D3.2.1.6
Requirements and methods for video summarization

Author(s): Henri Kivinen
Confidentiality: Consortium
Date and status: 08.03.2012

This work was supported by TEKES as part of the next Media programme of TIVIT (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT)
## Version history:

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>State</th>
<th>Author(s) OR Editor/Contributors</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.12.2011</td>
<td>draft</td>
<td>Henri Kivinen</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>31.1.2012</td>
<td>update</td>
<td>Henri Kivinen</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>29.2.2012</td>
<td>update</td>
<td>Henri Kivinen</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>9.3.2012</td>
<td>update</td>
<td>Henri Kivinen</td>
<td></td>
</tr>
</tbody>
</table>

## Participants

<table>
<thead>
<tr>
<th>Name</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henri Kivinen</td>
<td>Aalto SCI / Department of Media Technology</td>
</tr>
</tbody>
</table>
Executive Summary

Content-based video indexing and retrieval has still a number of challenges to be solved before being fully automatized and reliable, and therefore, these applications often fail. In these cases it is very practical that the retrieved videos can be browsed effortlessly and fast to evaluate their relevance and importance. Video summarization provides the means for such video browsing by reducing redundant content in video sequences.

In this work we will focus on the summarization of user-generated video content, which has its own distinct characteristics compared to other video types such as professionally produced video content. As a consequence, the summarization of user-generated content is more challenging than for more constrained video content. Due to its unconstrained nature the summarization techniques must be generic in such that they can cope with all the infinite number of variations in which user-generated content exists. In the report we propose that motion information–based video summarization is possible to be implemented on such a generic level, and therefore they are currently the most reliable techniques to be applied for user-generated video content.
Table of Contents

Executive Summary .................................................................................................................. 2
1 Table of Tables ........................................................................................................................ 3
2 Table of Figures ........................................................................................................................ 3
3 Introduction ............................................................................................................................... 4
4 Structural video analysis ......................................................................................................... 5
   4.1 Temporal video segments ................................................................................................. 5
   4.2 Methods for sub-shot segmentation .................................................................................. 6
5 Motion information –based video summarization ................................................................. 8
   5.1 Motion saliency ................................................................................................................ 8
   5.2 Camera motion characterization ....................................................................................... 13
6 Presentation of video summarization ..................................................................................... 14
   6.1 Static and dynamic presentation .................................................................................... 14
   6.2 Hierarchical presentation ............................................................................................... 15
7 Discussion ................................................................................................................................... 17

1 Table of Tables

Table 1. Comparison of video domains (adapted from Wang & Ngo 2011) ......................... 7

2 Table of Figures

Figure 1. Temporal video segments (adapted from Petershon 2007) ................................. 5
Figure 2. Visual skimming based on motion attention (Ma & Zhang, 2002) ....................... 9
Figure 3. Detection of keyframes based on turning points of dominant motion (Liu et al. 2003) ......................................................................................................................... 10
Figure 4. White area of the rectangle represents the motion parameter estimation template (Abdollahian et al. 2010) ........................................................................................................ 11
Figure 5. Total displacement of a trajectory compared to its cumulative length. (Guironnet et al. 2005) ................................................................................................................................. 14
Figure 6. Comic-like video summarization of a news sequence from the TRECVID 2006 corpus (Calic et al. 2007). ...................................................................................................................... 16
3 Introduction

Purpose of video summarization is to provide a compact but still informative representation of a video sequence. The need for it arises from the very nature of video content that is always packed with redundant information. For example, a video sequence that can be a movie, news story, commercial, home video, user-generated video clip, etc., has a duration that spans from anything below 30 seconds to hours. With a common frame rate such as 25 frames per second a one-minute video clip comprise 1500 still images, and usually one has to view them all before getting an overall gist of the content or making a judgment on its importance or relevance. This can be very time consuming, especially when the number of video sequences grows.

Video summarization is usually seen as a complementary to video retrieval, as it makes browsing of retrieved videos faster. Especially when content-based video retrieval fails and the user needs to browse through a large set of video sequences (Xiong et al. 2006, Truong & Venkatesh 2007). In these cases, the user could exploit the summarized video presentations of retrieved video sequences to locate the relevant videos within a set of retrieved videos (Hu et al. 2011). This requires a video summarization process that maintains all the salient information while removing all redundant data. After this, the process should output a composition of an abstract representation or summary of the extracted content, which is finally exhibited to users such that it facilitates browsing.

