Adaptive Keyframe Selection for Video Summarization

Shayok Chakraborty
Carnegie Mellon University
shayokc@andrew.cmu.edu

Omesh Tickoo and Ravi Iyer
Intel Labs
{omesh.tickoo,ravishankar.iyer}@intel.com

Abstract

The explosive growth of video data in the modern era has set the stage for research in the field of video summarization, which attempts to abstract the salient frames in a video in order to provide an easily interpreted synopsis. Existing work on video summarization has primarily been static - that is, the algorithms require the summary length to be specified as an input parameter. However, video streams are inherently dynamic in nature; while some of them are relatively simple in terms of visual content, others are much more complex due to camera/object motion, changing illumination, cluttered scenes and low quality. This necessitates the development of adaptive summarization techniques, which adapt to the complexity of a video and generate a summary accordingly. In this paper, we propose a novel algorithm to address this problem. We pose the summary selection as an optimization problem and derive an efficient technique to solve the summary length and the specific frames to be selected, through a single formulation. Our extensive empirical studies on a wide range of challenging, unconstrained videos demonstrate tremendous promise in using this method for real-world video summarization applications.

1. Introduction

The widespread emergence of inexpensive video cameras together with the advent of several social networking and video sharing websites (like Facebook, Flickr) has resulted in an unprecedented increase in the generation of video data in today's digital world. These videos are extremely unstructured and diverse in their contents (often with erratic camera motion, variable illumination and scene clutter) and can vary in duration from a few seconds to several hours. Further, most videos contain significant redundancy and only a small fraction of the frames are informative. Manually scanning through such videos to gain an understanding of their content is an expensive process in terms of time and human labor. Thus, while generating huge quantities of unstructured video is cheap and easy, navigating efficiently through them is a fundamental challenge. This has paved the way for the development of automated video summarization algorithms, which extract a short and informative summary of these videos to enable a more efficient and engaging viewing experience [20, 16].

Figure 1. Need for Adaptive Video Summarization

An end-to-end video summarization system can be conceptualized as consisting of two main steps: (i) deciding a summary length (number of frames to be selected in the summary) and (ii) selecting the specific exemplars from the video stream to produce the summary. Both these steps are of paramount importance in generating a compact and meaningful summary of the video in question. However, most of the existing work on automated video summarization has been static, that is they focus only on selecting the exemplar frames in the summary and assume the summary length to be specified as a user input. Given the inherent variability in video streams, it is difficult to decide a summary length at random and without any knowledge of the video being analyzed. The summary length should depend on the complexity of the video and should be larger for a video with more visual content. For instance, consider a one minute video depicting a person juggling a soccer ball against a one minute video showing soccer-ball juggling, a person playing golf, a person riding a horse and a person skiing, as shown in Figure 1. Evidently, we would expect the summary length to be larger for the second video (even though it is of the same length as the first) as it possesses significantly more visual content and information. Hence, there is a pronounced need for adaptive summary selection in video summarization algorithms.

In this paper, we propose a novel framework to address this problem. We develop an optimization-based frame-

978-1-4799-6683-7/15 $31.00 © 2015 IEEE
DOI 10.1109/WACV.2015.99

work for dynamic summary selection and propose an efficient strategy to solve for the summary length and extract the keyframes through a single integrated formulation. Although we focus on video summarization in this work, the proposed framework is generic and can be used in any application, where a subset of representative and informative samples needs to be selected from large amounts of redundant data, as in document summarization, exploratory data analysis and pre-filtering.

2. Related Work

Video summarization has been a fertile ground of vision research, especially over the last few years [22]. Existing approaches are primarily based on identification of the shot boundaries within a video and selection of salient frames from each shot using first/middle/last or simple random sampling. Zhang *et al.* [25, 24] and Gunsel *et al.* [9] used a thresholding scheme on frame differences to identify the keyframes in a shot. Motion-based sampling of frames has also been used for video summarization [4, 18]. The main drawback of these methods is that they capture local variations and may miss important segments while longer segments might appear multiple times with similar content.

To address this, methods which select the exemplars from all the available frames in the video, are needed. Clustering is an intuitive solution, where the entire data is segregated into a pre-defined number of clusters (based on the summary length) followed by the selection of exemplars from each cluster. Examples of this technique include [6, 26]. The clustering schemes demonstrate good performance in general; however, they may end up selecting summary frames only from the dominant clusters and may overlook interesting events which occur infrequently. Shroff *et al.* [20] used the diversity and coverage criteria to identify the salient frames in a video. Khosla *et al.* [12] used web-images as a prior to automatically summarize large scale user generated videos. Very recently, vision researchers have begun to explore egocentric video summarization [16, 14] to recognize important people, objects and the overall progress of the story in an egocentric video.

