Illumination Sensitive Dynamic Virtual Sets

Hideaki Nii Bert deDecker Yuki Hashimoto Dylan Moore Jay Summet Yong Zhao
Jonathan Westhues Paul Dietz John Barnwell Masahiko Inami Philippe Bekaert
Ramesh Raskar*
Mitsubishi Electric Research Laboratories (MERL)

1 Executive Summary

We will demonstrate new methods of flexible scene capture (including motion, orientation, and incident illumination), to create a dynamic 'virtual recording set.' We use tracking tags that are imperceptible under attire; inserted Computer Graphics elements can match the lighting on the presenter, making the technique ideal for real-time broadcast.

2 Vision

The goal of our project is to support accessible motion capture that is easy, and yet powerful, for the authoring and enhancement of visual effects for video production.

Our requirements to make this process accessible demand a tracking system that does not impose complicated set ups, while providing interactive feedback and maintaining incredibly accurate measurements, to the millimeter.

With the proliferation of digital video dissemination across media such as the world wide web, video authoring and animation are becoming an essential part of the online experience. In this YouTube empowered world, virtual sets (such as the one we propose) at home or school may become as routine as HTML editors of yesteryear.

While several matured techniques for motion-capture are available, they require work areas dedicated to esoteric systems. In addition, the data they obtain requires hours of post-processing to clean up.

Rapid advances in solid state lighting and sensing have made possible the exploration of new scene capture techniques for computer graphics and computer vision. The high speed and accuracy of recently developed light transmitters and photosensors enable very fast and robust attribute measurement even for highly dynamic scenes.

We combine the principles of optical data communication and the capture of scene appearance using the simplest possible optical devices – an LED with a passive binary mask used as the transmitter and a photosensor used as the receiver. We show how one can estimate geometric and photometric attributes of chosen scene points with speed and accuracy by strategically placing a set of optical transmitters to spatio-temporally encode the 3D space of interest.

This encoding is designed by exploiting the epipolar geometric relationship between the transmitters. Photosensors attached to scene points demultiplex the coded optical signals from multiple transmitters, allowing us to compute not only their location and orientation but also their incident illumination and the reflectance of the surfaces to which they are attached.

We use our wireless tag system to demonstrate methods of adding special effects to captured videos that cannot be accomplished using pure vision techniques that rely on camera images.

3 Technical Innovations

We demonstrate a new optical motion and lighting capture technique. Our system performs the same abilities as the best existing optical motion capture methods, but in addition, our method has X abilities above and beyond these traditional systems:

First, we can record orientation and incident illumination at the marker tags. For the motion capture portion, we can track the position of marker tags at a rate of 500 Hz, with 8 bit location precision, and with self-identifying tags. For the orientation, we strategically configuring a set of modulated light transmitters and use light modulation and demodulation techniques to estimate individual attributes at the locations of the receiving photosensors. Although these measurements are made at a sparse set of points in a scene, their richness allows extrapolation within a small neighborhood. These measured scene attributes can therefore be used to factorize a captured video sequence and manipulate the video based on the resulting attributes. In addition, such a factorization can be accomplished at a very high speed (much faster than a typical camera could achieve), allowing the manipulation of individual video frames at an intra-image level. All this is done using strikingly simple hardware components.

Second, since each tag records its own location, there are no issues of reacquisition in the case of occlusion–Therefore, our system can support an unlimited number of tags while maintaining the same fast capture rate.

Third, In a virtual set application, the flexibility of our tags becomes apparent: Not only can we capture motion and lighting conditions in their actual setting, but the tags worn by an actor are easily hidden by theatrical wardrobe as to not interfere with his or her performance or be visible in the video recording.

Fourth, a key advantage of our approach is that it is based on components developed by the rapidly advancing fields of optical communication and solid-state lighting. It allows us to capture photometric quantities without added software or hardware overhead. Marker-based techniques that use other physical media cannot capture photometric properties.

Finally, we must address a disadvantage of our approach to present a complete assessment of its ability. The tags must be in the line of sight of the transmitters (at least those that label the space it occupies). We face the usual challenges of dealing with limited dynamic range when the ambient lighting is very strong; dealing with loss of communication due to occlusions (shadows); and handling multi-path distortions due to secondary scattering of light. But our technique still allows much more freedom in tracking on location and in dynamic settings than current systems.

