Video Summarisation: A Conceptual Framework and Survey

ARTHUR G. MONEY and HARRY AGIUS
Brunel University, School of Information Systems, Computing and Mathematics, UK

Corresponding author: Harry Agius
Brunel University
School of Information Systems, Computing and Mathematics
St John's
Uxbridge
Middlesex
UB8 3PH
UK

E-mail: harryagius@acm.org
Telephone: +44 1895 265993
Fax: +44 1895 251686

Video Summarisation: A Conceptual Framework and Survey

Abstract:

Video summaries provide condensed and succinct representations of the content of a video stream through a combination of still images, video segments, graphical representations and textual descriptors. This paper presents a conceptual framework for video summarisation derived from the research literature and used as a means for surveying the research literature. The framework distinguishes between video summarisation techniques (the methods used to process content from a source video stream to achieve a summarisation of that stream) and video summaries (outputs of video summarisation techniques). Video summarisation techniques are considered within three broad categories: internal (analyse information sourced directly from the video stream), external (analyse information not sourced directly from the video stream) and hybrid (analyse a combination of internal and external information). Video summaries are considered as a function of the type of content they are derived from (object, event, perception or feature based) and the functionality offered to the user for their consumption (interactive or static, personalised or generic). It is argued that video summarisation would benefit from greater incorporation of external information, particularly user based information that is unobtrusively sourced, in order to overcome longstanding challenges such as the semantic gap and providing video summaries that have greater relevance to individual users.

Keywords: video summaries; video summarisation; video content; survey; conceptual framework; user based information; contextual information

1 Introduction

With the availability of digital video growing at an exponential rate, users are increasingly requiring assistance in accessing digital video [1]. Research into *video summarisation* helps to meet these needs by developing condensed versions of a full length video stream through the identification of the most important and pertinent content within the stream. The consequent video summaries may then be integrated into various applications, such as interactive browsing and searching systems, thereby offering the user an indispensable means of managing and effectively accessing digital video content [2, 3].

Video summarisation techniques produce summaries by analysing the underlying content of a source video stream, condensing this content into abbreviated descriptive forms that represent surrogates of the original content embedded within the video [4]. The multimodal nature of video, which conveys a wide range of semantics in multiple modes, such as sound, music, still images, moving image, and text [5], makes this task much more complex than analysing text documents. Furthermore, video summarisation research faces the challenge of developing effective techniques for abstracting useful and intuitive semantics from the video stream that are in step with the individual users' comprehension and understanding of video content [3, 6].

Video summaries incorporate several *audiovisual cues* for presenting the user with a condensed and succinct representation of the content of a video stream. The style and extent to which the various cues are used within summaries varies greatly since there is no set standard defining what should be included and excluded from a video summary. Four audiovisual cues may be identified as follows (illustrated in Figure 1):

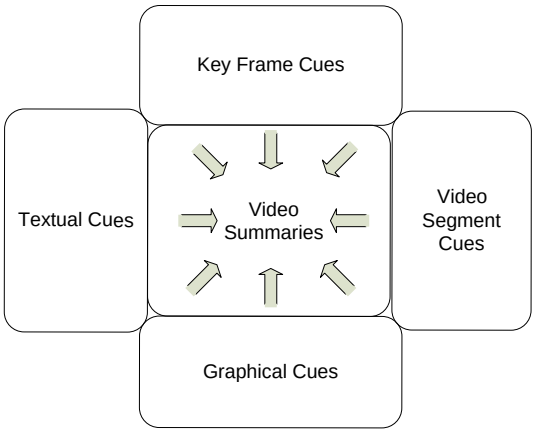

Figure 1: Audiovisual cues used for video summaries

- *Key frame cues* are still images extracted from the video stream and presented in temporal order. For example, Gianluigi and Raimondo [7] use this approach, providing key frame numbers as supplementary information within their video summaries.

- *Video segment cues* are a dynamic extension of key frame cues. Video summaries incorporating video segment cues may be composed of a series of video segments, rather than video key frames, which are often automatically selected and which represent the most important and pertinent segments of the original video stream. These video summaries generally preserve both the motion and audio element of the video and hence may be considered more attractive to the user. However, their disadvantage is that more time is required for the user to comprehend their content [8]. For example, *video skims* are a particular type of video summary that use video segment cues while preserving the temporal order of the original video stream. This is illustrated by Furini and Ghini [1] who remove segments of video that do not contain any audio detail (i.e. silent segments) such that the video summary becomes a condensed version of the original video, maintaining its temporal order, image detail and audio detail. The user subsequently spends less time viewing the video summary than if they had viewed the original video stream, but user comprehension is not as immediate as with key frame video summaries.
- *Graphical cues* present an additional level of detail by using visual cues and syntax as a substitute or supplement to the other cues. Video summaries incorporating graphical cues provide users with a detailed overview of the content of a video summary that would otherwise not be achievable via other methods. For example, the Movie Content Analysis System (MoCA) [9] presents a two-dimensional colour coded block map of the video stream which distinguishes moments of dialogue, explosions, on-screen text, and so on.
- *Textual cues* summarise the content of the video via textual descriptors. Luo et al. [10] present an example of this by automatically detecting the presence of text captions within the video image stream, which are then extracted and a video summary generated.

Through the presentation of a conceptual framework for video summarisation, this paper surveys the various techniques and resultant summaries that have been proposed within the research literature. Section 2 presents the conceptual framework and its components. Video summarisation techniques are split into three sub-types: internal (analyse information sourced directly from the video stream), external (analyse information not sourced directly from the video stream) and hybrid (analyse a combination of internal and external information). The video summaries that each technique produces are also categorised as a function of the type of content they summarise (object, event, perception or feature based) and the functionality offered to the user for their consumption (interactive or static, personalised or generic). Through the

conceptual framework, Sections 3, 4 and 5 survey the internal, external, and hybrid video summarisation techniques and the resulting video summaries presented in the research literature to date, respectively. Section 6 then recommends future video summarisation research directions. In particular, it is argued that video summarisation would benefit from greater incorporation of external information, particularly user based information that is unobtrusively sourced, in order to overcome longstanding challenges, such as the semantic gap and providing video summaries that have greater relevance to individual users. Section 7 concludes the paper.

2 A conceptual framework for video summarisation

In order to identify and extract the various audiovisual cues to be included within video summaries, the underlying content of a video must first be analysed [4]. Unlike textual information, the wide range of semantics that are apparent within a video are conveyed in multiple modes, which include sounds, music, still images, moving images and text [5]. Consequently, accurate and concise abstraction of video semantics still poses a difficult and ongoing research challenge for the video summarisation community. In response to this, a range of video summaries and enabling video summarisation techniques have been proposed. This section presents a conceptual framework for video summarisation derived from the techniques and summaries proposed in the research literature. The conceptual framework is shown in Figure 2 and is used as the basis for examining the research literature in subsequent sections.

2.1 Research method

The conceptual framework was developed based on a survey and analysis of contemporary video summarisation research literature. The literature included in the survey was identified by searching a range of full text databases for articles which proposed new video summarisation techniques and/or video summaries with a view to collecting a large sample of recent work within the field. To ensure the timeliness and manageability of the included literature, the search was initially limited to articles appearing within the last three years; however, a number of additional articles considered key within the field were also considered if published outside of the specified time frame.

The resulting selected literature was then considered as a dataset representative of contemporary video summarisation research. As a means of appropriately surveying and categorising the key video summarisation techniques and video summaries presented in the literature, a thematic analysis was performed on the dataset. Thematic analysis is an accepted qualitative research method for analysing textual datasets [11]. It provides a structured means of

identifying overarching themes (categories) and corresponding sub-themes (sub-categories) that occur within the dataset. In-depth descriptions of the process are available in [11-14]. Thematic analysis facilitates the effective abstraction of salient themes from a complex and detailed textual dataset [15], hence is particular suitable in this context and in line with the overall aims of the conceptual framework.

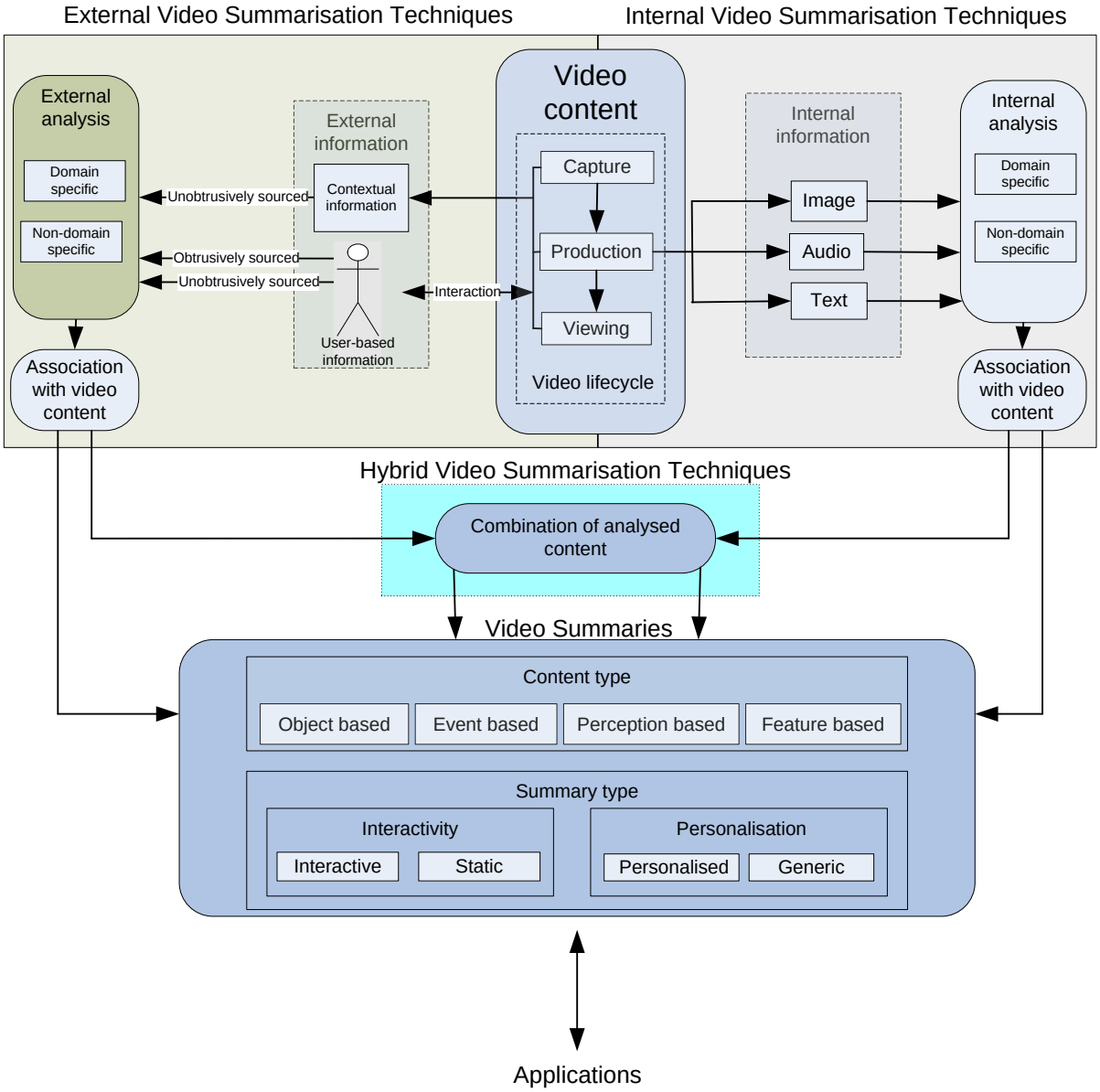

Figure 2: Conceptual framework for video summarisation

In order to carry out the thematic analysis, the entire dataset was initially perused to conceptualise the overarching categories existing within the literature at a high level. These high level concepts were noted in a coding frame, with each concept assigned a code name, a description and examples of text that fitted each concept. The dataset was then perused iteratively, enabling categories and sub-categories to be developed further. With each iteration, categories and sub-categories were spliced and linked together; text relating to each category

and sub-category were appropriately labelled and compared across publications to ensure consistency of categorisation.

When no further refinement of the categorisation could be derived, a group of final categories and sub-categories representative of the literature dataset were produced. These were then mapped visually, forming the basis of the conceptual framework for video summarisation presented in Figure 2. The conceptual framework distinguishes between *external summarisation techniques* which use information that is not sourced directly from the video stream, *internal summarisation techniques* which use internal information sourced directly from the image, audio and text features of the video content arising from the production stage of the video lifecycle, and *hybrid summarisation techniques* which combine the results of internal and external analysis. These are illustrated in the top half of the figure. The resulting *video summaries* produced by internal, external and hybrid summarisation techniques are illustrated in the bottom of the figure. They summarise content based on *content type* and differ in their levels of *interactivity* and *personalisation*. They are used within various *applications*. A more detailed explanation of the various components of the conceptual framework is now provided.