All the aforementioned approaches are static that is, they require the summary length to be pre-specified by the user. As mentioned previously, the inherent variability and complexity of video streams necessitate adaptive summarization schemes which adapt to the data being analyzed. Compared to the static approach, dynamic video summarization is considerably less explored. The few existing methods to address this problem are all heuristic in nature. They are mostly based on clustering strategies or some measures on successive frame difference values. The clustering strategies [15, 27, 7] attempt to determine the number of clusters in the data (based on heuristics) and select representative frame(s) from each cluster. The frame difference algorithms [23, 5, 8] derive the summary dynamically based on some properties of the successive frame differences. For instance, the approach proposed by Gianluigi and Raimondo [8] selects a variable number of keyframes from each video sequnce by detecting the curvature points within the curve of cumulative frame differences. However, these algorithms rely on the values of several parameters and thresholds which need to be set based on the results of extensive experiments on other video sequences and observation of the video clips and their frame difference graphs.

In this paper, we propose a novel algorithm for dynamic video summarization, where the summary length and frame selection criteria are integrated into a single framework. Contrary to the previous approaches, which are heuristic, our methodology is based on a concrete optimization framework and also enjoys nice theoretical properties. We now describe the mathematical formulation of our framework.

3. Proposed Framework

3.1. Problem Formulation

Due to the high frame rate of modern video cameras and the sparsity of interesting events, most real-world videos typically contain a significant percentage of redundant information. Thus, identifying the important events is a fundamental challenge in video summarization. On the other hand, the interesting events themselves follow a skewed statistical distribution and an action occurring more frequently in the video than the other interesting actions, can bias the sample selection towards this dominant action. This may result in the selection of overlapping / duplicate information in the summary. A good video synopsis is thus characterized by the extent to which it represents the essence of the original video and the selection of unique information from the video. We therefore quantify the quality of a video summary in terms of the following two criteria:

- **Representativeness**, which ensures that the summary consists of frames which represent a large portion of the spectrum of events in the video
- **Uniqueness**, which emphasizes that the frames in the summary should capture unique information; that is, they should all be distinct from each other.

A summarization framework driven by maximizing the representativeness and uniqueness conditions ensures that we represent the original video well (exhaustive) and we capture unique aspects of the video (mutually exclusive). This is illustrated in Figure 2. The representativeness and uniqueness criteria are commonly used in selective sampling algorithms like active learning [19].

Formally, consider a video $V = \{v_1, v_2 \dots v_n\}$ consisting of n frames and let S denote the set of frames selected

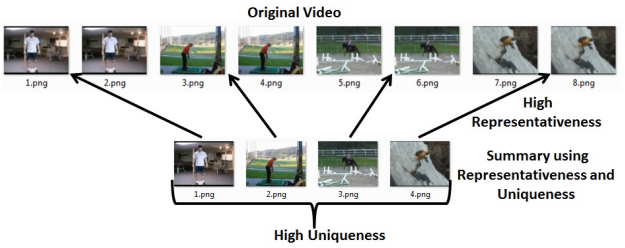


Figure 2. Representativeness and Uniqueness for Video Summarization

in the summary. The representativeness is a measure of the similarity of S to the entire set V. An efficient way to compute the representativeness of a set S is using the facility location function [2]:

$$R(S) = \sum_{i \in V} \max_{j \in S} w_{ij} \tag{1}$$

where $w_{ij} \geq 0$ denotes the similarity between two frames v_i and v_j in the video sequence. Maximizing $R(S)$ ensures that the summary covers most of the video in terms of global representation.

The uniqueness of a candidate frame is quantified as the minimum distance of the frame from the frames already selected in the summary. A greater value of the minimum distance denotes a more useful frame from the uniqueness perspective. This is conceptually similar to the Hausdorff distance between two point sets (commonly used for image matching [10]), which depicts the greatest of all distances from a point in one set to the closest point in the other. The uniqueness score of a set of frames is given by:

$$U(S) = \sum_{i \in S} \min_{j \in S: j < i} d_{ij} \tag{2}$$

where $d_{ij} \geq 0$ denotes the distance between two frames v_i and v_j. Maximizing $U(S)$ ensures that the elements in the summary are as distinct as possible. The overall quality of a summary can therefore be expressed as a weighted combination of the representativeness and uniqueness based terms:

