# Media Content and Type Selection from Always-on Wearable Video

Phil. Cheatle
Hewlett Packard Laboratories, Filton Rd. Bristol, BS34 8QZ UK
Email: phil.cheatle@hp.com

## Abstract

*A system is described for summarizing head-mounted or hand-carried "always-on" video. The example used is a tourist walking around a historic city with friends and family. The summary consists of a mixture of stills, panoramas and video clips. The system identifies both the scenes to appear in the summary and the media type used to represent them. As there are few shot boundaries in this class of video, the decisions are based on the system's classification of the user's behaviour demonstrated by the motion of the camera, and motion in the scene.*

## 1. Introduction

A limitation of consumer still and video photography is that capturing rich memories takes considerable care in selecting position, direction, and time of capture. This conflicts with the user's desire to participate in activities themselves, rather than record them. It is often not clear that an event should be recorded until after it has occurred, which is clearly too late if the camera is still in the pocket. For these reasons, consumers frequently fail to capture rich memories of their activities, often only capturing one or two images to summarise a complete day. Always-on, wearable cameras, combined with cheap storage, enable large amounts of material to be captured. The goal of this research is to turn raw wearable video into memories that are both representative of the user's interest, and pleasant to watch, with very little user effort.

Video summarization and keyframe extraction are established fields[1]. Much work is aimed at *produced* video where a manual editor has selected shots, typically shorter than 10 seconds, often linked by transition effects. Detecting shots and picking a keyframe provides a simple summary, though often with too many frames. Transition effects complicate shot boundary detection[2]. Some promising recent work uses models of user attention to aid the summarization[3]. The attention models are tuned to the video from skilled camera operators and editors.

In contrast, amateur video has far fewer shots. Semi-automatic tools have been developed to assist in editing home video [4], [5] by using a range of techniques such as time clustering, audio analysis, shot shortening, and rejection of unacceptable exposure.

Wearable video differs from home video in that there is not a conscious photographer. The user is concentrating on normal life, not on photography. The best that can be expected is that the video is switched off when the user believes there is no chance of seeing anything worth recording. In the extreme case, a single shot may last for hours. Ideally, a head mounted camera is used as this enables hands-free use. An alternative is to hand carry a camcorder, without using the viewfinder or zoom, pointing it in the direction the head is facing.

Un-edited wearable video is mostly unwatchable as much of it is motion blurred and jerky due to camera motion while the user is moving. It contains much uninteresting subject matter, and is repetitious when the user looks at a still scene. Nakamura[6] describes a method of segmenting head-mounted video into scenes. Two types of behaviour are identified: passive attention, where the camera wearer gazes at a scene; and active attention, where the camera tracks a moving object.

In many cases, segments of the raw material are best represented as a still image. When the user is looking at a wide scene, a panorama is more appropriate. This is especially true for wearable cameras as the user is typically close to the object of attention. Video clips are only relevant in cases of significant motion in the scene.

This paper describes a system that extends the ideas of [6] in a number of directions, in particular: 1) Additional patterns of user interest are detected such as "glancing" and "approach". 2) User behaviour and scene motion are interpreted to identify an appropriate media for an object: still frame, panorama formed from a sub-sequence of frames, or video clip.

## 2. User behaviour interpretation

### 2.1 Overview

Wearable video from a user walking around an area of interest provides strong clues as to what is of interest to the user, and how the interesting items can best be viewed. The system architecture is shown in Fig 1. The right hand

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

column illustrates the analysis process. The motion features extracted from the raw video are: *horizontal and vertical camera rotation*; *scene motion*; *magnification* – a measure of how much the scene is changing due to forward motion; and *fast motion* – rapid turns or motion towards nearby objects. The detectors produce a value for each of these features for each video frame processed.

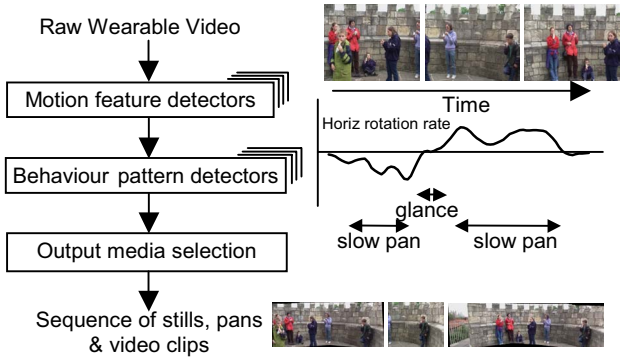


**Figure 1.** System architecture

The behaviour pattern detectors each analyse the time sequences of motion features to detect a specific behaviour pattern. A recognized instance of a particular behaviour pattern is marked by its start and stop time.

