WHAT ARE THE SALIENT KEYFRAMES IN SHORT CASUAL VIDEOS?
AN EXTENSIVE USER STUDY USING A NEW VIDEO DATASET


* Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
* Klagenfurt University, Institute of Information Technology, Klagenfurt, Austria
* University of Finance and Administration, Prague, Czech Republic

ABSTRACT

Understanding the saliency of keyframes in short casual/home-made videos containing redundant information is an important step towards the design of successful keyframe selection and summarization techniques for such videos. Therefore, we present an extensive user study focusing on saliency of keyframes in such short redundant videos. In our study, more than 200 users annotated 32 videos, altogether selecting more than 20,000 keyframes. We present the description of the user study, the utilized annotation tool and we discuss the results. We provide also a preliminary comparison of several popular keyframe selection techniques using the ground truth derived from the annotations.

Index Terms — Casual home-made videos, low-quality videos, keyframe selection, annotated video dataset

1. INTRODUCTION

Every year, the digital universe is reported to grow exponentially, where the volume of newly created multimedia data makes the most significant part of the reported growth [1]. The popularity of the multimedia data can be explained by the fact that every user can simply record and store/publish anything he/she wants. Especially video data are being recorded every day by millions of users and uploaded on various video sharing portals. Unlike professional video productions, the quantity of such casual home-made videos is an order of magnitude higher, while the quality of the videos is usually lower as the users do not spend much time on post-processing of the recorded videos. Casual home-made videos often contain shaking and blurred scenes, they have no or only a limited number of shot cuts and they typically show highly redundant contents. These characteristics represent an issue for video management and retrieval systems, because traditional approaches suitable for professional productions often fail for short casual home-made videos. Let us note that some of these characteristics can be observed also in other domains where the multimedia data are being employed, for example, in the endoscopic videos or in the car industry [2].

The highly redundant video contents require specific approaches for storage and retrieval. For example, it has been shown in the endoscopic video domain that highly redundant videos can be efficiently compressed, full HD quality is not needed [3]. The redundancy of the video contents and the lack of shot boundaries represent also a challenge for video summarization techniques that try to select the most representative key-frames from videos. Unfortunately, the research in this field is highly connected with many user studies that focus on questions like “what parts of videos are less important so their quality could be lower or what are the salient key-frames for a specific videos”.

To the best of our knowledge, there is no extensive study about keyframe selection in casual home-made videos including a publishable dataset together with ground truth obtained from a large number of user annotations. Furthermore, there are not many well annotated datasets of casual home-made videos with clearly specified ground truth for key-frame selection techniques. Therefore, in this paper we present a new available dataset of casual home-made videos (not just temporal links to YouTube) and the results of our study where each video was annotated by more than 120 users1.

2. RELATED WORK

In the general domain of multimedia research a lot of work has already been done in the area of keyframe extraction. Truong and Venkatesh [4] give a comprehensive overview of video abstraction methods. They summarize keyframe extraction techniques based on five different aspects: (i) the number of keyframes extracted, (ii) the temporal unit keyframes are extracted from, (iii) the representation scope of keyframes, (iv) the underlying computational mechanisms, and (v) the visualization of keyframes. Many keyframe selection methods sequentially compare all frames of a video and whenever a significant content change can be detected, a keyframe is selected.

The problem with such approaches is that they somehow preserve a clear structure in the videos, which is typically only existent in professional video content, like movies, TV shows and news videos. However, casual home-made videos have different characteristics than professional videos. Therefore, keyframe extraction techniques that pay attention to this fact are needed. Only a few approaches are focusing especially on casual home-made videos, but it must be distinguished between different types of casual home-made videos.

Early work focused on casual home-made videos recorded with camcorders. Such videos typically consist of different scenes sequentially recorded on a tape. Approaches for video summarization [5] [6] and video scene detection [7] have been introduced. As the videos consist of different scenes, all approaches have in common

1The dataset and results of the user study are available at http://www-itec.aau.at/ftp/datasets/short-casual-videos
that they first detect shots and then they proceed on keyframes selected from these shots.

In the last decade another type of casual home-made videos emerged: videos that are taken with handheld consumer devices (e.g., smartphones, tablets, flip video camera, etc.) and uploaded to video sharing platforms on the Internet. They are typically much shorter than camcorder videos and in most cases they only consist of a single shot. Therefore, shot detection is useless as basis for the keyframe selection and other methods must be applied.