$$Q(S) = R(S) + \lambda_1 U(S) \tag{3}$$

where $\lambda_1 \geq 0$ is a weight parameter to denote the relative importance of the two terms. It is evident that the objective function $Q(S)$ is monotonically non-decreasing, that is $Q(B) \geq Q(A)$ if $B \supseteq A$. Since there is no restriction on the summary length, maximizing $Q(S)$ will result in selection of *all* the frames in the summary, which defeats the basic purpose of summarization. To address this, we modify the objective by appending a penalty on the cardinality of the set S:

$$Q^{pen}(S) = R(S) + \lambda_1 U(S) - \lambda_2 |S| \tag{4}$$

where $|S|$ denotes the cardinality of the set S and its value increases with increasing summary size. The objective $Q^{pen}(S)$ ensures that all the frames are not selected in the summary; only the frames for which the representativeness and uniqueness terms outweigh the penalty term get selected. The penalty term denotes the total cost associated with the summary S. The cost is dependent on several factors like storage / power constraints of the system running the application (e.g. smartphone, wearable systems), time available to view the summary etc. The representativeness, uniqueness and penalty terms are all assumed to be measured in the same currency; different currencies can be transformed into a single utility using appropriate real-world conversions [11].

The parameter λ_2 is analogous to the cost of a single summary frame and is assumed to be known. This parameter has a direct effect on the length of the summary (smaller value of λ_2 results in larger permissible summary size and vice versa). However, due to the inherent variability of video streams, it is more appropriate to decide on the value of λ_2 (based on available system resources) and then compute the summary length dynamically for a given video, as opposed to deciding the summary length of a video without any knowledge of its variability and visual content and using the same length to summarize all videos (with different visual information). The assumption of a known cost coefficient does not impose significant constraints; similar assumptions are often used in machine learning applications (like active learning) where it is assumed that the cost of revealing the label of an unlabeled sample is known apriori [11].

To ensure non-negativity of the objective, we add a constant term $\lambda_2|V|$ to derive our final objective function $\widehat{Q}(S)$ (note that the optimal value of S is not affected due to the introduction of the constant term):

$$\widehat{Q}(S) = R(S) + \lambda_1 U(S) + \lambda_2 |V - S| \tag{5}$$

The adaptive summary selection problem therefore reduces to solving the following optimization:

$$\max_{S \subseteq V} \widehat{Q}(S) \tag{6}$$

Solving this problem yields the summary length (the cardinality of the set S) and the salient frames (the elements of S) through a single formulation. However, since the search space is exponentially large, exhaustive techniques are infeasible. In the following section, we propose an efficient strategy to solve this optimization problem.

3.2. Submodularity of the Objective Function

Consider a set of elements $Z = \{z_1, z_2 \ldots z_n\}$ and a function $f : 2^Z \rightarrow \Re$ that returns a real value for any subset

$S \subseteq Z$. Let $A \subseteq B \subseteq Z$ be two subsets of Z and consider an element $x \in Z \backslash B$. The function f is submodular if

$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B) \quad (7)$$

That is, a function is submodular if adding an element to a set increases the functional value by at least as much as adding the same element to its superset. This property is called the diminishing returns property [13].

Theorem 1. *The objective function $\widehat{Q}(S)$ defined in Equation (5) is submodular.*

Proof. By definition,

$$\begin{aligned}\widehat{Q}(S) &= R(S) + \lambda_1 U(S) + \lambda_2 |V - S| \\ &= \sum_{i \in V} \max_{j \in S} w_{ij} + \lambda_1 \sum_{i \in S} \min_{j \in S: j<i} d_{ij} + \lambda_2 |V - S|\end{aligned}$$

For an element $x \in V \backslash S$, we have

$$\begin{aligned}\widehat{Q}(S \cup \{x\}) - \widehat{Q}(S) = \sum_{i \in V} \max\{0, w_{ix} - \max_{j \in S} w_{ij}\} \quad (8) \\ + \lambda_1 \min_{i \in S} d_{ix} - \lambda_2\end{aligned}$$

Now, consider two arbitrary summary sets S_1 and S_2 such that $S_1 \subseteq S_2$. We then have,

$$\max_{j \in S_2} w_{ij} \geq \max_{j \in S_1} w_{ij} \quad (9)$$

$$\Rightarrow \sum_{i \in V} \max\{0, w_{ix} - \max_{j \in S_2} w_{ij}\} \leq \sum_{i \in V} \max\{0, w_{ix} - \max_{j \in S_1} w_{ij}\}$$

We also have,

$$\min_{i \in S_2} d_{ix} \leq \min_{i \in S_1} d_{ix} \quad (10)$$

$\forall x \in V \backslash S_2$. The inequality in Equation (9) holds since S_2 is a larger summary, it is possible to have an element in $S_2 \backslash S_1$ which is more similar to the current frame i in the video V. Equation (10) holds for a similar reason as it is possible to have an element in the superset which is closer to the sample x. λ_1 and λ_2 are positive scalars. From Equation (8) we thus have

$$\widehat{Q}(S_1 \cup \{x\}) - \widehat{Q}(S_1) \geq \widehat{Q}(S_2 \cup \{x\}) - \widehat{Q}(S_2)$$

$\forall S_1, S_2, S_1 \subseteq S_2$. Hence, $\widehat{Q}(S)$ is submodular. □

It is to be noted that even though $\widehat{Q}(S)$ is submodular, it is not monotonic due to the penalty term $|V - S|$.