The summarization process itself can be implemented with a multitude of different methods and techniques. Recent surveys (Truong & Venkatesh 2007; Money & Agius 2008) on video summarization provide a comprehensive classification and reviews of various summarization techniques. In addition, Hu et al. (2011) has made an extensive round up of the developments during past four years on content-based video indexing and retrieval in which video summarization has been included. These state-of-art techniques have been usually developed for a specific type of video content, such as movies, documentaries, commercials, sports, news, and video surveillance. In such domain-dependent video summarization it is common that specific features exploited, which are identified and available for that particular kind of video domain. In this report we will, however, focus on summarization of consumer videos, which are missing the temporal structure, and technical quality of professionally produced and edited video content.

Compared to domain-dependent abstraction of professional video content, such as news (Albiol et al. 2003, Choi & Jeong 2000) or sports (Tan et al. 2000, Rui et al. 2000, Babaguchi 2000) where the structure can be pre-defined and exploited in the analysis, summarization of consumer videos is much more ambiguous. Methods that we can apply on user-generated video content have to be more generic, suitable for any kind of scenes regardless of their content. Implementation of machine vision algorithms that generalize for a wider domain is still very challenging. There are in fact a multitude of researchers working with the problems that are inherited from the core challenges of finding a fundamental
solution for semantic interpretation of visual information. Fortunately according to Snoek and Smeulders (2010), broad categorization of image content as the first step for further analysis is within reach.

In the next chapters we will start from the first step of making a video summarization, structural video analysis and present the techniques that can be utilized in user-generated video content. After structural analysis, we present some of the most promising methods for representing the actual video summarization. Finally we present our recommendations for implementing a video summarization tool that can be effortless and efficiently used for monitoring of user-generated content for news gathering and publishing.

4 Structural video analysis

4.1 Temporal video segments

On its lowest level a video is a sequence of still images that are perceived as changing and moving visual sequence when the frames are rapidly projected together. On a syntactic and semantic level video sequence is, however, much more complex aggregate than a sequence of still images. To understand the structure of video sequence more analytically, it can be divided into different temporal segments, which represents the video as a hierarchical composition of syntactic and semantic units. These units are frame, shot, scene, and program, which comprise a hierarchy of video segments (in Figure 1). A video program (film, commercial, news program) is composed of various scenes that are based on video shots, which in turn are a composition of adjacent frames.

![Temporal video segments diagram](Figure 1. Temporal video segments (adapted from Petershon 2007))

According to Petershon (2010) temporal video segment can be defined as an unbroken part of a video stream for which a given set of properties is consistent. Thus, adjacent temporal video segments are always mutually similar. The similarity is just affected by varying properties that can be visual, structural, or semantic. Similarity properties can be also related to how the video was produced, for example, a video shot has been traditionally defined as a temporal segment of
consecutive frames that have been captured during a single camera operation. Due to the strong correlations between adjacent frames within a shot, it has been commonly used as a fundamental unit when a video sequence is analyzed for further processing.

In computational shot boundary detection the boundaries have been generally classified with video cuts, where transition between successive shots is abrupt, or with gradual transitions including dissolve, fade in and out, and wipe, which is stretching over a number of frames (Hu et al. 2011). Hence, the shot boundary detection, as a research topic, has been directed into analyzing of edited video sequences as a pre-processing step for content-based video retrieval and browsing. But the shot detection of more common transition effects is no longer seen as challenging problem. Thus, according to Smeaton et al. (2009) the average of 79% for precision and recall is enough, and the computational shot detection has therefore noted as solved. This conclusion was made based on notices taken on the latest novel approaches to shot boundary detection that did not improve their performance against state-of-art methods in annual TREC Vid challenge\(^1\). Due to these achievements it was decided that shot boundary detection task was no longer needed and it was discontinued from year 2008.

However, as was noted earlier, traditionally shot boundaries are not defined by edited frame transitions but camera operations. Thus, shot boundary detection algorithms are rarely useful for raw, non-edited user-generated video sequences that usually comprise only one shot. Therefore, the structural analysis of user-generated video content has to focus on a sub-shot level. Similarly, as Petershon (2007) has noted, structural hierarchy should be extended with a sub-shot level, because scene changes can occur during a continuous camera act due to camera movement. According to Petershon (2007), a sub-shot is defined as “a sequence of consecutive frames showing one event or part thereof taken by a single camera act in one setting with only a small change in visual content”. User-generated video sequence should be segmented into sub-shots for extended analysis.

