Streaming Mobile Augmented Reality on Mobile Phones

David M. Chen [1], Sam S. Tsai [1], Ramakrishna Vedantham [2], Radek Grzeszczuk [2], Bernd Girod [1]
[1] Information Systems Laboratory, Stanford University
[2] Nokia Research Center, Palo Alto

Abstract

Continuous recognition and tracking of objects in live video captured on a mobile device enables real-time user interaction. We demonstrate a streaming mobile augmented reality system with 1 second latency. User interest is automatically inferred from camera movements, so the user never has to press a button. Our system is used to identify and track book and CD covers in real time on a phone's viewfinder. Efficient motion estimation is performed at 30 frames per second on a phone, while fast search through a database of 20,000 images is performed on a server.

1 Introduction

Mobile augmented reality (MAR) is a wide class of applications where mobile devices augment users' perception of the world. Many mobile phones that capture video or still images of a scene can automatically recognize and annotate objects in the scene. Existing MAR systems include landmark recognition [4][9], product logo recognition [5], and CD/DVD cover recognition [8][10].

Real-time augmentation on a phone remains difficult because a MAR system has delays in (1) extraction of query data on the phone, (2) transmission of the query data from the phone over a wireless network to a server hosting an image database, and (3) search through the database. Real-time recognition requires small delays in all three stages, while ensuring high recognition accuracy. The MAR systems presented in [4][9][8][10] all require delays of at least 3 seconds to recognize a newly appearing object. For continuous augmentation of live video, the recognition latency must be reduced to about 1 second. The system in [5] achieves around 1 second delay, but at the expense of continuously streaming a video from the mobile phone to the server.

In this paper, a novel MAR system is presented for continuous recognition of book and CD covers in live video captured by a mobile phone, an innovation we refer to as streaming MAR. The user can point the camera at a book or CD and see the identity in the viewfinder in around 1 second. The boundary of the object is displayed and accurately tracked in real time. Both the object's identity and geometry are quickly retrieved from a server hosting a database of 20,000 book and CD images. As the user pans across the scene, the system automatically recognizes new objects that come into view, without the user ever having to press a button. Unlike [5], our system performs motion analysis on the phone and selectively decides when to send new query data, rather than continuously transmitting video over a wireless network.

The paper is organized as follows. Sec. 2 presents the design and implementation of the streaming MAR system. Then, Sec. 3 shows the results of recognition tests in which our streaming MAR system is used to recognize many books and CDs in cluttered settings.

2 System Design

An overview of the streaming MAR system is presented in Fig. 1. On the phone, an efficient motion estimator is used to determine camera movement, which enables us to selectively send query data to the server only when needed and to track an object after initial recognition. Details on phone operations are given in Sec. 2.1. On the server, local features are extracted from an uploaded query frame, classified through a k-d tree to find the best matching database image, and geometrically verified. A thorough description of server-side image matching is given in Sec. 2.2.

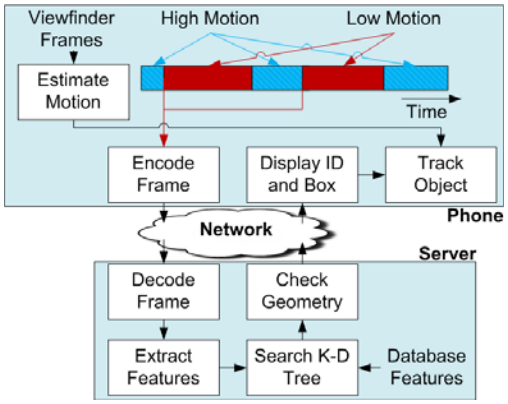


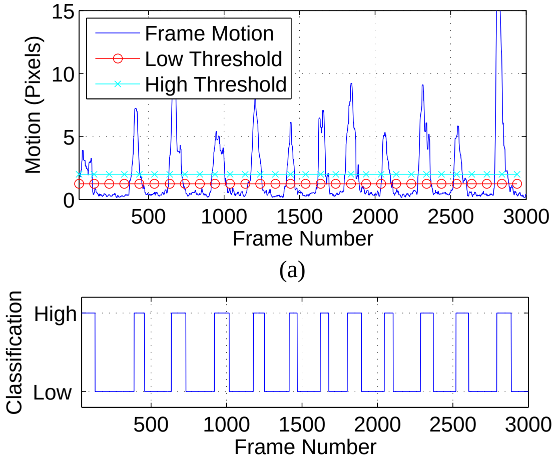
Figure 1: Design of the streaming MAR system.

2.1 Client on the Phone

Our user attention model assumes that when camera movement is low, the user is interested in the objects shown in the viewfinder. Based on this model, our system intelligently selects when to upload a new query frame to the server, rather than continuously sending frames as in [5]. Camera motion is estimated by aligning successive viewfinder frames [1]. Unlike conventional motion estimators which are too slow for mobile computing, viewfinder alignment is an optimized operation which can be performed 30 frames per second on a Nokia N95.

Fig. 2(a) plots the motion for a video captured on a Nokia N95, during which a user pans between multiple CDs. Because the motion values are noisy, classifying the sequence into low-motion and high-motion intervals using a single threshold is error-prone. We use two techniques to combat the noise. First, our system employs a Schmitt trigger with two different thresholds [7]. Only when the motion falls below a low threshold (rises above a high threshold) is a low-motion (high-motion) interval declared. Second, runlength thresholding is used. The runlength is defined to be the number of consecutive frames for which the classification stays constant. Our system requires the runlength to surpass a threshold before the classification can change. Rapid temporal oscillations in the classification are thus suppressed. After applying these two techniques, our system's motion classification is shown in Fig. 2(b).

