Keyframe-based Video Summarization with Human in the Loop

Antti E. Ainasoja, Antti Hietanen, Jukka Lankinen and Joni-Kristian Kämäräinen
Signal Processing Laboratory, Tampere University of Technology, PO Box 527, FI-33101 Tampere, Finland

Keywords: Video Summarization, Visual Bag-of-Words, Region Descriptors, Optical Flow Descriptors.

Abstract: In this work, we focus on the popular keyframe-based approach for video summarization. Keyframes represent important and diverse content of an input video and a summary is generated by temporally expanding the keyframes to key shots which are merged to a continuous dynamic video summary. In our approach, keyframes are selected from scenes that represent semantically similar content. For scene detection, we propose a simple yet effective dynamic extension of a video Bag-of-Words (BoW) method which provides over segmentation (high recall) for keyframe selection. For keyframe selection, we investigate two effective approaches: local region descriptors (visual content) and optical flow descriptors (motion content). We provide several interesting findings. 1) While scenes (visually similar content) can be effectively detected by region descriptors, optical flow (motion changes) provides better keyframes. 2) However, the suitable parameters of the motion descriptor based keyframe selection vary from one video to another and average performances remain low. To avoid more complex processing, we introduce a human-in-the-loop step where user selects keyframes produced by the three best methods. 3) Our human assisted and learning-free method achieves superior accuracy to learning-based methods and for many videos is on par with average human accuracy.

1 INTRODUCTION

Video summarization is a key technology to manually browse through multiple long videos – a user can quickly decide whether a video is interesting or not by viewing its summary. Summaries may also have a more predominant role beyond retrieval since many users have started to produce video data of their everyday life, hobbies and even for semi-professional purposes (e.g., "How to change a tire"). In the light of this, the original idea of informative summary must be expanded to more generic summaries that are visually plausible and entertaining as such. The main challenge is how to retain necessary visual and temporal structure to "tell the original story" within the requested duration (Figure 1).

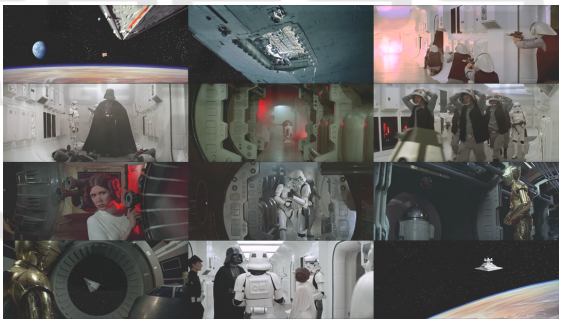


Figure 1: A static storyboard summary of the opening scene of Star Wars which can be converted to dynamic video summary by temporally expanding the static keyframes.

There have been many video summarization approaches with various objectives (see the Truong et al. survey (Truong and Venkatesh, 2007)), but their evaluation or comparison is difficult as users have various subjective preferences depending on their background, personal relation to the content, age, gender etc. Most of the works report only qualitative results or interview results from committees who have graded or ranked the generated summaries. Recently, Gygli et al. (Gygli et al., 2014) introduced the SumMe dataset of three different types of YouTube uploaded user videos (egocentric camera, moving camera, static camera) and asked several authors to summarize the videos manually to establish quantitative ground truth.

In this work, we adopt the popular keyframe-based approach to video summarization (Guan et al., 2013; Ma et al., 2005), and with the help of the SumMe benchmark experimentally evaluate important parts of the keyframe summarization pipeline. We investigate the basic workflow and introduce its human-in-the-loop variant to understand the semantic gap between unsupervised summarization and supervised manual summarization. Local features gener-

Ainasoja, A., Hietanen, A., Lankinen, J. and Kämäräinen, J-K.
Keyframe-based Video Summarization with Human in the Loop.
DOI: 10.5220/0006619202870296
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 4: VISAPP*, pages 287-296
ISBN: 978-989-758-290-5

ally perform well in detecting important scenes, but motion analysis provides complementary cue providing better keyframes. Weak supervision in the form of user-selected keyframes provides substantial performance improvement with minimal manual work. Our findings indicate that the both unsupervised and human assisted approaches are needed as the first can provide automatic summaries online and the latter can serve users who wish to convey their personal preferences or artistic view.

2 RELATED WORK

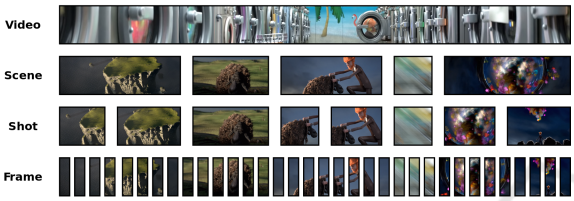


Figure 2: Temporal hierarchy of keyframe-based video summarization: from whole video (top) to single frames.

The starting point of many summarization methods is the hierarchical video structure illustrated in Figure 2 - still frames merge to continuous shots, shots merge to scenes and scenes merge to a full video. A popular approach to summarization is the keyframe-based summarization (Guan et al., 2013; Ma et al., 2005) where important keyframes are automatically detected from the input video and then a summary is constructed by temporally expanding keyframes to "key shots". The simplest method is to uniformly assign keyframes, but that will over-emphasize long shots that are not necessarily interesting. Therefore the higher level information pieces, shots and scenes (Figure 2), are more preferred starting points for keyframe detection, e.g., the middle frame of each shot/scene. Surveys for shot (Smeaton et al., 2010; Duan et al., 2013) and scene detection (Fabro and Böszörmenyi, 2013) provide overviews of the available methods. The definition of a video shot is more technical (Smeaton et al., 2010) and therefore video scenes that describe semantically and temporally coherent content are more meaningful data pieces for summarization. There are many video applications with similar objectives to summarization, for example, video thumbnail generation (Liu et al., 2015), video synopsis (temporal and spatial mixture) (Rav-Acha et al., 2006), video-to-comics (Hong et al., 2010; Herranz et al., 2012), and video-to-animated-gif (Gygli et al., 2016), but we consider only the recent video summarization and video scene detection methods. We also omit the methods that utilize metadata in summarization (Wang et al., 2012; Zhu et al., 2013; Bian et al., 2015).

