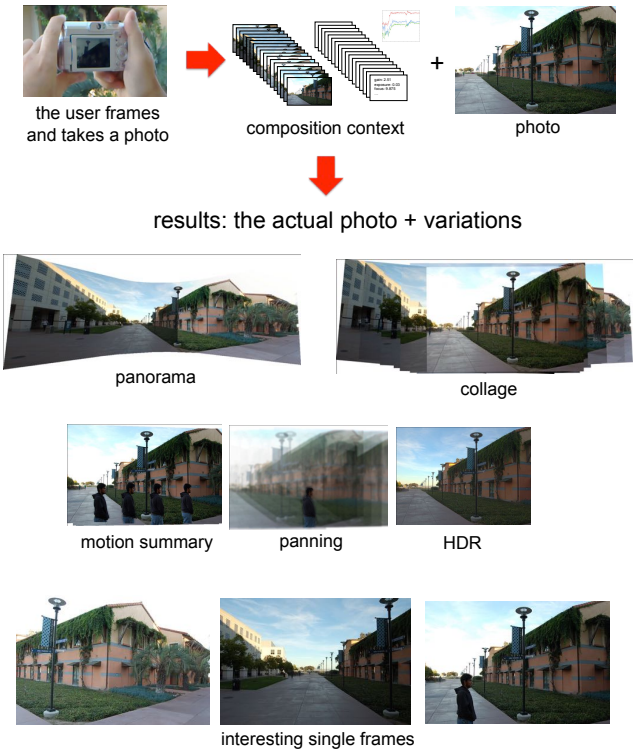


Figure 2. Using the composition context to extend the range of photographic possibilities. The user frames and takes a picture as with a point-and-shoot camera, but a collection of photo variations, computed using information from the composition context, may be obtained in addition to the actual photograph.

we created a capture application based on the Nokia N900 Frankencamera [1], modified to support digital zoom. We attached an experimental sensor box with accelerometers and gyroscopes (Figure 3(a)). At the time of data collection for our study, the N900 Frankencamera was the only device that allowed fine-grained control over capture parameters on a per-frame basis. Later on, the Nokia N9 and the NVIDIA Tegra Development Tablet also included the FCam API. As of now (November 2014), the Android 5.0 ("Lollipop") API has just been released, and it introduces a new camera API to facilitate fine-grain photo capture and image processing [12].

Our camera application (Figure 3) continuously captures and displays images (320x240, 25 fps) on the viewfinder. Every frame is added to a circular buffer before being displayed. The buffer holds the viewfinder frames and their capture parameters for the most recent 18 seconds, limited by the memory capacity of the N900. Autofocus is triggered by pressing the shutter button halfway. Once the shutter button is fully pressed, a 5 MP photograph[2] is captured with automatic settings (exposure time, gain, focus, and flash), and the buffer of viewfinder frames is saved.

We recruited 45 student volunteers from various disciplines, 24 male and 21 female, with ages ranging from 18 to 30 years old, to participate in our study. On average, the participants have used digital cameras for 5.75 years. We conducted study sessions at different times of the day and weather conditions, with a single participant at a time. Each session lasted about 45 minutes, and each user received $5 for their participation. The participants were provided with our prototype camera and requested to take pictures at a university campus. They were instructed to try their best to capture compelling images, relying on their own sense of what makes a good picture. We did not tell them that the camera was recording composition context information.

We requested three different photographs for each of seven categories, selected to represent common photography scenarios: an office environment, a close-up scene, a building, a sign, an open area, a posing person or group of people, and a moving subject. The participants were free to choose the scenes to be photographed, as long as they fit within these categories. Once the 21 pictures had been taken, the users were then asked to take at least five additional pictures of scenes chosen at their discretion.

