Towards Distributed Video Summarization

Shayok Chakraborty[1], Omesh Tickoo[2] and Ravishankar Iyer[2]
[1]Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA
[2]Intel Labs, Oregon, USA

ABSTRACT

Video summarization is a fertile topic in multimedia research. While the advent of modern video cameras and several social networking and video sharing websites (like YouTube, Flickr, Facebook) has led to the generation of humongous amounts of redundant video data, video summarization has emerged as an effective methodology to automatically extract a succinct and condensed representation of a given video. The unprecedented increase in the volume of video data necessitates the usage of multiple, independent computers for its storage and processing. In order to understand the overall essence of a video, it is therefore necessary to develop an algorithm which can summarize a video distributed across multiple computers. In this paper, we propose a novel algorithm for distributed video summarization. Our algorithm requires minimal communication among the computers (over which the video is stored) and also enjoys nice theoretical properties. Our empirical results on several challenging, unconstrained videos corroborate the potential of the proposed framework for real-world distributed video summarization applications.

Categories and Subject Descriptors

I.4.9 [**Image Processing and Computer Vision**]: Applications; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video Analysis*

Keywords

Video Summarization, Submodular Functions

1. INTRODUCTION

The progress of technology in leaps and bounds has led to the emergence and widespread deployment of inexpensive video cameras. These cameras have a high frame rate (typically, $25 - 30$ frames per second) and thus capture a humongous amount of data in a short duration. The captured data has high redundancy (due to the high frame rate) and also is also extremely unstructured and diverse (both in terms of contents and duration). Scanning through these staggering number of images manually entails significant human labor. This has set the stage for research in the field of video summarization, which automatically extracts the salient and informative frames from a video stream and enables a more efficient and engaging viewing experience [15]. It is extensively used as a preprocessing step in many video-based applications like indexing, retrieval and browsing. Common approaches of video summarization include shot boundary detection [17], motion-based sampling [12] and clustering [3]. Shroff *et al.* [14] proposed an iterative algorithm to select the exemplar frames from a given video sequence. Egocentric video summarization has recently gained attention to recognize important people and objects in a video [9].

Video data volumes are increasing faster than the ability of individual computers to store and process them. Consider a surveillance system where a video camera is installed in a public place to detect suspicious behaviors. The high frame rate and long duration of operation of the camera result in the generation of an enormous number of frames, which need to be stored across multiple machines. In order to detect unusual activities, it is necessary to generate a summary of the entire video stream. Further, it may be necessary to quickly sift through months of security video footage, stored across multiple computers, in order to recognize suspicious/anomalous behavior. Medical video analysis is another application where the informative frames need to be identified from vast amounts of a patient's video data, spread across multiple machines. Thus, there is a pressing need for an algorithm to summarize a video, which is distributed across multiple machines, with minimal interaction among them. Even though video summarization is extensively studied, distributed video summarization is much less explored. Ioannis *et al.* [8] proposed an unsupervised data reduction algorithm based on non-negative matrix factorization (NMF) to identify the summary frames in a distributed setting. The NMF algorithm is run separately on each machine to identify the representative frames and the results are merged to generate the final summary. To the best of our knowledge, this is the only published algorithm to summarize a video distributed across multiple computers. However, this method is heuristic in nature and also lacks any theoretical guarantees on the quality of the obtained solution. Other related efforts in this domain are all focused on speeding up the summarization process using multi-core CPUs, GPUs and parallelization schemes [2, 1], but the entire video is assumed to be present in a single computer.

In this paper, we propose a novel algorithm for distributed video summarization. We develop an optimization-based framework to identify the salient and informative frames distributed across multiple independent computers. Our algorithm requires minimal communication among the machines and the obtained solution can be theoretically guaranteed to be competitive with the centralized so-

MM'15 October 26-30, 2015, Brisbane, Australia
Copyright 2015 ACM ISBN 978-1-4503-3459-4/15/10 ...$15.00
DOI: http://dx.doi.org/10.1145/2733373.2806355.

lution (obtained when the entire video is present in a single computer). Although we focus on video summarization in this work, the proposed framework is generic and can be used in any application where the salient exemplars need to be identified from large amounts of redundant data distributed across multiple machines.