Liu et al. [8] use a motion-based approach to detect actions or events in videos. Kelm et al. [9] select keyframes based on visual quality, presence of text and faces and color diversity. Jiang et al. [10] first detect changes in the audio stream to segment casual home-videos and then they select keyframes from these segments based on visual quality and the presence of faces. All three approaches were evaluated with user studies, where users had to rate the quality of the keyframe selection. They have in common that these studies were performed with a small amount of participants. Only 12 or even less people rated the results.

Subjective user ratings of keyframe selections are somehow problematic. It must be ensured that the number of participants is high enough to be able to make conclusions based on the results. But even if this is the case, it is still difficult to compare the ratings of two studies based on the ratings of different people. Because of this reason, Jiebo et al. [11] create a ground truth for a subset of the publicly available Kodak consumer video dataset [12] in advance. For each video, the person who took the video and three further people were asked to select keyframes. As a consequence, this ground truth is used for the evaluation of a motion-based keyframe extraction method.

However, in our perspective four people are not sufficient to generate a solid ground truth. Therefore, we created an own dataset and performed a large user study with 212 participants in order to create a ground truth that can be used for keyframe evaluation tasks on this dataset (see Section 3.1 for details). In contrast to the Kodak consumer video dataset [12], we do not only provide links to YouTube videos, but the videos themselves and in contrast to Jiebo et al. [11], we also provide the results of our keyframe selection study to the research community.

3. USER STUDY

In the following, we describe our dataset of casual home-made videos and the user study that we conducted to obtain a large corpus of user annotations. An analysis of the annotations is provided in the next section.

3.1. Video Dataset

The video dataset consists of 32 authentic casual home-made video clips which are publicly available. All of them are single-shot videos, meaning that they were recorded in one continuous shot, as it is typical for casual home-made videos. Most of the clips have a duration of about one minute (average 84 seconds). The shortest clip lasts 42 seconds while the longest clip has a duration of 219 seconds. The clips have a resolution of 640x360 or 640x480 (depending on the original aspect ratio) and a framerate of 25 fps. The content can roughly be divided into three classes: animals, landscapes/nature and city panoramas. Figure 1 illustrates a few example frames from the videos. To avoid potential privacy violations we made sure that no individuals and especially no recognizable faces appear in the video clips.

3.2. Annotation Task

In the user study, the participants were asked to select representative keyframes from each video, which “describe what is happening in the video”. It is important to note that the number of keyframes to be selected was not specified in advance and thus the users had to decide themselves about the number of interesting moments in the videos. The only clue for the users was the information that the selected keyframes should summarize the video content in order to help users during search tasks in huge collections of such videos by showing a good preview.

3.3. Annotation Tool

The user study was conducted using a custom web-based annotation tool. Figure 2 depicts a screenshot of the tool. The timeline allows for standard navigation like jumping to an arbitrary position by clicking or dragging. Additionally, it displays a preview image of the according temporal position in the video when the mouse is positioned over the bar. Framewise navigation (back and forth) is possible by clicking the according buttons. The playback speed of the video can be changed (0.5x-7x) by simply scrolling the mouse wheel in order to speed up the review process and quickly find the relevant scenes in the video. The current playback speed is visualized by a speedometer to the right of the video window. It can easily be reset to normal speed by clicking on the corresponding button.

To select a keyframe, the user just needs to click the snapshot icon or press the “space” key. Thumbnails of the selected keyframes are displayed below the video window. Their temporal position is indicated on the timeline. By clicking on the thumbnail, users can directly jump to the according position in the video. They can also jump from one keyframe to the next by clicking the according buttons. The keyframe selection can be refined by temporally shifting the image framewise with the +/- button. Moreover, a keyframe can be deleted and a comment can be stated for every selected keyframe.

3.4. Participants

We asked students of several high schools and universities in Austria (Klagenfurt University) and Czech Republic (University of Finance and Administration in Prague, Secondary technical college, Kostanova 4, Olomouc) to participate in the online study. In total, 212 different users (Considering unique session IDs) participated in the
study, alltogether selecting 20,464 keyframes. Not all of them annotated each and every video, but each video was annotated by at least 120 different users.

4. ANALYSIS OF THE RESULTS

In this section, we provide a basic analysis of our extensive user study. Since the number of users and annotation tasks were high, it was impossible to control every user during each task and thus also some misleading annotations appeared in the results. For example, there were 12 annotations with abnormally high number of selected keyframes (> 30). These annotations were removed from the corpus. Some users (about 10) also probably did not understand the task and selected always just one most representative keyframe from each video, even though there were clearly more visually different scenes.

Each video was annotated by at least 120 users, where each user selected on average about 5 keyframes per video.