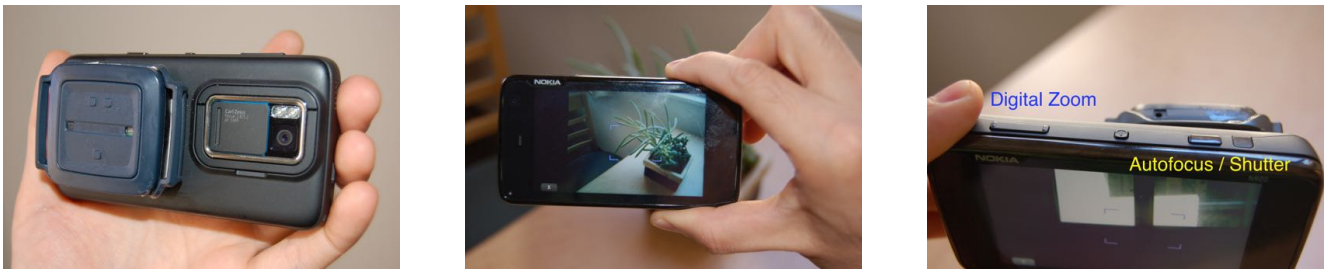
Users rated their level of experience as a photographer; the average self-assigned experience level was 4 on a 1-7 scale. Users were also asked to rate the similarity of the interface of our camera to digital point-and-shoot, cameraphone, and digital SLR cameras. On a 1-7 scale, the average ratings were 5.37, 5.57, and 2.94, respectively; these are good indication that our capture prototype successfully mimics the user interface of point-and-shoot cameras and cameraphones.

3.2. Dataset Statistics and Analysis

The data collection sessions resulted in a database of 1213 pictures with associated composition context data. For each set of framing images, we manually annotated the time intervals where the user was framing the photograph. This is hard to define since we do not have access to information about the user's intentions, but for this purpose we considered the frames for which the user seemed to be deliberately pointing the camera at a scene. We also annotated the time intervals whose views overlap with the one in the final picture.

We then computed statistics on the duration of the framing procedure. Figure 4(a) shows a distribution of framing times for the pictures in the dataset, beginning at the first frame for which the user seemed to be framing, and ending when the final picture was taken. Due to the memory limitations of the implementation platform, durations longer than 18 seconds were recorded as 18 seconds. The median framing duration was of 8.92 seconds, and 16 seconds were enough to record the entire framing procedure in 80% of the

[2]From now on, we refer to this photograph as the *final picture* or *final image*, as it is the final output of the point-and-shoot photography process.



(a) (b) (c)

Figure 3. Our composition context capture prototype. (a) Nokia N900 with attached sensor box; (b) Point-and-shoot camera application, including a viewfinder; (c) Camera buttons used for digital zoom and autofocus / shutter trigger.

cases. We also analyzed the amount of framing time during which the viewfinder image overlapped with the final image (Figure 4(b)). The median framing duration while the camera displayed viewfinder frames that overlapped with the final picture was of 8.04 seconds.

For the viewfinder frames that overlap with the final picture, we computed the range of variation of the following parameters: focus, exposure (as the multiplication of the exposure time by the sensor gain), and zoom, in diopters, stops and magnification change, respectively. The changes in focus come from the autofocus procedure, which can vary focus from 5 cm to infinity (20 to 0 diopters); variations in exposure come from the auto-exposure algorithm; and zoom variation comes from the digital zoom controls, which either increase or decrease the magnification by 0.067x. Figure 4(c-e) shows the distribution of the variations in our dataset.

Figure 4 indicates that a large number of frames is highly correlated to the final picture, due to the overlaps between views. The variations in capture parameters can be understood as imaging the scene in the final picture under different parameters. Also, the framing duration may inform the choice of how much contextual information to capture given hardware constraints.

Qualitatively, we observed that the viewfinder streams contain intervals of framing, typically characterized by periods of stillness, and movement for composition adjustment; and intervals of erratic imaging, where the camera records random scenes that are not of interest, such as the user's feet, ground or sky at times when the user is simply walking between photo shots.