2.2 Techniques for summarising video

Video content progresses through three stages of a video lifecycle: *capture*, where the video is captured or recorded, *production*, where the video is edited and transformed into a format that conveys a story or message appropriately, and *viewing*, where the video is viewed by a target audience. Consequently, video summarisation techniques analyse a range of information sources at various stages of the video lifecycle in order to abstract the semantics relating to the content of a video stream and subsequently extract the various audiovisual cues. The various techniques presented in the literature fall into three categories based on the information sources they analyse:

1. *Internal summarisation techniques*, which analyse internal information from the video stream produced during the production stage of the video lifecycle.
2. *External summarisation techniques*, which analyse external information during any stage of the video lifecycle.
3. *Hybrid summarisation techniques*, which analyse both internal and external information during any stage of the video lifecycle.

Each of these techniques carries out *domain specific* or *non-domain specific* analysis. The former refers to techniques that cater specifically for a domain in which video content is summarised. Common content domains include sports, music, news and home video. Focusing on a particular domain helps to reduce levels of ambiguity when analysing the content of a video stream by applying prior knowledge of the domain during the analysis process. Although there are domain specific examples for all three of the above categories of summarisation technique, internal summarisation techniques most commonly benefit from domain-specific analysis. This is because inferring meaning from low level video stream cues is a notoriously ambiguous process, the ambiguity of which can be reduced by applying prior domain knowledge during analysis. Non-domain specific techniques are the opposite of domain specific techniques, in that they present solutions for summarising video content in any content domain.

By definition, internal summarisation techniques apply to video content produced during the production stage of the video lifecycle. These techniques automatically analyse low-level *image, audio and text* features of the video stream directly, to abstract semantics appropriate for a video summary. Wang et al. [16] provides a detailed review of some of the more common techniques used. Internal summarisation techniques are the most common form of summarisation technique and maintain their focus on the video stream by adhering to three basic assumptions [17], which we apply within the context of our conceptual framework for video summarisation as follows:

- *Analysis is disconnected in space and time from the capture and viewing stages of the video lifecycle*. Analysis must not affect the video capture or video viewing process in any way, and hence no additional information can be deliberately collected at the capture or viewing stages that may assist during the analysis process.
- *All analysis and abstraction must be concentrated and sourced from the video stream alone*, hence the focus is on abstracting semantics from the video content as it is naturally produced in the video lifecycle (from the production stage). In addition, the semantics abstracted must be sourced from the information that exists within the video stream.
- *Analysis must be of an automated nature and avoid relying on user participation to support and enhance the process*. This avoids inconveniencing the user by requiring little or no manual input from them, to ensure that the analysis process does not disrupt the natural conclusion of the video lifecycle. Techniques do not require the user to provide any form of description of the video content.

Image features can include changes in colour, texture, shape and object motion sourced from the image stream of the video. These can be used to segment a video into shots by identifying shot boundaries, such as cuts which are represented by sharp changes in image features or fades which are identified by slower changes in image features. Specific objects can also be identified and analysing image features for video with a known structure can help to improve the depth of summarisation achieved. Sports video is particularly well suited to event detection given its rigid structure. For example, Ekin et al. [18] rely on the fact that the majority of coverage of soccer games generally conforms to long, medium and close-up view shots and carry out shot type classification for soccer video to infer these shots based on the grass coloured pixel ratio. The structure of soccer games relating to the typical shot types and the duration of breaks after a goal is scored are used to detect key events that occur within the game. Shih and Huang [19] detect events within baseball videos by using structural knowledge related to the sequence of events in a baseball game and knowledge about specific objects that appear in the image stream. Specific events such as running, defence, and pitching are identified. The rigid structure of other content domains can also be used to achieve object and event detection, such as in news video, which normally starts with an overview of headlines, run for a set duration, followed by a series of reports which end in a return to the anchor person. Knowledge of this content domain can be mapped onto parameters for analysing image features to aid the abstraction process, resulting in more effective identification of objects within the video (e.g. anchor person) and events (e.g. the start of a news report or the news headlines).

Audio features appear in the audio stream associated with the video stream and include speech, music, sounds and silence. These prove useful for identifying candidate segments to be included in a video summary and knowledge from particular domains can be used to enhance the depth of summarisation achieved. For example, audio features occurring within sports videos prove useful for supporting event detection since they are often more consistent and distinguishable than other content domains. For example, long-whistling, double-whistling, excited commentator speech and excited audience sounds can be used to discern a number of potential events such as the start of a free kick, penalty kick, foul, goal and so forth [20]. Audio features can also be used in other content domains; for example, we may take high energy dialogue in a drama to suggest an argument which, depending on the objective of the video summary, may be considered a candidate for inclusion in the video summary. Specific semantic detail can also be inferred from audio features by using speech recognition techniques. These can be used to convert segments of dialogue into textual representations.

Text features appear in the video in the form of subtitles or text captions, the latter being embedded or 'burnt in' to the image stream of the video rather than being available as a separate stream. Text is a valuable information source for video summarisation as it can contain detailed information about the video content, e.g. captions during a live broadcast boxing match tend to display the names of the fighters and the location and date of the fight. Events can also be detected from text. For example, Zhang and Chang [21] use changes in onscreen game stats to identify important events within baseball games such as 'score' and 'last pitch'.

External summarisation techniques analyse information that is not sourced directly from the video stream. There are two types of external information:

- *User-based information*, which is additional information about the content of a video that is sourced directly from the user.
- *Contextual information*, which is additional information that is not sourced directly from the user or the video stream.

The capture and production stages of the video lifecycle have traditionally been carried out by professionals. However, the proliferation of low cost digital video recording equipment coupled with a growing range of software applications that enable the user to analyse, annotate, edit and view video has meant that home users are increasingly involved with video at every stage of the lifecycle. Consequently, external summarisation techniques analyse information at every stage of the video lifecycle. This challenges the assumptions of internal summarisation techniques focusing only on analysing the video stream.

User-based information typically includes information that can be related to the user's behaviour towards, interactions with and comprehension of the content of a video stream at the capture, production and viewing stages of the video lifecycle. In addition, user-based information can include information relating to the user's preferences, e.g. the types of video content the user may be interested in viewing. User-based information can be *obtrusively sourced*, such that the user has consciously provided information to aid the summarisation process, or *unobtrusively sourced*, such that the information has been obtained without requiring conscious input from the user. *Obtrusively sourced* information is advantageous since, in many cases, it can be as detailed as the user wants or requires it to be. This is often achieved by requiring the user to provide detailed graphical and textual annotations for the video content. However, obtaining information obtrusively can be costly to the user in terms of the time it takes

to create the annotations and the overall human effort involved in accurately describing the content. As a result, manual annotation is often not considered a feasible solution, particularly when internal summarisation techniques are used. An example of obtrusively sourced user-based information is detailed manual annotations provided at the production stage which enable detailed event and object level summarisation for sports video content, such as player names, penalty takers, goals, shots on goal, corner kicks and fouls for soccer videos [22]. Other examples of obtrusively sourced user-based information include the user providing spoken descriptions of video content during capture [23], manual annotations in the form of interactive browsing and annotation via graphical and textual descriptions [24], and user profiles used to specify and filter content that is of most interest to the user [25, 26]. *Unobtrusively sourced* information has the advantage of not requiring conscious input from the user. However, unobtrusive sources are limited in terms of the level of detail that can be inferred from them. Examples include monitoring changes in a user's brainwaves while capturing video content (via wearable sensors) which can be used to later identify pertinent video segments [27] and monitoring the naturally occurring browsing behaviour of a user while viewing video content in order to identify video segments of interest [28].

Contextual information is always unobtrusively sourced because it is not sourced directly from the user. Hence, the benefit is that it does not require conscious user input. Some examples of contextual information include recording the geographical location in which a video has been captured via GPS (global positioning system) sensors attached to the video camera [29] and sourcing information from the environment in which the video is captured by integrating pressure based floor sensors into a purpose built home to monitor movement activity [6]. The Internet is another contextual information source which can be used to achieve event detection in sports video. For example, soccer websites hold detailed information about a particular soccer game which can be associated with the video stream of that game and used to summarise the key events [30] and webcast text can be used to achieve real time event detection of live soccer coverage, with events such as goal, shot and save identifiable from analysing the text [31]. Similar approaches to those above have been used for summarising news programmes and music videos [32].

Hybrid summarisation techniques analyse both internal and external information, and hence can apply at any stage of the video lifecycle. They can be made up of any combination of internal and external summarisation techniques and aim to maximise the effectiveness of video summaries by capitalising on the strengths of each of these approaches while minimising their weaknesses. Hybrid approaches have proven useful for domain-specific techniques, such as

detailed event level summarisation in sports video. For example, the rigid structure of tennis video can be used to segment the content based on shot change detection, dominant image motion analysis and audio analysis to identify ‘applause’ and ‘ball hits’, and then the segmented video manually browsed by the user to select example events such as ‘serves’ and ‘rallies’ that act as a training set for automatically finding similar segments in the rest of the video [33]. In another example, specific events such as ‘red card’, ‘yellow card’, ‘corner kick’ and ‘on pitch fight’ are automatically identified from an analysis of image features in soccer video and a specialised human operator applies priority functions to each event which are reflected in the content of the video summary [34]. Hybrid approaches have also proven useful for non-domain specific techniques, initially reducing the amount of manual effort required to summarise generic video content by employing internal analysis techniques to automatically segment video content into shots. Automatically abstracted shots may then be presented to the user for browsing and this browsing activity monitored so as to identify candidate shots to be included in the video summary [35].

2.3 Video summaries produced by video summarisation techniques

The resulting video summaries produced by internal, external and hybrid summarisation techniques are illustrated in the bottom of Figure 2. They may be distinguished by the content types included, which may be object based, event based, perception based and/or feature based, depending on the type of analysis and the criteria used to determine the content to be included in the video summary. Video summaries are also distinguished by their interactivity and personalisation.

Object based summaries focus on specific objects that occur within the video, such as a chair, table, key characters and players and so forth. Video summaries are considered as object based if objects are identified during analysis of the video and are subsequently used as a basis for formulation of the video summary. Object based summaries may optionally include text or graphics to explicitly represent the objects within the video summary or to assist the process of selecting key frames or video segments best summarising the video content based on the detection of specific objects.

Event based summaries are based on specific events that occur within the original video stream. For sports videos, the events may be goals, free-kicks match points, and so forth, whereas for a popular movie, a fight or a love scene may be a more common event. Video summaries are considered to be event based if events are identified during analysis and subsequently used for creating the video summary. Optionally, text or graphics may also be used

to explicitly represent events within the video summary or to assist in the selection of suitable key frames or video segments based on the detection of specific events indicated within the text or graphics.

Perception based summaries are derived from high-level concepts representing the ways in which the user perceives, or may perceive, the content of a video. Some examples include the degree of excitement the user may experience while viewing the content, the level of importance the user may associate with the content, the intensity or types of emotion the user may perceive from the content, or the perceived amount of attention video segments may demand from the user. High level abstraction is often achieved by applying theories from other research domains as a means of interpreting the semantics of the video content, such as human perception theory, semiotics, and user attention theories. Perception based summaries therefore focus on inferring the way in which the video content is or may be perceived by the user, rather than identifying tangible objects and events that are conveyed within the video content, as is the case for object and event based video summaries.

Feature based summaries are not based on specific object, event or perception based semantics, but rather are objectively generated from an analysis of low level features, such as changes in colour, texture and motion in the image steam or segments of silence or speech in the audio stream. Although object, event and perception based summaries may also analyse low-level features, the key differentiator of feature based summaries is that no attempt is made to infer object, event or perception based semantics from these low level features. That is, they are used in their 'raw' state. Commonly, feature based summaries are not accompanied by any textual descriptors of the summarised content, since it is not possible to specifically associate explicit semantic descriptors with the summarised content. As a result, summaries are typically presented as a series of key frame images or video segments that are not accompanied by any description to clarify the meaning to the user or to clarify why the segments have been selected to represent the summarised video content. Due to the lack of semantic inference achieved, feature based summaries produce the least semantically meaningful video summaries.

As well as content type, video summaries may also be distinguished by whether they are interactive or static and whether they are generic or personalised. *Interactive summaries* typically offer the user various means to interact with the summary, such as navigating the summary or retrieving summaries according to specified semantic criteria expressed through queries. Queries may be in terms of key word searching, in terms of drop down lists of textual descriptions of objects and events the video is summarised by, or in terms of the user providing example images or video segments that can be used to retrieve matching key frames or video

segments. As will be seen in Sections 3 to 5, while the majority of video summarisation literature places great emphasis on developing analysis techniques, considerably less explores the ways in which summarised content can be presented to the user. Consequently, the majority of summaries are *static*, e.g. presented as a series of static key frames or video segments or as a series of time stamps that represent the pertinent segments of a video. However, some studies [e.g. 19, 36, 37] do focus a degree of attention on developing interactive summaries.