Video Scene Detection – Video scene detection methods analyze both temporal and spatial structures to identify video scenes that are semantically and/or temporarily coherent. Unsupervised scene detection exploits standard unsupervised techniques such as clustering (Gatica-Perez et al., 2003; Odobez et al., 2003), Markov models (Shai and Shah, 2006) and graphs (Gu et al., 2007; Ngo et al., 2005), and can combine audio and video (Kyperountas et al., 2007). Early scene detection methods were reviewed in the video summarization survey by Truong and Venkatesh 2007 (Truong and Venkatesh, 2007). These methods were based on various imagery features and suitable for offline processing since they process the whole input video at once. Offline scene detection can be formulated as a structure optimization problem (Gu et al., 2007; Ngo et al., 2005; Han and Wu, 2011) and online detection as a reconstruction problem (Zhao and Xing, 2014) where a threshold defines whether a current scene can be reconstructed using the previous scenes.

Video Summarization – The early attempts of video summarization are surveyed in Truong and Venkatesh 2007 (Truong and Venkatesh, 2007) and these methods often use simple rules for keyframe selection and skimming. More recently, Shroff et al. (Shroff et al., 2010) optimize trade-off between coverage and diversity. Han et al. (Han et al., 2010) proposed an assisted approach similar to our work, but in their system user needs to browse through the original video and manually select keyframes.

SIFT descriptors have been used in (Lu et al., 2014) where the descriptors are weighted based on their "importance" (a Bag-of-Importance model) and importance mapping is learned from data. One of the few online methods is by Zhao and Xing (Zhao and Xing, 2014) who dynamically construct a codebook from local image and optical flow features similar to us. However, their method selects keyframes using the rule of whether new scene can be reconstructed using codes of all previous scenes - that works for informative summaries rather than for entertaining home video summaries. Lee and Graugman (Lee and Grauman, 2015) proposed a summarization for egocentric camera that optimizes summarization based on visual content and metadata (location and time). Meng et al. (Meng et al., 2016) replace keyframes with "key objects" which are found by first selecting object region proposals and then clustering these into key objects that then help to select keyframes. Gong et al. (Gong et al., 2014) and Zhang et al. (Zhang et al., 2016) introduce summarization as a supervised

problem where human made summary in the training set is transferred to unseen video. Potapov et al. (Potapov et al., 2014) use a supervised approach where user needs to define one of the pre-defined video categories and category specific summarization is then executed.

Gygli et al. (Gygli et al., 2014; Gygli et al., 2015) introduced a new video summarization benchmark SumMe and a learning based method for unsupervised summarization. Their method is based on detection of "superframes" which are similar to key shots constructed by keyframe expansion in our work and merging of the superframes by optimizing various motion and visual content based features.

Contributions – Our main contributions are:

- We propose a learning free method for video summarization using local feature (SIFT) based scene detection and motion feature (HOOF) based keyframe selection. In the human-in-the-loop setting suitable keyframes are selected by users.
- We introduce a high recall scene detection method by extending our previous static video Bag-of-Word (BoW) (Lankinen and Kämäräinen, 2013) to dynamic video BoW that is an online method and provides over segmentation to scenes.
- We quantitatively evaluate variants of SIFT local regions (Lowe, 2004) and HOOF motion descriptors (Chaudhry et al., 2009) in keyframe selection for keyframe-based summarization.

3 OUR METHOD

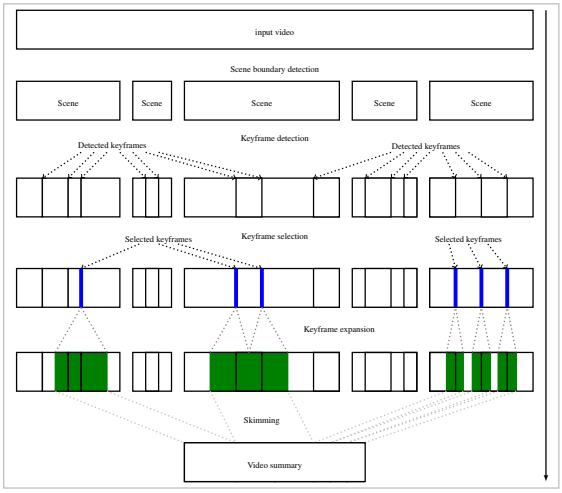


Figure 3: The summarization workflow used in this work.

The overall workflow of our method is illustrated in Figure 3. At the first stage, video scenes are detected by our dynamic video Bag-of-Words (Section 3.1). At the second stage, keyframes for each scene are selected using either local feature or motion cues (Section 3.2). In the unsupervised mode the detected keyframes are expanded to key shots proportional to the scene length (Section 3.4) and combined to form a final summary. In the human-in-the-loop mode the keyframes are presented to a user (Section 3.3) who selects the final keyframes for the summarization. Otherwise the two modes are identical.

