RedTag: Automatic Content Metadata Capture for Cameras

Tom Bartindale, Daniel Jackson, Karim Ladha, Sebastian Mellor, Patrick Olivier, Peter Wright
Culture Lab, Newcastle University
Newcastle upon Tyne, UK
{tom.bartindale, d.g.jackson, karim.ladha, s.j.i.mellor, p.l.olivier, p.c.wright}@ncl.ac.uk

ABSTRACT
RedTag is an optical tagging system that provides time based identification of objects, people or devices via small low cost infrared transmitters and receivers. We have developed RedTag as a cheap and flexible method of augmenting existing video capture equipment with an additional temporal metadata output of content based information. In this note, we describe the technology behind RedTag and demonstrate the interaction opportunities that arise through access to temporal metadata.

Author Keywords
Film; metadata; infrared; DTMF; editing; production; electronics.
ACM Classification Keywords
I.4.8 IMAGE PROCESSING AND COMPUTER VISION: Scene Analysis (Tracking)

INTRODUCTION
There are many uses for a low cost and flexible system for tagging physical objects and people for remote identification, specifically within a specific spatial frame of reference (such as that dictated by a camera's field-of-view). Within video production, temporal, content based information about each video clip is vital metadata that it is currently not possible to automatically generate, especially at the point of capture. The identification of actors, participants or objects within a video clip is vital for later categorization and identification during post-production. Professional teams often employ a Script Supervisor to manually gather this data, which is reconciled in post-production, but this is often not possible for smaller, non-professional teams. A system that integrates with existing camera equipment and workflow tools would avoid unnecessary expense and change to existing workflows. RedTag can capture content based metadata about the visible infrared tags which is then recorded onto existing camera equipment using an unused audio recording channel, providing content based metadata about the frame context with implicit temporal synchronization. No additional or non-standard camera equipment is required except a small RedTag receiver and audio cable. RedTag components are small, wireless, cheap to manufacture and deploy, robust, flexible and offer sensing within a cameras field-of-view.

VIDEO METADATA
New forms of media delivery, user interaction, branching narrative, multi-format content and second-screen content are emerging as key outputs of a production. Rich and accurate metadata on source footage is key to creating value in such content. Metadata allows footage to be re-purposed, re-edited and matched with additional content without direct human intervention. It is also key for making content accessible, providing triggers for audio-description, subtitling and re-mastering of content. Video files already support native attachment of metadata through MXF [3] and many video cameras automatically record static technical information about shot and camera setup, however, these formats do not generate or store temporal context or content identification. RedTag augments objects, props, scenery and actors to automatically recording when these items enter and exit frame. The practice of accompanying 'rushes' (dumps of footage from the cameras) with notes on shot content from a Script-Supervisor (or other member of the crew) is common practice, and this data is used by the post-production team to more easily interpret footage. The automatic availability of this metadata facilitates a number of processing opportunities with the captured footage for post-production and additional content which are not possible with the data provided by manual annotation techniques both for professional and amateur video producers:

Editing: Video metadata can be used to either automatically mark in and out points in clips depending on the entry, exit or presence of specific objects or persons, or be used as an indicator for the editor as a time-based overlay on the editing surface. Rough or 'rush' edits can then be created with little effort ready for clip review or further editing.

Continuity Checking: Continuity checking across multiple scenes, shots and locations is key to producing believable content. Content based metadata regarding props in shot, costumes or extras within each shot can support automatic highlighting of inconsistencies and possible problems.

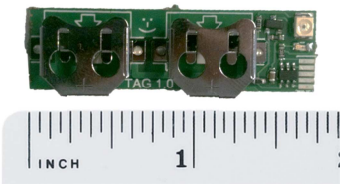

Figure 1 A RedTag Transmitter

Segmentation: Content based metadata supports the segmentation of video into semantically meaningful chunks, and subsequent efficient identification and contextualizing of footage. Temporal metadata allows sub-segments of clips to contain meta-information without segmenting clips into individual files, supporting increased temporal granularity for capturing contextual metadata and allowing for time based tagging of content within the scene. This can be used for providing time-based second screen content during broadcast.

Searching: Often individuals and organizations produce and archive all of their footage for audit or future use. Time-based content based metadata supports searching large archives for specific objects, locations or persons as well as retrieving the specific point within a clip that the search term was found.

Annotation: Clips that are augmented with time-based metadata of the appearance of objects, people and locations can be used to automatically segment clips for on screen overlays, subtitling and audio description. In live broadcasts, metadata could provide automatic titling of presenters or interviewees, whilst simultaneously generating live audio description about the presence of objects, locations or people who are silent in the scene.