### 4.2 Methods for sub-shot segmentation

There are usually three main steps in general shot boundary detection: extraction of features from adjacent frames, similarity measurements between features, and detection of shot boundaries based on interpretation of feature similarities. Color histogram has been commonly used as a baseline for evaluating more advanced features. According to Hu et al. (2011) color histograms are robust to small camera motion, but detection of shots that belong to same scene is challenging for them, and they are sensitive to large camera and object movement. Especially changes and disturbances in illumination causes problems for color information based methods because they cause variation in visual content that is falsely interpreted as shot boundary (Yuan et al. 2007). Compared to color histograms edge features are more robust for video sequences with illumination changes and motion, whereas motion-compensated or actual motion features can characterize object and camera motion for extended analysis.

Due to the very nature of dynamic videos motion information is very important for content analysis as it describes the temporal dimension of the visual content (Ma et al. 2011). Compared to static image features, such as color histogram, motion information based features are closer to actual semantic concepts.

Motion information of a video sequence includes both camera motion and object motion. Camera motion is visible as background motion, whereas moving objects causes foreground motion. According to Ma et al. (2011), motion-based features are therefore usually either camera- or object based. Camera-based features, such as in recent study by Wang and Ngo (2011), extract the basic types of camera movement such as panning left and right, tilting up and down, zoom in and out, and parallax movement of camera itself (dollying).

The study by Wang and Ngo (2011) focused in summarizing rush video sequences, i.e. raw video footage produced in professional broadcasting and filmmaking industries, which are used to compose the final video product. Compared to professional movie production consumer videos are obviously very different, especially in camera work, professional editing, and lack of redundancy, which partly comes along with the editing. Compared to rushes, consumer videos are also raw footage with a lot of redundancy. But it is caused by different factors. The comparison of rushes videos to consumer videos and movie productions is presented in Table 1.

**Table 1. Comparison of video domains (adapted from Wang & Ngo 2011).**

<table>
<thead>
<tr>
<th></th>
<th>Consumer video</th>
<th>Movie product</th>
<th>Rushes video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera work</td>
<td>Amateur</td>
<td>Professional</td>
<td>Professional</td>
</tr>
<tr>
<td>Editing</td>
<td>None or little</td>
<td>Professional</td>
<td>None</td>
</tr>
<tr>
<td>Redundancy</td>
<td>A lot</td>
<td>No</td>
<td>A lot</td>
</tr>
</tbody>
</table>

Redundancy in the rushes videos is due to repetitive camera work, such as taking multiple shots of a scene when actor forgets his lines or for other reasons (Wang & Ngo 2011). Consumer videos, on the other hand, contain redundancy that is caused by lack of pre-planned camera work, which occurs as unnecessary camera movements. In addition, both consumer and rushes videos contain unintentional camera movement. Unintentional camera movement, such as camera shakiness, is much stronger in consumer videos compared to rushes as they commonly recorded with hand-held cameras. Rushes also contain shaky camera movement, but due to different reasons, as for example, when a cameraman changes camera parameters between shots. In professional production these shaky footages, in addition to other redundant content, are replaced with stock footages in a final product. Wang and Ngo (2011) do pose a relevant question on how to classify useful footages for potential reuse. This is a relevant question for consumer videos as well, and especially for publishing purposes. Though explicit retakes are not
that common in consumer videos there are still similar redundancy concepts such as back and forth panning of the same scene.

Understanding of the camera motion is an important step before higher-level semantic analysis of video sequence. After segmenting a video sequence according to camera motion we can analyze and describe these segments further with other methods. In addition, as the camera motion reflects the intentions of a camera user, and therefore the content of interest in a video sequence, the camera actions provides a subjective foundation for further objective analysis. This is especially important, as the quality and efficiency of the video summarizations are usually subjectively biased.

For consumer videos that have exceptionally heterogeneous content, we propose that the most generic feature with lowest content-dependency is camera and object motion. Extraction and analysis of motion information including both camera and object motion should be the first step in analyzing consumer videos.