3.1 Dynamic Video BoW

We extend our previous static video BoW (Lankinen and Kämäräinen, 2013) to dynamic video BoW. The static version uses a fixed BoW codebook for a whole input video while our dynamic video BoW constructs a new codebook for every detected scene. A static codebook can be global (fixed globally) or video specific (codes computed from an input video). Video specific codebooks were clearly superior in (Lankinen and Kämäräinen, 2013), but they require offline processing while our algorithm is online (video can be processed simultaneously with uploading). Moreover, in our experiments our dynamic BoW performs better in the high-recall region (recall ≥ 0.90) of the precision-recall curve.

Our descriptor of choice is the dense SIFT that performed well in video event detection (Tamrakar et al., 2012), but can be replaced with any local feature detector-descriptor pair in OpenCV with negligible loss in accuracy. We process the input video V as frames $V = \{F_i\}_{i=0,\ldots,N}$ where N together with the frame rate (fps) define the video length in wall time. We further divide the frames into non-overlapping blocks (short processing windows W_j) of N_W frames each $W_j = \{F_{j*N_W}, F_{j*N_W+1}, \ldots, F_{j*N_W+N_W-1}\}$ where the number of frames $|W_j| = N_W$ corresponds to one second. Our dynamic video BoW processes the windows and either merges them together to a single scene or assigns a scene boundary.

Each frame is rescaled to 300×300 image and dense SIFT descriptors of the size 20 pixels with 10 pixels spacing are extracted. In the initial stage and after each scene boundary detection a new BoW codebook of 100 codes is generated using the standard c-means algorithm. A BoW histogram $H(W_c)$ is computed for the current window W_c using the codes from all frames. Then processing advances to the next window W_{c+1} for which the BoW histogram $H(W_{c+1})$ is computed and compared to $H(W_c)$. For histograms we adopt the L1-normalization and for histogram comparison we use the L2 distance. If the distance is $\geq \tau_{BoW}$ then a scene boundary is detected, otherwise the windows are merged into the

same continuous scene and the method jumps to the next window (W_{c+1}). Whenever a scene boundary is detected the codebook is recomputed and the process continues until the whole video has been processed. In our experiments we fixed the threshold to $\tau_{BoW} = 0.76$ that was found to provide high recall with the TRECVid 2007 dataset used in (Lankinen and Kämäräinen, 2013).

3.2 Keyframe Detection

The output of the previous step in Section 3.1 is a set of adjacent scenes $\{S_k\}$ that consist of one or multiple windows which again consist of N_W frames each, i.e. $S_k = \{W_j, W_{j+1}, \ldots\} = \{F_{j*N_W}, F_{j*N_W+1}, \ldots, F_{(j+1)*N_W}, \ldots\}$. The goal of this step is to select one or multiple "keyframes" for each scene, $F_{S_k,1}, F_{S_k,2}, \ldots, F_{S_k,i}$, that describe the spatio-temporal content of the scene S_k (Figure 1).

The keyframe detection serves two purposes; Finding the frame which best describes the content of the scene and finding the temporal location around which the most interesting things in the scene happen. We tested several techniques to detect these frames. In its simplest form, "baseline", we picked the middle frame from each scene. In order to better analyze the video content we experiment dissimilarity given by local region descriptors (SIFT) during scene detection or optical flow based motion analysis (HOOF motion descriptors (Chaudhry et al., 2009)).

The dense SIFT performed well in scene detection and since the descriptors are available from that stage it is justified to adopt them for keyframe selection as well. We can compute a scene BoW histogram using various methods (average, median) and compare each window histogram to that. For keyframe selection also various strategies exist: the most similar window to the scene histogram or the least similar (Figure 4). In the experimental part, we tested these variants and effect of the frame rate (15 fps, 24 fps and original fps).

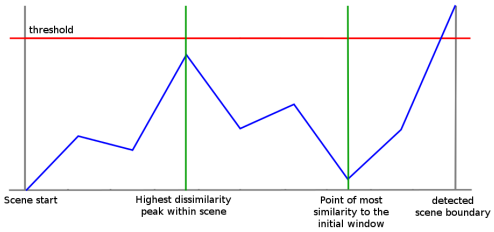


Figure 4: Video BoW dissimilarity within a scene.

Motion Analysis – We based our motion analysis on the popular Farnebäck's (Farnebäck, 2003) dense optical flow (Figure 5). We tested four global criteria in keyframe selection: frame with the minimum flow, the maximum flow and the frames having the amount of flow nearest to the average and the median flow of the scene. Moreover, to account for the direction of the motion, we also used motion histograms (Figure 5 bottom). We computed motion histograms by assigning the optical flow vectors to 8 discrete bins according to their direction and summing up the magnitudes within each bin (Chaudhry et al., 2009). The use of the motion histograms enabled us to find the frames that most or least resemble the average motion of a scene.

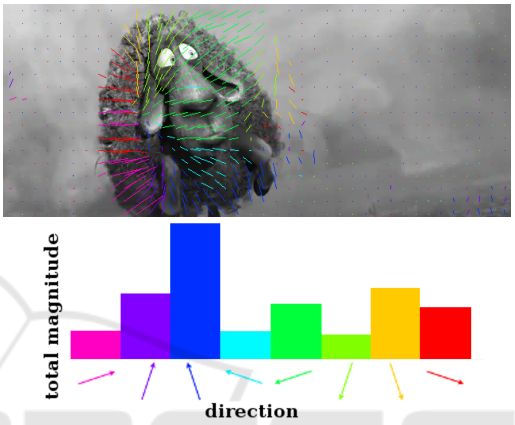


Figure 5: Computed optical flow vectors in a frame and the corresponding optical flow histogram (below).

