An Interactive and Multi-level Framework for Summarising User Generated Videos

Saman H. Cooray
CLARITY: Centre for Sensor Web Technologies
Dublin City University, Ireland
coorays@eeng.dcu.ie

Hervé Bredin
Institut de Recherche en Informatique de Toulouse
Toulouse, France
herve.bredin@irit.fr

Li-Qun Xu
BT Research, British Telecommunications Plc
Adastral Park, Ipswich, UK
li-qun.xu@bt.com

Noel E. O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University, Ireland
oconnorn@eeng.dcu.ie

ABSTRACT

We present an interactive and multi-level abstraction framework for user-generated video (UGV) summarisation, allowing a user the flexibility to select a summarisation criterion out of a number of methods provided by the system. First, a given raw video is segmented into shots, and each shot is further decomposed into sub-shots in line with the change in dominant camera motion. Secondly, principal component analysis (PCA) is applied to the colour representation of the collection of sub-shots, and a content map is created using the first few components. Each sub-shot is represented with a "footprint" on the content map, which reveals its content significance (coverage) and the most dynamic segment. The final stage of abstraction is devised in a user-assisted manner whereby a user is able to specify a desired summary length, with options to interactively perform abstraction at different granularity of visual comprehension. The results obtained show the potential benefit in significantly alleviating the burden of laborious user intervention associated with conventional video editing/browsing.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*abstracting methods*

General Terms

Algorithms, Experimentation, Design

1. INTRODUCTION

The emergence and proliferation of video capture devices, such as digital cameras, camcoders and smartphones, have resulted in many of us becoming non-professional content producers capturing large volumes of video to record special occasions in our lives. Despite this, many of us do not actually later browse or search our collections due mainly to the fact that the video is not structured or indexed. Although tools exist to assist in indexing UGV (e.g. iMovie and similar), this is typically a laborious process that is not particularly user-friendly. Furthermore, users tend to not bother to watch long video clips as most contain substantial amounts of irrelevant or less interesting content. Consequently, automatic summarisation is an important functionality for managing home video archives. This process must, however, ensure that users not only benefit from the decreased browsing time but are also able to retain all the content of the source video that is particularly interesting or valuable to them.

Conventional video editing/browsing applications require numerous repetitions of user interactions until a satisfactory version of the video is produced, which is obviously a time consuming process. Thus, new approaches targeting UGV have been presented in the literature with varying degrees of success. Lienhart [1] proposed an automatic home-video abstraction technique using shot clustering based on time/date information, followed by shot-shortening based on audible noises. In [2], a combined approach using audio and video features was presented for home video abstraction. The authors of [3] presented an integrated approach, comprising fast-pan elimination and face-shot detection techniques. In the home-video summarisation system proposed by Kender and Yeo [4], a zoom-and-hold filter based approach was employed to describe the structural backbone of home videos. More recently, Mei *et al.* [5] proposed a novel approach for home video summarisation by exploiting the intention of the user at the time of capture. A detailed review of generic video abstraction techniques can be found in [6].

In this paper, we propose an interactive and multi-level framework for video summarisation to meet the challenges of managing exponentially growing home video archives. The paper is organised as follows: Section 2 presents the proposed framework building upon the approach proposed by Bredin *et al.* [7] for selecting the more informative and less redundant clips in rushes (a.k.a. raw or unedited videos). We describe our camera motion estimation technique, which is used for sub-shot detection, in Section 2.1. A description of computing sub-shot footprints for content representation is then given in Section 2.2. The summarisation criteria studied in this framework are described in Section 2.3. Section 3 is devoted to a description of the system functionality, describing how the user can quickly and easily create multi-

MM'09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10 ...$10.00.

ple short synopses of the raw video in order to select the one of most interest. The paper concludes in Section 4, pointing out the main contributions and future research directions.

2. PROPOSED FRAMEWORK

The proposed framework for UGV summarisation comprises several content analysis modules, namely shot-cut detection, keyframe extraction, camera motion estimation, and dimensionality reduction. Specifically, we take into account one notable property of home videos, which makes them very different from other forms of content. Videos taken by home users generally comprise one long single shot that, in many instances, depicts different types of camera motions. Thus, in order to make our analysis useful, we employ a sub-shot segmentation module based on camera motion estimation.