Analysis of motion information has however many challenges. One is the issue with unintentional camera motion, which should be separated from intentional camera movements. Unintentional motion, such as shakiness is very common in consumer videos that are usually recorded with hand-held cameras. This motion noise must be compensated before we can extract motion features. Computational video stabilization methods that can be utilized for motion compensation have recently evolved on a new level where they can effectively remove unintentional motion. For example, the recent subspace video stabilization method incorporates promising techniques for modeling unintentional camera motion and content-preserving frame warping for its compensation between video frames (Liu et al. 2011).

5 Motion information –based video summarization

5.1 Motion saliency

Motion features have been extensively studied in various researches. In Ma and Zhang (2002) they modeled generic motion attention for video skimming. The attention model was based on an assumption that when motion vectors are processed through three motion inductors they will form a saliency map: Intensity inductor, Spatial Coherence inductor, and Temporal Coherence inductor. The saliency map can be then used for reducing the gap between low level visual features and human perception, and thus, facilitate the computational video attention detection. The motion estimation is carried out using the motion compensation vectors in MPEG video stream, where they are used to encode P- and B-frames out of intra-frames (I), which are can be decoded independently other frames in video sequence. Motion compensation vectors are useful for estimating motion since their spatial layout can be formed into a motion vector field. The motion vector field is used as the input for computation of inductor responses.
The first motion inductor, intensity inductor induces motion energy or activity, which is computed as a magnitude of motion vector that is normalized with maximum magnitude. The second inductor, spatial coherence inductor induces spatial phase consistency of motion vectors, which can infer to belonging to same moving object having high attention value. The spatial coherence of motion vector is computed as the entropy of probability distribution of spatial phase histogram. Finally, the temporal phase coherency is computed similarly as spatial coherence, as the entropy of probability distribution of temporal phase histogram. These three inductors are then integrated into a motion saliency map that can be utilized in visual video skimming by indicating salient sub-sequences in a video shot (Figure 2). (Ma & Zhang 2002)

Estimation of motion in

![Figure 2. Visual skimming based on motion attention (Ma & Zhang, 2002)](image)

Another motion analysis-based video abstraction technique has been presented by Liu et al. (2003), in which they use a triangle model of perceived motion energy to model motion patterns in video sequence for abstraction. The triangle model of extracted motion energy segments is first used segment a video sequence into meaningful action events based on motion patterns. The motion energy is calculated as a product of average magnitude of motion vectors and ratio of dominant motion direction. Consequently, the perceived motion energy is a metric that gives adaptive weight for motion intensity according to its distinct characteristics with more emphasis on dominant motion. In addition, by using the triangle model, which is illustrated in Figure 3, a video sequence can be segmented into sub-segments of different motion patterns in terms of acceleration and deceleration. These turning points of dominant motion, which can be also defined as motion states, are selected as keyframes that are included in video abstract.
In more recent work by Abdollahian et al. (2010) a new concept known as camera view was presented for the temporal segmentation of user-generated video. The authors’ define the camera view as the basic unit of user-generated video structure due to its one-shot characteristics. The concept itself builds on a notice that user-generated videos are commonly edited while recording. As in professional production there are usually multiple shots that are cut and edited to include only an interesting segment, in user-generated video on the other hand, it is common that camera is zoomed or panned to follow an interesting object. Authors argue that this kind of content-based camera motion is not typical in professionally produced video where the same effect is obtained through video editing.

The motion estimation was carried out using the Integral Template Matching algorithm (Lan et al. 2003), which was developed for fast extraction of dominant motion in video sequence. The dominant motion model is defined as

\[
\begin{bmatrix}
  x' \\
  y'
\end{bmatrix} = R \begin{bmatrix}
  x \\
  y
\end{bmatrix} + \begin{bmatrix}
  H \\
  V
\end{bmatrix}
\]  

(1)

where \((x,y)\) and \((x',y')\) are the matched pixel location in current and reference frame, respectively. Parameters of the model, horizontal (H), vertical (V), and radial (R), are estimated using the Integral Template Matching algorithm, which uses a template \(T\) in the current frame is matched against the previous frame using a full search in a predefined search window (in Figure 3). By excluding the central area of a frame algorithm can exclude object motion that is commonly located in the center of the frame. In addition to the central area, also the border areas of a frame are excluded, as pixel value errors due to video compression and camera aberration is common in these areas. (Abdollahian et al. 2010)
Novelty of the camera view-based segmentation is in how it structures a video sequence into different views that a camera is “observing” during panning of a scene. The segmentation technique is based on measuring a displacement of camera with eight-dimensional motion feature. The motion feature is based on work by Wu et al. (2005) in which they extracted feature vectors that are defined by three basic properties: speed, direction, and acceleration. These properties were formed into four dimensions on x- and y-axes: average speed of translational motion, average acceleration, variance of acceleration, and frequency of direction change. In Wu et al. they used Support Vector Machines to classify video quality properties of a video sequence.