It is becoming increasingly recognised that video summaries should be personalised to reflect the individual users' comprehensions of video content in a way that is intuitive and in step with individual needs [3, 6]. Consequently, with *personalised video summaries*, the video summary is tailored to the individual user's comprehension or understanding of the video content. In contrast, *generic video summaries* are not tailored in any way to the individual user's comprehension or understanding. Summaries can be personalised in a number of ways; for example, filtering video content before it is presented to the user via user profiles [38], summarising content based on a user's behaviour and movements while capturing the video [29], and summarising video based on the user's individual browsing and access patterns [35]. Since a personalised summary must incorporate the user in some way during analysis, prior to the summary being presented for consumption and query by the user, summaries that allow a user to query video content do not automatically qualify as personalised summaries. For this reason, despite a number of video summaries, e.g. [39-41], including a function that allows the user to specify a required duration (time) of the video summary, the resulting video summaries are regarded as generic rather than personalised, unless the criteria used for inclusion or exclusion of content is derived directly from the user's understanding or comprehension of the video content. Many techniques that offer the time constraint as a user defined parameter include and exclude content via generic, pre-determined criteria that do not reflect an individual user's preferences.

2.4 Applications of video summarisation

Traditional methods of video access are uninterrupted and serial: the user must playback the video stream inclusively in order to experience its content. This is costly in terms of the time it takes for users to do this and the resources required, particularly when considering the ever growing numbers of users who access content via resource-limited mobile devices. Video summarisation enables content to be presented alternatively to traditional methods, offering opportunities to reduce the time and resources required for the user to experience the content. Video summarisation therefore has numerous applications, each seeking to exploit these two

benefits to differing extents and to meet varying aims. The use of video summarisation by applications is indicated at the base of Figure 2. The two major applications of video summarisation are video adaptation and video navigation.

Video adaptation seeks to transform one or more input videos into either a video or an augmented multimedia form via manipulations at signal, structural and/or semantic levels, to meet diverse resource constraints and user preferences while optimising the overall utility of the video [42]. Video summarisation therefore serves to support video adaptation processes by providing a transformation of a given source video that is less resource intensive and less time consuming by only retaining the most important elements of the original content (which may be presented using the original or alternative cues). For example, Tseng et al. [38] demonstrate a system that adapts video content according to user's personal preferences. Video content is passed through an adaptation engine and subsequently made available via a web portal to a user's individual mobile device. Similarly, Lin and Tseng [25] demonstrate how video content can be filtered according to user specific characteristics (such as gender) and the environment in which the content is being accessed.

Video navigation seeks to improve and enhance the means by which users browse a video stream. The conventional approach, proliferated by media players such as Windows Media Player and QuickTime, is based on serial browsing using the VCR metaphor (such as fast forward and rewind) [43]. The application of video summarisation to video navigation provides a means by which video content can be browsed visually in a non-linear fashion, enabling the user to 'move through' the video more efficiently and effectively. For example, the MyInfo system [26] offers access to news content via a browser designed to be presented on consumer televisions, whereas Otsuka et al. [37] present a user interface that allows enhanced access to sports video via the use of a handheld remote control.

In the following sections, the conceptual framework for video summarisation that was presented in this section is used as a means of presenting the surveyed research literature. Section 3 reviews internal techniques and summaries, Section 4 reviews external techniques and summaries, and Section 5 reviews hybrid techniques and summaries. Table 1 is a list of abbreviations used from this point onward.

3 Internal summarisation techniques and summaries

Internal summarisation techniques are by far the most common form of video summarisation technique presented in the literature. Table 2 presents a summary of key internal summarisation techniques and the video summaries they produce, which form the focus of this section.

Table 1: List of abbreviations.

Abbreviation	*Meaning*
OS	Obtrusively Sourced
US	Unobtrusively Sourced
C	Capture stage
P	Production stage
V	Viewing stage
I	Interactive
S	Static
E	pErsonalised
G	Generic

3.1 Techniques for analysing internal information sources

Image features are the sole source of analysis for [7, 36, 41, 44-49] and are all *non-domain specific*, seeking to identify representative key frames for inclusion in the video summary based on techniques such as assessing distortion [48] or difference levels [7, 44] between frames, and determining differences between shot fade-ins and fade-outs from object and camera motion [47].

The internal video summarisation techniques presented in [18, 19, 50-53] also solely analyse *image features* from the video stream, but are *domain specific*. Since certain types of domain content conform to explicit structures, all of the techniques exploit this to identify specific objects or events for inclusion in the video summary. Four of these six techniques are specific to the sports content domain: Benjamas et al. [50] analyse boxing videos, Bezerra and Lima [53] and Ekin et al. [18] analyse soccer videos, and Shih and Huang [19] analyse baseball videos. For example, Benjamas et al. [50] use structural knowledge about fighting sports videos to identify segments that contain shots of the fighters for inclusion in the video summary, specifically identifying fighters by skin colour, shots where fighters are in close proximity to each other and the presence of flash lights. The remaining techniques work within the television drama [51] and home video domains [52]: Jung et al. [51] automatically abstract shots and

Table 2: Internal video summarisation techniques and summaries

	Internal video summarisation techniques				Internally-produced video summaries					
	Internal information			Internal analysis	Content type				Summary type	
Technique/summary	Image	Audio	Text	Domain specific	Feature-based	Object based	Event based	Perception based	Interactivity	Personalisation
Benjamas et al. [50]	X			X		X			S	G
Bezerra and Lima [53]	X			X			X		S	G
Cai et al. [54]		X		X			X		S	G
Cernekova et al. [47]	X				X				S	G
Cheng and Xu [44]	X				X				S	G
Ciocca and Schettini [49]	X				X				S	G
Ekin et al. [18]	X			X		X	X		S	G
Ferman and Tekalp [41]	X				X				S	G
Furini and Ghini [1]	X	X			X				S	G
Gianluigi and Raimondo [7]	X				X				S	G
Girgensohn [45]	X				X				S	G
Hanjalic [55, 56]	X	X		X				X	S	G
Hanjalic and Xu [57]	X	X		X				X	S	G
Haubold and Kender [58]	X	X	X	X		X	X		I	G
Jung et al. [51]	X			X				X	S	G
Kim et al. [59]		X	X	X		X	X		I	G
Kopf et al. [39]	X	X		X	X				S	G
Li et al. [48]	X				X				S	G
Liang et al. [60]	X		X	X			X		S	G
Lienhart et al. [9]	X	X		X		X	X		I	G
Lu et al. [61]	X	X		X		X	X		S	G
Luo et al. [10]			X			X	X		S	G
Ma et al. [62]	X	X						X	S	G
Mei et al. [52]	X			X	X				S	G
Ngo et al. [46]	X				X			X	S	G
Otsuka et al. [37]		X		X				X	I	G
Rui et al. [63]		X		X			X	X	I	G
Shao et al. [64]	X	X		X			X		S	G
Shih and Huang [19]	X			X		X	X		I	G
Smith and Kanade [65]	X	X	X			X	X		S	G
Sugano et al. [40]	X	X		X			X		S	G
Sundaram et al. [66]	X	X		X	X			X	S	G
Tjondronegoro et al. [67]	X	X	X	X			X		S	G
Xu et al. [20]		X		X			X		S	G
Wu et al. [68]			X			X	X		S	G
Zhang and Chang [21]	X		X	X			X		I	G
Zhu and Wu [36]	X				X				I	G

scenes and then apply a narrative abstraction model (NAM) to unravel typical editorial techniques used in television drama to identify dramatic incidents within the video content, while Mei et al. [52] exploit the set structure conformed to by home video, such as relatively low levels of editing (lack of edited shot cuts) and frequent jerkiness, infidelity, blurring and brightness of image, incorporating these criteria into a quality metric that is temporally linked to the video content and used to include or exclude segments of video for the summary.

Audio features are the sole source of analysis for [20, 37, 54, 63] and all are *domain specific* techniques, specific to sports video content. For example, Rui et al. [63] analyse the speech track in baseball videos and apply a model derived from human speech processing theory (in the form of phoneme-level feature extraction) to abstract exciting segments and events such as baseball hits from baseball games, while Otsuka et al. [37] look for content such as applause, cheering, excited speech, normal speech and music to plot a level of importance curve to represent the summarised content of the video which the user may then use to navigate to the most important segments of video. Cai et al. [54] is also specific to entertainment and home videos, detecting and keeping a temporal record of laughter, applause and cheers which is used to identify highlighted video segments.

The techniques proposed in [1, 9, 39, 40, 55-57, 61, 62, 64, 66] all analyse *audio* and *image features* in the video stream. An example of a *non-domain specific* technique is that of Furini and Ghini [1] who summarise video content based on whether noise or silence is detected in the corresponding audio stream. In silent segments, half (X2) or two-thirds (X3) of image and audio frames are removed from the original video stream. Therefore, the resulting video summaries are essentially the original video with these frames removed. The X2 video summaries produce less temporally condensed video summaries but better maintain the original content, and particularly the transition between scenes, whereas the X3 video summaries require less time to view, but offer the user a more 'jerky' summary of the original content. Another example is that of Mah et al. [62] who apply theories of user attention to image and audio features to probabilistically summarise video into segments corresponding to the user's attention (visual, aural and linguistic). The vast majority of the other techniques, specifically [9, 39, 40, 55-57, 61, 64, 66], are all *domain specific*. A number of techniques [40, 55-57] are specific to sports video content. For example, Sugano et al. [40] use knowledge of the audio structure of soccer games to identify events such as goals and attempted goals in the summaries, while Hanjalic [55, 56] and Hanjalic and Xu [57] combine motion activity and cut density with a sound energy measure to probabilistically infer the extent to which the user is excited or aroused by the video content, producing excitement curves [55, 56], that represent the content of the

video, and affect curves [57], that represent the perceived changes in the user's affective state while viewing the sports videos. Other techniques summarise movie content [9, 39, 66]. For example, Lienhart et al. [9] summarise movies using feature based analysis for identifying shots and scenes, face detection for identifying main actors, and image and audio stream analysis for identifying key events such as gun fire and explosions, while Kopf et al. [39] produce video summaries for historical film, the majority of which is in black and white and hence generic measures of camera motion, object movement, shot duration and silence are used as criteria for inclusion of specific segments into the video summary. The remaining techniques summarise sitcoms using human face detection and background laughter detection [61] and summarise music videos by identifying specific events, such as choruses and song introductions [64].

Textual features in the form of text captions are the sole information source used for summarising video in [10, 68]. Whereas Luo et al. [10] do not undertake any analysis of the semantic meaning of the text captions, only incorporating key frames into the video summary that include text captions, Wu et al. [68] use OCR (optical character recognition) techniques to extract meaning where a thesaurus is used to assimilate the extracted words so as to identify important video segments.

The internal video summarisation techniques specified by [21, 58-60, 65, 67] also analyse *textual features*, but in combination with *image* or *audio features* or both. With the exception of [65] which prioritises the inclusion of video segments that include caption text in the final summary, all of these techniques are also *domain specific.* Three of the techniques [21, 60, 67] are specific to sports video content. For example, Liang et al. [60] combine analysis of image features with OCR of text captions for baseball videos, while Tjondronegoro et al. [67] detect the presence of full screen text captions in order to identify significant events within soccer videos. Other domains include news [59], where semantics about objects and events in the video are extracted from subtitles and used to aid speech analysis of the audio stream, all of which are stored in an MPEG-7-compliant description scheme, and presentation videos [58], where OCR-led identification of textual descriptors appearing in the slides of a presentation is combined with structural knowledge (regarding noticeable changes in the image stream and speaker tone changes when the next slide is presented) to derive shot boundaries.

3.2 Summaries produced by internal summarisation techniques

Since internal summarisation techniques do not analyse user-based information at all, all summaries produced by internal summarisation techniques are *generic*, not personalised. The summaries in [1, 7, 36, 39, 41, 44, 45, 47-49, 52] are all *static feature based summaries* derived

from the analysis of image or audio features. They include video skims [1], key frame summaries [7, 41, 44, 45], video segment summaries [39], and summaries that include both key frames and video segments [36, 48]. Cheng and Xu [44] and Ferman and Tekalp [41] produce some of the most effective summaries of this group since in addition to abstracting key frames, they also present summaries hierarchically, abstracting key frames at the scene level, to give the user a higher level representation of the video content, and at the shot level, to give the user more detailed representations of a specific scene. Similarly, Girgensohn [45] attempts to enhance the level of detail included in the summary by varying the key frame sizes based on the importance associated with each key frame, with larger frames deemed to represent more important information than smaller key frames.

Object based summaries are produced by Benjamas et al. [50] who use fixed inclusion criteria, based on the presence of objects (fighters) and flashlights in the crowd, to determine whether content is included or excluded from the summary on a frame by frame basis. The frames are then concatenated at the end of analysis and presented as a series of video segments that summarise the video content.