While shots are defined as a contiguous set of frames recorded by a single camera without a break, sub-shots are a temporal segment of video separated by specific types of camera motions, enabling a temporal structuring of the raw footage. Sub-shots detected from a given video can still show huge diversity, with the possibility of a sub-shot's duration varying from a few seconds to several minutes. Therefore, such long sub-shots must be shortened to the most interesting parts, which we term "segments" in this paper.

Addressing the inherent difficulties in assuring the expected visual comprehension in fully automatically created summaries, we provide user interaction functionalities to allow a user to select different summarisation criteria and, in turn, to make video summarisation an effective and less tedious task. Key enabling techniques of the framework are the camera motion estimation and content representation by sub-shot footprints, which are described in Section 2.1 and 2.2 respectively.

2.1 Camera Motion Estimation

Our camera motion estimation technique is based on the analysis of the type of dominant-motion transformation matrix between two consecutive frames in the video, which is pan, tilt and zoom in this study.

For this purpose, the SIFT feature extraction/matching algorithm [8] is used. SIFT features are extracted for every frame in the video. For each pair of consecutive frames, feature matching is performed, which subsequently allows us to find the best affine transformation M (3×3 matrix). Denoting by (x, y) and (u, v) the coordinates of corresponding features in two consecutive frames, we get:

$$\begin{aligned} \begin{pmatrix} x' \\ y' \\ w' \end{pmatrix} &= M \times \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} && (1) \\ (u, v) &= \left(\frac{x'}{w'}, \frac{y'}{w'} \right) && (2) \end{aligned}$$

Using matrix M, it is possible to classify, for each frame, the corresponding camera motion into categories: pan left/right, tilt up/down and zoom in/out. We apply the following criteria to decide on the nature of the camera motion:

$$M_p = \begin{pmatrix} 1 & 0 & t_x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \; M_t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{pmatrix} \; M_z = \begin{pmatrix} r_z & 0 & 0 \\ 0 & r_z & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

More specifically, we say that camera is panning if $M \simeq M_p$, tilting if $M \simeq M_t$, and zooming if $M \simeq M_z$.

For UGV summarisation, we make the assumption that there are specific types of camera motions that are useful to be retained in the summary while the other types of camera motions are not meaningful to the user. For instance, a relatively high-speed camera pan can be considered as a useful operation to capture an interesting activity taking place in the video. Given the nature of home videos, we are particularly interested in retaining video clips containing slow and medium levels of camera tilting, and clips corresponding to camera zooming are not considered to be useful for our analysis. Based on the observation of matrix M, we have empirically identified suitable threshold values for panning, tilting and zooming parameters t_x, t_y and r_z accordingly. Thus, based on the intrinsic camera motion, a set of *relevant* sub-shots is built while the rest of the video is discarded.

2.2 Computing Sub-shot Footprints for Content Representation

We apply the concept of footprints to each sub-shot as opposed to the entire shot discussed in [7]. Each frame f_t is processed individually and described by a 3-dimensional RGB colour histogram, with 8 bins per channel, thus leading to a $D = 512$-dimensional feature vector $\vec{x_t}$.

PCA is then applied to the whole set of descriptors, $\vec{x_t}$, extracted from the full collection of sub-shots. This allows us to obtain descriptors that are unique and specific to given content:

$$\vec{x_t} = \vec{\mu} + \sum_{k=1}^{k=D} \alpha_{kt} \cdot \vec{\lambda_k} \tag{3}$$

where $\left(\vec{\lambda_k}, \epsilon_k\right)$ are eigenvector/eigenvalue pairs sorted in decreasing order of eigenvalues ($1 \leq k \leq D$).

Subsequently, only the first few principal components are kept, which explain at least 90% of the data variance, to approximate $\vec{x_t}$:

$$\vec{x_t} \approx \vec{\mu} + \sum_{k=1}^{k=\mathbf{K}} \alpha_{kt} \cdot \vec{\lambda_k} \tag{4}$$

where $K = \underset{1 \leq K \leq D}{\operatorname{argmin}} \left\{ \sum_{k=1}^{K} \epsilon_k \geq 90\% \sum_{k=1}^{D} \epsilon_k \right\}$.