We experimented with several histogram distance functions (1)-(5). For example, chi-square is weighted so that distance is relational to the overall magnitude and the amount of difference is considered more meaningful when magnitudes are low. Bhattacharyya as well weights distances, but it is done based on the mean value of all histogram bins rather than per bin basis. Alternatively, correlation and intersection distances compare the shapes of the histograms disregarding the magnitudes.

$$d(H_a,H_f) = \sqrt{\sum_{n=1}^{N}(H_a(n)-H_f(n))^2} \quad \text{L2} \quad (1)$$

$$d(H_a,H_f) = \sum_{n=1}^{N}\frac{(H_a(n)-H_f(n))^2}{H_a(n)} \quad \text{chi-square} \quad (2)$$

$$d(H_a,H_f) = \sqrt{1-\frac{1}{\sqrt{\bar{H}_a\bar{H}_f}}\sum_{n=1}^{N}\sqrt{H_a(n)H_f(n)}} \quad \text{Bhattacharyya} \quad (3)$$

$$d(H_a,H_f) = \frac{\sum_{n=1}^{N}(H_a(n)-\bar{H}_a)(H_f(n)-\bar{H}_f)}{\sqrt{\sum_{n=1}^{N}(H_a(n)-\bar{H}_a)^2\sum_{n=1}^{N}(H_f(n)-\bar{H}_f)^2}} \quad \text{correlation} \quad (4)$$

$$d(H_a,H_f) = \sum_{n=1}^{N} min(\frac{H_a(n)}{|H_a|_1},\frac{H_f(n)}{|H_f|_1}) \quad \text{intersection} \quad (5)$$

3.3 Keyframe Selection

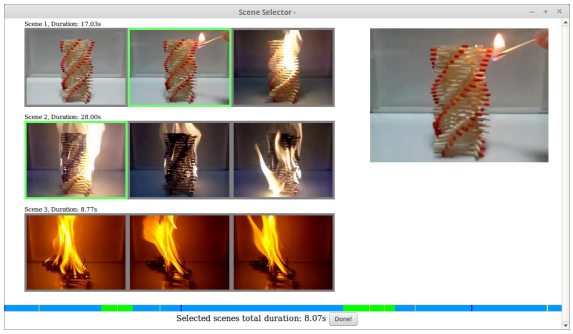


Figure 6: Screenshot of our GIU for keyframe selection in human-in-the-loop video summarization.

For human-in-the-loop video summarization, we chose the three best performing unsupervised keyframe detection methods (one for each video type in SumMe). One to three (removing replicates) keyframes were presented to test subjects with a simple graphical user interface (Figure 6) that resembles a static storyboard. Users were able to see a larger preview of a keyframe by clicking the frame. Test subjects selected their preferred keyframes by double-clicking them (highlighted with green borders in Figure 6). A timeline of the original video and the summary based on the keyframe selections were shown at the bottom of the user interface window.

3.4 Expansion of Keyframes

We expanded the selected keyframes into key shots by taking video content around them. We divided the target duration into each scene that had at least one keyframe selected. The duration allocated to each scene was proportional to the duration of the scene. We distributed the allocated duration of each scene uniformly to the selected keyframes forming shots around each keyframe. We then shifted and combined the acquired shots as necessary to avoid overlapping and to ensure each shot stays within the boundaries of the originally detected scene. Finally, we combined the shots to create a dynamic summary (Figure 7).

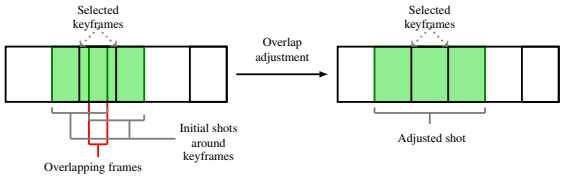


Figure 7: Overlapping shots around the keyframes are combined into a single shot in the keyframe expansion.

4 EXPERIMENTS

4.1 Data and Performance Measures

There are many benchmark datasets available for video processing and analysis tasks, for example, Open Video Project (OVP, open-video.org), Kodak Consumer Video (Kodak (Yanagawa et al., 2008)), Youtube (de Avila et al., 2011a), CUS (Comparison of User Summaries (de Avila et al., 2011b)), Columbia Consumer Video (CCV (Jiang et al., 2011), an extension of Kodak), TRECVid (Over et al., 2015) and SumMe (Gygli et al., 2014). OVP, Kodak, Youtube and CUS provide user-selected keyframes as ground truth and can be used in benchmarking static storyboard summarization methods (Luo et al., 2009; Gong et al., 2014). Potapov et al. (Potapov et al., 2014) constructed a MED-summaries dataset from TRECVid 2011 where importance scores were introduced for different video events (e.g., a kid blowing birthday cake candles has high importance score in the birthdays category) and used it in the evaluation of dynamic summaries. Only the SumMe dataset by Gygli et al. provides user made video summaries as ground truth and therefore we selected SumMe for our experiments. Moreover, since multiple (10-15) summaries are provided for each video this allows also to investigate the effect of subjective variance.