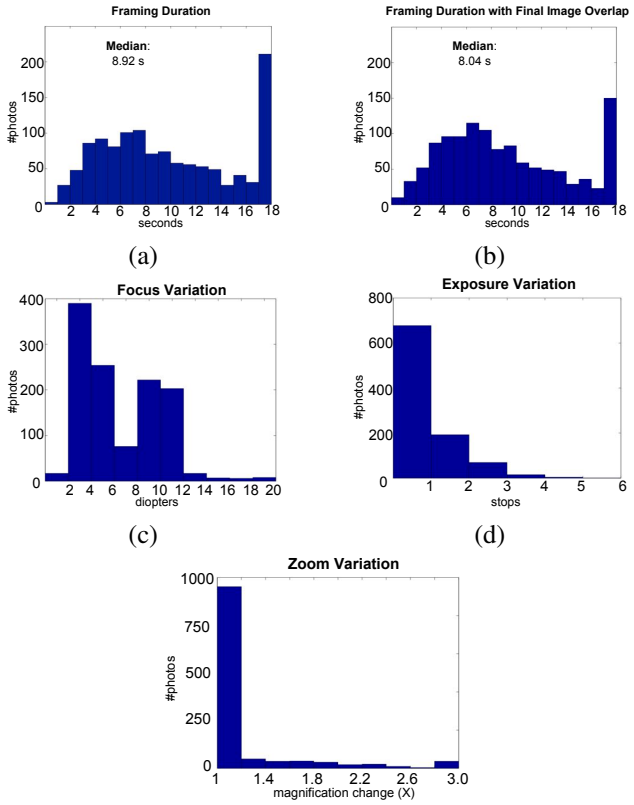


Figure 4. Statistics from the framing procedure. (a) Distribution of framing times. (b) Distribution of framing times considering only periods when the viewfinder overlaps with the final image. (c-e) Variations of capture parameters for viewfinder frames that overlap with the final image during the framing procedure: (c) variations in focus; (d) variations in exposure (exposure time multiplied by sensor gain); (e) variations in zoom.

3.3. Generation of Photo Suggestions

Given the collected dataset, we developed a system to create additional photo suggestions using composition context data. The goal was not to achieve pixel-perfect results, but to explore the possibilities that the composition context brings for extending the output of the photo taking process. Therefore, we integrated simple techniques or used off-the-shelf implementations of the computer vision and computational photography algorithms involved. For more details on specific algorithms, please refer to the first author's PhD dissertation [23].

The first step was to align the viewfinder frames to the final picture, so that pixel correspondences can be established. We estimated transformations that spatially align

each viewfinder frame to the final picture. To perform alignment between pairs of frames, we used the algorithm from Wagner et al. [24] that showed good results for images taken by mobile devices.

The following step was to find the image regions that correspond to moving objects or subjects in the scene. We implemented a simple method with workarounds to address practical issues from our scenario. The general idea was to align the viewfinder frames to the final image and perform median-based background subtraction while disregarding detections near intensity edges (to avoid false positives due to small misalignments), after discarding frames with large camera motion (using the inertial sensor outputs) and focus and exposure time that considerably differed from the final image's. This resulted in mask estimates for moving areas.

By using the aligned frames and the detected motion masks, we generated image composites created from multiple composition context frames. As previously discussed, parts of the composition context can be understood as imaging the scene with different capture parameters (Figure 4(c-e)), and can be used as input to computational photography algorithms, which explore variations in capture parameters to create images that could not have been captured with a single shot. We only used frames likely to exhibit insignificant blur due to camera motion, by filtering out frames by gyroscope energy.

We created panoramas [21], collages [18], all-in-focus images [2], high dynamic range images [8], synthetic long exposure images [22], and flash/no-flash composites [10, 20]. We also simulated the panning technique [19], by aligning moving objects across multiple frames and averaging the images, creating background blur while keeping the moving object sharp. We depicted motion in a static image by repeating moving subjects at multiple places in the scene (stroboscopic images). To remove undesired moving subjects, we cut them out and used viewfinder frame information to fill the holes left, a process known as inpainting [3].

To generate the composites, we used the OpenCV panorama library, which implements [6], and the *Enfuse 4.0* software [17], which implements the Exposure Fusion [16] algorithm for extended dynamic range, flash/no-flash and all-in-focus composites. For the remaining composite types, we implemented simple methods as a proof-of-concept.

Given the large number of frames available, it was important to select relevant images as input to the algorithms. Using the entire composition context may require a large amount of processing time and memory, and poor registration can accentuate artifacts. We automatically identified groups of frames to be provided as inputs to compositing algorithms. Inspired by [5], which searched image collections to identify groups of frames to be stitched as panoramas, we extended this idea to find inputs for other image composites within the composition context.

Table 1. Number of generated image composites using our dataset, and average number of identified input frames per composite type.