*e-mail:raskar(at)merl.com,web:http://www.merl.com/people/raskar/

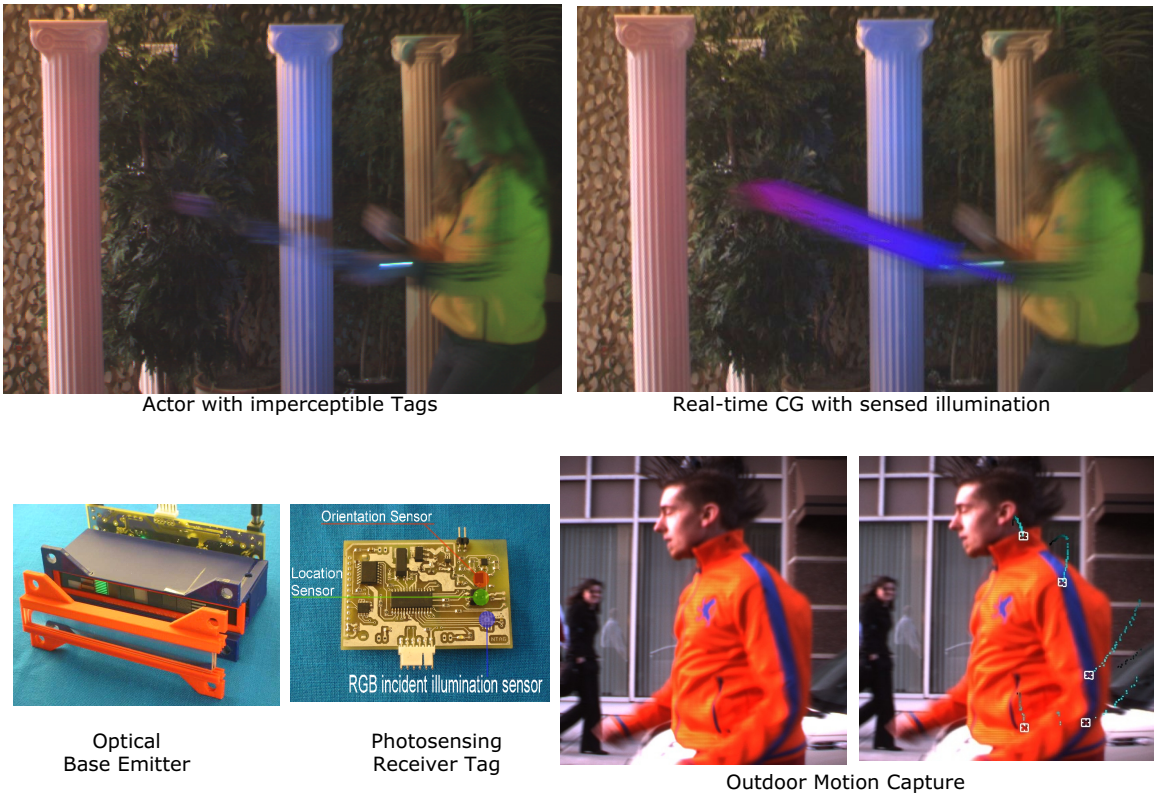

Figure 1: *(Top) We track the tags on the sword, insert CG sword and change the appearance of the CG sword to match the spotlight colors. (Bottom Left) Our hardware is made of off-the-self components and costs 10s of US dollars. (Bottom Right) The man is wearing visually imperceptible photosensing tags under the jacket. He is recorded from a moving car to capture his motion at 500 Hz in daylight.*

4 Future Impact

To summarize the technical advancements and impact thereof, consider this: The idea of an accessible motion capture system is quite underdeveloped; the majority of mature systems require high speed cameras, which elevates the cost of the system dramatically. Our system foregoes these expensive parts, which greatly reduces the cost and allows the technology to enter realms of study and creative generation where previously, motion capture could not be afforded. To put it bluntly: No longer does motion capture have to occur in a specially designated space, with special lighting, and cost five to seven figures.

Fields that would immediately see the benefits of an accessible motion capture system could include independent biomedical research centers, rehabilitation clinics, and independent research scientists in many fields including physics, anthropology, and sociology. Furthermore, fields beyond human research would benefit as well: Even veterinary clinics could capitalize on the accessibility of motion tracking in examining animal gaits and behaviors for diagnosis. Finally, this would be a boon to artists everywhere who have been kept out of motion capture because of price tag and technological complexity alone.

The one predicable impact of opening exclusive techniques to a wider audience has always been the unpredictability of innovation. We expect our system to evolve, and the usefulness of such a system to evolve with it.

5 User Experience

Observers will arrive to see a mock television studio, complete with an actor (an active participant) and a mobile camera (controlled by another active participant). They will see, in real-time, the video feed enhanced with synthetic computer graphic (CG) elements, which are added to the video on the actor, or on the props in the set.