The summaries in [40, 53, 54, 60, 64, 67] are all *static event based summaries* and are presented in the form of video segment summaries where the segments have been chosen as a result of containing specific events. For example, Cai et al. [54] present sound effect attention curves, labelled with audio specific events from sports videos such as laughter, applause, and cheers, which gives the user an additional level of understanding of the nature of the content included in the summary.

The summaries presented in [9, 10, 18, 19, 58, 59, 61, 65, 68] are both *object* and *event based*, of these [9, 19, 58, 59] are *interactive* and the remainder are *static* video summaries. The static video summaries include static key frames linked to segments of video [61], soccer video summaries based on objects such as the referee and penalty box and events such as goals [18], and key frame summaries that include text captions within each key frame describing specific objects and events [10, 68]. Interactive video summaries vary in the degree of interactivity they provide. For example, in Shih and Huang's [19] baseball video summaries, individual player objects (e.g. batter and defender) and events (e.g. running and pitching) are identified, with shaded circles indicating recognised event semantics from the selected video segment such as 'overview' and 'defence', and the user is given the option of querying the summarised segments. Conversely, Lienhart et al. [9] present graphical interactive summaries, specific to the movie content domain, presented as two-dimensional colour coded block maps of the video stream, distinguishing moments of dialogue, explosions and appearances of text, which the user is able

to browse so as to retrieve video segments based on the selected information. Haubold and Kender [58] provide an interactive interface that allows the user to browse through the summarised content, where key frames representing video segments are presented horizontally across the top of the screen and a visual and audio segmentation graph, which allows the user to identify video segments containing high or low levels of audio and visual activity, and an index of key phrases and actual textual phrases identified within the image stream, which enables the user to obtain an overview of the content of each video segment without having to watch the video segment in full, are provided alongside.

Perception based summaries [37, 46, 51, 55-57, 62, 63, 66] are all *static* apart from [37, 63] which are *interactive* summaries. In addition, the vast majority [46, 51, 55-57, 62, 63, 66] apply theories from other domains to enable the abstraction of content in terms of high level concepts such as the end users' affective state, and perceived levels of user excitement and attention. A typical static perception based summary is that of Ngo et al. [46] which presents key frames that are perceived to demand the highest levels of attention from the user. Some examples of interactive perception based summaries include Otsuka et al.'s [37], which is based on a level of importance curve derived from audio in baseball videos (e.g. applause, cheering, excited speech, normal speech, and music) overlaid onto the original video such that the user may browse or retrieve summarised segments, the interactive summaries of Rui et al. [63], which identify events such as 'ball hit' and 'hitter running' for baseball videos and summarise as a function of the perceived levels of interest that each segment elicits in the user such that users can specify the level of interest they require to retrieve the summarised content, and Hanjalic [55] and Hanjalic and Xu [56, 57], who produce graphical summaries of video content in the form of arousal or excitement curves.

3.3 Discussion

Internal summarisation techniques produce a wide range of video summaries that are presented in various formats. Many promising techniques have been presented in the literature and achieve a wide spectrum of semantic summarisation to yield feature, object, event and perception based summaries. However, internal summarisation techniques still face a number of challenges, not least the *semantic gap*, which is the disparity between the semantics that can be abstracted by analysing low-level image audio and text features and the semantics that the user associates with and primarily uses to understand and remember the content of a video. This is exemplified by the fact that none of the internal summarisation techniques reviewed yield personalised video summaries reflecting the user's individual comprehension of the video content. Moreover, only

a relatively finite range of semantics can be abstracted from a video stream, compared with the wealth of semantics they represent, reflected in the fact that many techniques produce feature based summaries.

Techniques often need to be highly domain specific in order to control for the possible meaning of low level features in order to achieve object and event based summaries. The literature reflects this, in that 23 out of 37 internal summarisation techniques were domain specific. Sports video content also seems to be popular for internal video summarisation techniques, with 14 out of the 23 domain specific techniques being sports domain specific [18-21, 37, 50, 53-57, 60, 63, 67]. The highly structured nature of sports games coupled with the inferences that can be made from audiovisual cues, such as crowd noise and the sound of a whistle, make sports video the most popular domain for internal video summarisation techniques. In addition, all techniques that achieved object and event based summaries were domain specific, apart from [10, 65, 68], although the reliance on textual features by Luo et al. [10] and Wu et al. [68] means that these techniques cannot be applied to content where text does not feature and work best for textually rich content.

Techniques that are non-domain specific seem to lack precision or detail, due to the lack of domain knowledge to help frame low-level features within a specific context, reducing the application of these techniques in real world settings. This was reflected in the literature, where the majority of non-domain specific techniques produced (exclusively) feature based video summaries [1, 7, 36, 41, 44, 45, 47-49]. One exception was Ma et al. [62] who managed to produce perception based summaries through the use of applied theories during analysis.

Due to the challenge of the semantic gap, and the assumptions that focus analysis on the video stream, internal summarisation techniques rarely have the luxury of asking which semantics should be extracted based on the needs of the user, but rather due to the limited range of semantics that can be abstracted from the video stream alone ask which semantics can feasibly be extracted. The resulting summaries therefore are unable to summarise video based on individual user requirements, though the importance of personalisation is increasingly being stressed [3, 6, 69]

Identifying suitable criteria for evaluating the impact and value of video summarisation research is a current topic of debate. Chang [70] gives a valuable account of current issues relating to this debate and outlines a set of impact criteria that may serve as a means of maximising the value of future research in this domain. Evaluation methods of the video summaries presented in this section can be considered in two broad categories: *subjective* and *objective*. Subjective evaluation methods incorporate the judgement of the user in evaluating the

effectiveness of a video summary. Subjective measures are used by [39, 46, 61, 66]. Lu et al. [61] asked users to rate video summaries based on perceived levels of coherence, according to the number of key actors and key events included in the video summary. Sundaram et al. [66] used similar criteria but also asked the user to rate the extent to which the video summary enabled them to grasp when and where the video content was filmed. Ngo et al. [46] measured the extent to which users rated video summaries in terms of enjoyability and informativeness, while Kopf et al. [39] obtained feedback from 17 users by means of informal discussions regarding the quality of the video summaries and the content included in them. Objective methods do not incorporate user judgement into the evaluation criteria, but evaluate the performance of a given technique based on, for example, the extent to which specific objects and events are accurately identified in the video stream and included in the video summary. Commonly they are measured in terms of precision (percentage of accurate identifications of specific objects and events as a function of the total number of such events that are apparent in the full length video) and recall (number of specific objects and events recalled as a function of the total number recalled) [18, 50, 54]. For example, Cai et al. [54] measured precision and recall for three events detected within the audio stream (laughter, applause, cheering), while Ekin et al. [18] measured precision and recall rates for three events in soccer videos (yellow/red cards, penalties, free kicks). Other objective measures include the coherence or jerkiness of a video summary based on the average duration of video segments included in the video summary [1], or the summary compression ratio, that is, the extent to which the temporal duration of the original video has been reduced as a result of applying a given video summarisation technique [60].

4 External summarisation techniques and summaries

Although internal summarisation techniques benefit by minimising the amount of human effort required to produce video summaries, it is becoming recognised that the video source in isolation does not provide sufficient information to support a full range of semantic abstraction [3, 4, 17]. As a result, research is now looking to make use of information outside of the video stream, which could be a potentially valuable source for developing more effective video summaries but has not yet been sufficiently explored [5]. As shown earlier in Figure 2, external information consists of contextual information and user-based information. External summarisation techniques use external information in an attempt to summarise video content at enhanced levels of semantic abstraction, in order to better reflect the user's personal

comprehension of video content. External information is collected by capitalising on new scenarios in which users view and interact with video content. Information is collected at the capture, production and viewing stages of the video lifecycle and then analysed to develop video summaries based on this information. External summarisation techniques exemplify how valuable contextual and user-based information can help to overcome the semantic gap, thereby producing more personalised video summaries. Techniques that summarise video based *purely* on external information are rare, however. In this section, we review key external techniques and the video summaries they produce (Table 3).

Table 3: External video summarisation techniques and video summaries

	External Video Summarisation Techniques					**Video Summaries**					
	Contextual information		**User-based Information**		**External analysis**	**Content Type**				**Summary Type**	
Technique/summary	Lifecycle stage	Type	Lifecycle stage	Type	Domain specific	Feature based	Object based	Event based	Perception based	Interactivity	Personalisation
De Silva et al. [6]	C	US			X			X		I	E
Jaimes et al. [71]			P/V	OS	X			X	X	S	E
Takahashi et al. [72]			P	OS	X		X	X	X	I	G

4.1 Techniques for analysing external information sources

De Silva et al. [6] propose a *domain specific* video summarisation technique that collects *unobtrusively sourced contextual information* at the capture stage in the form of inhabitants' movements around the home. Fixed-position video cameras and integrated pressure-based floor sensors track the location of user activity within the home. Users are not required to input any information consciously, hence the information is categorised as being unobtrusively sourced. Summarisation of the captured video content is carried out by analysing data relating to footstep activity, taking into account measures such as distance between steps, overlap of durations between pairs of footsteps, and footstep direction changes.

Takahashi et al. [72] and Jaimes et al. [71] both analyse manually annotated (*obtrusively sourced*) information provided at the production stage and stored in MPEG-7 description schemes to produce video summaries. In both cases, videos are assumed to have a basic level of manually annotated content descriptions and neither analyse the video stream directly. Takahashi et al. [72] also require the user to provide the desired summary duration, which is

used to inform the criteria used to include or exclude video content in the final video summary. In both cases, the MPEG-7 description schemes are detailed and *domain specific*, both specifying sports video content: baseball and soccer content respectively. Takahashi et al. [72] produce an AudioVisualSegment annotation from a baseball video which includes object based information in the form of the player name as well as event based information regarding an 'at-bat' event within the video. Annotations are temporally linked to the video content via the MediaRelTimepoint tag. Jaimes et al. [71] have two modes of analysis. In the first phase, videos with a basic level of manual annotations are analysed in more detail and users are asked to provide annotations related to individual level of interest for each segment of video. The second phase takes in new video also with a basic level of manual annotation and propagates more detailed descriptions from already annotated video to the new video by matching the similarity of basic annotations via learning algorithms and a classifier module. This adds a level of automation that attempts to reduce human effort. Video summaries are produced based on analysis of annotations and matched to user profiles provided by individual users.

4.2 Summaries produced by external summarisation techniques

Benefiting from contextual information, De Silva et al. [6] produce *interactive personalised, event based* summaries based on user movements within their homes. Although the user does not consciously specify the semantics that are abstracted, video segments and key frames relevant to the individual's movements around the home are effectively abstracted and represent a personalised representation of their individual movements within the home environment. In addition, event level semantics are summarised and associated with the content via textual descriptors such as entering or leaving the house and the various rooms in the house. An interactive browsing and querying 'event summariser' interface is developed to exemplify the information summarised via this technique, which allows the user to search video based on the time the video was captured, the location in which it was captured, and the individual who featured in the video. The user can also browse key frames that represent significant events within the content, such as the subject entering or leaving a room, and subsequent video segments relating to each of these key frames can also be viewed.

Takahashi et al. [72] achieve *static object, event* and potentially *perception based* summaries of baseball videos. While emphasis is placed on player names, and events such as innings, plays of the ball, and 'at-bats', perceptual semantics relating to the user's enjoyment of video content can also be captured in the form of free text annotations. Textual annotations at all levels are used to supplement key frame and video segment summaries of the content which are

displayed through the interactive browsing and querying interface. At the highest level, content is summarised at the level of a game; at the next level, users can browse each innings within each game, and each at-bat, and each play within each innings. Selection of games, innings, at-bats or plays results in the presentation of a series of key frames on screen which can be selected to view the corresponding video segment. Users can also carry out key word searches on the summarised content.

Jaimes et al. [71] achieve *interactive personalised event* and *perception based* summarisation for soccer videos. Users manually identify and annotate specific events, as well as the level of personal interest they associate with an annotated event. A 'video digest' of the video content is then automatically generated, which is essentially a series of video segments.

4.3 Discussion

External summarisation techniques capture and analyse valuable information related to context and user to aid the summarisation process. However, these techniques are rare. Nevertheless, the following conclusions can be drawn with regards to these techniques and the summaries they produce.

External techniques achieve a broad and detailed range of semantic summarisation, and none present feature based summaries. In the case of both Takahashi et al. [72] and Jaimes et al. [71], this is as a result of there being no limit to the amount of detail the user can provide in terms of manual or spoken annotations.

Two of the three techniques [6, 71] present personalised summaries. Jaimes et al. [71] require obtrusively sourced information in order to achieve this, which can be costly to the user in terms of the time and effort required to provide this information, whereas De Silva et al. [6] achieve personalisation via unobtrusively sourced contextual information, hence minimising the inconvenience to the user. However, achieving this in the latter case requires a purpose built home with embedded floor sensors and built in video cameras, which limits wide applicability in real world settings.