The following behaviours are detected: *Turns*: the user makes a significant, possibly fast, rotation. *Walking:* the user is translating forwards such that the view changes significantly. *Gaze:* the user stops to look at something. *Glance*: the user turns the direction of view, pauses in the new direction of interest, then turns again, usually reverting to the original direction. *Approach:* the user is looking ahead while moving forward. As a destination gets nearer, the views show increasing detail, eg walking towards a group of people. *Pan:* the user turns *slowly*, taking in a wide view. *Action:* the user is looking at a sustained sequence of motion in the scene.

*Gaze*, *glance*, *approach*, *pan* and *action* are indicative of a form of user attention that should be represented in the output. Different behaviour types are best represented by different media. *Gaze* and *glance* can be represented by a single still image. *Approach* can be represented by one or more still frames. Multiple frames are used when the users motion towards the destination causes significantly different views to be obtained, the selected frames provide increasing detail on the object of attention. This helps compensate for the lack of a zoom lens in a wearable camera. *Pan* can be used to generate a panorama. *Action* is best represented with a video clip. Parts of the raw data that do not exhibit one of the recognized behaviours are ignored as junk.

The output media selection stage combines the responses of each of the behaviour pattern detectors to form a sequence of stills, panoramas and video clips. During this stage the output from some detected instances of behaviour patterns are suppressed to avoid repetitions. A browser application views the output sequence by indexing stills, and video clips from the source video and inserting panoramas at the correct points in the sequence.

## 2.2 Motion feature detectors

In typical wearable video from a sightseeing tourist, objects are fairly distant from the camera and frame-to-frame camera motion is small. As a result the 8 parameter perspective model has proved adequate to relate adjacent frames. Feature based frame registration is used. Locally maximal features are detected using a Harris detector [7] and then matched in a later frame using a correlation search. (To facilitate motion in scene detection 1 in N frames is processed; N is set to 5.) The perspective model is fitted using the RANSAC method [8] to give robustness to outliers caused by moving objects in the scene, minor parallax of nearby objects if the camera is translating, and mismatching features. The mean difference of 2D correspondence points provides horizontal and vertical camera rotation metrics. Temporally smoothed versions are used as the feature detector outputs $H_{rot}(t)$, $V_{rot}(t)$.

The system needs to detect when the user is moving forward, or, more precisely, how much the view is changing due to forward motion. This is a function of both speed of camera translation and distance of objects from the camera. Robustly recovering accurate camera translation together with a scene depth map in the presence of significant motion in the scene is a difficult problem. Instead a magnification measure is used as an approximation of how much the scene is changing due to forward motion. If $d^t_{ij}$ is the distance between a pair of correspondence points $c_i$ and $c_j$ in frame t, the magnification is $m(t) = \text{Median}(d^{t+1}_{ij} / d^t_{ij})$, for all $c_i$, $c_j$ where $d^t_{ij}$ is greater than a minimum, D. The magnification feature, $mag(t)$, is a temporally smoothed version of $m(t)$.

Very fast motion generates motion-blurred frames that are never the subject of attention. The fast motion feature, $fast(t)$, has a binary value 1 if $mag(t) >$ threshold $M_{fast}$, $\text{Max}( H_{rot}(t), V_{rot}(t) ) >$ threshold $R_{fast}$, or the global registration failed, and 0 otherwise.

Scene motion is determined by an edge correlation method of three frames where the previous and next frames are motion compensated using the perspective model, Fig 2(a) to (c). This technique was found to be preferable to the more familiar background subtraction or motion differencing methods as a clear background image is not available due to the moving camera and scene objects. Motion differencing is highly subject to the speed of moving objects across the image. A small motion may generate a halo around the moving object, whereas fast motion (such as a football) generates an entire silhouette in both its old and its new position. See [9] for a review.

Edge pixels are matched using a weighted sum of edge strength, edge direction and colour on each side of the edge. Edge pixels that do not match are potentially moving edges. A further stage avoids insignificant small motion caused by parallax due to camera translation, or minor scene motion, (such as wind blowing trees or swaying people). Each potentially moving edge pixel is matched against others in a direction perpendicular to the edge. If a matching edge pixel is found within a search window, the pixel is no longer considered to be moving.