2. PROPOSED FRAMEWORK

2.1 Problem Formulation

Consider a distributed video summarization set-up, as shown in Figure 1. The video captured by the camera is discharged in computer C_0, which broadcasts the frames to m computers C_1 to C_m. Each computer (C_1 to C_m) processes its own chunk of the video stream and the results are broadcasted to a backend server (it is not possible to store the entire contents of C_1 to C_m in the backend, due to resource limitations). The objective is to select k summary frames from the entire video.

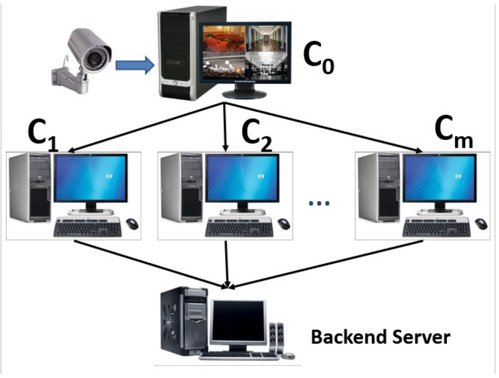


Figure 1: Distributed Video Summarization Setup

Summary Selection Criteria: Video streams have an inherent redundancy among them due to the high frame rate of video cameras. Thus, a good summarization technique should focus on identifying the critical events in a given video. However, if two summary frames separately furnish important, but duplicate information, then the net information gained by selecting both of them is not maximal. A metric computing the redundancy among a set of selected frames is therefore of paramount importance. We hence quantify the utility score of a set of summary frames in terms of the following two criteria:

- **Exhaustive**, which ensures that the summary captures a large portion of the events in the video
- **Mutually Exclusive**, which enforces the summary frames to be non-redundant (distinct from each other)

A summarization framework based on maximizing the exhaustive and mutually exclusive criteria ensures that the essence of the original video is captured well in the summary (exhaustive) and that there is minimal duplication of information (mutually exclusive). Such selection criteria are commonly used in selective sampling applications like active learning [13].

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 in the summary. The exhaustive-ness of the summary can be quantified as:

$$E(S) = \frac{1}{|V|} \sum_{i \in V, j \in S} w_{ij} \tag{1}$$

where $w_{ij} \geq 0$ denotes the similarity between two frames v_i and v_j in the video sequence. Maximizing $E(S)$ ensures that the summary set S has maximal similarity to the ground set V.

The mutually exclusive-ness score of a frame is computed as its minimum distance from the set of already selected summary frames:

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

where $d_{ij} \geq 0$ denotes the distance between two frames v_i and v_j. This is conceptually similar to the Hausdorff distance metric, commonly used for image matching [7]. Maximizing $M(S)$ avoids selection of duplicate information in the summary S. The net utility score of a summary set can be expressed as a weighted combination of $E(S)$ and $M(S)$:

$$Q(S) = E(S) + \lambda M(S) \tag{3}$$

where $\lambda \geq 0$ denotes the weight parameter. The summary selection can therefore be posed as the following optimization (k denotes the summary length):

$$\max_{S \subseteq V} Q(S)$$

$$\text{s.t.} \quad |S| = k \tag{4}$$

The search space being exponentially large, exhaustive search techniques are computationally prohibitive. We solve this problem using submodular optimization algorithms.

2.2 Submodular Optimization

Let N be a finite ground set and consider a function $f : 2^N \rightarrow \Re$, that returns a real value for any subset $S \subseteq N$. The function f is called submodular if for all $A \subseteq B \subseteq N$ and $x \in N \backslash B$,

$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B) \tag{5}$$

This is called the diminishing returns property [6]. Further, f is called monotonically non-decreasing if $f(A) \leq f(B)$ whenever $A \subseteq B$.