In essence, each frame f_t is described by a K-dimensional vector $\vec{\alpha_t} = \{\alpha_{1t} \ldots \alpha_{Kt}\}$. An additional step of quantisation is performed, leading to a discrete representation in the form of $\vec{\beta_t} = \{\beta_{1t} \ldots \beta_{Kt}\}$ where N is the number of bins per dimension and

$$\beta_{kt} = i \text{ iff } \left(\frac{\alpha_{kt} - m_k}{M_k - m_k} \right) \in \left[\frac{i-1}{N}, \frac{i}{N} \right] \text{ with } i \in [\![1, N]\!]$$

and $m_k = \min_t \{\alpha_{kt}\}$, $M_k = \max_t \{\alpha_{kt}\}$.

Given a sub-shot s, its footprint is defined as:
$\mathbb{FP}_s : [\![1, N]\!]^K \mapsto \{0, 1\}$ with

$$\mathbb{FP}_s\left(i_1, \ldots, i_K\right) = \begin{cases} 1 \text{ if } \exists f_t \in s / \forall k \in [\![1, K]\!], \beta_{kt} = i_k \\ 0 \text{ otherwise} \end{cases}$$

Figure 1 graphically illustrates this concept with $K = 2$ (for ease of illustration) and $N = 8$. In practice, however, K is dependent on the chosen threshold on data variance and the user-generated video itself.

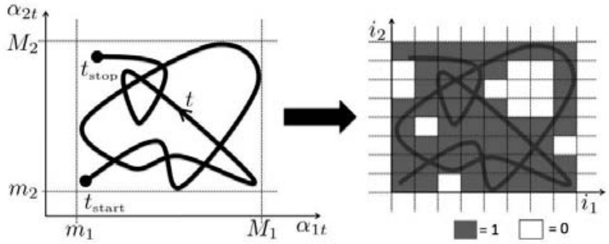

Figure 1: Sub-shot footprint with $K = 2$ and $N = 8$.

2.3 Summarisation Criteria

The proposed summarisation system offers the possibility to create summaries at different granularities of visual comprehension. The summarisation criteria studied are derived from different combinations of the levels in our multi-level approach. Shot and sub-shot detection techniques collectively refer to the first stage of our summarisation system where we remove irrelevant parts of the video through camera motion estimation. In the second stage, as in [7], we use the concepts of coverage, union and relative intersection of footprints.

The **coverage** of a footprint $|\mathbb{FP}|$ is defined as its L1 norm:

$$|\mathbb{FP}| = \sum_{i_1,\ldots,i_K} \mathbb{FP}\left(i_1, \ldots, i_K\right) \tag{5}$$

where K is the number of principle components used in the footprint analysis.

In order to measure how similar two sub-shots are, we also define the **intersection** $\mathbb{FP}_1 \cap \mathbb{FP}_2$ (and **union** $\mathbb{FP}_1 \cup \mathbb{FP}_2$) of two footprints as follows:

- $\{\mathbb{FP}_1 \cap \mathbb{FP}_2\}(\bullet) = \min\left(\mathbb{FP}_1(\bullet), \mathbb{FP}_2(\bullet)\right)$
- $\{\mathbb{FP}_1 \cup \mathbb{FP}_2\}(\bullet) = \max\left(\mathbb{FP}_1(\bullet), \mathbb{FP}_2(\bullet)\right)$

The higher the value of $\dfrac{|\mathbb{FP}_1 \cap \mathbb{FP}_2|}{|\mathbb{FP}_1 \cup \mathbb{FP}_2|}$, the more similar the sub-shots are.

Clearly, a user should have the flexibility to create variable length summaries. The third stage of our summarisation system refers to the level of compaction that allows the user to create a final summary to a pre-defined length. We use a similar sliding window method as in [7] to select the most dynamic segment from each of the chosen sub-shots. In this task, we are looking for a segment $(T \rightarrow T + \delta)$ that is the **most representative segment** of duration δ of a sub-shot. It can be seen as solving the following equation:

$$T = \max_{t} |\mathbb{FP}_{t \rightarrow t+\delta}| \tag{6}$$

2.3.1 *Abstraction Criteria*

An abstraction criterion determines which of the sub-shots are chosen to be included in the summary. The following three methods are investigated in this paper:

- Footprint coverage: The user specifies a threshold value for footprint coverage and accordingly includes a subset of the full set of sub-shots in the summary. Since the sub-shots are pre-sorted based on the principles of maximum coverage and minimum intersection [7], the selected sub-shots can be considered as the most interesting ones in the video. Appropriate segments from each sub-shot need to be selected so that their aggregate satisfies the total length of the summary.