	Generated	Avg. #input frames
Panorama	127	137
Collage	127	137
HDR	13	68
Synthetic Long Exposure	1073	59
All-in-Focus	29	70
Flash/No-Flash	29	N/A
Stroboscopic	295	18
Synthetic Panning	263	13
Moving Object Removal	213	62

We leveraged the capture parameter metadata to recognize stacks of frames for high dynamic range, all-in-focus imaging, synthetic shutter speed, and moving object composites. For panoramas and collages, we use an algorithm similar to [5], and eliminate viewfinder frames whose capture parameters significantly differ from the final picture parameters or include moving objects. For the other composites, we first find a stack of images that approximately aligns with the final picture, and then analyze the capture parameters and the presence of moving objects in the stack to determine groups of frames to be provided as inputs to different image composites. Only composites for which enough variation of the required capture parameters (*e.g.*, exposure for HDR, field of view for panorama, etc.) is detected are created.

The method identified input frames for at least one type of image composite in 1105 of the 1213 videos. Table 1 shows the number of created composites by type, as well as the average number of identified input frames for each type. Figure 5 shows some of the generated composites along the corresponding final pictures. The composites provide wider field of view, extended dynamic range, motion-blur effects, better photos in low-light conditions, interesting motion effects, and moving object removal.

In addition to image composites, our system generated suggestions of individual frames, aiming to provide interesting alternative moments or views of the scene. The goal was to select frames from areas in which the user indicated interest while framing, which were not too similar to the final picture, and which maximized measures of quality.

Inspired by eye gaze tracking [9], we created a map that associated scene locations to attention scores, which indicated how long the user spent pointing the camera at a particular location. In this way, it was possible to suggest only viewfinder frames that corresponded to areas of high interest. To compute the map, we aligned all viewfinder frames to the final picture, and then added the regions occupied by each aligned frame onto an accumulator.

Only frames with attention scores greater than 50 were considered as candidates for suggestions, as this corresponds to areas at which the user spent at least two sec-