THEOREM 1. *The objective function $Q(S)$ defined in Equation (3) is submodular.*

PROOF. From Equation (3), we get:

$$\begin{aligned} Q(S) &= E(S) + \lambda M(S) \\ &= \frac{1}{|V|} \sum_{i \in V, j \in S} w_{ij} + \lambda \sum_{i \in S} \min_{j \in S : j < i} d_{ij} \end{aligned}$$

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

$$Q(S \cup \{x\}) - Q(S) = \frac{1}{|V|} \sum_{i \in V} w_{ix} + \lambda \min_{i \in S} d_{ix} \tag{6}$$

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

$$\min_{i \in S_2} d_{ix} \leq \min_{i \in S_1} d_{ix} \tag{7}$$

This holds since S_2 being a larger summary, it is possible to have an element in this superset, which is closer to the element x. The first term on the right side of Equation (6) does not depend on the set S; λ is a non-negative scalar. We therefore have,

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

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

Further, since both the distance and similarity-based terms are non-negative, the objective $Q(S)$ is monotonically non-decreasing as addition of elements to the set S can only increase the value of $Q(S)$. The optimization in Equation (4) is therefore the maximization of a monotonically non-decreasing submodular function, subject to a cardinality constraint. This can be efficiently solved using the greedy algorithm proposed by Nemhauser *et al.* [11], which produces a solution guaranteed to be within $1 - \frac{1}{e}$ of the optimal.

2.3 Distributed Submodular Maximization

In the distributed video summarization setup (outlined in Figure 1), however, the entire video V is not present in a single machine, but is distributed across multiple machines. Our summary selection problem therefore reduces to maximizing a monotone submodular function in a distributed setting. Mirzasoleiman *et al.* [10] proposed the GREEDI framework for distributed submodular maximization under cardinality constraints. The algorithm first distributes the ground set uniformly at random across the m machines. It operates in two stages, where in the first round, each machine separately runs the standard greedy algorithm [11] on its local data samples and selects k exemplars. In the second step, the selected mk elements are merged in the backend server (Figure 1) and k final summary frames are selected using the same greedy algorithm on the mk samples.

The computation of the term $E(S)$ in our algorithm (Equation 1) requires the complete ground set V, while evaluation of $M(S)$ (Equation 2) requires only the elements of the set S. In the first step, $E(S)$ is evaluated in a particular machine with respect to the local data samples in that machine. According to our definition, $E(S)$ satisfies the definition of decomposable functions [10] and thus, in the second step, it can be evaluated with respect to a randomly chosen subset of the ground set, of size $\lceil n/m \rceil$, where n is the size of the ground set V. This distributed maximization algorithm has a concrete mathematical guarantee on the quality of the obtained solution with respect to that of the optimum centralized solution (the solution obtained when the entire video is present in a single machine). This is formalized in the following theorem [10] (we take $\kappa = l = k$, and follow the notations in the paper); f is the submodular function being used for sample selection ($f \equiv Q$, in our case):

THEOREM 2. *Let $A^c[k]$ denote the optimum centralized solution and $A^{gd}[m,k]$ denote the solution obtained using the distributed maximization algorithm. Let $\delta > 0, \epsilon < 1/4$ and let n_0 be an integer such that for $n \geq n_0$ we have $\ln(n)/n \leq \epsilon^2/(mk)$. For $n \geq \max(n_0, m\log(\delta/4m)/\epsilon^2)$, we have, with probability at least $1 - \delta$,*