PREVIOUS WORK

Systems such as iBand [4] nTag, GroupWear [1], SpotMe[1] and Poken[2] all provide wireless technologies for tagging people or objects, but these systems are design for enabling social interaction between participants through facilitating social connections within shot distances. Longer range wireless tagging has been successful for indoor localization: The ActiveBadge [6] system uses Infrared (IR) beacons worn by people in a building which transmit a unique code at 15s intervals. Transmissions are detected by sensors placed throughout the building to determine the wearer's position. Similarly, Digital Assistant [5] IR badges can identify people to specific interaction points by transmitting unique codes to worn receivers, and Meme Tags can exchange short messages using IR for two way-communication but these solutions do not take into account the camera frame as a reference. Similarly Intellibadge [2] is a system for adding value to social interaction at live events through augmenting attendees with RFID aware device and fixed RF beacons in specific locations in the venue, combining the personalization of Poken with the indoor localization and contextual control of media demonstrated by the Active Badge and Digital Assistant. Although these technologies provide temporal identification of subjects in a location, they do not integrate with the field-of-view of a camera or recording equipment within the space. These technologies make use of IR technology for identification of people and objects from a distance but without considering camera field of view or recording of this data without additional hardware. Optical technologies such as QR codes or AR tags however have proved successful in tagging objects within video, but have the obvious disadvantage of visibility within the scene, negating their use in broadcast media.

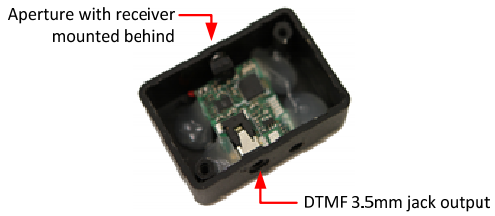

Figure 3 RedTag Receiver (in a camera mounted enclosure with an aperture matching the camera lens)

THE REDTAG SYSTEM

The RedTag system consists of multiple low-power infrared emitters, and rechargeable receivers (see Figure 2) leveraging robust proven IR technology. Each small transmitter is programmed with a unique identifier (ID) and affixed to a person, object or at a fixed location. Receivers are mounted on camera equipment orientated in the same direction as the lens and connected to the secondary audio recording input. Tags regularly 'chirp', transmitting their ID via infrared. These codes reach the receiver only when within the camera field-of-view and are output as DTMF tones representing the visible tag's ID and are recorded onto the camera audio feed. Simple DTMF decoding software is used to return the ID and relative timestamp of each tag in the recording.

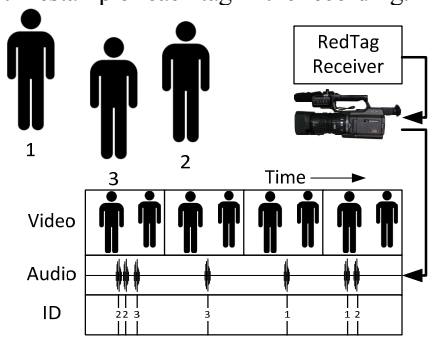

Figure 2 RedTag System Functional Workflow

RedTag transmitters use a modulated IR signal, similar to TV remote controls. Emitters are designed to be low cost, small consumables which can be embedded into objects. In one implementation, each emitter is powered by two replaceable coin-cell batteries, giving it a lifespan of around 3 weeks. In the most common configuration, a RedTag receiver emits each code that it receives from multiple RedTag transmitters as a stream of DTMF tones, which is recorded onto a cameras additional audio channel.

Transmitter

A RedTag transmitter (Figure 1) may consists of just seven or less components on a single layer PCB, including: a 6-pin PIC 8-bit microcontroller, Infrared emitter (850nm wavelength, modulated at 455 kHz), appropriate resistors and one or two coin-cell batteries. Even in small scale production, the unit cost for each transmitter when mass produced is under $3, allowing transmitters to be used as non-returnable consumables in large scale deployments. During operation, the tag waits a pseudo-random interval (around 1 second)