SumMe contains 25 user videos with little or no editing. The durations of the videos range from 30 to 240 seconds totaling 1 hour, 6 minutes and 18 seconds. 4 of the videos are recorded using egocentric, 4 using static and 17 using moving cameras. Each video has 15 to 18 ground truth summaries with lengths of 5% to 15% of the original duration. The ground truth summaries were manually created by human test subjects in a controlled psychological experiment. The dataset is used to evaluate the quality of a summary by computing per-frame pairwise f-measure

$$F_s = \frac{1}{N}\sum_{i=1}^{N} 2\frac{p_{is}r_{is}}{p_{is}+r_{is}}, \tag{6}$$

where N is the number of ground truth summaries, p is the precision and r the recall of the summary s being evaluated. The precision p is computed according to

$$p = \frac{|n_{gt} \cap n_s|}{|n_s|} \tag{7}$$

and recall r is computed as

$$r = \frac{|n_{gt} \cap n_s|}{|n_{gt}|}, \tag{8}$$

where n_{gt} is the number of frames in the ground truth summary and n_s the number of frames in the summary

Table 1: Comparison of approaches for video summarization using SumMe videos (per-frame pairwise f-measure performance in (6)): average human performance computed against the SumMe ground truth, Gygli et al., random frame selection, and our keyframe-based summarization (Section 3.2). We also mark the set of parameters for which the best result was achieved.

[a](Gygli et al., 2014)

	Video	*Human-avg*	*Gygli*[a]	*Rand*	*Our*	*Motion*	*SIFT*	*Dist*	*Hist*	*Crit*	*fps*
Egocentric	Base jumping	0.646	0.304	0.362	**0.729**	✓		L2	-	med	orig
	Bike polo	0.640	**0.708**	0.266	0.553		✓	Int	avg	min	24
	Scuba	0.561	0.475	0.357	**0.651**	✓		corr	med	max	orig
	Valparaiso_downhill	0.637	0.567	0.333	**0.698**	✓		Int	avg	max	orig
Moving	Bearpark_climbing	0.630	0.358	0.445	**0.818**	✓		Int	med	max	orig
	Bus_in_Rock_Tunnel	0.552	0.376	0.376	**0.588**	✓		Bhat	med	max	15
	Car_railcrossing	0.693	**0.703**	0.272	0.553	✓		L2	-	avg	orig
	Cockpit_Landing	0.630	0.388	0.307	**0.596**	✓		L2	avg	min	24
	Cooking	0.718	**0.608**	0.275	0.602	✓		χ	avg	max	orig
	Eiffel Tower	0.668	**0.632**	0.278	0.561	✓		L2	avg	min	orig
	Excavators river crossing	0.737	**0.460**	0.350	0.450	✓		L2	-	med	orig
	Jumps	0.791	**0.699**	0.244	0.429	✓		L2	med	min	15
	Kids_playing_in_leaves	0.734	0.226	0.353	**1.000**	✓		corr	avg	max	15
	Playing_on_water_slide	0.574	0.588	0.394	**0.668**	✓		χ	avg	min	15
	Saving dolphines	0.601	0.463	0.460	**0.930**	✓		Int	avg	max	orig
	St Maarten Landing	0.795	0.502	0.229	**0.537**		✓	Int	avg	min	orig
	Statue of Liberty	0.554	0.578	0.367	**0.572**		✓	Int	avg	max	24
	Uncut_Evening_Flight	0.692	0.536	0.259	**0.573**		✓	Int	avg	max	24
	paluma jump	0.769	0.273	0.210	**0.648**	✓		χ	med	min	15
	playing_ball	0.672	0.432	0.360	**0.779**	✓		L2	avg	max	15
	Notre_Dame	0.642	**0.653**	0.381	0.528	✓		L2	med	med	orig
Static	Air_Force_One	0.678	0.649	0.294	**0.755**	✓		L2	avg	max	orig
	Fire domino	0.767	0.253	0.282	**0.527**	✓		L2	avg	max	orig
	car_over_camera	0.706	**0.759**	0.273	0.563	✓		L2	-	max	orig
	Paintball	0.725	**0.582**	0.231	0.527	✓		L2	-	avg	orig

being evaluated. The higher f-measure values imply better performance in comparison to human annotation based ground truth summaries.

4.2 Unsupervised Summarization

In the first experiment, we executed our scene detection (Section 3.1) to all SumMe video clips and then tested variants of the SIFT and motion analysis based keyframe selection (Section 3.2). The results for the clips are collected into Table 1. The first finding is that our simple keyframe detection and video expansion perform surprisingly well being on par to average human and outperforming the state-of-the-art learning based method by Gygli et al. (Gygli et al., 2014). However, it is noteworthy that the best keyframe selection method varies from one video to another and their average performances remain between the random and state-of-the-art learning-based methods (Figure 8). The second and more important finding is that motion features generally perform better than region features - motion features provide the best result for 21 out of 25 clips. The local feature based keyframes were better for the following four videos: Bike polo (0.553), St Maarden Landing (0.537), Statue of Liberty (0.572) and Uncut evening flight (0.573), but the best motion based keyframes achieved similar accuracies: 0.495 (10% worse), 0.460 (14% worse), 0.563 (2% worse) and 0.555 (3% worse), respectively.