Overcoming the challenge of the semantic gap therefore appears to be more achievable using external information than through internal information due to the personalisation that can be achieved. However, there are implications in terms of cost to the end user.

Video summaries produced by external summarisation techniques tend to be evaluated using subjective measures. De Silva et al. [6] evaluate their video summaries by showing users a number of video summaries and asking them why they find one summary better than others and in what ways the summaries could be improved. Some more frequent responses included

keeping the number of redundant key frames to a minimum and including content that involved human-human interaction. Jaimes et al. [71] required users to manually select video highlights by browsing through the metadata representing the content of the video. The precision and recall levels of the selected segments were then compared against a video summary defined by a user who manually selected the segments to be included the summary after watching the entire video content. Similarly, Takahashi et al. [72] compared video summaries produced by their technique with video summaries produced manually by television broadcasters, the results again being measured in terms of precision and recall.

5 Hybrid summarisation techniques and summaries

In order to capitalise on the strengths of both internal and external summarisation techniques, hybrid summarisation techniques analyse a combination of internal and external information. Table 4 summarises the key hybrid summarisation techniques and summaries in the literature.

5.1 Techniques for analysing hybrid information sources

A small number of techniques [32, 73, 74] use *unobtrusively sourced contextual* and *internal information* to abstract context based semantics for video summaries. All of the techniques are *domain specific* but vary in the particular domain of application. Lienhart [73] produces summaries of home video by analysing the time and date that the video content was captured and the clarity of the audio stream, while Agnihotri et al. [32] summarise music videos by analysing colour features and classifying audio to determine song start and finish boundaries, abstracting semantics from on screen text captions via OCR, and using this text to search the Internet for further song specific information such as song title and artist name. Yang et al. [74] produce news video summaries by using user textual queries to search for additional contextual text from news websites which is then compared with text derived by speech recognition from the audio stream. Relevant audio segments are then identified, together with associated video segments (selected according to specific criteria that minimises the number of anchor person shots and ensures that the image content included in the summary is significantly varied), for inclusion in the video summary.

Initial shot detection and segmentation is combined with knowledge of typical user behaviour to abstract video segments representing typical day-to-day user behaviour such as glance, gaze and walking. In the latter technique, a wearable video camera and microphone built into a pair of sunglasses captures video while a brainwave sensor built into a wearable headband unobtrusively monitors the user's brainwave activity. Brainwave data and video are temporally

Table 4: Hybrid summarisation techniques and video summaries

	Hybrid video summarisation techniques								Hybrid-produced video summaries					
	Internal information			Contextual information		User-based Information		Analysis	Content type				Summary Type	
Technique/summary	Image	Audio	Text	Lifecycle stage	Type	Lifecycle stage	Type	Domain specific	Feature based	Object based	Event based	Perception based	Personalisation	Interactivity
Agnihotri et al. [32]	X	X	X	C	US			X		X	X		G	I
Agnihotri et al. [75]	X	X	X			P/V	OS			X	X	X	E	S
Aizawa et al. [27]	X					C	US	X				X	E	I
Aizawa et al. [29]		X		C	US	C	OS	X		X	X		E	I
Babaguchi et al. [76]	X		X	P	US	V	OS	X		X	X		E	S
Cheatle [77]	X					C	US	X			X		E	S
Coldefy et al. [33]	X	X				P	OS	X			X		G	S
Fayzullin et al. [78]	X					P	OS	X		X	X		G	S
Fayzullin et al. [34]	X					P	OS	X		X	X		G	S
He et al. [79]	X	X				V	US	X				X	E	S
Lee et al. [80]	X					V	OS		X				E	I
Lienhart [73]		X		C	US			X	X				G	S
Lienhart [81]	X	X		C	US	C	OS	X		X	X		E	S
Lin and Tseng [25]	X	X				P/V	OS			X	X		E	S
Moriyama and Sakauchi [82]	X	X				P	OS	X				X	G	S
Rui et al. [83]	X					P/V	OS			X	X		E	I
Shipman et al. [24]	X					V	OS		X				E	I
Syeda-Mahmood and Poncelon [28]		X				V	US					X	E	S
Tjondronegoro et al. [22]	X	X	X			P	OS	X		X	X		G	I
Tseng et al. [38]	X					P/V	OS			X	X	X	E	I
Tseng and Lin [84]	X	X				P/V	OS			X	X		E	I
Xu et al. [31]	X		X	P	US	P/V	OS	X		X	X		E	S
Yu et al. [35]	X					V	US					X	E	I
Yu et al. [85]	X					P/V	OS			X	X		E	S
Yang et al. [74]	X	X		P	US			X		X	X		G	I
Zimmerman et al. [26]	X		X	P	US		OS	X		X	X		E	I

linked and high levels of brainwave activity used to indicate interesting segments of video for inclusion in the video summary. The remaining techniques [28, 35, 79] analyse unobtrusively sourced user-based information at the *viewing stage* rather than the capture stage. The techniques proposed by Yu et al. [35] and Syeda-Mahmood and Poncelon [28] are *non-domain specific*, while He et al.'s [79] technique is *domain specific* for presentation videos. All three techniques capture valuable user-based information from the interactions that users have while

browsing and viewing video content. The browsing information collected by these techniques is considered unobtrusively sourced since user browsing behaviour is captured as a result of the user's natural browsing activity, as opposed to requiring the user to browse with the specific goal of generating browsing information.

A larger number of techniques [27, 28, 35, 77, 79] use *unobtrusively sourced user-based* information together with *internal information* for producing video summaries. Cheatle [77] and Aizawa et al. [27] both analyse unobtrusively sourced *domain specific* user based information collected at the *capture stage*, both capturing video life experiences. In the former technique, an 'always on' head mounted video camera records the user's day to day activity which serves both as unobtrusively sourced user based information and as internal information for summarisation.

A range of techniques [22, 33, 34, 78, 82] analyse *obtrusively sourced user-based information* collected at the *production stage*, and all are *domain specific* techniques, almost all of which are specific to sports video content [22, 33, 34, 78]. For example, Fayzullin et al. [34, 78] use obtrusively sourced information to summarise soccer video content. In one approach [78], a domain expert is required to specify parameters such as subset-average and colour continuity functions specific to image features for soccer videos so that summarised segments within the video source can be extracted. Two techniques [22, 82] rely on obtrusively sourced information in the form of manual annotation to aid the identification of appropriate video content for summarisation. For example, Moriyama and Sakauchi [82] ensure obtrusively sourced information is kept to a minimum by automatic analysis of image and audio cues; however, obtrusively sourced information from the user is required to supervise the detection of start and end points of speech in television drama content, enhancing the accuracy of automated abstraction. Theories relating to the psychological unfolding of drama are also applied to assist the process of semantic abstraction.

Shipman et al. [24] and Lee et al. [80] analyse *obtrusively sourced user-based information* collected at the *viewing stage*, along with *internal information* in the form of image features. Both techniques *are non-domain specific*. The automated summaries of Shipman et al. [24] are further enriched by enabling interactive browsing and manual annotation by the user, such that automatically abstracted segments can be graphically linked and textually annotated as the user sees fit. This is a conscious effort on the user's behalf and so is considered as obtrusively sourced information. Lee et al. [80] automate the abstraction of summaries and store the semantics derived form image features such as colour layout, dominant colour, colour structure and motion within an MPEG-7 description scheme. Summaries are then generated and presented to the user as story units, which can be supplemented with further manual annotation.

A range of techniques [25, 38, 75, 83-85] analyse *obtrusively sourced user-based information* collected at both the *production* and *viewing stages* all are *non-domain specific*. For example, Rui et al. [83] use an initial training set of manually annotated content from a specified video at the production stage, which is then used to propagate semantic annotations to the remainder of the video. At the viewing stage, the user is presented with summarised video segments and is required to further enrich the summary by using custom built interactive annotation tools. Tseng et al. [38] also use custom built annotation tools, but focus purely at the production stage. In order to minimise manual effort, manual descriptors are automatically propagated to other, similar, untagged content. This technique also uses unobtrusively sourced user profiles provided by the user to personalise video summary content. Profiles are represented in MPEG-7 and MPEG-21 compliant format. Similar personalisation using MPEG-7 and MPEG-21 has been discussed by Lin and Tseng [25], but with video content streamed and filtered in real-time.

Some *domain specific* techniques [26, 29, 31, 76, 81] utilise both *unobtrusively sourced contextual* and *obtrusively sourced user-based information*. Domains include sports video [31, 76], news video [26] and life/home video [29, 81]. A good proportion of techniques [26, 29, 31, 76] use the Internet to obtain additional information about video content as a means of obtaining the unobtrusively sourced information. For example, in the news video domain, Zimmerman et al. [26] use image features for shot and scene segmentation as well as obtaining object, event and general descriptors from the web for detailed descriptions of the video content. The user is also required to provide their zip code, used to search for news content specific to the user's area of residence, and their video content preferences, used to develop personalised content summaries at the viewing stage. In the sports video domain, Babaguchi et al. [76] summarise videos of American Football games by using OCR to abstract descriptors from on screen text captions which are then used to search the web for further information such as pre-play and post-play statistics. In addition obtrusively sourced user profiles are required from the user, enabling the technique to produce personalised summaries of the content. In the life/home video domain, Aizawa et al. [29] use a wide range of contextual and user based information in conjunction with video captured from wearable devices. At the capture stage, a built in gyro for motion sensing and a global positioning system for location information records information including the user's speed, acceleration direction and geographical location. Wireless web access provides additional information such as weather and news information for the day. This information is used to abstract summarised video segments based on standardised pre-defined analysis rules. In addition, the information can be offered to the user in the form of 'contextual

retrieval keys' for specifying particular constraints for video summarisation (or video retrieval). At capture stage, the user can add obtrusively sourced voice annotations and send SMS keyword tags to the capture system.

5.2 Summaries produced by hybrid summarisation techniques

Two techniques [24, 80] produce *feature based* video summaries that are *interactive* and *personalised*. As an example, Lee et al. [80] present an interactive custom browsing interface that allows users to manually annotate automatically abstracted video segments at the viewing stage into story units that reflect their personal understanding of the video content. This results in personalised summaries. The browsing interface presents the user with a simple graphical representation of the automatically abstracted shot segments, clusters and story units. The user is given the option to split or group together the segments, clusters and story units on-screen via a move and insert button on a handheld remote control device, which can be used within the home. Lienhart [73] also presents *feature based* video summaries however these are *static* and *generic*. Summaries of home video are presented as a series of video segments in which the temporal order of the selected video segments is preserved. Video summary durations vary between one to five minutes for two-and-a-half hours of video.

Two techniques [33, 77] produce *static event based summaries*. Cheatle [77] present personalised video summaries as a series of key frames with corresponding video segments representing specific events, such as walking, gazing and glancing, that relate to an individual user's behaviour while capturing video, whereas Coldefy et al. [33] summarise events in tennis videos, such as serves and rallies which are presented as static video segment summaries that identify 6-13 percent of the original full length video for inclusion in the video summary.

A large number of hybrid video summarisation techniques [22, 25, 26, 29, 31, 32, 34, 74, 76, 78, 81, 83-85] yield *object* and *event based summaries*, just over half of which [26, 29, 31, 76, 81, 83-85] are *personalised*, and [22, 26, 29, 32, 74, 83, 84] all produce *interactive* summaries. For example, Babaguchi et al. [76] and Zimmerman et al. [26] achieve personalisation by matching user profiles provided by the user with unobtrusively sourced detail from the Internet. In the case of Zimmerman et al. [26], summaries are also interactive. Based on user profiles and contextual information sourced from the Internet, video content relating to daily news, sports news and weather news is summarised. An unusual feature is the personalisation of video based on the user's postcode, sourced from the user profile, which enables the summary to identify content that is most relevant to the area in which the user lives. For example, the 'headlines screen' provides the user with a high level textual overview of the

key news stories for that day, each textual overview being accompanied by a key frame image taken from the video corresponding to each key news story. Should any of the key news stories be of interest to the user, there is the option of selecting it from the list of key frames and textual descriptions, which results in the relevant video segment being played back to the user. Similarly, Aizawa et al. [29] present a multimedia interactive browsing and summarisation interface that can output key frame images, video segment summaries and graphical summaries of the content in the form of geographical maps indicating the areas in which the video content was captured. As an example of a *static* video summary, Lienhart [81] presents video segment-based summaries in which the temporal order of the selected video segments is preserved. The summaries are similar to [73], but the video summaries are personalised for the individual who captured the video content, by using structured voice annotations that are added by the user at the point of capture. The voice annotations, provided in a structured order, include names and descriptors of people and objects that appear within the video, and the actions, and events that are being recorded. Unlike many of the summaries produced by internal techniques, non-domain specific techniques are able to produce object and event based summaries [25, 83-85]. For example, Rui et al. [83] produce web-browser-based interactive video summaries that allow the user to browse key frame representations of the content which are organised within a scene-group-shot-frame hierarchy, presented as a video table of contents. The user is also able to browse key frame representations of the video content organised into visual, semantic and camera motion indices. The table of contents and indices are interrelated, and the user is free to browse, query and retrieve video segments from the summary, based on any combination of the table of the contents and indexed criteria.