Similarly in Abdollahian et al. (2010) they used trained SVM classifiers with eight-dimensional motion features to label set of frames as camera zooming, motion blur caused by fast camera motion, and shakiness due to unintentional camera motion. After motion labeling segment borders are defined such that the frame before the segmented camera view and successive frame are not correlated, i.e. they are not overlapping each other. The correlation in this case is defined as a magnitude of a displacement vector between two frames. The displacement vector between frames $i$ and $j$ is defined as $D_{ij} = [D_{ij}^x, D_{ij}^y, D_{ij}^r]$, where $D^x$, $D^y$, and $D^r$ are total horizontal, vertical and radial inter-frame displacements. The camera view boundary frame is identified whenever a displacement vector between current frame and previously detected boundary frame is larger than $T_d=1$.

After the detection of boundaries of video segments corresponding keyframes are identified according to simple rules that were formulated based user study. In the study (Abdollahian et al. 2010) a number of users were asked to select a set of representative frames from several user-generated video clips that where segmented with their the camera view algorithm, while being able access to individual video frames as well. In the test setup they gave following statement to the users, “If you wanted to summarize the important content of the video segment in minimum number of frames, which frame(s) would you choose?” The user also had the option not to select any frames if they thought the segment had no significant information. Users were also asked to provide the reason why they selected each of the keyframes.

![Figure 4. White area of the rectangle represents the motion parameter estimation template (Abdollahian et al. 2010)](image)
According to the results they made the following conclusions. User's preferred to select frames when there was a "nice" view of the object of interest, a new action occurred, after a zoom-in (close-up view), after a zoom-out (overall view), a new semantically significant object entered the camera view (text, people, buildings), there was a pause after a continuous camera movement or a significant change in the background. In addition, in cases when there was high blurring or shaky movements, or when the camera motion did not change and the scene was static (e.g. camera pans to the left on a homogenous background).

According to Abdollahian et al. (2010) following rules were used to select a keyframe:

- Frame after a zoom-in, which is usually a close-up view of interesting object
- Frame after a large zoom-out that gives an overview of the scene
- Frame after a camera stops after movement, and when camera is on hold, which are both indicative patterns
- Those segments where a camera has constant motion, all frames are equally as important, and therefore the frame closest to the middle of a segment with least amount of motion is selected as the keyframe

In addition to camera motion–based keyframe detection Abdollahian et al. (2010) presented also a novel location-based saliency map for analyzing the keyframes for their region of interest (ROI). The location-based saliency map is a combination of color contrast saliency map, moving objects saliency map, and highlighted faces. The color contrast saliency map is based on work by Ma & Zhang (2003), in which they used the color components in LUV-colorspace as a stimulus for calculating saliency maps for images. According to Abdollahian et al. (2010) the RGB color space yielded to better results as it combines the luminance and color contrast. The technique begins by using K-nearest neighbor algorithm in clustering color vectors of pixels into 32 color clusters. After color clustering, the image is down-sampled by a factor of 16 in each dimension in order to reduce the computational complexity. Finally, the contrast saliency map, including each pixel in the down-sampled image, is calculated with a sum of weighted distance between the pixel and its neighborhood.

The moving objects saliency map is generated due to assumption that regions with significant motion in relation to background motion attract more attention in viewer. Generation of the map follows the same principles as in previously discussed motion inductors by Ma & Zhang (2002).

For keyframe selection Abdollahian et al. (2010) argue that even previously presented simple selection strategy, such as selecting one keyframe per detected segment, will make it possible to cover all the camera views. Hence, a video summary will include all the scenes captured by the camera.
5.2 Camera motion characterization

In work by Guironnet et al. (2007) Transferable Belief Model (Smets & Kennes 1994) is applied for characterizing motion states into different camera operations: pan and tilt, zoom and static camera motion. The belief model itself builds on a structure that can process imprecise data by combining various sources of information and managing the conflict between the sources. Analysis of camera motion comprises three steps, motion parameter extraction, camera motion classification, and description of motion.