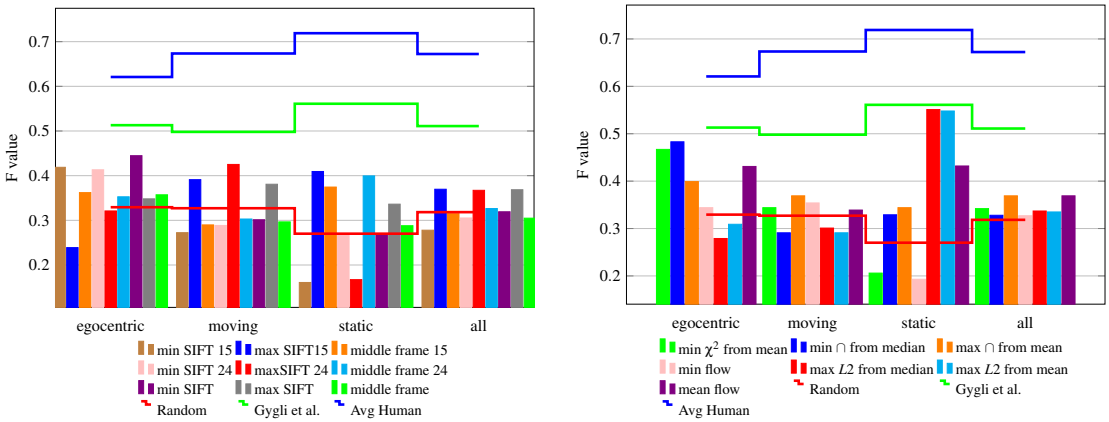
The average performances over all videos for the local feature and motion based keyframes are comparable except for egocentric camera and static camera cases for which motion analysis is clearly better (see Figure 8). However, the average results indicate that no single keyframe selection method can succeed and therefore we need to use multiple of them.

4.3 Motion based Keyframe Selection

From the previous experiment, we found that the motion analysis based keyframes provide better results, which can be explained by their complementary cue

Table 2: Highest ranked methods for videos recorded with different types of cameras when using various frame rates.

fps	Camera type *egosentric*	*moving*	*static*
15	min correlation with median histogram	min flow	min. L2 distance to avg histogram
24	min intersection with median histogram	min correlation with median histogram	max L2 distance to median histogram
orig.	min Bhattacharyya distance to mean histogram	min. flow	max L2 distance to median histogram



(a) (b)

Figure 8: Overall performance using variants of (a) local feature (SIFT) and (b) motion based keyframe selection.

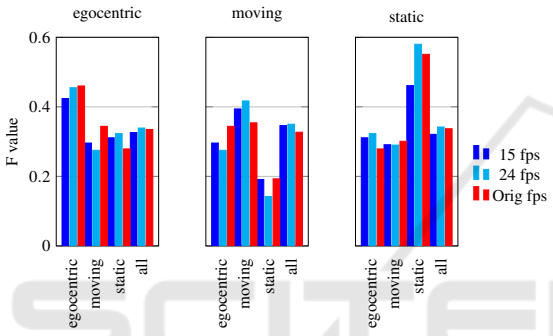


Figure 9: Comparison of the best motion based keyframes for the different video types and frame rates.

(motion) to the local regions that were used in scene detection. From the results in the previous experiment (Table 1, Figure 8) we selected the best motion analysis methods from each action type: egocentric, moving and static camera. The best methods with different parameter settings are shown in Table 2 and a comparison between the highest ranking methods in Figure 9. Using different frame rates did not yield significant differences in the results and therefore using the original frame rate of each video was used in the remaining experiments as it simplifies the preprocessing step and the composition of the final skim.

4.4 Human-in-the-loop

The previous experiments verified that dynamic video BoW based scene detection and motion analysis based keyframe selection provide a powerful processing pipeline for video summarization and they can be run in online mode. However, the best motion based keyframes varied from one video to another and therefore the online method needs to be run using several best of them. We selected the three best methods providing 1-3 keyframes for each scene (less than three

Figure 10: Our weakly supervised and all learning-based methods fail to detect the subtle movement of the diver cluttered by the water fall motion field.

if two or more detect the same frame). By the simple selection GUI in Section 3.3 users can quickly select their preferred keyframes. In this experiment, we tested how well this assisted setting works. We collected annotations from 5 independent annotators and report the results in Table 3. It is noteworthy, that the best human annotator is comparable to the average human annotator in the SumMe groundtruth where annotators carefully watched the whole video and dedicatedly selected parts of the full video to be included to the final summary. Their supervised summarization was significantly more time consuming (hours) than ours (from seconds to minutes).

Assisted summarization is clearly superior to the state-of-the-art learning-based methods by Gygli et al. (Gygli et al., 2014) and Ejaz et al. (Ejaz et al., 2013). The average performance of Ejaz et al. is clearly inferior to ours and SumMe. There was only one video for which all methods are clearly below average human performance, *Paluma jump*, and only one additional video for which our method is clearly worse than SumMe *Uncut Evening Flight*. The main reason for lower performance is obvious - static keyframes cannot represent dynamic contents and, in particular, subtle "motion inside motion" that happens in Paluma jump where there is a large motion field (water fall) and a distant person jumping along the water fall (Figure 10). For the evening flight scene our annotators reported that "the whole video is

Table 3: Weakly supervised keyframe-based video summarization results including average time used for keyframe selection. Avg human corresponds to average performance of all SumMe annotators (ideal performance), ours (best) is the best summary achieved by one of our annotators and ours (avg) is average performance of our annotators.