Generic summaries are seen in [22, 25, 32-34, 73, 74, 78, 82], of which [22, 32, 74] are *interactive*. Despite the lack of personalisation, some of these produce highly detailed *object* and *event based summaries* that are highly descriptive. For example, Agnihotri et al. [32] unobtrusively source additional information about music videos from the Internet, producing interactive key frame summaries typically containing a facial image of the artist and a detailed textual description of the video content including artist name, song title, track duration and chorus lyrics. The summary is presented as a browseable webpage that allows the user to click on key frames of interest to play the respective music video. In another example, Tjondronegoro et al. [22] demonstrate the high level of detail that can be achieved through manual annotations, which includes the names of the two teams playing and the name of the sport being played. Manual annotations are provided by a skilled expert annotator of sports videos at the production stage. Multiple objects and events are hierarchically stored and presented in a custom built

interactive browsing interface which is split into four main sections: a video hierarchy consisting of drop down boxes from which the user can select sports games, a list of key breaks and events within the selected game (video), a video player, and an annotation box to display the annotations and descriptors relevant to the video segment being played.

Perception based summaries are also apparent in some of the literature [27, 28, 35, 38, 75, 79, 82] of which [27, 35, 38] are *interactive*. Unlike the perception based summaries yielded by internal video summarisation techniques, almost all of these summaries, with the exception of Moriyama and Sakauchi [82], are *personalised*. For example, Yu et al. [35] present summarised video content based on users' browsing behaviour within an interactive customised browsing environment that contains a story-tree hierarchy of videos, story units and scenes. Key frames corresponding to the various shots within each story unit can be selected which results in the corresponding video segment being played. The user also has the option of viewing rank-based skims of the video content of any requested duration, ranking being based on the perceived level of importance of each shot derived from the user's previous interaction and browsing behaviour with the video content. As another example, Aizawa et al. [27] summarise video relating to an individual user's day to day activity as a measure of the user's interest (derived from the user's brainwave activity while capturing the video content). The video summaries consist of a series of key frames representing the most pertinent video segments, based on a given inclusion threshold value related to the significance of brainwave activity associated with the video content, and a custom built interactive browsing environment where the user may view the video content and the brainwave activity corresponding to the video content. Brainwave activity is presented graphically as a series of temporally indexed bar graphs, which the user can navigate to find the most important video segments based on the significance of their brainwave activity. Tseng et al. [38] is unusual in that it achieves *object*, *event* and *perception based summaries*; however, the summaries are achieved purely by requiring the user to describe the content manually. A number of interactive video browsing environments are presented. For example, a web portal retrieves individual's user profiles at log in and generates video summaries based on current interests and on a target duration stated by the user, while an interactive video on demand environment presents video hyperlinks in synchronisation with the profile preferences as well as allowing the user to carry out keyword searches within a subset of video matching their preferences and constrained by a user-defined target duration.

5.3 Discussion

Hybrid summarisation techniques combine the analysis of internal and external information sources in order to produce video summaries. As a result of the multiple information sources, the summaries produced by such techniques are often detailed and in many cases personalised to reflect the individual user's comprehension of the video content. The following conclusions can be drawn.

Hybrid summarisation techniques produce detailed and specific semantic summarisation. In part this has been achieved if we consider that only three hybrid techniques produce feature based video summaries [24, 73, 80] compared with internal summarisation techniques, in which twelve techniques [1, 7, 36, 39, 41, 44, 45, 47-49, 52, 66] produced purely feature based video summaries. It seems that this is due to the added level of detail that can be achieved by combining the analysis of both internal and external information sources. Having said this, it is still necessary for hybrid techniques to maximise the level of detail by means of specifying the content domain. 16 out of 26 techniques are domain specific [22, 26, 27, 29, 31-34, 73, 74, 76-79, 81, 82] in which the technique can summarise video content, thus compromising the versatility of the respective techniques. Having said this, only 6 of the 16 domain specific techniques [22, 31, 33, 34, 76, 78] were specific to the sports video content domain, which is a significantly lower proportion than was seen for internal techniques, which may suggest that hybrid techniques are less reliant on the rigid structure and reliable image and audio cues that sports video is known to provide. This may be as a result of gaining additional information from external sources to help aid the summarisation process.

Contextual information is collected unobtrusively in all cases, and adds another level of detail to the resulting video summaries. However, techniques that do not supplement contextual information with user-based information [32, 73, 74] do not achieve personalised levels of summarisation.

In the case of [27, 28, 35, 77, 79], user-based information is collected unobtrusively and, as a result, personalised summaries are achieved. However, often this is only achieved for the individual who captured the video content. Syeda-Mahmood and Poncelon [28] and Yu et al. [35] are examples of techniques that only achieve personalisation for the individual who carries out browsing and viewing of video content using custom browsing and viewing software applications, which is a less common scenario in the consumption of video content. Aizawa et al. [27] and Cheatle [77] both unobtrusively source user-based information at the capture stage, achieving personalised summaries for the individual who captured the content. Thus, once again,

the versatility of these approaches is reduced, especially considering the smaller number of users who are actively involved in capturing video compared with the larger number who watch video.

The remainder of techniques that achieve personalised summaries [24-26, 29, 31, 38, 75, 76, 80, 81, 83-85] all require obtrusively sourced information from the user in some form. Obtrusively sourced information is considered inconvenient and costly to the user due to the time and effort involved in providing the information. Conversely, many techniques require obtrusively sourced user information, but still do not achieve personalised summaries [22, 33, 34, 78, 82]. This is as a result of the video lifecycle stage at which user-based information is collected. All techniques that collect user-based information purely at the production stage [22, 33, 34, 78, 82] do not achieve personalised summaries at the viewing stage, despite in all cases requiring obtrusively sourced information. This is because these techniques disconnect the production and viewing stages in space and time, hence the user who provides the user-based information used for the video summary is not the same user who views the video summary.

Video summaries produced by hybrid techniques were evaluated by both subjective and objective evaluation methods. Subjective evaluations include [34, 77, 79, 82, 83, 85]. For example, hand picked key frame images representing the user's optimum video summary may be compared with key frame images selected by the summarisation technique and the accuracy of the key frame images selected by the technique are then measured as a percentage match against the user selected images [77]. Similarly, the percentage match of video segments manually selected by users against video segments identified by the technique may be used to evaluate the effectiveness of the produced video summaries [82]. Other approaches include output from informal discussions with users as a means of evaluation [83], or large surveys regarding the perceived level of quality of video summaries [34]. Some evaluation methods [79, 85] require users to rate video summaries in terms of clarity, conciseness, and coherence; in addition users have been asked to evaluate video summaries in terms of overall choppiness and quality [79]. Objective evaluation methods are used by [33, 74, 76, 80]. For example, Coldefy et al. [33] evaluate the extent to which video summaries identify specific events, ball hits and applause within baseball videos, while Lee et al. [80] use precision and recall values for the number of correctly identified scenes in the video summary and compare the accuracy rate achieved for abrupt and gradual scene changes. Babaguchi et al. [76] use both subjective and objective measures: the objective measures involve identifying 45 events within a full length video and then evaluating how accurately the produced video summaries identify each of those events, whereas the subjective evaluation involves comparing the produced video summaries with man-made video summaries.

6 Challenges and recommendations for future directions

It is recognised that users are increasingly requiring personalised video summaries that reflect their individual comprehension and understanding of video content [3, 6, 69]. From the techniques and summaries reviewed here, it is apparent that internal summarisation techniques do not provide such personalised video summaries. This is due to a lack of external information being incorporated into the summarisation process. Although there are undoubtedly significant benefits and value in summarising video using internal techniques, not least in terms of the convenience to the user, overcoming the challenge of the semantic gap and providing personalised video summaries is still a significant challenge for internal summarisation techniques, even for those techniques that are highly domain specific. It seems that this is due to the lack of additional external information being incorporated into the video summarisation process. Having said this, the application of theories from other domains appears to be valuable in that they can be used to enhance the effectiveness of internal summarisation techniques by summarising video content at a higher semantic level without requiring high domain specificity.

In response to the challenges posed to internal summarisation techniques, external summarisation techniques source information from outside the video stream in order to abstract a range of semantics that otherwise would not be able to be abstracted from the video stream alone. Some external techniques are also able to produce personalised video summaries, often as a result of incorporating user-based information into the video summarisation process. However, although two-thirds of the external summarisation techniques reviewed in Section 4 achieve personalised video summaries [6, 71], Jaimes et al. [71] require detailed, obtrusively sourced textual information (provided by the user), while de Silva et al. [6] require a custom built video capture home environment. Hence, there is a cost involved in each case.

Hybrid summarisation techniques attempt to achieve enhanced levels of semantic summarisation by combining the analysis of internal and external information. The majority of techniques [22, 24-26, 29, 31, 33, 34, 38, 75, 76, 78, 80-85], however, still require obtrusively sourced user-based information in order to achieve enhanced levels of summarisation. Many techniques achieve personalised video summaries; however, the majority of these [24-26, 29, 31, 38, 75, 76, 80, 81, 83-85] achieve this by means of obtrusively sourced user-based information. The few techniques [27, 28, 35, 77, 79] that achieve personalised video summaries as a result of analysing unobtrusively sourced information are promising solutions. However, in the case of He et al. [79], Syeda-Mahmood and Ponceleon [28] and Yu et al. [35], unobtrusively sourced

user-based information is assumed to be so since it is collected on the assumption that users interact with video content by means of browsing content using custom video browsing applications that record the user's interactions with the video content; whereas most users still view video content in more traditional ways that require no explicit interaction, such as viewing content on their home television sets. Aizawa et al. [27] and Cheatle [77] achieve personalisation by capturing and analysing unobtrusively sourced user-based information at the capture stage, relating to the individual who captured the video content. However, the number of users that capture versus view video content is considerably smaller, and the majority of video content that users view is seldom captured by themselves.

We therefore conclude that video summarisation research faces the following challenges:

- *Internal summarisation techniques abstract a limited range of semantics for inclusion in the video summary.* Often internal techniques need to be domain specific in order to maximise the range of semantics they can abstract. Even so, the resulting summaries are often produced based on the semantics that are able to be abstracted as opposed to the semantics that are most useful or desired by the user. As a result, it is still difficult for internal summarisation techniques to overcome the semantic gap or produce personalised video summaries. Having said this, internal video summarisation techniques are valuable in that they do not inconvenience the user in any way, and hence achieve a level of functionality that many external and hybrid techniques continue to strive for.
- *External and hybrid techniques incorporate user-based information as a valuable resource that can enable the production of personalised video summaries.* The majority of summarisation techniques achieve personalised summaries by means of obtaining obtrusively sourced user-based information via detailed textual descriptions and spoken annotations. Although video summaries produced using this information tend to be better aligned with the user's understanding of the video, obtrusively sourced information is obtained at a cost to the user in terms of the time and effort required to provide it.
- *The small number of video summarisation techniques that produce personalised summaries via unobtrusively sourced information recorded at the point of video capture are only personalised for the user who captured the content.* A much larger number of users actively watch video content, compared with the relatively small numbers that capture it.
- *The small number of video summarisation techniques that produce personalised summaries at the viewing stage via unobtrusively sourced information do so by requiring users to*

browse video content via bespoke browsing applications. The reality is that commonly users take a much more passive role when viewing video content, most commonly sitting at home watching video content on their televisions.

- *Evaluations of video summarisation techniques and video summaries are difficult to compare.* A range of subjective and objective evaluation methods are used within the research literature which have been applied to quite disparate video summarisation techniques and video summaries that incorporate a combination of audiovisual cues for varying aims. This makes authoritatively comparing results across systems difficult.

In light of the challenges to existing video summarisation techniques and the achievements reviewed in this paper, the following recommendations have been identified. We propose these guidelines can be used to scope and focus future research in the video summarisation domain in order to maximise the utility and value of such research:

1) *Focus on incorporating user based sources of information into video summarisation techniques*. In order for video summaries to increase the range of semantics that can be included in a video summary, and to enable summaries to be personalised and better aligned with the user's understanding of the video content, it is important that new techniques exploit external information during the video summarisation process. In particular, it is clear from the literature that techniques that do not incorporate external user based information into the video summarisation process cannot produce personalised video summaries. In an era where the user is becoming more demanding, personalised video summaries are set to become an expected, standard requirement of contemporary users.
2) *Identify and incorporate new, previously untapped external sources of information into existing video summarisation techniques.* From existing research, it appears that future video summarisation research would benefit greatly from identifying new sources of contextual and user-based information, and developing techniques that incorporate this information into the video summarisation process. Furthermore, there is a wealth of valuable research that focuses on developing internal summarisation techniques. In order to capitalise on the knowledge gained from such research, new techniques incorporating user based and contextual information should be developed and used in conjunction with existing internal summarisation techniques. New research in this area could hold the key to overcoming the challenge of the semantic gap and would enhance the utility of the

video summary to the end user, particularly in presenting video summaries with an added level of personalisation.