Table 1: Summarisation criteria studied.

Summarisation Criterion	Abstraction Criterion / Segment Selection
A	Footprint Coverage / Length
B	Prominent Subshots / X
C	All Subshots / Length
D	Footprint Coverage / Footprint Coverage
E	All Subshots / Footprint Coverage

- Prominent sub-shots: This refers to a subset of sub-shots whose aggregate satisfies the requested length of the summary. Since the sub-shots are pre-sorted according to their relative importance, the selected sub-shots correspond to the most interesting sub-shots in the video.
- All sub-shots: The user requests that segments from all the sub-shots should be included in the summary. Appropriate lengths for segments in each of the sub-shot need to be computed for this abstraction criterion.

2.3.2 *Segment Selection*

Having determined the desired set of sub-shots to be included in the summary, the final summary should be constructed using the most informative segments from each of these sub-shots. For example, relatively long static scenes can be generally viewed in a short span of time. However, given a number of sub-shots depicting different types of scenes with varying lengths, the proportion of each segment that should be presented in the final summary can be quite subjective. In this context, we study two different methods, namely *Length of Sub-shot* and *Footprint Coverage of Sub-shot*, for computing the length of each segment for a given total length of the summary.

- Length of sub-shot: In this method, the user chooses to extract segments for which the length is proportional to the total length of its respective sub-shot. This is based on the assumption that the user is not concerned about prioritising the importance of segment selection based on the type of activity in the respective sub-shot.
- Footprint coverage of sub-shot: This method allows the user to select the length of segment as proportional to the coverage value of its respective sub-shot. This is based on the assumption that a measure of the variability of a given sub-shot can be regarded as how well it *covers* the various scenes of the whole video.

Table 1 shows a list of different summarisation criteria, i.e. A, B, C, D, and E, that are studied based on the combination of abstraction criterion and segment selection method described above. It should be noted that, "X" corresponding to segment selection method in summarisation criterion B indicates a property of mutual exclusiveness.

3. SYSTEM FUNCTIONALITY

An example screenshot of the GUI for the system is shown in Figure 2. It allows the user to load a raw video and create video summaries of different lengths and visual comprehension. The user can also playback the raw video (top left pane), summarised video (top middle pane) or any selected segment in the summary (top middle pane). Additionally, the GUI includes a storyboard (top right panel) so that the user will be able to preview the loaded video in terms of a collection of keyframes.

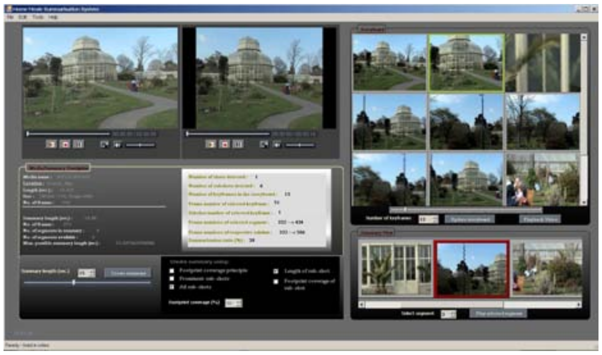

Figure 2: A GUI for home video summarisation.

The proposed summarisation system is designed to be executed using the button, “Create summary”, provided in the GUI. The length of the summary can be easily set using the slider-bar provided in the bottom left panel. The summarisation criterion described in Section 2.3 can be selected from the options provided in the bottom middle panel of the GUI. Once a raw video is processed, the system stores all the metadata related to that video for promptly creating summaries during subsequent calls. Details of the raw video and summary are shown to the user in the left middle panel of the GUI. The user also has the option to playback any segment in the summary provided in the right bottom panel, where each segment is represented with a thumbnail image corresponding to the middle frame of that segment. The storyboard can be easily re-created having initially processed the raw video. The user can specify the number of keyframes and press the button “Update storyboard”. Additionally, we also provide a functionality so that clicking on any keyframe in the storyboard allows the user to playback the raw video at its sub-shot’s occurrence.