before transmitting its 16-bit payload (ID). This jitter prevents any tags' chirp from falling into phase with another and repeatedly colliding which would prevent successful reception. To prevent spurious or misidentification of tags, Manchester coding is used to transmit the payload, which helps to identify collisions with other transmitters. As an additional measure, the 16-bit payload consists of a 10-bit unique identifier and a 6-bit CRC (using the ITU 6-bit CRC) to protect against ID corruption. Given the 22.75 kbaud (455 kHz carrier, 20 cycles per bit), each 16-bit ID takes 1.76ms to transmit. Given perfect synchronization between transmitters this allows 568 transmitters to be detected a second. However, as no synchronization (or two-way communication) exists between transmitters or receivers, the collision rate increases with the number of transmitters visible. With 100 transmitters in use, all transmissions with 1.76ms duration are randomly allocated within a 1s intervals, and the probability of any one transmission avoiding collision with 99 others in one second is $P(no\ collision|T=1) = (1-0.00176)^{99} \approx 0.84$. Thus the probability of one transmission avoiding collision at least once in 5 seconds (a contextually useful time window) is $P(no\ collision|T=5) = 1-(1-0.84)^5 \approx 0.9999$. A statistical simulation of collisions shows that we can reliably ($P(success) \approx 0.99$) receive the tag ID of 60 transmitters within a 5 second window.

Receiver

Each RedTag receiver (Figure 3) features a 16-bit microcontroller, infrared receiver, a re-chargeable battery and combination 3.5mm audio jack and USB connector. Receivers are mounted in an enclosure of IR blocking material with an aperture directly in front of the IR receiver. The aperture size is calculated to provide the same field-of-view to that of a camera (see Figure 4). When a transmission is received, the RedTag code received is output via the audio jack as DTMF audio tones which are still identifiable through any compression used. Each code is emitted as four DTMF[3] symbols: a '#' symbol, used as a delimiter between records, followed by a three digit number. Each tone is played for 40ms with a 50ms interval, allowing three newly seen devices to be identified each second. The receiver operates a memory queue to buffer incoming codes, and ensures that newly observed codes are reported in a first-in-first-out order as soon as possible and duplications in the queue have lower precedence than unique entries. In this way, the receiver maintains a current list of observed transmitters, and only fails to pass on this data through DTMF if the queue overflows. Although using DTMF codes means that there may be a slight temporal delay between the receiver observing the transmitter, and emitting the DTMF code for it, this queue method ensures that all of the observed devices will be recorded within a temporally relevant period. By attenuating the transmitters in firmware or optically, the effective range of the transmitter can be adjusted from 5m to

[3]http://en.wikipedia.org/wiki/Dual-tone_multi-frequency_signaling

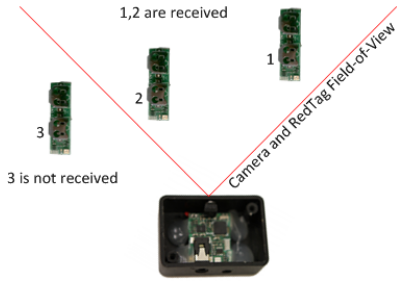

Figure 4 RedTag Receiver with a Camera Apeture

100m, depending on the sensitivity required. Transmitter range is particularly directional and subject to multi-path reflection to receivers however this can be advantageous when applying apertures to RedTag receiver units to match a camera's field-of-view. To retrieve RedTag data from a camera recording, the relevant audio channel is extracted and passed through a software DTMF decoder which outputs the ID and timestamp of each detected RedTag. Before post-production, software is used to batch process clips, separating the additional audio channel and detecting the tones (tags id's) alongside a timestamp within the stream. Each ID is replaced with a description of where the tag was located (provided by the crew) and the information is saved back to the source file as XMP metadata. If required, RedTag data (as audio) is maintained through subsequent manipulation (cutting, editing and mixing) and stripped out later.

EVALUATION

We tested RedTag in a controlled setting to determine its spatial and technical limitations. In a single tag test, the effective range of a RedTag was tested to be 6±0.1m, and the effective angle of rotation from the receiver before losing signal was 185° from the horizontal. In two standard camera shots (close shot and mid-shot), with an $f/4$ lens, RedTags are received within +=3° outside of the camera frame bounding box, whereas in a wide shot, RedTags are only received within with center 40% of the camera frame, due to the fixed sensor aperture. We accepted this limitation of a fixed aperture due as we had configured the receiver for close, interview style filming. To evaluate the effectiveness in a real world scenario, we setup a 3 key use-cases as controlled tests: an interview or presentation scenario with 1 or 2 people in frame and a fixed camera; social coffee break scenario, with multiple small groups of people and a moving camera; a film acting style scenario, with a fixed camera, and acting towards the camera. For each one, we analyzed: the **correctness** of generated metadata (false positives for transmitters not present); **temporal instability** (is the transmitter data recorded within a useable timeframe of sensing the transmitter); **maximum tags**; **range constraints** and **data accuracy** (e.g. missing people from clip). For the test, a single RedTag receiver was mounted underneath the lens of a Panasonic AF101 camera and plugged into the secondary audio input channel, and each of 11 participants wore a transmitter inside in a badge on a lanyard. 10 videos were captured in various filming scenarios. The transmitter