The camera looks like store-bought stock, except for a small addition along the top: One 10cm wide black box that doesn't appear to be adding anything to the scene–or at least, to the naked eye. The actor is wearing normal clothing, but with embedded tags which are otherwise unnoticeable. As the actor moves from one colored spotlight to another colored spotlight in the real studio, the enhanced video show that added CG elements also are affected by the color of the spotlights. (Included figure shows a woman with a sword, demonstrating this principle). These visuals, including the color augmentation, are done in real-time. Note that no visually distracting markers are visible and the video is shot in natural lighting.

Two attendees can participate in the demonstration, one as the camera-person and one as the actor. Active participants will be presented with a choice of different action shots, CG elements, lighting, and costume options. Once the duo is ready to produce in our virtual set, they will be given 30 seconds to record. With preparation time, the two participants will spend less than five minutes in the studio. We will encourage actors to choose fast actions that would normally cause motion blur, such as wildly gesticulating arms and hands, and quick, sudden changes in motion. We hope to surprise the observers with high-speed motion tracking in such a dynamic setting.

6 Technical Summary

We describe an economical and scalable system where the light transmitters are space-labeling beamers and beacons. Each beamer

is simply an LED with a passive binary film (mask) set in front. The light intensity sequencing provides a temporal modulation, and the mask provides a spatial modulation. We use an array of such beamers, called projectors, where the binary masks of individual units are carefully chosen to exploit the epipolar geometry of the complete beamer arrangement. Each unit projects invisible (infrared) binary patterns thousands of times per second. Tags with photosensors attached decode the transmitted space-dependent labels that enable us to compute their locations in 3D. The tags also decode the received intensities from the set of beacons, which are used to compute their orientations. The location and orientation data for each unit are computed hundreds of times per second. Attaching the tags to scene points therefore yields the locations and orientations of these scene points at a very high frequency. In addition, the tags measure their incident ambient illumination. When the tagged scene points are imaged with an external camera, we can factor in real time the radiances measured at the corresponding camera pixels into the incident illuminations and the intrinsic reflectance of the corresponding scene points. Because the entire working volume of the system is optically labeled, the speed of the system remains constant, regardless of how many tags (scene points) are tracked.

6.1 Receiver Tag

Although various combinations are possible, we suggest the simplest possible optical receiver, a photosensor. Our photosensor excludes the customary lens so that when the sensor is tilted, the resulting photocurrent exhibits a cosine falloff. In addition, the tag is equipped with flash memory for storing data and an RF antenna for wireless communication. The receiver can report three quantities. It can decode binary data from a light source in a given time slot by tuning into 455 kHz carrier signal, compute the signal strength by low-pass filtering the 455 kHz carrier signal and sense the 'ambient' reading, i.e. total DC irradiance. To simplify the design and electronics, our current prototype, however, uses three distinct photosensors, one for each task. (Please see a detailed description in the supplementary material.)

6.2 Optical Transmitter

Our transmitter, a beamer, is similarly very simple, a light emitting diode (LED) with a passive film set in front. The LED is temporally modulated at 455 kHz. The binary film achieves fixed spatial modulation–the optical signal is transmitted in parts where the film is transparent and blocked where the film is opaque. The binary film represents one bit position of the binary Gray code and represents one of the patterns of the conventional binary coded structured light illumination. The system also includes a RF antenna for wireless communication. Since we are using time division multiplexing, each LED beamer is assigned a time slot. The receiver decodes presence or absence of modulated light in this time slot with synchronization. The transmitters run in an open loop. The tag stores the decoded data in onboard flash memory or transmits via the RF channel. A camera is optionally used in the system to compute reflectance. (Please see a detailed description in the supplementary material.)

The optical tags offer the ability to directly compute the location, surface orientation, and the aggregate incident illumination from multiple ambient light sources.

6.3 Location

We use the traditional Gray coded patterns. However, each bit of the pattern is projected by a different beamer. In traditional Gray coding projection, all patterns originate from a single projector. Our goal to is to achieve binary codes with different code from each of the non-colocated beamers. The collection of N light sources behind a common lens and mask create a compact array for coding one dimension of the geometry. The tags decode the presence and absence of a 455 kHz carrier as 1's and 0's to directly compute the projector space coordinate. By using 3 or more such beamer arrays, we compute the 3D location of the tag.

6.4 Orientation

It is possible to estimate tag orientation by sensing relative location of 3 or more sensors rigidly mounted on the tag. However, this becomes unreliable as the distance between the mounted sensors approaches resolvable location resolution. We describe a solution that fits within our framework of a single photosensor tag to estimate instantaneous orientation. Our method estimates photosensor orientation by exploiting the natural cosine falloff due to foreshortening and employing the known instantaneous location estimation.