4.1. Discussion

One of the goals of this user study was to determine, which keyframes are salient and suitable as a good representation of the utilized videos. As many users in this study have tried to select salient keyframes for the videos, we have obtained a rough approximation of keyframe saliency. For each video, the selected keyframes were aggregated into a normalized histogram that was used as keyframe saliency distribution for the video. In this annotation histogram one bin typically corresponds to one second of the video and the sum of all bins is equal to one. In the following text we discuss, if such saliency distributions can be simply used as a ground truth for keyframe selection techniques.

Theoretically, there are two extreme cases of keyframe saliency distributions – uniform distribution of selected keyframes and distribution where the selected keyframes form clear Gaussian peaks with low variance. In the first case, the uniform distribution of selected keyframes means that there is no agreement between users on what are salient keyframes in the video. This also means that it is unclear what the ground truth annotation for such a video should be (all keyframes or a uniformly sampled subset?). The second case represents situations, where the users always agree on several clear salient moments in the videos, which means that the ground truth keyframes can be simply identified and analyzed.

As the videos in our dataset often contain highly redundant content, we have expected a high ambiguity of the selected keyframes by users. Surprisingly, even in the most redundant videos the selected keyframe distribution was not fully uniform. In most cases, the distribution obtained some clear peaks and also some shallow hills, separated by empty bins or connected in a mixture of Gaussians (see black and orange rectangles in Figure 3). In other words, in most of the videos there were few moments where many users agreed on the clear keyframe saliency, while there were also videos where users selected different moments which resulted in shallow peaks in the histogram. The ambiguity of the keyframe selection (especially the shallow peaks) can be demonstrated also by the fact that there is on average 25% difference between normalized saliency histograms for first 60 users per video and second 60 users per video. On the other hand, both corresponding histograms often share few clear peaks representing salient moments in the video.

4.2. Observations

In the last part of this section, we present also several observations for all keyframe saliency histograms where each bin corresponds to one second of video.

- Whereas the thin peaks have often indicated a unique action with a high entropy (something that has not happened yet), the shallow hills often indicated either long lasting actions (e.g., cougar playing with box) or constantly changing signal without significant keypoints (e.g., panorama of the resident district of San Francisco).
- Many peaks had also tails on the left and/or on the right (reminiscent of the normal distribution) as users often did not agree on the exact moment of the actions.
- In the results, the users often tried to select the most distinct moments of each video, regardless the complexity of the videos. Even if the video contained monotonously repeating actions (sea waves, stream and blurred fish), the users have adaptively focused on unique actions.
- Longer videos were annotated with less keyframes per minute and in many annotations, there are still empty bins.
- The first and last keyframes of the videos were not selected often, mainly in cases where a salient moment is shown.
• The video category often did not determine, if the annotation was or was not highly ambiguous.

5. AUTOMATIC KEY-FRAME SELECTION

We have implemented a few keyframe selection methods and evaluated them with the ground truth data described above. In the following we describe these methods, how we evaluated them, and discuss the achieved performance.

5.1. Motion Flow

The first keyframe selection method is based on optical flow estimation according to the pyramidal implementation of the Lucas-Kanade feature tracker [13]. Our implementation starts with dense sampling of keypoints \( P \) for frame \( f_i \) and tracks them over successive frames \( f_{i+k}, f_{i+k+1}, \ldots \) over the current segment of length \( s \). Segment boundaries are detected at heavy motion changes, as this method assumes that important keyframes are located at those positions where the content significantly changes. Due to motion in the video some keypoints eventually move out of the frame and cannot be tracked further (we call these keypoints non-trackable keypoints \( P_N \)). As soon as the number of non-trackable keypoints reaches a certain threshold (e.g., \( \frac{|P_N|}{m} \geq 0.4 \)), the frame \( f_{i+k} \) is selected as a keyframe.

Together with the keyframe we store the priority of the keyframe through comparison with the previously selected keyframe \( f_{i-1} \) (if there is one). This priority is computed as the histogram intersection between 256-bin color histogram (over all three RGB channels) for both frames \( f_{i-1} \) and \( f_{i+k} \). We use the priority in our evaluation to test different numbers \( k \) of keyframes (see below in Section 5.5).