(a) (b) (c) (d)

**Figure 2.** Scene motion detection: a) Previous frame warped to current b) Current frame c) Next frame warped to current d) Large moving area

A pixel filling routine similar to [10] fills rows and columns of pixels between moving edges. Erosion and dilation remove small noise areas, Fig 2(d). The scene motion feature, *mot*(t), is the area of moving pixels.

### 2.3 Behaviour pattern detectors.

The behaviour pattern detectors are empirical models applied to the motion feature sequences, $H_{rot}(t)$, $V_{rot}(t)$, *mag*(t), *fast*(t) and *mot*(t).

*Turns* are defined as sequences where the sign of $H_{rot}(t)$ remains the same and $\sum_{turn}|H_{rot}(t)| > R_{turn}$ where $R_{turn}$ determines the minimum rotation of turn. Vertical *turns* are defined similarly.

*Walking* is a binary sequence, *walk*(t), which is 0 when $|\,mag(t) - 1| < Z_{walk}$ and 1 otherwise, $Z_{walk}$ is a constant.

*Gazes* are identified as sequences longer than $T_{gaze}$, where all frames have $\max(|H_{rot}(t)|, |V_{rot}(t)|) < R_{gaze}$ and $|mag(t)-1| < Z_{gaze,}$, $T_{gaze}$, $R_{gaze}$ and $Z_{gaze}$ are constants ensuring stability during the gaze.

A *glance is* defined as a sequence shorter than $T_{glance}$ between two adjacent *turns*, during which *fast*(t) = 0.

*Approach* is a sequence longer than $T_{glance}$ between two adjacent *turns* during which *fast*(t) = 0 and with significant walking, defined by $\sum_{approach} walk(t) > W_{approach.}$

*Pan* is a *turn* subsequence that has duration > $T_{pan.}$ All frames have $H_{rot}(t) < P_{pan}$, $V_{rot}(t) < P_{pan}$, and *fast*(t) = *walk*(t) = *mot*(t) = 0. For the whole pan, $\sum_{pan} |H_{rot}(t)| > R_{pan}$. $P_{pan}$ is a rotation speed limit set to 1 frame width per second. $R_{pan}$ ensures a minimum total rotation. These parameters ensure that the *pan* is sufficiently large and free from motion blur or moving objects.

*Action* is a sequence of frames with *fast*(t) = 0 and containing at least one subsequence with *mot*(t) > 0, of duration > $T_{act\ seed}$. It is further delimited by a sequence of frames with *mot*(t) = 0, of duration > $T_{act\ gap.}$

### 2.4 Output media selection

The final stage generates output from each of the detected behaviour patterns, pruning duplicate outputs if multiple behaviours are detected for the same events.

*Gazes* are very reliable indicators of user interest. Multiple gazes, not separated by turns or walking are suppressed; only the longest out of the group is retained. In the period between *turns*, the presence of a *gaze*, suppress *glances* and *approaches*.

Output frames are selected for *gazes* and *glances* by picking the frame within the behaviour sequence which contains the parts of the scene most commonly viewed. This is found by warping the frames to a common plane, counting the samples for each pixel on the common plane. The counts are back-projected to each source frame and summed. The frame with the highest sum is identified as the frame containing the most commonly viewed area. For *approach,* the same algorithm is applied to subsets of the sequence separated by W frames with *walk*(t) > 0. W is set so that 10 seconds of walking elapse between successive output frames.

An output panorama is automatically generated from detected *pans*, centered on the frame containing the part of the scene that was viewed longest.

*Action* instances in which over 20% of the frames have *walk*(t) > 0 are suppressed unless they contain a detected *glance* or *gaze*. *Action* instances separated by less than time $T_{vidgap}$ are merged. The resulting sequences generate output video clips. The frame with maximum *mot*(t) value within each non-suppressed *action* is used as a keyframe.

Excess keyframes within a video clip are limited by suppressing any output still within a video that is not separated from the next by a gap containing $\sum_{gap} H_{rot}(t) > R_{keygap}$ or $\sum_{gap} mot(t) > M_{keygap}$. $R_{keygap}$ and $M_{keygap}$ ensure that keyframes are significantly different in either camera direction or scene motion.

The system thus generates a summary of the source video consisting of a combination of still frames, panoramas and video clips. The user behaviour and the motion in the scene determine the choice of media.

## 3. Experimental results.

The system has been tested on several videos from tourist outings of family groups looking round historic cities. Videos were captured on a hand-held camcorder, pointed to approximate the head motion. The camcorder was left running for periods of potential interest as the tourist walked around. Output examples are shown in Fig 2, together with a typical junk frame.