3.3. Efficient Optimization

The problem of adaptive summary selection therefore reduces to the unconstrained maximization of a non-negative, non-monotone submodular function. Recent work on submodular optimization [1] has proposed a tight, linear time approximation algorithm to address this problem. The algorithm maintains two solutions X and Y with initial values $X_0 = \phi$ and $Y_0 = N$. The elements in the ground set are randomly permuted to derive an arbitrary sequence $u_1, u_2, \ldots u_n$. In the i^{th} iteration, the element u_i is either added to X_{i-1} or removed from Y_{i-1}. This decision is made randomly, with probabilities derived from the values a_i and b_i (Algorithm 1). Thus, for every $1 \leq i \leq n$, X_i and Y_i are random variables denoting the sets of elements in the two solutions generated by the algorithm at the end of the i^{th} iteration. After n iterations, both solutions coincide and we get $X_n = Y_n$, which is the output of the algorithm. The pseudo-code is depicted in Algorithm 1.

Algorithm 1 Randomized Algorithm for Unconstrained Submodular Maximization

1: Start with $X_0 = \phi$ and $Y_0 = N$
2: **for** $i = 1 \rightarrow n$ **do**
3: $a_i = f(X_{i-1} \cup \{u_i\}) - f(X_{i-1})$
4: $b_i = f(Y_{i-1} \backslash \{u_i\}) - f(Y_{i-1})$
5: $a_i^{'} = \max\{a_i, 0\}, b_i^{'} = \max\{b_i, 0\}$
6: With probability $a_i^{'}/(a_i^{'} + b_i^{'})$, assign $X_i = X_{i-1} \cup \{u_i\}, Y_i = Y_{i-1}$
7: Else, with complement probability $b_i^{'}/(a_i^{'} + b_i^{'})$, assign $X_i = X_{i-1}, Y_i = Y_{i-1} \backslash \{u_i\}$
8: **end for**
9: **return** X_n or equivalently Y_n

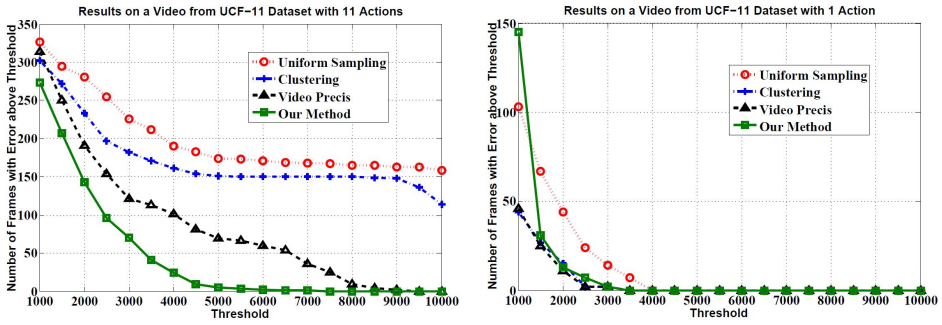
This algorithm has linear time complexity and also has a theoretical guarantee on the quality of the solution, as formalized in the following lemma [1]:

Lemma 1. *Let X^* be the optimal solution. With the random variables X_i and Y_i defined above, the final solution returned by the algorithm is bounded as $\mathbb{E}[f(X_n)] = \mathbb{E}[f(Y_n)] \geq f(X^*)/2$.*

Further, the approximation bound of $1/2$ on the solution quality is tight [1]. Due to its promising theoretical guarantee, we use this algorithm to maximize $\widehat{Q}(S)$ in our work.