3) *Focus on unobtrusively sourced user based information.* Although user based information should be incorporated into the summarisation process, it is important to qualify the ways in which the user-based information is obtained. Techniques that require user based information to be manually input often represent solutions that are too costly and inconvenient to the end user, and hence it may be argued are not feasible solutions. Consequently, new user based sources of information are required that do not demand detailed and conscious information to be provided by the user. An example of such a technique already exists in the literature, such as collecting user's brainwaves while capturing video content [27], although the brainwave activity is collected at the point of video capture and thus resulting video summaries are personalised only for the individual who captured the video content. If summarisation techniques are to output video summaries that are relevant to the individual user, alternative sources of analysis must be incorporated. New user based sources of information should be identified that lend themselves to collecting user-based information at little or no cost to the end user; that is, collected and analysed unobtrusively.

4) User based information sources that have wide applicability should be focused at the viewing stage. Although rare, existing techniques that achieve personalised summarisation from user based information have shown promising results. However, the drawback of such techniques often tends to be that personalised summaries are developed as a result of constraints on the conditions in which the video content is initially captured. This is not an ideal solution, since video content is often not restricted to those who captured it. Consequently, in order to maximise the utility and applicability of a video summarisation technique, future video summarisation techniques that incorporate user based information should strive to cater for every user who may view the content. This implies that techniques that focus on collecting user based information at the capture and production stages may not be present solutions that ensure wide and feasible application of the proposed solution. Therefore, it is suggested that future research should focus on developing techniques that collect user based information at the viewing stage. User based information may be obtained in various forms; for instance, already user browsing behaviour seems to have demonstrated some value for video summarisation. However, techniques may find value in focusing on the user directly; for example, monitoring user behaviour and physiology while viewing video content may help to identify candidate

content for inclusion in a video summary. Indeed, due to developments in sensor technology, physical changes in posture and facial expressions can all be unobtrusively recorded and automated techniques for analysis of such data are being developed with promising results. The recent focus of affective video content analysis [86] and summarisation [57] has demonstrated the value of abstracting affective qualities of a video. However, to the best of our knowledge no external or hybrid video summarisation techniques have been developed that summarise video from such external information. Hence, there is a need to explore the potential of affective video content summarisation based on external information and this is likely to be sourced directly from the user.

5) *Extend the application of theories from other domains to user based information.* Theories from a wide range of other domains have been shown to have significant value, particularly for summarising video at a perceptual level. However, such theories have been used to aid analysis of the video stream and not applied to external information. For example, hybrid techniques that use such theories to infer meaning by direct analysis of the video [79, 82]. However, such theories could be applied to user based data to infer meaning. Indeed, future techniques incorporating user based information may well use such theories as a key component to decipher user based interactions with video content, such as changes in posture, facial expressions and physiological responses. Without such theories, user based information may remain removed from the development of video summarisation techniques.
6) *Develop a set of standard evaluation methods and content for video summarisation.* Standard information retrieval measures, such as precision, recall and fallout, allow for basic comparisons of video summarisation techniques and video summaries. However, developing more extensive metrics, which are more specific to video summarisation techniques and video summaries and which have equal applicability across domains, will enable a standardised, comparable set of evaluation results. The ready availability of reference video streams, which researchers can use as test sources, is paramount to this. The recent inclusion of rushes summarisation submissions at TRECVID 2007 [87] is a step in the right direction in this regard.

7 Concluding discussion

In this paper we have presented a conceptual framework for video summarisation which encompasses the rich range of techniques and summaries presented within the research literature. The conceptual framework was used to survey the various video summarisation

techniques and resultant video summaries that have been proposed. Video summarisation techniques were split into three sub-types: internal (analyse information sourced directly from the video stream), external (analyse information not sourced directly from the video stream) and hybrid (analyse a combination of internal and external information). The video summaries that each technique produce were also categorised as a function of the type of content they summarise (object, event, perception or feature based) and the functionality offered to the user for their consumption (interactive or static, personalised or generic).

Despite there being valuable contributions in all categories, it was revealed that internal summarisation techniques abstract a limited range of semantics for inclusion in video summaries and lack the ability to produce personalised video summaries. This is due to the lack of contextual and user based information used to inform the summarisation process. External summarisation techniques are rare, but there does not appear to be a limit to the range and detail that such techniques can include in a video summary, largely due to such techniques requiring detailed obtrusively sourced information from the user. As a result, external summarisation techniques do achieve personalised video summaries, but the level of detail and personalisation tends to come at a cost in terms of the time and effort required from the user in providing this information. Hybrid techniques combine internal and external techniques, but despite there sometimes being no limit to the level of detail that can be included in the summary and the summaries often being personalised to some extent, this is achieved by requiring the user to manually input information, or by placing constraints on the ways in which the video can be captured. Hence, personalised summaries can only be achieved for the user who captured the content.

With a view to overcoming some of the challenges faced by internal, external and hybrid techniques, this paper has proposed a number of recommendations for future video summarisation techniques. We proposed that future video summarisation research would benefit from focusing more attention on user based sources of information, but in a way that minimises the level of effort required by the user in capturing this information. To this end, new sources of user based information must be identified that better lend themselves to unobtrusive capture and analysis. Analyses of user's physiological responses, facial expressions and posture have been suggested as potentially fruitful areas of focus. Furthermore, techniques would benefit from concentrating on developing solutions that can be applied at the viewing stage of the media lifecycle so that the number of users that benefit is maximised.

References

[1] M. Furini, V. Ghini, An Audio-Video Summarisation Scheme Based on Audio and Video Analysis, in: Proc. IEEE Consumer Communications and Networking Conference (CCNC '06), Vol. 2, Las Vegas, NV, USA, 8-10 January, 2006, pp. 1209-1213.

[2] Y. Li, S. Lee, C. Yeh, C. Kuo, Semantic retrieval of multimedia, IEEE Signal Processing Magazine, 23 (2) (2006) 79-89.

[3] M. S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-Based Multimedia Information Retrieval: State of the Art and Challenges, ACM Transactions on Multimedia Computing, Communications and Applications, 2 (1) (2006) 1-19.

[4] M. Barbieri, L. Agnihotri, N. Dimitrova, Video Summarization: Methods and Landscape, in: Internet Multimedia Management Systems IV: Proceedings of SPIE, vol. 5242, J. R. Smith, S. Panchanathan, T. Zhang (Eds.), Bellingham, WA, USA, 2003, pp. 1-13

[5] N. Dimitrova, Context and Memory in Multimedia Content Analysis, IEEE Multimedia, 11 (3) (2004) 7-11.

[6] G. de Silva, T. Yamasaki, K. Aizawa, Evaluation of video summarization for a large number of cameras in ubiquitous home, in: Proc. 13th Annual ACM International Conference on Multimedia, Singapore, 2005, pp. 820-828

[7] C. Gianluigi, S. Raimondo, An Innovative Algorithm for Key Frame Extraction in Video Summarization, Journal of Real-Time Image Processing, 1 (1) (2006) 69-88.

[8] X. Zhu, X. Wu, J. Fan, A. Elmagarmid, W. Aref, Exploring video content structure for hierarchical summarization, Multimedia Systems, 10 (2) (2004) 98-115.

[9] R. Lienhart, S. Pfeiffer, W. Effelsberg, Video abstracting, Communications of the ACM, 40 (12) (1997) 55-62.

[10] B. Luo, X. Tang, J. Liu, H. Zhang, Video caption detection and extraction using temporal information, in: Proc. IEEE International Conference on Image Processing (ICIP 2003), Vol. 1, Thessaloniki, Greece, 2003, pp. 297-300.

[11] H. Joffe, L. Yardley, Content and thematic analysis, in: Research methods for clinical and health psychology, D. F. Marks,L. Yardley (Eds.). Sage Publications, London, 2004.

[12] R. E. Boyatzis, Transforming Qualitative Information. Sage Publications, 1998.

[13] S. J. Taylor, R. Bogdan, Introduction to qualitative research methods: The search for meanings. John Wiley & Sons, New York, 1984.

[14] D. Silverman, Doing Qualitative Research : Second Edition Sage Publications, London, 2005.

[15] I. Dey, Qualitative data analysis: A user-friendly guide for social scientists. Routledge, London, 1993.

[16] Y. Wang, Z. Liu, J. Huang, Multimedia content analysis: using both audio and visual clues, IEEE signal processing magazine, 17 (6) (2000) 12-36.

[17] M. Davis, S. King, N. Good, R. Sarvas, From Context to Content: Leveraging Context to Infer Media Metadata, in: Proc. 12th Annual ACM International Conference on Multimedia, Vol. October 10-16, New York, NY, USA, 2004, pp. 188-195.

[18] A. Ekin, A. M. Tekalp, R. Mehrotra, Automatic soccer video analysis and summarization, IEEE Transactions on Image Processing, 12 (7) (2003) 796-807.

[19] H. Shih, C. Huang, MSN: statistical understanding of broadcasted baseball video using multi-level semantic network, IEEE Transactions on Broadcasting, 51 (4) (2005) 449 - 459.

[20] M. Xu, C. Maddage, C. Xu, M. Kankanhalli, Q. Tian, Creating audio keywords for event detection in soccer video, in IEEE International Conference on Multimedia and Expo Baltimore, USA: IEEE, 2003.

[21] D. Zhang, S. Chang, Event Detection in Baseball Video Using Superimposed Caption Recognition, in ACM international conference on Multimedia. Juan-les-Pins, France: ACM Press, 2002.

[22] D. Tjondronegoro, Y. Chen, B. Pham, Highlights for more complete sports video summarization, IEEE Transactions on Multimedia, 11 (4) (2004) 22-37.

[23] J. Pinzon, R. Singh, Designing an Experiential Annotation System for Personal Multimedia Information Management in: Proc. Human-Computer Interaction: IASTED-HCI, Phoenix, AZ, USA, 11/14/2005 - 11/16/2005, 2005.

[24] F. Shipman, A. Girgensohn, L. Wilcox, Creating navigable multi-level video summaries, in: Proc. IEEE International Conference on Multimedia and Expo (ICME '03), Vol. 2, Baltimore, MA, USA, 2003, pp. 753-756.

[25] C. Lin, B. L. Tseng, Optimizing user expectations for video semantic filtering and abstraction, in: Proc. IEEE International Symposium on Circuits and Systems (ISCAS '05), Vol. 2, Kobe, Japan, 23-26 May, 2005, pp. 1250-1253.