[a](Gygli et al., 2014), [b](Ejaz et al., 2013)

	Video	Avg Human	Gygli[a]	Ejaz[b]	Rand	Our (best)	Our (avg)	Avg time
Egocentric	Base jumping	0.646	0.304	0.487	0.362	**0.505**	0.413	01:27
	Bike polo	0.640	**0.708**	0.151	0.266	0.594	0.278	01:10
	Scuba	0.561	**0.475**	0.517	0.357	0.403	0.318	00:47
	Valparaiso_downhill	0.637	0.567	0.541	0.333	**0.677**	0.492	00:18
Moving	Bearpark_climbing	0.630	0.358	**0.688**	0.445	0.630	0.571	00:28
	Bus_in_Rock_Tunnel	0.552	0.376	0.312	0.376	**0.613**	0.451	00:54
	Car_railcrossing	0.693	**0.703**	0.124	0.272	0.596	0.339	00:55
	Cockpit_Landing	0.630	0.388	0.262	0.307	**0.779**	0.493	01:12
	Cooking	0.718	0.608	0.223	0.275	**0.705**	0.348	00:26
	Eiffel Tower	0.668	**0.632**	0.291	0.278	0.606	0.400	00:55
	Excavators river crossing	0.737	0.460	0.100	0.350	**0.937**	0.546	01:46
	Jumps	0.791	0.699	0.398	0.244	**0.791**	0.351	00:27
	Kids_playing_in_leaves	0.734	0.226	0.213	0.353	**0.876**	0.628	00:18
	Playing_on_water_slide	0.574	**0.588**	0.365	0.394	0.515	0.486	00:37
	Saving dolphines	0.601	0.463	0.492	0.460	**0.530**	0.467	00:16
	St Maarten Landing	0.795	0.502	**0.671**	0.229	0.651	0.346	00:20
	Statue of Liberty	0.554	**0.578**	0.250	0.367	0.467	0.386	00:56
	Uncut_Evening_Flight	0.692	0.536	**0.591**	0.259	0.249	0.228	00:34
	paluma jump	0.769	**0.273**	0.042	0.210	0.231	0.121	00:09
	playing_ball	0.672	0.432	0.347	0.360	**0.479**	0.382	00:37
	Notre_Dame	0.642	**0.653**	0.383	0.381	0.533	0.431	01:13
Static	Air_Force_One	0.678	0.649	0.439	0.294	**0.755**	0.704	00:19
	Fire domino	0.767	0.253	0.490	0.282	**0.687**	0.342	00:22
	car_over_camera	0.706	0.759	0.410	0.273	**0.824**	0.656	00:34
	Paintball	0.725	0.582	0.511	0.231	**0.589**	0.419	00:35

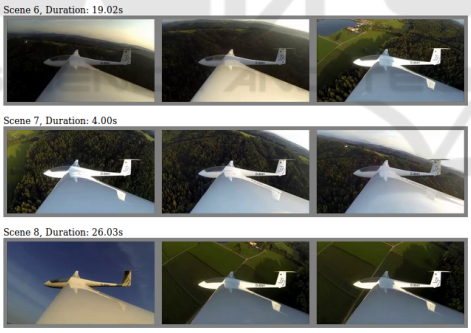


Figure 11: Keyframes-based summarization cannot properly capture temporal saliency that might play more crucial role in videos that are "boring to watch" for users without personal interest to the content.

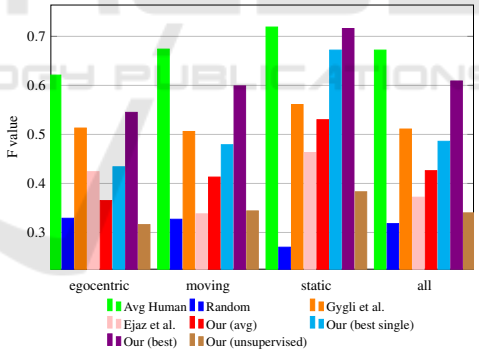


Figure 12: The average performances of state-of-the-art learning-based methods, various variants of our unsupervised and weakly supervised methods and average human video summaries for SumMe dataset. Our (best single) corresponds to the best single user performance.

boring", and since no visual cues exist, the temporal saliency perhaps plays a more important role (Figure 11). The average results are shown in Figure 12 where single person performance is always superior to Ejaz et al. and comparable or better to Gygli et al. - the best human-in-the-loop summary is always superior to all other methods indicating that there is still significant subjective variation between users.

5 CONCLUSIONS

In this work we evaluated the contribution of each processing stage in keyframe-based video summarization: scene detection, keyframe selection and human supervision in selecting the best keyframes. For scene detection we proposed a dynamic video BoW method which provides high recall and there-

fore over segmentation to scenes that capture subtle changes in the visual content. For keyframe selection, we found that motion descriptors are superior over region features used in scene detection which can be explained by their complementary information. However, we also found that average performance of keyframe selection methods is substantially lower than with learning-based state-of-the-arts. We also found that original frame rate provides good results. We introduced a GUI for fast (from 9 seconds to less than two minutes) human-in-the-loop keyframe selection which provides superior/on par performance to state-of-the-art learning-based methods while retaining user control over personal preferences.

REFERENCES

Bian, J., Yang, Y., Zhang, H., and Chua, T.-S. (2015). Multimedia summarization for social events in microblog stream. *IEEE Trans. on Multimedia*, 17(2).

Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. (2009). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

de Avila, S., Lopes, A., and et al. (2011a). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68.

de Avila, S. E. F., Lopes, A. P. B., da Luz Jr., A., and de Albuquerque Araajo, A. (2011b). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1).

Duan, X., Lin, L., and Chao, H. (2013). Discovering video shot categories by unsupervised stochastic graph partition. *IEEE Trans. on Multimedia*, 15(1).

Ejaz, N., Mehmood, I., and Baik, S. W. (2013). Efficient visual attention based framework for extracting key frames from videos. *Image Commun.*, 28(1):34–44.

Fabro, M. D. and Böszörmenyi, L. (2013). State-of-the-art and future challenges in video scene detection: a survey. *Multimedia Systems*, 19:427–454.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA)*.

Gatica-Perez, D., Loui, A., and Sun, M.-T. (2003). Finding structure in home videos by probabilistic hierarchical clustering. *IEEE Trans. on Circuits and Systems for Video Technology*, 13(6).