4. Experiments and Results

Datasets and Feature Extraction: To depict the generalizibility of our approach, we conducted experiments on a wide range of videos from several application domains. The following datasets were used in our work: **The UT Egocentric Video Dataset** [14], which contains egocentric videos captured by subjects under uncontrolled natural settings, using a wearable camera (we used the video



(a) Results on a video containing 11 actions (b) Results on a video containing 1 action

Figure 3. Dynamic vs Static Video Summarization on the UCF-11 Dataset. Best viewed in color.

captured by user 3), **Unconstrained Office Video** which depicts a person touring an office and meeting various employees (similar to [20]) and **UCF-11, UCF-50 and UCF-101 Actions Datasets**, containing challenging realistic action videos (collected from YouTube) with large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination conditions [21]. Based on the purpose of an experiment, appropriate datasets were selected, which best convey the essence of the experiment. Each video was split into images and the histogram of oriented gradients (HOG) feature [3] was extracted from each image. To compute the distance d_{ij} between two image frames v_i and v_j, we used the chi-squared distance between their color histograms [14]; to compute the similarity, we simply used the cosine similarity measure. We set $\lambda_1 = 1$ and $\lambda_2 = 5$ in our experiments, based on priliminary empirical studies.

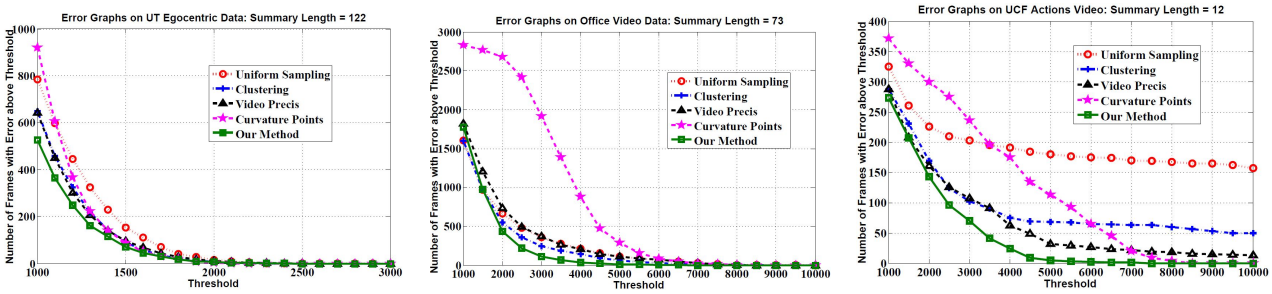
Baselines: We compared our approach against 3 static summarization techniques which need the summary length k as an input: (1) **Uniform Sampling**, which selects k summary frames that are uniformly spaced in the video sequence, (2) **Clustering**, which performs k-means clustering on the data and selects the sample closest to the centroid of each cluster in the summary and (3) **Video Precis**, proposed by Shroff *et al.* [20]. We also compared our algorithm against the Curvature Points (CP) based dynamic summary selection method [8]. This method finds the high curvature points within the curve of cumulative frame differences and then extracts keyframes by taking the midpoint between two consecutive high curvature points. This algorithm was shown to outperform other dynamic summarization techniques [8].

Evaluation Metric: To evaluate a summary objectively, we compute the count of frames whose reconstruction error, using the summary, is above a given threshold γ. The main intuition behind this metric is that a good summary is one where all the frames in the video lie in the space spanned by the linear combination of the exemplars; fewer the number of frames in the null space of the exemplars, better the summary. Please refer [20] for more details.

4.1. Experiment 1: Static vs. Dynamic Summarization

In this experiment, we performed a comparative study of static against dynamic video summarization (since the objective was to compare the proposed dynamic approach against the static approaches, the CP based dynamic strategy was not used in this study). The UCF-11 actions dataset was used here. The three static algorithms and the proposed dynamic algorithm were applied on two video streams (containing about 3000 frames each) where the first one contained images of all 11 actions and the second one contained the images of just a single action. The static summary size was fixed at 30. The summary lengths predicted by the proposed framework for the two videos were 47 and 6 respectively. The results are depicted in Figure 3 where the x axis denotes the threshold γ and the y axis denotes the number of frames with error above this threshold.

We note that for the first video (Figure 3(a)), the proposed approach comprehensively outperforms the static techniques, as for this approach, the number of frames with reconstruction error above the threshold drops at the fastest rate with increasing values of the threshold. This is due to the fact that the proposed adaptive technique selects a proper summary length to aptly capture the visual content of the video. The static frameworks suffer as the selected summary length is too less to capture the visual content. For the second video with a single action (Figure 3(b)), the clustering and the video precis static algorithms perform marginally better than dynamic selection; however, this comes at the cost of storing a much greater number of images for the summary ($30/6 = 5$ fold in this case). The static algorithms need a summary length to be pre-specified without any knowledge of the video stream being analyzed. They sometimes select too few frames which results in poor summarization of the original video; sometimes they select



(a) UT Egocentric Dataset: S.L. = 122 (b) Office Video Dataset: S.L. = 73 (c) UCF 101 Actions Dataset: S.L. = 12

Figure 4. Comparative Performance for a given Summary Length (S.L.). Best viewed in color.

too many frames (incuring a considerable cost) to achieve a marginal increment in summary quality. The dynamic selection framework, on the other hand, identifies the summary length by appropriately finding a trade-off between the summary quality and the associated cost and is thus a more sound basis for video summarization.