id's from the resulting video were retrieved and compared against the same footage which had been annotated by hand to provide a ground truth of people in the camera frame. Overall, no false positives were experienced. In clips with a static camera and a single person in frame, their RedTag was received on a regular schedule (2s) throughout the clip. In clips with more action and moving cameras we see 90% and 84% discovery for film style and coffee break style respectively – we sample only the transmitters visible for at least 2 seconds and require the detection to occur while it is visible. When allowing for delays in detection we detect 87% and 99% of transmitters within 1 minute of visibility. This is appropriate for identification of people within multiple clips but negates specific temporal identification. It was noted that for 2 people in the scene, no transmissions were received, suggesting hardware failure. In most cases, including those with larger groups of people, there were usually only 2 or 3 people facing towards the camera while others faced away or side on. A slow sweeping motion or several steady shots into the group generally captured all participants. With our placement of transmitter (in a neck worn badge holder) they are relatively low on the body but were often still be detected as the receiver was mounted below the lens and provides a non-rectangular field of view, slightly larger in the vertical direction. In practice, 10% of cases where the a person was visible in shot resulted in failure to detect their transmitter due to orientation to the camera, in addition to brief appearances (less than 1s) and transmitters occluded by arms, hands, or objects, as expected. Whether deliberate or not, this prevented identification within the scene and is a critique on the lanyard mounting method.

DEPLOYMENT

We have deployed RedTag during three academic conference events with ~200 attendees each. At each event, attendee badges contained a transmitter pre-linked to their event registration information. RedTag was used to augment video footage taken of talks and interviews to segment and label footage for rapid editing. Eight cameras used throughout the venue were equipped with a receiver. During editing the associated metadata was retrieved and mapped onto attendee registration information and displayed as a visual overlay on the editing timeline aiding rapid segmentation and identification of specific clips, as well as automatic tagging of interviewees and speakers. This information was also used to overlay associated captions and information on all playback and output streams.

DISCUSSION

Our technical evaluation and subsequent deployment has highlighted practical considerations for RedTag in practice which are key to discuss:

Reliability: In scenarios with large numbers of transmitters, RedTag may not receive temporally correct data, but depending on the length of capture will receive the transmitter at some point. This is useful for identifying if a transmitter exists in a video, but negates the use of time-based data to locate it. In choreographed capture scenarios however, transmitters can be mounted to face the camera and thus can be used to reliably determine time based frame entry and exit information about the transmitter. Some materials reflect IR light, as such false positives are a possibility. Due to the hardware auto-gain control and receiver aperture however, this scenario is unlikely and did not occur in either our study or deployment.

Occlusion: Although operating with line-of-sight IR, transmitters can be hidden in clothing or objects to avoid their appearance in shot as long as the material is IR transparent and many different materials have such properties. Occlusion by other objects however will prevent reception, thus production crew must be aware of this property when mounting transmitters. In some cases however, this can be advantageous, such as for generating audio description of items visible in a scene. In our deployment, the primary cause of occlusion was the user covering their badge accidentally, and this could be controlled by alternative mounting of the transmitters.

Human Effort: The addition of RedTag to the production workflow requires management of transmitter IDs and associated metadata. As managing cast, props and locations are already performed by roles within a production team, this responsibility could be shared amongst the crew, resulting in little extra overhead. In conclusion, the RedTag system provides a method of automatically capturing content based temporal metadata, storing this data as part of the video media for later use in post-production.

REFERENCES

1. Borovoy, R., Martin, F., Resnick, M., and Silverman, B. GroupWear. *In Proc. CHI '98*, ACM Press (1998), 329–330.
2. Cox, D., Kindratenko, V., and Pointer, D. IntelliBadge TM: towards providing location-aware value-added services at academic conferences. *In Proc. UbiComp '03*, (2003).
3. Devlin, B. What is MXF? *EBU Technical Review*, (2002).
4. Kanis, M., Winters, N., Agamanolis, S., Gavin, A., and Cullinan, C. Toward wearable social networking with iBand. *In Proc. CHI '05*, ACM Press (2005), 1521.
5. Sumi, Y. and Mase, K. Digital assistant for supporting conference participants: An attempt to combine mobile, ubiquitous and web computing. *In Proc. Ubicomp '01*, (2001).
6. Want, R., Hopper, A., Falcão, V., and Gibbons, J. The active badge location system. *ACM Transactions on Information Systems 10*, 1 (1992), 91–102.