5.2. Color Histogram

The second keyframe selection method uses color histogram comparison only. It computes the 256-bin histogram \( H \) (over all three RGB color channels) for every frame \( f_i \) and the previous frame \( f_{i-1} \) and computes the histogram intersection \( D \) between them. If \( D \) is sufficiently different to the average histogram intersection distance \( D_{\text{avg}} \), computed over the last 25 frames as given in Equation 1, a keyframe is selected. In our measurements we use Formula 2 and \( \alpha = 0.2 \) for this check. The absolute difference between \( D \) and \( D_{\text{avg}} \) is also used as priority of the selected keyframe.

\[
D_{\text{avg}} = \frac{1}{25} \sum_{k=1}^{25} D(H_{i-k}, H_{i-k-1}) \quad (1)
\]

\[
|D - D_{\text{avg}}| > \alpha \quad (2)
\]

5.3. Motion Areas

The motion area method is an algorithm used to divide videos into small segments [14]. The frames are separated spatially into \( m \times m \) similar sized areas. For each area feature points are tracked with the pyramidal implementation of the Lucas-Kanade feature tracker [13]. Afterwards, average motion values are calculated for the areas. These average motion values are averaged again and the standard deviation is computed. The resulting average value and the standard deviation value are observed within a time window. If the values are above a certain threshold a keyframe is selected, which is taken from the center of the segment.

5.4. Uniform Sampling

The last method can be considered as the baseline and uses no content analysis at all. It just uniformly selects keyframes at a predefined step size according to the tested number of keyframes \( k \) (see subsection below). Uniform frame sampling can be computed very quickly and is often used in practice due to its simplicity. Moreover, it has been proven to provide remarkable results for keyframe selection [15] and interactive video search [16].

5.5. Evaluation Method

To evaluate the four keyframe selection methods described above, we used the annotation results described in Section 4. Since the duration of the 32 videos ranges from 42 seconds to a few minutes, there is no clear optimal number of keyframes. Hence, we tested different numbers of keyframes by selecting only the top-\( k \) keyframes for every video (except for uniform sampling where we simply extracted \( k \) frames from equidistant positions). Due to the rather short duration of the videos we tested \( k = 1 \ldots 9 \).

For every keyframe selection method and every tested value of \( k \), we compute the \( \text{Score} \) as sum of the corresponding values in the normalized annotation histogram (see Section 4.1) and compute the average from all selected keyframes, i.e., from all \( k \) considered values.

5.6. Results

Figure 4 shows the resulting performance for the different values of \( k = 1 \ldots 9 \), based on the annotation counts averaged over all 32 videos. As obvious in the diagram there is no clear winner from the tested keyframe selection methods. Although motion flow performs best for \( k = 1 \) and \( k = 2 \), it is outperformed by uniform sampling for \( k = 3 \) and \( k = 4 \). For larger values of \( k \) the performance of all tested methods is similar, though the one of motion areas is slightly lower.

However, when computing the average score over all nine values of \( k \) and comparing the scores of the four tested keyframe selection method for every video (see Figure 5), we can see that for each method the performance heavily varies with the video. For example, in average motion flow performs worst for video 14, 15, 17, 23, and 29 and color histogram is worst for videos 4, 7, 12, 20, 21, 24, 25, and 31. Motion area is last for videos 5, 6, 9-11, 13, 18, and 27 and uniform sampling achieves lowest performance for videos 1, 8, and 19.

In Figure 6 we can see example results for \( k = 3 \) and six selected videos. As obvious in the figure the advantage of uniform sampling is the fact that usually quite different frames – from different locations – will be selected (e.g., for videos 6, and 15). However, on the other hand content-based methods may select much more relevant keyframes, since content changes are considered (e.g., with motion flow for video 32 or with color hist for video 6). Also for the panorama of video 15 the keyframes provided by the motion area are more relevant than the highly different frames provides by uniform sampling (the reason why only one keyframe is shown for motion area under video 6 is the fact that this method could only provide one keyframe for the whole video). However, as apparent from the diagrams above and the examples in Figure 6, unfortunately there is no one single method with the best performance.
6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an extensive user study where 32 short casual home-made videos were annotated by 212 different users resulting in 20,464 selected keyframes. We have described the tool and methodology used for the study as well as a basic analysis of the results. Our user study has demonstrated that short casual home-made videos represent a true challenge not only for keyframe selection and video summarization techniques, but also for understanding the saliency at all. We have observed different saliency for different actions and also context-dependent saliency patterns. We believe that such saliency patterns have to be studied also from the psychological point of view.

In our preliminary results, we have used the saliency histograms to compare several different keyframe selection techniques. From this preliminary comparison, none of the techniques came as a clear winner selecting optimal keyframes based on the user annotations. In the future, we would like to investigate if the ground truth definition should be reconsidered or if the utilized keyframe selection techniques should be improved. For example, the keyframe selection techniques could measure entropy of visual/motion information in the videos and focus only on the events with the highest corresponding entropy.

Acknowledgment

This work was funded by Grant Agency of Charles University project 201515, by project SVV-2015-260222, and by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF - 20214 22573 33955.

7. REFERENCES