4.2. Experiment 2: Comparison against Baselines for the Same Summary Length

The purpose of this experiment was to study the performances of all the video summarization algorithms, for the same summary length. For a given video, the proposed adaptive framework was applied to derive the summary length; this summary length was used as an input to the CP-based dynamic approach and the static selection techniques, for fair comparison. We used the UT Egocentric (7500 frames), the Office Video (7500 frames) and the UCF-101 actions (1000 frames containing images of 10 random actions [1] sampled non-uniformly) datasets for this experiment. The results are presented in Figure 4. It is evident that the proposed framework depicts the best performance among all the summarization techniques; at any given threshold, it has the lowest number of frames with error above that threshold, among all the methods. The video precis method also depicts good results. The other strategies are not consistent in their performance.

Figure 5 shows the images selected in the summary by each algorithm for the UCF-101 dataset. Our algorithm captures all 10 actions, which explains its best performance in Figure 4(c). Video precis covers 9 of the 10 actions while the other strategies capture much lesser visual content.

4.3. Experiment 3: Proposed vs. CP-based Dynamic Video Summarization

The objective of this experiment was to compare the proposed algorithm against the CP-based approach for dynamic

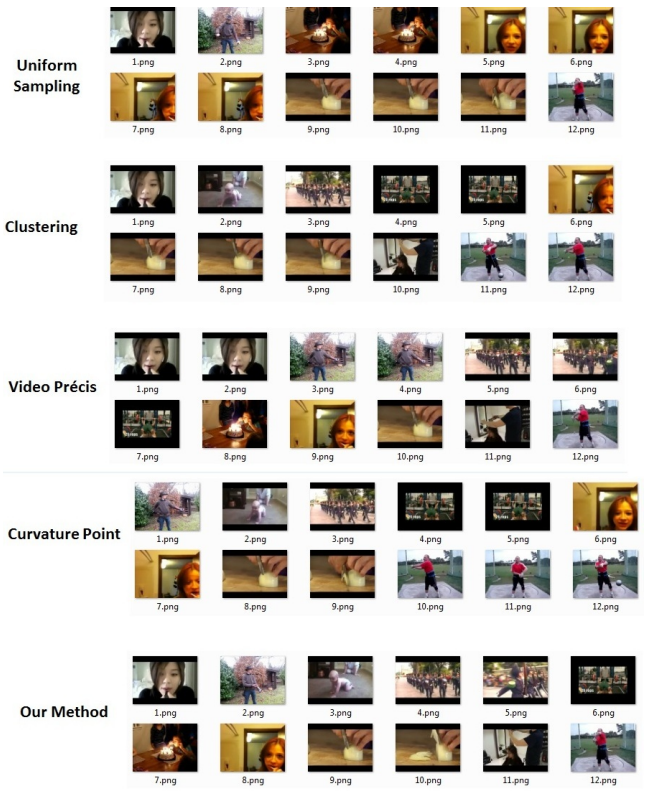
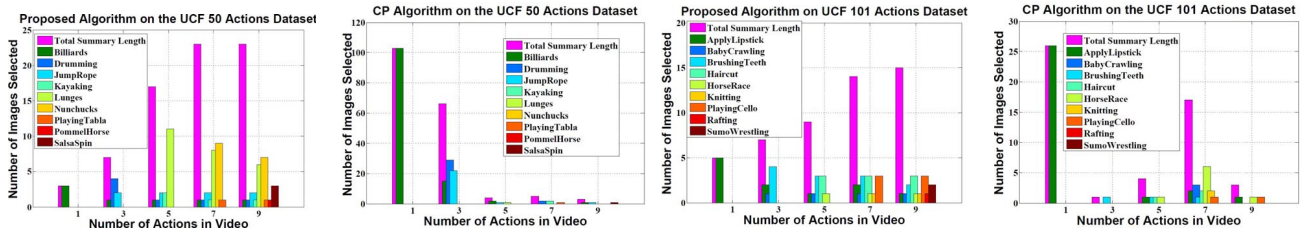


Figure 5. Images Selected by the Algorithms for the UCF-101 Dataset. Best viewed in color.

video summarization. For this purpose, the algorithms were first applied on a video ($\approx$ 1000 frames) from the UCF actions dataset, containing a single action. The number of actions in the video was gradually increased keeping the total length of the video constant ($\approx$ 1000 frames). However, the algorithms were incognizant about the composition of the videos. The summary lengths predicted by the two strategies were noted together with the number of images from each action that got selected in the summary.