We assume the photosensor (without lens) is attached parallel to the surface. We then find the surface normal–i.e. orientation up to two degrees of freedom. We cannot estimate the rotation around the optical axis of the photosensor. The incident angles between the incident rays from distinct light sources and the surface normal attenuate the received strengths from those sources sensed at the photosensor by the cosine falloff factor. By observing the intensity from four or more light sources with known location and brightness, one can determine the surface normal associated with the sensor.

6.5 Illumination

We measure the flux arriving at the photosensor using the photocurrent which is a measure of the ambient irradiance multiplied by the sensor area. The ambient flux is measured in the visible range and hence, it is not affected by near-IR emission of beamers or beacons. Since the area of the detector is fixed, the photocurrent is proportional to the integration of the irradiance over the hemisphere. However, to sense color, we used a separate triplet of sensors, for red, green and blue wavelength. Note that irradiance integrated over the whole hemisphere includes direct as well as global illumination.

6.6 Implementation

The tags use off-the-shelf IR decoding modules and PIC18 microcontrollers, and hence they are very inexpensive. We designed and built the projectors using CAD software and printed with a 3D printer. Each beamer is on for only 33 microseconds with 33 microsecond delay at 455 kHz. A set of 10 beamers can complete the frame in less than 800 microseconds including guard delays. A single beamer, thus, has a bandwidth of approximately 10,000 bits per second. A set of beamers on a single axis can project patterns at 1000 Hz. Thus the spatiotemporal labeling bandwidth is 10^3 Hz * 10^3 locations = 10^6 bits per second. When orientation beacons are used, the additional timeslots reduce the update rate to 124 kHz.

The extent of working volume of the system is 2 meter in each dimension at approximately 3 meters from the location beamers. The latency of our system is less than 1 millisecond. The optical packet length is 600 microseconds. The PIC microcontroller on the tag has a delay of 100 microsecond. Our bottleneck is wireless communication. Our wireless modules work at 78 kbits/sec but the bandwidth is shared between the tags. A wired version can work at full rate. The accuracy of location is 4 millimeters at 2 meters.

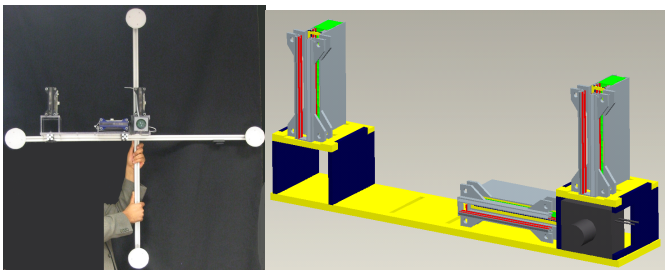

Figure 2: *Out setup with 3 projectors and four beacons.*

7 Context

As noted, our project is a direct extension upon previous work in the field of motion capture, which has a rich history of contiguous developments and improvements [Welch and Foxlin 2002]. This demo is related to a paper appearing at Siggraph 2007 on the same tracking technology, which goes into further detail about the roots of our work [Raskar et al. 2007]. We showed a preliminary version at Siggraph 2006 under constrained setting. The demo was 'Instant Replay' where the two participants play the game of Air Hockey. We tracked a puck in 2D in real time and displayed its trail by projecting from an overhead projector. Our virtual set demonstration this year involves real time 3D tracking from a dynamically moving camera, sensing orientation and incident illumination to create appealing visual effects.

Let us consider optical location tracking with tags. Motion capture systems used in movie studios commonly employ high-speed cameras to observe passive visible markers or active light emitting diode (LEDs) markers [ViconPeak 2006; Motion Analysis Corporation 2006; Phase Space Inc 2007; PTI Inc 2006]. For example, the Vicon MX13 camera can record 1280x1024 full-frame Grayscale pixels at speeds of up to 484 frames per second with onboard processing to detect the marker position. These devices provide highly reliable output data and have been developed over the last three decades. However, the expensive high-speed cameras pose several issues in terms of scalability. Bandwidth limits the resolution as well as the frame-rate. Higher frame-rate (shorter exposure time) means brighter controlled scene lights for the passive markers or the use of power hungry LED markers. To robustly segment the markers from the background, these systems also use methods for increasing marker contrast. This usually involves requiring the actor to wear dark clothes in a controlled lighting situation. In our case, the use of photosensing allows capture in **natural settings**. Since the photosensors are barely discernible, the character can wear natural clothing with the photosensing element of the tag poking out of the clothing. The ambient lighting can also be natural because the photosensors receive well-powered IR light. The power is comparable to IR emission from TV remote controls. So, in studio settings, the actor may wear the final costume, and he/she can be shot under theatrical lighting.