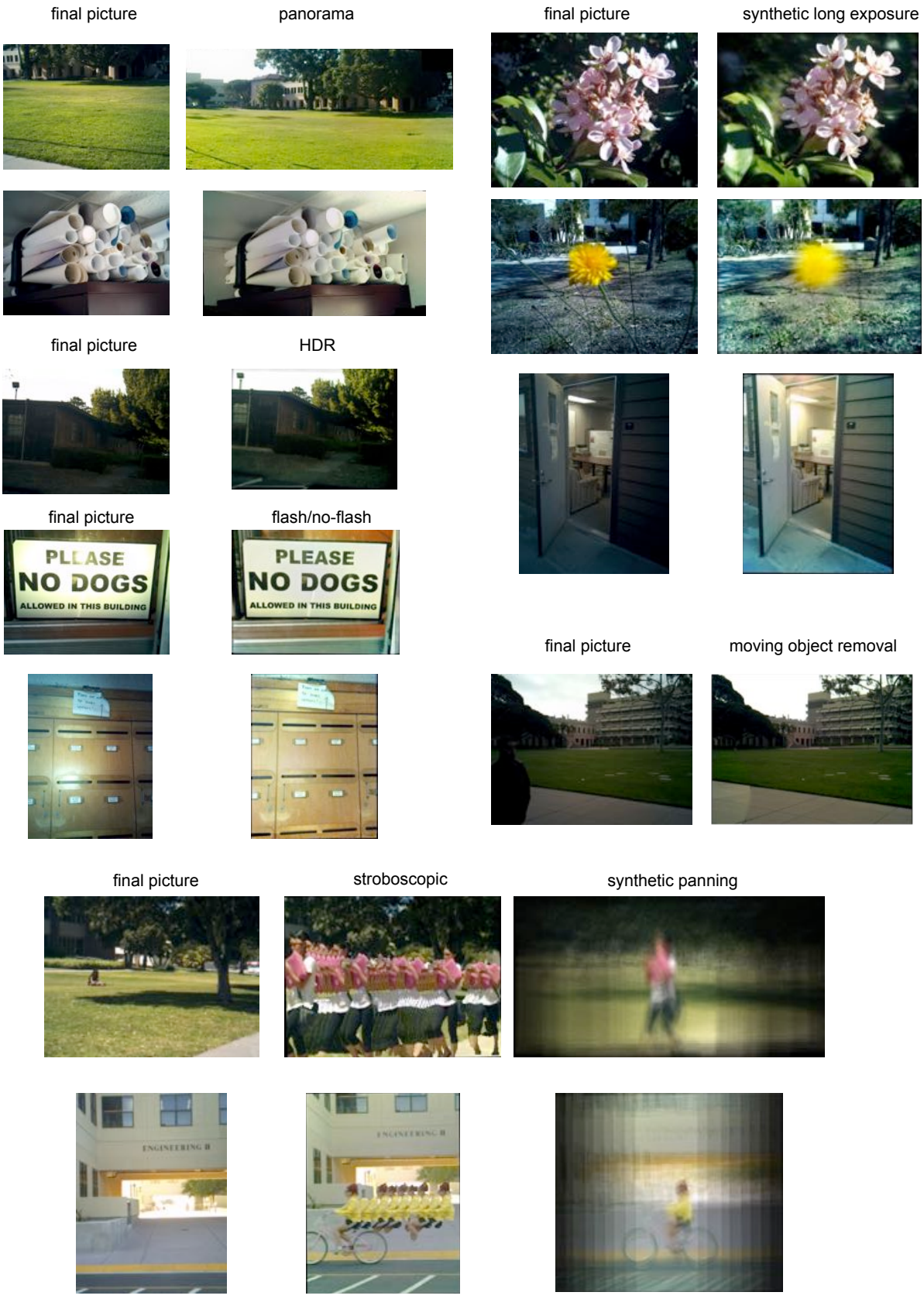


Figure 5. Examples of photo suggestions created from the composition context by combining multiple frames.

onds viewing. To avoid suggestions too similar to the final image, we filtered out frames with roughly the same viewpoint, focus, zoom and exposure as the final image, and which did not contain moving objects. We clustered the remaining frames into groups by proximity to the peaks of the attention map. For each group, we then suggested optimal frames according to different quality measures, such as amount of camera motion, moving object area, and computational aesthetics criteria [15].

Figure 6 shows a few examples of images from our dataset and useful suggestions recommended by our algorithm. The examples include alternative views (first row), alternative orientation (second row, left), different exposure and moving objects (second row, right), focus at different distances (flowers), and moving subjects at different instants (bicycles).

4. Research Agenda

Our exploration of the composition context and the recent commercial interest indicate that contextual photography is a promising area for continuing investigation. While our initial simple methods and the market applications are not free of visual artifacts, they provide enhanced experiences to photographers, by creating images that could not have been captured before. This is done in a computational fashion, with in-camera processing helping to gather and make sense of contextual data.

Our results motivate several topics for further research and development in the area.

- Additional research with users is required to quantitatively analyze the benefit provided by the generated suggestions. More detailed research on the human factors involved in the process of framing photographs would be useful, as well as studying how a composition context camera impacts user behavior.
- Our solution does not alter the point-and-shoot photography process; this has the advantage of expanding the output of the process without requiring additional effort from the user. However, it would be interesting to introduce active variation of capture parameters to increase variability in the composition context, possibly using additional cameras or custom hardware, while still trying to minimize changes to the capture process.
- Additional types of contextual information, such as GPS tied to geolocated image databases, could be explored. In fact, recent work [25] has already begun to address these issues.
- Different output modalities such as 3D models, higher resolution images synthesized by superresolution techniques, denoised images, and cinemagraphs, could be provided.
- Reliably performing image alignment and detecting moving objects in scenarios with uncontrolled camera motion and variations in capture parameters are still challenging problems being addressed in the computer vision community.
- Optimizations in context capture and generation of photo suggestions would allow the entire pipeline to run on-camera. Advances in the underlying camera architecture and middleware would enable higher-quality results, by capturing higher resolution images during the framing process.
- On-the-fly compression algorithms for the composition context would also allow more efficient recording.

Acknowledgments

The authors would like to thank Cha Lee and Tobias Höllerer for the help with the exploratory study design and Steffen Gauglitz for the image alignment software; Kari Pulli for advice and providing the hardware used in this work; the Computational Photography team at the Nokia Research Center for fruitful discussions; and the UCSB students who volunteered to participate in our data collection procedure.