Camera motion extraction (Guironnet et al. 2006) is accomplished with estimation of an affine parametric model for dominant motion between successive frames. The affine parameters are then converted into symbolic values in classification process that builds on rule-based system that is divided into three stages. In the first stage previously gathered data is combined for obtaining frame-level belief masses on camera motion, which are filtered according to transferable belief model to ensure the temporal coherence of the belief masses.

In the second stage, static and dynamic sub-segments are separated, and in the third stage, more detailed analysis of motion is carried out with a temporal integration, which allows the preservation of frames with significant motion magnitude and duration. Thus, it allows motions to be studied on segment level by gathering frames with a certain belief in a type of motion. The camera motion description phase is finally carried out with the extraction of different motion features on each video segment, which contain characterized camera motion. According authors’ evaluations performance of the TBM-based method is very good (precision 100% and recall > 92%) for single camera motions, such as translation, zoom and static motion. Performance for composed motions where the motions can be superimposed (zoom and translation) or successive in the same extract was reduced to 79 % (Guironnet et al. 2006)

After the classification of different camera motion in video sequence, the keyframe extraction for video summary is performed with three techniques. First, the keyframes selection is done according to succession of camera motions based on heuristic rules. The rules imply that only two frames are selected to describe the succession of two camera motions. In addition, if one of the segments is static, the two frames are selected at the beginning and at the end of the segment that has camera motion. One of these frames will then represent also the static segment. (Guironnet et al. 2006)

The second technique for selecting keyframes is magnitude of camera motions. This rises from an intuitive assumption that, for example, a translation motion with a strong magnitude will require more keyframes to be represented than a static segment, because the visual content changes more rapidly. The magnitude is defined as a total displacement of motion between first and last frame of a segment. The displacement itself can be defined as a total length of motion vector trajectory or as absolute total displacement between first and last frame. Here, Guironnet et al. (2005) uses a definition of rectilinear, which measures the cumulative change of direction of motion in a trajectory. In case the length of trajectory is longer than its absolute total displacement, the trajectory is not
rectilinear and the motion changes directory, such as in Figure 5. In addition, if the total displacement is large, the frames of the beginning, the middle, and the end of the segment are selected as keyframes. In case the magnitude is weak, only the last frame of the segment is selected. Whereas if the motion trajectory is considered as rectilinear, i.e. the total displacement is closer to the length of the motion trajectory, and the total displacement is large, then the first and the last frames of the segment are selected as keyframes. If the total displacement is weak, only the last frame is selected. The rectilinear coefficient is defined as,

$$c_r = \frac{(dt - td)}{dt}$$

where $dt$ is the total length of a trajectory and $td$ is the total displacement.

![Figure 5. Total displacement of a trajectory compared to its cumulative length. (Guironnet et al. 2005)](image)

The third technique combines the first and second techniques for detecting static segments. If the magnitude of camera motion in a segment is weak or it has a short duration the segment is defined as static. In addition, if the translation related to its duration is weak and the duration is longer that a specific limit, the current segment is defined as static.

### 6 Presentation of video summarization

#### 6.1 Static and dynamic presentation

Video summarizations are commonly (Truong & Venkatesh 2007, Ma et al. 2011) presented as a set of static keyframes or dynamic videos skims. After extracting the keyframes of video sequence there are different options for presenting them to the user. One of the most common video summarization presentation techniques is a storyboard, which is usually a static grid of extracted keyframes. According to a recent study on evaluation of video summarization techniques (Westman 2010), the storyboard has a capability to give an informative summary of the original video content. However, according to the user studies the storyboards lacked in their representativeness and ability to replace the original video content.
Dynamic video skimming is a technique that condenses the original video into a shorter version, while preserving important content with its time-evolving properties. Hence, video skims are practically short video clips cut from the original video sequence. Preservation of motion information is one of greatest advantages of video skims, in addition to aural information, which can both enhance the expressiveness of the video summary. (Truong & Venkatesh 2007, Ma et al. 2011)

Compared to static storyboards, dynamic videos skimming also support the recognition of objects in the content, and their representativeness is enough even for replacing the original video content. According to the user study (Westman 2010) the dynamic video skims were liked especially due to the clarity in presentation and the normal pace of the moving imagery.