[26] J. Zimmerman, N. Dimitrova, L. Agnihotri, A. Janevski, L. Nikolovska, Interface Design for MyInfo: a Personal News Demonstrator Combining Web and TV Content., in: Proc. INTERACT: IFIP International Conference on Human-Computer Interaction, Zurich, Switzerland, 1-5 September, 2003.
[27] K. Aizawa, K. Ishijima, M. Shiina, Summarizing wearable video, in: Proc. IEEE International Conference on Image Processing (ICIP '03), Vol. 3, Thessaloniki, Greece, 7-10 October, 2001, pp. 398 - 401.
[28] T. Syeda-Mahmood, D. Ponceleon, Video Retrieval and Browsing: Learning video browsing behavior and its application in the generation of video previews, in: Proc. 9th ACM International Conference on Multimedia, Ottowa, Ontario, Canada, 30 September-5 October, 2001, pp. 119-128.
[29] K. Aizawa, D. Tancharoen, S. Kawasaki, T. Yamasaki, Efficient retrieval of life log based on context and content, in: Proc. 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences, New York, NY, USA, 15 October, 2004, pp. 22-31.
[30] N. Babaguchi, Y. Kawai, T. Kitahashi, Generation of Personalized Abstract of Sports Video, in: Proc. IEEE Conference on Multimedia and Expo (ICME 2001), Tokyo, Japan, 22-25 August, 2001, pp. 800-803.
[31] C. Xu, J. Wang, K. Wan, Y. Li, L. Duan, Live Sports Detection Based on Broadcast Video and Web-casting Text, in Proceedings of the 14th annual ACM international conference on Multimedia (MM '06). Santa Barbara, CA: ACM Press, 2006.
[32] L. Agnihotri, N. Dimitrova, J. R. Kender, Design and evaluation of a music video summarization system, in: Proc. IEEE International Conference on Multimedia and Expo (ICME '04), Vol. 3, Taipei, Taiwan, 27-30 June 2004, pp. 1943 -1946.
[33] F. Coldefy, P. Bouthemy, M. Betser, G. Gravier, Tennis video abstraction from audio and visual cues, in: Proc. IEEE 6th Workshop on Multimedia Signal Processing, Siena, Italy, 29 Sept.-1 Oct. 2004, 2004, pp. 163-166
[34] M. Fayzullin, V. S. Subrahmanian, M. Albanese, A. Picariello, The priority curve algorithm for video summarization, in: Proc. 2nd ACM International Workshop on Multimedia Databases, Arlington, VA, USA, 13 Nov., 2004, pp. 28-35.
[35] B. Yu, W.-Y. Ma, K. Nahrstedt, H.-J. Zhang, Video summarization based on user log enhanced link analysis, in: Proc. 11th Annual ACM International Conference on Multimedia, Berkeley, CA, USA, 2-8 November, 2003, pp. 382-391
[36] X. Zhu, X. Wu, Sequential association mining for video summarization, in: Proc. IEEE International Conference on Multimedia and Expo (ICME '03), Vol. 3, Baltimore, MD, USA, 6-9 July, 2003, pp. 333-336.
[37] I. Otsuka, K. Nakane, A. Divakaran, K. Hatanaka, M. Ogawa, A highlight scene detection and video summarization system using audio feature for a personal video recorder, IEEE Transactions on Consumer Electronics, 51 (1) (2005) 112-116
[38] B. L. Tseng, C.-Y. L., J. R. Smith, Using MPEG-7 and MPEG-21 for personalizing video, IEEE Transactions on Multimedia, 11 (1) (2004) 42-52.
[39] S. Kopf, T. Haenselmann, D. Farin, W. Effelsberg, Automatic generation of video summaries for historical films, in: Proc. IEEE international conference on Multimedia and Expo (ICME '04), Vol. 3, Taipei, Taiwan, 27-30 June, 2004, pp. 2067-2070.
[40] M. Sugano, Y. Nakajima, H. Yanagihara, Automated MPEG audio-video summarization and description, in: Proc. IEEE International Conference on Image Processing, Vol. 1, Rochester, NY, USA, 2002, pp. 956-959.
[41] A. M. Ferman, A. M. Tekalp, Two-stage hierarchical video summary extraction to match low-level user browsing preferences, IEEE Transactions on Multimedia, 5 (2) (2003) 244-256.
[42] S.-F. Chang, A. Vetro, Video adaptation: Concepts, technologies, and open issues, Proceedings of the IEEE, 93 (1) (2005) 148-158.
[43] C. Crockford, H. Agius, An empirical investigation into user navigation of digital video using the VCR-like control set, International Journal of Human-Computer Studies, 64 (4) (2006) 340-355.
[44] W. Cheng, D. Xu, An approach to generating two-level video abstraction, in: Proc. 2nd IEEE International Conference on Machine Learning and Cybernetics, Vol. 5, Xi-an, China, 2-5 November, 2003, pp. 2896-2900.
[45] A. Girgensohn, A fast layout algorithm for visual video summaries, in: Proc. IEEE International Conference on Multimedia and Expo (ICME '03), Vol. 2, Baltimore, MD, USA, 6-9 July, 2003, pp. 77-80.
[46] C. Ngo, Y. Ma, H. Zhang, Video summarization and scene detection by graph modeling, IEEE Transactions on Circuits and Systems for Video Technology, 15 (2) (2005) 296-305.

[47] Z. Cernekova, I. Pitas, C. Nikou, Information theory-based shot cut/fade detection and video summarization, IEEE Transactions on Circuits and Systems for Video Technology, 16 (1) (2006) 82-91.
[48] Z. Li, G. M. Schuster, A. K. Katsaggelos, MINMAX Optimal Video Summarization, IEEE Transactions on Circuits and Systems for Video Technology, 15 (10) (2005) 1245-1256.
[49] G. Ciocca, R. CSchettini, Dynamic Storyboards for Video Content Summarization, in 8th ACM SIGMM International Workshop on Multimedia Information Retrieval. Santa Barbara, CA: ACM Press, 2006.
[50] N. Benjamas, N. Cooharojananone, C. Jaruskulchai, Flashlight and player detection in fighting sport for video summarization, in: Proc. IEEE International Symposium on Communications and Information Technology (ISCIT 2005), Vol. 1, Beijing, China, 12 - 14 Oct 2005, 2005, pp. 441 - 444.
[51] B. Jung, T. Kwak, J. Song, Y. Lee Narrative abstraction model for story-oriented video, in: Proc. 12th Annual ACM International Conference on Multimedia New York, NY, USA, 10-15 October, 2004.
[52] T. Mei, C. Zhu, H. Zhou, X. Hua, Spatio-temporal quality assessment for home videos in: Proc. 13th Annual ACM International Conference on Multimedia Singapore, 6-11 November, 2005.
[53] F. N. Bezerra, E. Lima, Low Cost Soccer Video Summaries Based on Visual Rhythm, in: Proc. Proceedings of the 14th annual ACM international conference on Multimedia (MM '06), 2006, pp. 71 - 77.
[54] R. Cai, L. Lu, H. Zhang, L. Cai, Highlight sound effects detection in audio stream, in: Proc. IEEE International Conference on Multimedia and Expo (ICME '03), Vol. 3, Baltimore, MD, USA, July, 2003, pp. 37-40.
[55] A. Hanjalic, Adaptive extraction of highlights from a sport video based on excitement modeling, IEEE Transactions on Multimedia, 7 (6) (2005) 1114-1122.
[56] A. Hanjalic, Generic Approach to Highlight Extraction in a Sport Video, in: Proc. IEEE International Conference on Image Processing (ICIP 2003), Vol. 1, Barcelona, Spain, September, 2003, pp. 1-4.
[57] A. Hanjalic, L. Q. Xu, Affective Video Content Representation and Modeling, IEEE Transactions on Multimedia, 7 (1) (2005) 143-154.
[58] A. Haubold, J. R. Kender, Augmented segmentation and visualization for presentation videos, in: Proc. 13th Annual ACM International Conference on Multimedia, Singapore, 6-11 November 2005, 2005, pp. 51-60.
[59] J. Kim, H. Chang, K. Kang, M. Kim, H. Kim Summarization of news video and its description for content-based access, International Journal of Imaging Systems and Technology, 13 (5) (2004) 267-274.
[60] C. Liang, J. Kuo, W. Chu, J. Wu, Semantic units detection and summarization of baseball videos, in: Proc. 47th Midwest Symposium on Circuits and Systems (MWSCAS '04) Vol. 1, Hiroshima, Japan, 2004, pp. 297-300
[61] S. Lu, M. R. Lyu, I. King, Video summarization by spatial-temporal graph optimization, in: Proc. International Symposium on Circuits and Systems (ISCAS '04), Vol. 2, Vancouver, Canada, 2004 pp. 197-200.
[62] Y. Ma, X. Hua, L. Lu, H. Zhang, A generic framework of user attention model and its application in video summarization, IEEE Transactions on Multimedia, 7 (5) (2005) 907-919.
[63] Y. Rui, A. Gupta, A. Acero, Automatically extracting highlights for TV Baseball programs in: Proc. 8th ACM International Conference on Multimedia, Los Angeles, CA, USA, 30 October, 2000, pp. 105-115.
[64] X. Shao, C. Xu, M. S. Kankanhalli, K. M. Inkpen, Automatically generating summaries for musical video, in: Proc. IEEE International Conference on Image Processing (ICIP 2003), Vol. 3, Barcelona, Spain, 2003, pp. 547-50
[65] M. A. Smith, T. Kanade, Video skimming and characterization through the combination of image and language understanding, in IEEE International Workshop on Content-based Access of Image and Video Databases (ICCV98). Bombay, India: IEEE, 1998.
[66] H. Sundaram, L. Xie, S. Chang, A utility framework for the automatic generation of audio-visual skims in: Proc. 10th ACM International Conference on Multimedia, Juan Les-Pins, France, 2002, pp. 189-198.
[67] D. W. Tjondronegoro, Y. Chen, B. Pham, Classification of self-consumable highlights for soccer video summaries, in: Proc. IEEE International Conference on Multimedia and Expo (ICME '04), Vol. 1, Taipei, Taiwan, 2003, pp. 579-582.
[68] Y. Wu, L. Y., C. C., VSUM: summarizing from videos, in: Proc. IEEE 6th International Symposium on Multimedia Software Engineering, Miami, FL, USA, 2004, pp. 302-309.

[69] B. L. Tseng, J. R. Smith, Hierarchical Video Summarization Based on Context Clustering, in: Internet Multimedia Management Systems IV: Proceedings of SPIE, vol. 5242, J. R. Smith, S. Panchanathan, T. Zhang (Eds.), Bellingham, WA, USA, 2003, pp. 14 - 25.
[70] S. Chang, The Holy Grail of Content-Based Media Analysis, IEEE Multimedia, 9 (2) (2002) 6 - 8.
[71] A. Jaimes, T. Echigo, M. Teraguchi, F. Satoh, Learning personalized video highlights from detailed MPEG-7 metadata, in: Proc. IEEE International Conference on Image Processing, Vol. 1, New York, NY, USA, 22-25 Sept., 2002, pp. 133-136.
[72] Y. Takahashi, N. Nitta, N. Babaguchi, Video Summarization for Large Sports Video Archives, in: Proc. IEEE International Conference on Multimedia and Expo (ICME 2005), Amsterdam, The Netherlands, July 2005, 2005, pp. 1170-1173
[73] R. Lienhart Abstracting home video automatically, in: Proc. 7th ACM International Conference on Multimedia, Orlando, FL, USA, 30 October - 5 November, 1999, pp. 37-40.
[74] H. Yang, L. Chaisorn, Y. Zhao, S. Neo, T. Chua, VideoQA: question answering on news video, in: Proc. 11th Annual ACM International Conference on Multimedia (ACMM'2003), Berkeley, CA, USA, 2-8 November, 2003, pp. 632-641.
[75] L. Agnihotri, J. Kender, N. Dimitrova, J. Zimmerman, Framework for Personalized Multimedia Summarization, in 7th ACM International Workshop on Multimedia Information Retrieval (MIR '05). Singapore: ACM Press, 2005.
[76] N. Babaguchi, Y. Kawai, T. Ogura, T. Kitahashi, Personalized abstraction of broadcasted American football video by highlight selection, IEEE Transactions on Multimedia, 6 (4) (2004) 575-586.
[77] P. Cheatle, Media content and type selection from always-on wearable video, in: Proc. IEEE 17th International Conference on Pattern Recognition (ICPR '04), Vol. 4, Cambridge, UK, 23-26 August, 2004, pp. 979-982
[78] M. Fayzullin, V. S. Subrahmanian, A. Picariello, M. L. Sapino, The CPR model for summarizing video, in: Proc. 1st ACM International Workshop on Multimedia Databases, New Orleans, LA, USA, 2003, pp. 2 - 9
[79] L. He, E. Sanocki, A. Gupta, J. Grudin, Auto-summarization of audio-video presentations, in: Proc. 7th ACM International Conference on Multimedia, Orlando, FL, USA, October 30 - 5 November 1999 1999.
[80] J. Lee, G. Lee, W. Kim, Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder, IEEE Transactions on Consumer Electronics, 49 (3) (2003) 742 - 749.
[81] R. W. Lienhart, Dynamic video summarization of home video, in: Storage and Retrieval for Media Databases: Proceedings of SPIE, vol. 3972, M. M. Yeung, B. Yeo, C. A. Bouman (Eds.), 2000, pp. 378-389
[82] T. Moriyama, M. Sakauchi, Video Summarization Based on the Psychological Unfolding of Drama, Systems and Computers in Japan, 33 (11) (2002) 1122-1131.
[83] Y. Rui, S. X. Zhou, T. S. Huang, Efficient access to video content in a unified framework, in: Proc. IEEE International Conference on Multimedia Computing and Systems, Vol. 2, 7-11 June, 1999, pp. 735-740.
[84] B. L. Tseng, C. Lin, Personalized video summary using visual semantic annotations and automatic speech transcriptions, in: Proc. IEEE Workshop on Multimedia Signal Processing, Virgin Islands, USA, 9-11 Dec. , 2002, pp. 5-8.
[85] J. C. S. Yu, M. S. Kankanhalli, P. Mulhen, Semantic video summarization in compressed domain MPEG video, in: Proc. IEEE International Conference on Multimedia and Expo (ICME '03), Vol. 3, Baltimore, MA, USA, 6-9 July, 2003, pp. 329-332.
[86] H. Kang, Affective Content Detection using HMMs, in: Proc. 11th ACM International Conference on Multimedia, Berkeley, CA, USA, 2003, pp. 259-262.
[87] NIST, DRAFT Guidelines for the TRECVID 2007 Evaluation, http://www-nlpir.nist.gov/projects/tv2007/tv2007.html, 2007.