The system output was analysed over 21 minutes of video comprising 25 shots up to 3 minutes long. The system picked 268 still images, including 37 panoramas, and keyframes for 60 short video clips. This output was compared against i) a set of 87 still frames manually identified by the tourist as being representative of the event by careful inspection of the raw video; and ii) a *null hypothesis* which selects the same number of frames as the system, evenly spaced in time. This is a stringent, but important comparison for keyframing systems that many authors fail to make.

(a) (b) (c) (d) (e)

**Figure 2.** Example frames: a) *Gaze* frame; b) Keyframe from an *action*; c) *Glance* frame; d) Pan; e) REJECTED junk frame.

**Table 1. Evaluation results**

| | Hand-picked frames selected | Match | Dupe | Acceptable non-match | Junk |
|---|---|---|---|---|---|
| System | 72 (83%) | 27% | 10% | 20% | 43% |
| Null | 56 (64%) | 21% | 5% | 9% | 65% |

The results are summarized in Table 1. The 2nd column shows the recall rate of the handpicked frames, (or very similar frames). The 83% recall rate is encouraging. The lower recall of the null hypothesis is surprisingly good due to a combination of the relatively long duration of *gazes* and the large number of frames generated by the system. The remaining columns give the fraction of all 268 output frames that: match the handpicked frames; duplicate handpicked frames; are acceptable frames that were not handpicked; or are unacceptable (junk).

The system currently generates duplicates if the user turns towards a target of interest multiple times, separated by turns in a different direction. The large number of junk frames is mostly the result of *glance* detection when the target is uninteresting, (looking both ways when crossing a road for example).

Over half the video clips correctly identify a motion sequence that adds atmosphere to the associated still frame. The remainder are caused by: apparent scene motion due to large parallax resulting from walking past nearby obstructions; uninteresting motion of pedestrians in front of, or passing the camera; motion of obstructions not tracked by the camera, eg branches of nearby trees.

All but 5 of the 87 panoramas add an interesting wide-angle perspective that helps to compensate for the poor viewpoint typical of wearable camera images.

## 4. Conclusions and further work.

Always-on video is a difficult media to interpret unaided, however, the results obtained in this work show significant advances, both in the 83% recall rate, and in the automatic selection of media type (still, pan or video).

It is anticipated that future work to reduce duplicates and inappropriate video clips will minimize excess output, accentuating the improvement over the null hypothesis.

Although the source video used is from a sight-seeing tourist scenario, the techniques described here are likely to be applicable to other wearable video situations.

## References


[1] Y. Li, T Zhang, D Tretter. "An Overview of Video Abstraction Techniques", *Hewlett-Packard Labs Technical Report* HPL 2001-191, 2001.

[2] A. Hanjalic. "Shot-Boundary Detection: Unraveled and Resolved?", *IEEE Trans Circuits and Systems for Video Technology*, vol. 12, no. 2, 2002, pp. 90-105.

[3] Y. Ma, L Lu, H.J. Zhang, M Li. "A User Attention Model of Video Summarization", *Proc. ACM Multimedia* 2002 pp.533-42

[4] R. Lienhart, "Dynamic video summarization of Home Video", *Proc. IS&T/SPIE*, vol. 3972, 2000, pp. 378-389.

[5] A. Girgensohn, et. al. "A Semi-automatic Approach to Home Video Editing", *Proc. ACM UIST*, 2000, pp.81-89.

[6] Y. Nakamura, et. al. "Structuring Personal Activity Records based on Attention - Analysing Videos from Head-mounted Camera", *IEEE Proc ICPR,* vol. 4, 2000, pp. 222-224.

[7] C. Harris and M. Stephens. "A combined corner and edge detector." In *Alvey Vision Conference*, 1988.

[8] M. Fischler and R. Bolles. "Random sample consensus" *Comm. ACM*, vol. 24, 1981, pp. 381-395.

[9] P. Salembier, F. Marques. "Region-based representations of image and video: Segmentation tools for multimedia services", *IEEE Trans Circuits and Systems for Video Technology*, vol. 9. no. 8 1999, pp. 1147-1167.

[10] T. Meier, K. N. Ngan. "Automatic Segmentation of Moving Objects for Video Object Plane Generation", *IEEE Trans Circuits and Systems for Video Technology*, vol. 8. no. 5 1998, pp. 525-538.