Table 2 shows some statistics of the summaries created for 4 different sample videos. The first column indicates that all 4 videos are single shot but lead to varying numbers of sub-shots. The second column gives the length of the raw video and summary measured in seconds. The rest of the description corresponds to the details of the summaries created using the 5 different summarisation criteria in terms of the start and ending frame numbers of the segments in the summary. The number of rows in each result indicates the resulting number of segments in the summary. Observing these results, it is clear that different summarisation criteria result in different visual expressions in the summary. Some sample summarisation results given at: `http://elm.eeng.dcu.ie/~coorays/UGV-summarisation.html` show the level of quality that can be achieved from the proposed summarisation criteria. A special scenario can be, however, observed in the last video having one single sub-shot, where 4 of the 5 summarisation criteria lead to the same level of visual comprehension. This is due to the fact that the requested length of the summary is shared within only one sub-shot and the most dynamic segment is identified at the same position in the sub-shot, illustrating that it makes no difference what the summarisation criterion is used for single sub-shot videos. Nevertheless, the occurrence of such type of content is very unlikely in practice.

4. CONCLUSION

Table 2: Details of example summaries created using different summarisation criteria.

Name of raw video (no. of shots /subshots)	Raw video / summary length (sec.)	Details of the summaries created using different criteria (Start -- Ending frame number of segment)				
		A	B	C	D	E
M2U01859.MPG (1 / 5)	38.8 / 9	295 - 388 711 - 838 930 - 960	295 - 518	23 - 31 73 -132 295 - 363 708 - 800 938 - 960	295 - 438 712 - 800 942 - 960	22 - 31 59 - 135 295 - 389 710 - 768 945 - 957
CLIP_2008_07_26_21_19_57.MPG (1 / 3)	31.8 / 8	219 - 358 460 - 521	61 - 261	268 - 361 480 - 521 577 - 643	217 - 358 462 - 521	247 - 361 474 - 521 577 - 616
M2U01932.MPG (1 / 3)	46.1 / 11	132 - 213 560 - 754	506 - 781	132 - 201 318 - 358 576 - 741	142 - 201 718 - 934	163 - 213 316 - 358 559 - 742
M2U01939.MPG (1 / 1)	32.2 / 8	378 - 578	14 - 214	378 - 578	378 - 578	378 - 578

In this paper, we have presented an interactive and multi-level framework for UGV summarisation, addressing the inherent difficulties in creating good summaries. We provide a number of options to create summaries interactively at different granularities of visual comprehension. The proposed framework shows the potential of a significant reduction in the burden of laborious user intervention associated with conventional video browsing and editing operations. However, one drawback of our system is the time-consuming sub-shot detection, which strongly relies on SIFT feature extraction. We will investigate other efficient and effective features for this task in the future, in addition to providing various other user interactive features to the user interface. Using audio information and combining with image analysis are another interesting research area. Finally, we will perform a comprehensive user evaluation to examine which of the proposed criteria are more interesting to the user in summarising UGV content in general.

5. ACKNOWLEDGMENTS

This work is supported by *BT Group plc* and the *Irish Research Council for Science, Engineering and Technology*: Funded by the *National Development Plan*, and *Science Foundation Ireland* under grant 07/CE/I1147.

6. REFERENCES

[1] R. Lienhart. Abstracting Home Video Automatically. ACM Multimedia, pages 37–40,Orlando, FL, USA, 1999.

[2] M. Zhao, J. Bu, and C. Chen. Audio and video combined for home video abstraction. *IEEE Intl. Conf. on Acoustic, Speech and Signal Processing (ICASSP)*, (5):620–623, 2003.

[3] S.-H. Huang, Q.-J. Wu, *et al.* Intelligent home video management system. Intl. Conf. on Information Technology: Research and Education, pages 176-180, June 2005.

[4] J. R. Kender and B.-L. Yeo. On the Structure and Analysis of Home Videos. *In Proc. of ACCV'2000*, 2000.

[5] T. Mei, X.-S. Hua, H.-Q. Zhou, and S. Li. Modeling and Mining of Users' Capture Intention for Home Videos. *IEEE Transactions on Multimedia*, 9:66-77, 2007.

[6] B. T. Truong and S. Venkatesh. Video Abstraction: A System Review and Classification. *ACM TOMCCAP*, 2007.

[7] H. Bredin, D. Byrne, H. Lee, N. O'Connor and G. Jones. Dublin City University at the TRECVid 2008 BBC Rushes Summarisation Task. TRECVID BBC Rushes Summarization Workshop, ACM Multimedia, Canada, 2008.

[8] R. Hess and A. Fern. Improved Video Registration using Non-Distinctive Local Image Features. *In Proc. of IEEE CVPR'2003*, pages 1-8, June 2007.