Recently, [Nii et al. 2005] [Raskar et al. 2004] [Lee et al. 2005] have presented the idea of locating photosensing RFID tags with a traditional data projector. Their goal was pose estimation and augmented reality. One may be tempted to use a TI Digital Light Processing (DLP) system or Gated Light Valves (GLV) for high-speed binary coding [Mark Bolas and Ian McDowall 2004] [Cotting et al. 2004] [Wenger et al. 2005] [Silicon Light Machines 2006]. By limiting ourselves to a specific goal, we are able to use a more common light source – an LED – making our system scalable, compact, and inexpensive.

The UNC HiBall system [Welch 1996] uses a group of 6 rigidly fixed position sensitive detectors (PSD) to find location and orientation with respect to actively blinking LEDs. Each LED provides a single under-constrained reading at a time. The system requires a large ceiling installation. We use a single photosensor in place of the 6 PSDs and we do not require active control of the LEDs; they run in an open loop. Thus, multiple receivers can simultaneously operate in the same working volume.

Systems such as Indoor GPS [Kang and Tesar] use low-cost photosensors and two or more spinning light sources mounted in the environment. The spinning light sources sweep out distinct planes of light that periodically hit the optical sensors, and the system uses the timing of the hits to derive the position at 60 Hz. However, nothing moves faster than electrons – fast-switching, solid-state emitters have the potential to achieve extremely high rate at a very low cost and form-factor compared to mechanically rotating devices.

Let us now consider the update rate. Systems that use cameras are limited by their frame-rate. Active beacons must use time division multiplexing [PTI Inc 2006] so that only one LED can be turned on at a time. Passive markers need to resolve correspondence to avoid the "marker swapping" problem. Existing systems cannot sense orientation and incident illumination. Although our current prototype lacks the engineering sophistication to compare with state-of-the art motion capture systems, it is worth exploring what may be possible in the future.

References

COTTING, D., NAEF, M., GROSS, M., AND FUCHS, H. 2004. Embedding imperceptible patterns into projected images for simultaneous acquisition and display. In *International Symposium on Mixed and Augmented Reality*.

KANG, S.-H., AND TESAR, D. Indoor gps metrology system with 3d probe for precision applications. *The University of Texas at Austin*.

LEE, J. C., HUDSON, S. E., SUMMET, J. W., AND DIETZ, P. H. 2005. Moveable interactive projected displays using projector based tracking. In *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*, ACM Press, New York, NY, USA, 63–72.

MARK BOLAS AND IAN MCDOWALL, 2004. Snared illumination. Emerging Technologies, Siggraph.

MOTION ANALYSIS CORPORATION, 2006. Hawk-i digital system.

NII, H., SUGIMOTO, M., AND INAMI, M. 2005. Smart light ultra high speed projector for spatial multiplexing optical transmissio. *Procams*.

PHASE SPACE INC, 2007. Impulse camera. http://www.phasespace.com.

PTI INC, 2006. Visualeyez vz 4000.

RASKAR, R., BEARDSLEY, P., VAN BAAR, J., WANG, Y., DIETZ, P., LEE, J., LEIGH, D., AND WILLWACHER, T. 2004. Rfig lamps: interacting with a self-describing world via photosensing wireless tags and projectors. *ACM Transactions on Graphics 23*, 3 (Aug.), 406–415.

RASKAR, R., NII, H., DE DECKER, B., HASHIMOTO, Y., SUMMET, J., MOORE, D., ZHAO, Y., WESTHUES, J., DIETZ, P., INAMI, M., NAYAR, S., BARNWELL, J., NOLAND, M., BEKAERT, P., BRANZOI, V., AND BRUNS, E. 2007. Luminetra: High speed scene point capture and video enhancement using photosensing markers and multiplexed illumination. *ACM Transactions on Graphics 26*, 3 (Aug.).

SILICON LIGHT MACHINES, 2006. Gated light valve.

VICONPEAK, 2006. "camera mx 40".

WELCH, G., AND FOXLIN, E. 2002. Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Comput. Graph. Appl. 22*, 6, 24–38.

WELCH, G. F. 1996. Scaat: Incremental tracking with incomplete information. Tech. rep., Chapel Hill, NC, USA.

WENGER, A., GARDNER, A., TCHOU, C., UNGER, J., HAWKINS, T., AND DEBEVEC, P. 2005. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Trans. Graph. 24*, 3, 756–764.