

## Video Content Detection and Tracking for Mobile Augmented Reality

Gabriel T, [gtakacs@stanford.edu](mailto:gtakacs@stanford.edu), Vijay C, [vijayc@stanford.edu](mailto:vijayc@stanford.edu), Louis C, [louistw@stanford.edu](mailto:louistw@stanford.edu)

Over the past decade mobile phones have steadily increased in capability. They have evolved from simple wireless telephones into ubiquitous, hand-held, multimedia computers. This new generation of phones is fully-networked, audio/video-input/output devices with localization technology.

We believe that Mobile Augmented Reality (MAR) has a great potential to synergize these technologies. MAR promises to leverage the vast amount of information available over the Internet to augment the users experience of reality [1]. A commonly used paradigm for MAR is for a camera-phone to sample the world from the user's perspective, extract pertinent information from the images, and re-display the images to the user with overlaid information.

Our method of extracting information from the images relies on computing unique, identifiable size-invariant image-features (SURF, SIFT) [2]. These features are then matched against a database to discover the contents of the image. The information from the database is then localized in the image for re-display to the user. Our goal is to perform these operations many times per second to give the user a smooth, real-time update of information.

We have developed a framework for feature extraction and matching for still-images. Feature matching with the database is computationally expensive and is the bottleneck. It is infeasible and inefficient to perform feature extraction and matching for each frame in the video sequence. Therefore, temporal redundancy within the frames in the video sequence should be exploited to reduce the frequency of feature extraction and matching. By tracking the content of the frames, we can determine when significant new content has appeared in the video. Only after there is new image content should we re-extract features and query the database.

Various object tracking algorithms have been proposed, many of which are computationally complex. State-of-the-art video compression standards produce motion vectors at the macro-block level in real-time using dedicated hardware. These motion vectors can be directly derived from the encoder without any additional computation. However, there are several challenges in using motion vectors for tracking purposes [3].

The macroblock motion vectors point to blocks in previous or future frames and are optimized for compression using a rate-distortion Lagrangian metric. Therefore, the motion vectors do not represent the true motion of the scene. Also, not all macro-blocks have motion vectors associated with them. For example, the macro-blocks that are intra-coded do not have motion vectors. As a result, the macroblock mode-information needs to be incorporated in the tracking process.

The macroblock motion vectors can be decomposed into two parts: the flow caused by the camera motion, and the flow caused by object motion in the video. The scene background is more constant across time, and therefore better for content matching. Segmenting the frames into foreground and background based on the motion can allow us to preferentially track robust features in the background [4][5][6][7]. Additionally, rate-distortion optimal motion-vectors have information irrelevant to camera and object movement. The erratic motion vectors need to be properly filtered to remove unnecessary information. By segmenting the image and filtering the motion vectors we hope to increase their reliability for tracking.

We plan to develop a framework for reliably tracking background features that have been matched up with our database. We will do so by exploiting the motion vectors of a simultaneously encoded video stream. Our tracking algorithm must be aware of geometric

constraints, as well as the reliability of the motion vectors. With our framework for video MAR, we hope to provide a smooth interactive experience for the MAR user.

## REFERENCES

- [1] S. Feiner, B. MacIntyre, T. Höllerer, and T. Webster, "A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment", *Proc. ISWC '97 (First IEEE Int. Symp. on Wearable Computers)*, October 13-14, 1997, Cambridge, MA.
- [2] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", *Proceedings of the ninth European Conference on Computer Vision (ECCV 2006)*, May 2006, Graz, Austria.
- [3] Sung-Mo Park, Joonwhoan Lee, "Object Tracking in MPEG Compressed Video Using Mean Shift Algorithm", *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia*, vol 2, pp 748-752, Singapore, Dec 2003.
- [4] C. Stiller, J Konrad, "Estimating Motion in Images Sequences," *IEEE Signal Processing Magazine*, pp. 70, July 1999.
- [5] M. Harville, "Foreground Segmentation Using Adaptive Mixture Models in Color and Depth," *IEEE Workshop on Detection and Recognition of Events in Video*, pp. 3-11, 2001.
- [6] C. Stauffer, W. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999.
- [7 ] J. Wang, E. Adelson, "Spatio-Temporal Segmentation of Video Data," *Proceedings of the SPIE: Image and Video Processing II*, vol. 2182, San Jose, Feb 1994.