Gong, B., Chao, W.-L., Grauman, K., and Sha, F. (2014). Diverse sequential subset selection for supervised video summarization. In *Conference on Neural Information Processing Systems (NIPS)*.

Gu, Z., Mei, T., Hua, X.-S., Wu, X., and Li, S. (2007). Ems: Energy minimization based video scene segmentation. In *ICME*.

Guan, G., Wang, Z., Liu, S., Deng, J. D., and Feng, D. (2013). Keypoint-based keyframe selection. *IEEE Trans. on Circuits and Systems for Video Technology*, 24(4).

Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. (2014). Creating summaries from user videos. In *European Conference on Computer Vision (ECCV)*.

Gygli, M., Grabner, H., and Van Gool, L. (2015). Video summarization by learning submodular mixtures of objectives. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gygli, M., Song, Y., and Cao, L. (2016). Video2gif: Automatic generation of animated gifs from video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Han, B., Hamm, J., and Sim, J. (2010). Personalized video summarization with human in the loop. In *IEEE Workshop on Applications of Computer Vision(WACV)*.

Han, B. and Wu, W. (2011). Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In *ICME*.

Herranz, L., Calic, J., Martinez, J., and Mrak, M. (2012). Scalable comic-like video summaries and layout disturbance. *IEEE Trans. on Multimedia*, 14(4).

Hong, R., Yuan, X.-T., Xu, M., and Wang, M. (2010). Movie2comics: A feast of multimedia artwork. In *ACM Multimedia (ACMMM)*.

Jiang, Y.-G., Ye, G., Chang, S.-F., Ellis, D., and Loui, A. (2011). Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ACM International Conference on Multimedia Retrieval (ICMR)*.

Kyperountas, M., Kotropoulos, C., and Pitas, I. (2007). Enhanced eigen-audioframes for audiovisual scene change detection. *IEEE Trans. on Multimedia*, 9(4).

Lankinen, J. and Kämäräinen, J.-K. (2013). Video shot boundary detection using visual bag-of-words. In *Int. Conf. on Computer Vision Theory and Applications (VISAPP)*.

Lee, Y. and Grauman, K. (2015). Multimedia summarization for social events in microblog stream. *Int J Comput Vis*, 114:38–55.

Liu, W., Mei, T., Zhang, Y., Che, C., and Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *Int J Comput Vis*, 60(2):91–110.

Lu, S., Wang, Z., Mei, T., Guan, G., and Feng, D. (2014). A bag-of-importance model with locality-constrained coding based feature learning for video summarization. *IEEE Trans. on Multimedia*, 16(6).

Luo, J., Papin, C., and Costello, K. (2009). Towards extracting semantically meaningful key frames from personal video clips: from humans to computers. *IEEE Trans. on Circuits and Systems for Video Technology*, 19(2):289–301.

Ma, Y.-F., Hua, X.-S., Lu, L., and Zhang, H.-J. (2005). A generic framework of user attention model and its application in video summarization. *IEEE Trans. on Multimedia*, 7(5).

Meng, J., Wang, H., Yuan, J., and Tan, Y.-P. (2016). From keyframes to key objects: Video summarization by representative object proposal selection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ngo, C.-W., Ma, Y.-F., and Zhang, H.-J. (2005). Video summarization and scene detection by graph modeling. *IEEE Trans. on Circuits and Systems for Video Technology*, 15(2).

Odobez, J.-M., Gatica-Perez, D., and Guillemot, M. (2003). On spectral methods and the structuring of home videos. In *Int. Conf. on Image and Video Retrieval (CIVR)*.

Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A. F., Quenot, G., and Ordelman, R. (2015). Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA.

Potapov, D., Douze, M., Harchaoui, Z., and Schmid, C. (2014). Category-specific video summarization. In *European Conference on Computer Vision (ECCV)*.

Rav-Acha, A., Pritch, Y., and Peleg, S. (2006). Making a long video short: Dynamic video synopsis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shai, Y. and Shah, M. (2006). Video scene segmentation using Markov chain Monte Carlo. *IEEE Trans. on Multimedia*, 8(4).

Shroff, N., Turaga, P., and Chellappa, R. (2010). Video précis: Highlighting diverse aspects of videos. *IEEE Trans. on Multimedia*, 12(8).

Smeaton, A. F., Over, P., and Doherty, A. (2010). Video shot boundary detection: Seven years of TRECVid activity. *Computer Vision and Image Understanding*, 114(4).

Tamrakar, A., Ali, S., Yu, Q., Liu, J., Javed, O., Divakaran, A., Cheng, H., and Sawhney, H. (2012). Evaluation of low-level features and their combinations for complex event detection in open source videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Truong, B. T. and Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1).

Wang, M., Hong, R., Li, G., Zha, Z.-J., Yan, S., and Chua, T.-S. (2012). Event driven web video summarization by tag localization and key-shot identification. *IEEE Trans. on Multimedia*, 14(4).

Yanagawa, A., Loui, A., Luo, J., Chang, S.-F., Ellis, D., Jiang, W., Kennedy, L., and Lee, K. (2008). Kodak consumer video benckmark data set: concept definition and annotation. Technical report, Columbia University. ADVENT Technical Report 246-2008-4.

Zhang, K., Chao, W.-L., and Grauman, K. (2016). Summary transfer: Exemplar-based subset selection for video summarization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, B. and Xing, E. (2014). Quasi real-time summarization for consumer videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, X., Loy, C., and Gong, S. (2013). Video synopsis by heterogeneous multi-source correlation. In *Int. Conf. on Computer Vision (ICCV)*.