Generation of dynamic and static video summaries have been usually carried out differently, but it is still possible to transform from one form to another. Whereas video skims can be created from keyframes by joining fixed-size segments, sub-shots, or the whole shots they are included in, the set of keyframes for static storyboard can be created from a video skim by uniform sampling or selecting one frame from each skim. (Truong & Venkatesh 2007)

### 6.2 Hierarchical presentation

In addition to static and dynamic video abstracts also a technique for hierarchical abstraction has been studied (Geng et al. (2006). The hierarchical abstract is a multilevel video summarization that is based on the importance of structural units in video sequence that is defined by visual and aural attention levels. According to Geng et al. (2006) the hierarchical video summarization approach can provide different levels of granularities for video abstracts: scene level, shot level, and sub-shot level. The granularity of a video structure is defined with the importance rank, which controls the skim ratio and the keyframe ratio that are used for constructing a structure-level video summary. According to Geng et al. (2006), utilization of an importance rank on structural level provides more comprehensible and compact video summaries, which are also flexible in such that users can also tune the skim and keyframe ratios.

Calic et al. (2007) has also presented a novel approach for user-centered video abstraction in which they exploit the universally familiar narrative structure of comics to enhance the intuitiveness and readability of generated video abstracts. The abstract representation follows the narrative structure of comics by linking the temporal flow of video sequence with the spatial positions in a comic strip, while the significance of video structure unit is expressed with the size of its representation. The importance ranking is defined with an application-depended cost function.

In Calic et al. (2007) their keyframe extraction method is based on spectral clustering approach where keyframes are clustered based on their similarity, which is defined with HSV color histogram –based visual feature. The cost function is used to weight keyframes based on their distance to cluster centre. As
a result, those keyframes that are concentrated around cluster centre are treated as repetitive, redundant content, whereas cluster outliers are presented as more important. According to Calic et al. (2007) these outliers are usually cutaways, establishing shots, and other content that should be highlighted in order to generate video summaries for professional video content. Comic-like video summary is then assembled with a layout structure that follows the values of the cost function using panel templates that describe the frames sizes based on their significance. In the layout structure repetitive content is always presented by the smallest frame size, whereas outliers are present with the largest frame size. An example of comic-like video summarization is presented in Figure 6.

![Image of comic-like video summarization](image)

**Figure 6.** Comic-like video summarization of a news sequence from the TRECVID 2006 corpus (Calic et al. 2007).
7 Discussion

Many of the video summarization techniques are trying to remove redundant content in a video sequence. Therefore they are also depended on the definition of redundant content compared to significant and interesting content. This definition is the formulated into a similarity metric that measures either visual similarity or something more generic similarity. While the similarity, or distance of data points in a space with multitude of dimensions such as space that defines an image or a video, is actually one of the greatest challenges of any application that is based on computational vision. From a computational perspective visual similarity is highly corrupted by context-depended semantic similarity, which has highly subjective characteristics depending on the type of application and content.

Whereas the user-generated video content differs from professionally produced in many aspects. Low technical quality makes it challenging for low-level machine vision algorithms, and on the other hand its unconstrained temporal structure due to the lack of higher-level quality (camera work, narration, unedited, etc.) makes it difficult for creating semantically more valid summarizations. For this reason, utilization of spatio-temporal information, or motion, is useful as it provides the subjective aspect to the content by modelling the camera motion, and therefore, the intentions of the person who recorded the video sequence. In addition to camera motion, another subjective aspect rises with the exploitation motion saliency. Modelling of the motion attention in a video sequence provides us the information for highlighting perceptually significant content that has high saliency. Hence, it is possible to make more meaningful video summarizations when these salient sub-segments are processed with a higher weight.

The use of user-generated video content will become more and more popular as a part of crowdsourcing and social media. But there is a lot of useless content that should be filtered, especially when the purpose is to locate newsworthy content to be facilitated in media. This will require laborious and time consuming monitoring before the content can be published or used as a news source. Video summarization can be exploited especially as a tool for monitoring the content by making it fast and effortless to examine the importance of a video sequence. While as it is important that the monitoring is done based on all the relevant information with a guarantee that no information gets lost before examination, we propose that motion-based techniques are currently the most reliable option for summarization. Hence, we can track any objects or areas in a video scene and make analysis on generic level based on that without knowing what we are exactly tracking of. Thus, the complete video summarization is still depended of object and scene recognition, but it is also very likely that the exploitation of spatio-temporal information is required for general visual recognition. Thus, it is very important to understand motion information in video sequences.
References