References

[1] A. Adams et al. The Frankencamera: an experimental platform for computational photography. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 29(4):1–12, 2010.

[2] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 23(3):294–302, 2004.

[3] M. Bertalmío, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. *ACM Trans. Graph. (Proc. SIGGRAPH)*, pages 417–424, 2000.

[4] S. Bourke, K. McCarthy, and B. Smyth. The social camera: a case-study in contextual image recommendation. In *Proc. of the Intl. Conf. on Intell. User Interfaces*, pages 13–22, Palo Alto, California, 2011.

[5] M. Brown and D. G. Lowe. Recognising panoramas. In *Intl. Conf. on Comp. Vision*, pages 1218–1225, Nice, France, 2003.

[6] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *Int. J. Comput. Vision*, 74(1):59–73, 2007.

[7] M. F. Cohen and R. Szeliski. The moment camera. *Computer*, 39(8):40–45, 2006.

[8] P. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *ACM Trans. Graph. (Proc. SIGGRAPH)*, pages 369–378, 1997.

[9] A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer-Verlag New York, Inc., 2007.

Figure 6. A few examples of suggested frames from our dataset. Each pair displays the final picture and one of the suggestions. The suggestions provide useful alternatives, such as different views, focus at different depths, different exposures, and moving subjects at different moments in time.

[10] E. Eisemann and F. Durand. Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 23(3):673–678, 2004.

[11] J. Fiss, A. Agarwala, and B. Curless. Candid Portrait Selection From Video. *ACM Trans. Graph. (Proc. SIGGRAPH ASIA)*, 30(6):128:1–128:8, 2011.

[12] Google. Android 5.0 APIs. `https://developer.android.com/about/versions/android-5.0.html`.

[13] M. Håkansson, L. Gaye, S. Ljungblad, and L. E. Holmquist. More than meets the eye: an exploratory study of context photography. In *4th Nordic Conf. on Human-Computer Interaction*, pages 262–271, Oslo, Norway, 2006.

[14] P. Holleis, M. Kranz, M. Gall, and A. Schmidt. Adding context information to digital photos. In *25th IEEE Intl. Conf. on Distrib. Computing Sys. Workshops*, pages 536–542, Columbus, Ohio, 2005.

[15] D. Joshi et al. Aesthetics and emotions in images. *IEEE Signal Process. Mag.*, 28(5):94–115, 2011.

[16] T. Mertens, J. Kautz, and F. V. Reeth. Exposure fusion. In *Proc. Pacific Graphics*, pages 382–390, 2007.

[17] A. Mihal et al. Enblend / enfuse 4.0. `http://enblend.sourceforge.net`.

[18] Y. Nomura, L. Zhang, and S. Nayar. Scene Collages and Flexible Camera Arrays. In *Eurographics Symposium on Rendering*, pages 127–138, Grenoble, France, 2007.

[19] B. Peterson. *Understanding Shutter Speed: Creative Action and Low-Light Photography Beyond 1/125 Second*. Amphoto Books, 2008.

[20] G. Petschnigg, M. Agrawala, H. Hoppe, R. Szeliski, M. Cohen, and K. Toyama. Digital photography with flash and no-flash image pairs. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 23(3):664–672, 2004.

[21] R. Szeliski and H. Shum. Creating full view panoramic image mosaics and environment maps. In *ACM Trans. Graph. (Proc. SIGGRAPH)*, pages 251–258, 1997.

[22] J. Telleen et al. Synthetic shutter speed imaging. *Comput. Graph. Forum (Proc. Eurographics)*, 26(3):591–598, 2007.

[23] D. A. Vaquero. *Composition Context Photography*. PhD thesis, University of California, Santa Barbara, USA, 2012.

[24] D. Wagner, A. Mulloni, T. Langlotz, and D. Schmalstieg. Real-time panoramic mapping and tracking on mobile phones. In *IEEE Virtual Reality Conf.*, pages 211–218, Waltham, Massachusetts, 2010.

[25] W. Yin, T. Mei, C. W. Chen, and S. Li. Socialized mobile photography: Learning to photograph with social context via mobile devices. *IEEE Trans. on Multimedia*, 16(1):184–200, 2014.