At the start of every low-motion interval, a 320×240 JPEG-compressed viewfinder frame is uploaded to the server. This selective upload policy is much more rate efficient than continuously sending frames to the server. The average query frame size is 15 KB and can be quickly transmitted over a WiFi or 3G connection. Subsequently, the server replies with the object's identity as well as its geometry within the viewfinder. A boundary for the object



(a)

(b)

Figure 2: (a) Motion caused by panning between different CDs, captured on a Nokia N95. (b) Classification of motion as low or high.

is drawn and tracked using viewfinder alignment, for as long as the object remains visible. After the phone contacts the server at startup, communications during the queries are fast because a persistent connection is maintained. The latencies for the operations on the phone are listed in Table 1.

2.2 Recognition Engine on the Server

On the server, accurate recognition is achieved through feature-based image matching. SURF features [3] are extracted from the received viewfinder frame. On average, 150 SURF features are found in each 320 × 240 query frame. These features can be reliably matched despite occlusions, clutter, geometric deformations, and photometric distortions. For fast search through the image database, the SURF features are quantized through a k-d tree with approximate nearest neighbor (ANN) search [2]. We avoid using the popular scalable vocabulary tree (SVT) [6], because with so few features extracted from each 320 × 240 frame, SVT classification becomes inaccurate. A 640 × 480 frame would give more accurate SVT results, but it would increase transmission and feature extraction delays by 4×. Finally, a geometric consistency check (GCC) is performed between the query frame and the database image that has the largest number of matching features found in the ANN search. GCC significantly reduces false positives and allows spatial localization of the object within the query frame. The latencies of the different operations on the server are shown in Table 1.

Table 1: Operations in the recognition pipeline.

Location	Operation	Latency (ms)
Phone	Align two frames	30
Phone	JPEG-encode frame	200
Network	Transmit frame	120 (WiFi), 400 (3G)
Server	Extract SURF features	100
Server	Search through k-d tree	500
Server	Check geometry	50

3 Recognition Tests

We test the streaming MAR system on a sequence of 16 books and a sequence of 13 CDs. The phone client is implemented on a Nokia N95, and the server is a dual-core Linux machine. A video of the live tests is included in the ISMAR submission and also available online.[1] The objects are placed in a cluttered environment. Poor lighting, reflections, camera noise in the frames, and random orientations of objects all create significant challenges for accurate recognition. Despite these obstacles, in each sequence, our system correctly recognizes all the objects appearing in the viewfinder. The average latency is 1 second over WiFi, which makes the user experience very interactive. After recognition, the identity and outline of each object is shown in the viewfinder, like in Fig. 3.

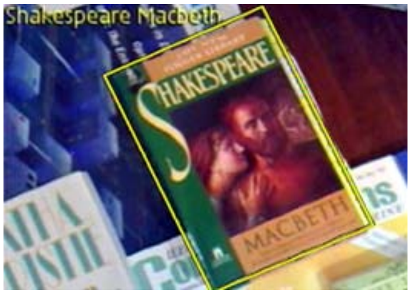


Figure 3: Identity and boundary of book shown in viewfinder.

4 Conclusion

Our streaming MAR system's low latency enables continuous augmentation of live video. Achieving a 1 second recognition latency is the result of careful optimization of the recognition stages on the phone and server. As user interest is automatically inferred from camera movements, there is no need to press a button, providing a very intuitive and compelling user experience.

Acknowledgements

The authors wish to thank Andrew Adams for his viewfinder alignment software and the reviewers for their helpful suggestions.

References

[1] A. Adams, N. Gelfand, and K. Pulli. Viewfinder alignment. *Computer Graphics Forum*, 27(2):597–606, April 2008.

[2] S. Arya and D. M. Mount. Algorithms for fast vector quantization. In *IEEE Data Compression Conference*, pages 381–390, 1993.

[3] H. Bay, T. Tuytelaars, and L. J. V. Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages I: 404–417, 2006.

[4] J.-P. Chevallet, J.-H. Lim, and M.-K. Leong. Object identification and retrieval from efficient image matching. *Information Processing Management*, 43(2):515–530, March 2007.

[5] Kooaba. Product logo recognition. http:// www.kooaba.com / technology / labs.

[6] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Computer Vision and Pattern Recognition*, pages II: 2161–2168, 2006.

[7] O. H. Schmitt. A thermionic trigger. *Journal of Scientific Instruments*, 15:24–26, January 1938.

[8] SnapTell. Media jacket recognition. http:// www.snaptell.com / demos / DemoLarge.htm.

[9] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W.-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod. Outdoors augmented reality on mobile phone using loxel-based visual feature organization. In *ACM International Conference on Multimedia Information Retrieval*, pages 427–434, 2008.

[10] S. S. Tsai, D. Chen, J. Singh, and B. Girod. Image-based retrieval with a camera-phone. In *IEEE Internal Conference on Acoustics, Speech, and Signal Processing*, 2009. Technical demo.

[1] http://mar3.tnt.nokiaip.net/cibr/ISMAR-2009-Demo.wmv