The results are presented in Figure 6 for the UCF-50 and

[1] ApplyLipstick, Archery, BabyCrawling, BandMarching, BenchPress, BlowingCandles, BrushingTeeth, CuttingInKitchen, Haircut and HammerThrow



(a) Proposed Dynamic Algorithm on UCF-50

(b) CP-based Dynamic Algorithm on UCF-50

(c) Proposed Dynamic Algorithm on UCF-101

(d) CP-based Dynamic Algorithm on UCF-101

Figure 6. Proposed vs. CP-based Dynamic Video Summarization on the UCF-50 and UCF-101 Datasets. Best viewed in color.

UCF-101 datasets. We note from Figures 6(a) and 6(c) that, as the number of actions in the video increases from 1 to 9, the proposed algorithm automatically decides on a larger summary length (denoted by the pink bars). This corroborates our intuition as, with increasing number of actions, the visual content of the video increases which necessitates a larger summary length to adequately summarize the video. Moreover, we also note that our algorithm selects exemplars from each of the actions in a given video. Thus, besides adapting to the complexity, our algorithm selects samples to encapsulate the important events in the video. The CP-based method (Figures 6(b) and 6(d)) selects the summary length and the specific samples using a heuristic measure on cumulative frame difference values. The summary length therefore does not bear any specific trend to the composition/complexity of the video. Also, it fails to capture samples from all actions, specially from videos with high visual content. Thus, while the Curvature Points based method summarizes a video based on a heuristic measure on successive frame differences, the proposed dynamic algorithm appropriately balances the summary quality and the cost through a concrete optimization framework and is thus a better strategy for adaptive keyframe selection. (We also note that the number of images selected from each action for a given video is not uniform due to the fact that each action has a different level of visual information). Since our algorithm aptly captures the complexity of a video through the predicted summary length, it generates much better quality summaries than the CP-based approach. Sample results are included in the Supplemental File due to lack of space.

4.4. Experiment 4: Effect of the Cost Parameter λ_2

In this experiment, we study the effect of the cost parameter λ_2 on the summary length and frame selection. We apply our algorithm on a video consisting of 5 random actions from the UCF-50 actions dataset (note that the specific actions selected from the UCF 11, 50 and 101 datasets were different in each experiment, to validate the generalizibility of our approach). We gradually increase the value of the parameter λ_2 from 1 to 10 and note the summary length

λ_2	A1	A2	A3	A4	A5	Summary Length
1	3	12	10	2	5	32
2	2	8	3	2	3	18
3	1	7	4	1	3	16
4	1	5	3	1	3	13
5	1	3	2	1	3	10
6	1	4	2	1	2	10
7	1	3	2	1	2	9
8	1	3	1	1	2	8
9	1	3	1	1	2	8
10	1	2	1	1	2	7

Table 1. Effect of the Cost Parameter λ_2 on a video from UCF-50 with 5 actions.

λ_2	A1	A2	A3	A4	A5	Summary Length
1	9	7	15	34	10	75
2	6	5	13	21	6	51
3	5	3	12	17	5	42
4	4	3	10	13	4	34
5	3	3	7	11	3	27
6	2	2	7	9	2	22
7	3	2	6	8	3	22
8	2	2	5	7	3	19
9	2	2	5	6	2	17
10	2	2	5	5	2	16

Table 2. Effect of the Cost Parameter λ_2 on a video from UCF-101 with 5 actions.

predicted by the algorithm, as well as the number of images Ai from each action i that gets selected in the summary (the other parameter λ_1 was fixed at 1). The results are reported in Table 1. We note that, as the value of λ_2 increases, the summary length depicts a decreasing pattern. This corroborates our intuition as increasing cost implies more restriction on the number of frames that can be used to generate the summary. However, we note that, even with high values of the cost parameter, the algorithm selects at least one image from each of the actions. Thus, even with severe budget restrictions, the algorithm generates the summary so as to capture the important aspects of the original video. A similar result is obtained on another video containing 5 other random actions from the UCF-101 dataset and is presented in Table 2. Further results on the generated summaries are included in the Supplemental File.

5. Discussion

In this paper, we proposed a novel approach for dynamic video summarization. Our algorithm integrates the summary length and keyframe selection in a single formulation and solves for the two parameters through a single optimization framework. The empirical results corroborate the efficacy of the framework in adapting to the complexity of a video stream and generating a succinct and condensed representation of the contents of the video. We formulated a generic objective function to summarize any arbitrary video without any assumptions about its contents. If domain knowledge is available about the video data in question (e.g. it contains a specific set of objects/people), then the objective function can be modified accordingly. The same principle based on a penalty on the set cardinality can still be used for adaptive summarization (for a submodular and non-negative objective).

The proposed framework can also be used for static video summarization with a pre-specified summary length k, by maximizing the objective $Q(S)$ in Equation (3) subject to the constraint $|S| = k$. This objective is submodular, monotonically non-decreasing and has a cardinality constraint. It can therefore be maximized using the greedy algorithm proposed by Nemhauser *et al.* [17] which gives a performance guarantee of $1 - \frac{1}{e} \approx 0.632$ on the solution quality and the guarantee is tight unless $P = NP$. Sample results are presented in the Supplemental File. As part of our future work, we plan to conduct extensive user studies for qualitative evaluation of the proposed algorithm.

References

[1] N. Buchbinder, M. Feldman, J. Naor, and R. Schwartz. A tight linear-time (1/2)-approximation for unconstrained submodular maximization. In *FOCS*, 2012. 4

[2] G. Cornuejols, M. Fisher, and G. Nemhauser. On the uncapacitated location problem. In *Studies in Integer Programming*, 1977. 3

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 5

[4] A. Divakaran, R. Radhakrishnan, and K. Peker. Video summarization using descriptors of motion activity: A motion activity based approach to keyframe extraction from video shots. In *Journal of Electronic Imaging*, 2001. 2

[5] W. Farag and H. Wahab. Adaptive key frames selection algorithms for summarizing video data. In *Technical Report, Old Dominion University*, 2001. 2

[6] A. Ferman and A. Tekalp. Multiscale content extraction and representation for video indexing. In *SPIE Multimedia Storage and Archiving Systems*, 1997. 2

[7] Y. Gao, W. Wang, J. Yong, and H. Gu. Dynamic video summarization using two-level redundancy detection. In *Springer Multimedia Tools and Applications*, 2009. 2

[8] C. Gianluigi and S. Raimondo. An innovative algorithm for keyframe extraction in video summarization. In *Journal of Real Time Image Processing*, 2006. 2, 5

[9] B. Gunsel, Y. Fu, and A. Tekalp. Hierarchical temporal video segmentation and content characterization. In *SPIE Multimedia Storage and Archiving Systems*, 1997. 2

[10] D. Huttenlocher, G. Klanderman, and W. Rucklidge. Comparing images using the hausdorff distance. In *IEEE TPAMI*, 1993. 3

[11] A. Kapoor, E. Horvitz, and S. Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *IJCAI*, 2007. 3

[12] A. Khosla, R. Hamid, C. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 2

[13] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *UAI*, 2005. 4

[14] Y. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2, 4, 5

[15] R. Lienhart. Dynamic video summarization of home video. In *SPIE Storage and Retrieval for Media Databases*, 2000. 2

[16] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 1, 2

[17] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions. In *Mathematical Programming*, 1978. 8, 10

[18] C. Ngo, Y. Ma, and H. Zhang. Automatic video summarization by graph modeling. In *ICCV*, 2003. 2

[19] D. Shen, J. Zhang, J. Su, G. Zhou, and C. Tan. Multi-criteria based active learning for named entity recognition. In *ACL*, 2004. 2

[20] N. Shroff, P. Turaga, and R. Chellappa. Video precis: Highlighting diverse aspects of videos. In *IEEE Transactions on Multimedia*, 2010. 1, 2, 5, 10

[21] K. Soomro, A. Zamir, and M. Shah. Ucf 101: A dataset of 101 human action classes from videos in the wild. In *Techical Report, UCF*, 2012. 5

[22] B. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. In *ACM TOMCCAP*, 2007. 2

[23] Y. Yuan, D. Feng, and Y. Zhong. A novel method of keyframe setting in video coding: Fast adaptive dynamic keyframe selecting. In *IEEE International Conference on Computer Networks and Mobile Computing*, 2003. 2

[24] H. Zhang, C. Low, S. Smoliar, and J. Wu. Video parsing, retrieval and browsing: An integrated and content-based solution. In *ACM Multimedia*, 1995. 2

[25] H. Zhang, J. Wu, D. Zhong, and S. Smoliar. An integrated system for content based video retrieval and browsing. In *Pattern Recognition*, 1997. 2

[26] D. Zhong, R. Kumar, and S. Chang. Real-time personalized sports video filtering and summarization. In *ACM Multimedia*, 2001. 2

[27] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *ICIP*, 1998. 2