$$f(A^{gd}[m,k]) \geq (1 - 1/e)^2 (f(A^c[k]) - 2\epsilon)$$

3. EXPERIMENTS AND RESULTS

Datasets and Feature Extraction: We conducted experiments on a wide range of videos from different application domains to validate the generalizability of our approach: (1) **The UT Egocentric Video dataset**, which contains videos captured by a subject under unconstrained natural settings, using a wearable camera [9], (2) **DARPA's VIVID video surveillance dataset**, which contains low resolution aerial videos captured by an unmanned aerial vehicle (UAV) [4] and (3) **The SFU Skating dataset**, which contains unconstrained images with real pan, tilt and zoom of the camera capturing rapid moves of a skater like jumps, spins, lifts and turns [16]. Each video was split into images and the histogram of oriented gradients (HOG) feature [5] was used as a descriptor of each frame. The distance d_{ij} between two frames was computed using the chi-squared distance between their color histograms; the cosine similarity metric was used to compute the similarity.

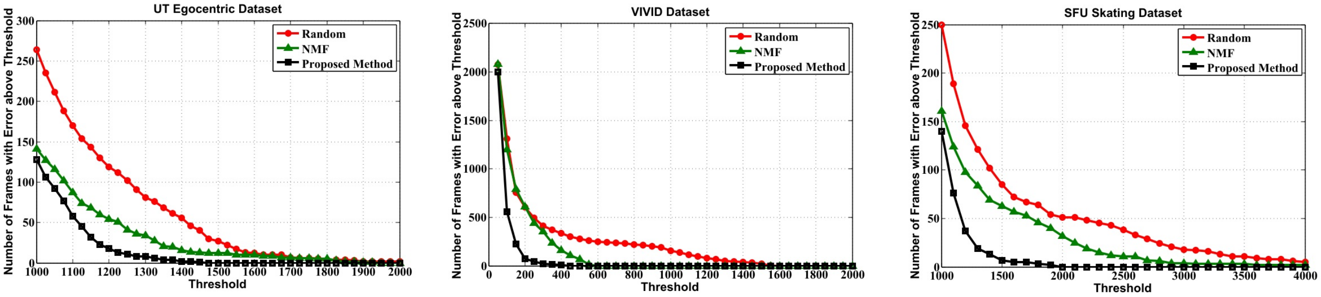
Experiment Set-up: The experimental set-up is outlined in Figure 1. Since this is our first effort to validate the proposed framework, we simulate the set-up in a single computer by partitioning the video data V into m segments (corresponding to the m machines) $V_1, V_2 \ldots V_m$, such that $V_i \cap V_j = \phi, \forall i, j, i \neq j$; no communication mechanism is assumed among the data in different segments (replicating the situation with m different machines). The outputs from the different segments are combined to select the final summary. Validation of the framework in a distributed infrastructure with multiple computers will be taken up as part of future research. The summary length k was taken as 50 and the number of machines m as 10 for the UT Egocentric and VIVID datasets; for the SFU Skating dataset, we take $k = 10$ and $m = 4$. The weight parameter λ was set as 5, based on preliminary experiments.

Comparison Baselines: We compare the proposed framework against the following two approaches for distributed summarization: (i) **Random**: in the first round, each machine selects k frames at random from its local contents and in the second round, k out of the mk frames are again selected at random in the final summary; (ii) **NMF**: the method proposed in [8] for distributed video summarization, where non-negative matrix factorization (NMF) is used in the first step to select k frames from each of the computers and the outputs are combined to select the final summary.

Evaluation Metric: To evaluate a video summary quantitatively, we compute the count of the number of frames whose reconstruction error using the summary is above a given threshold. The insight 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 [14] for details.

Experiment 1: Proposed against Baselines: The results comparing the proposed framework against the baselines are depicted in Figure 2, where the x-axis denotes the threshold and the y-axis denotes the number of frames with error above this threshold (lower the better). We note that for the proposed algorithm, the number of frames with reconstruction error above the threshold drops at the fastest rate with increasing values of this threshold; at any given threshold, it consistently has the lowest number of frames with error above that threshold. This corroborates the fact that our algorithm appropriately identifies the exemplar frames in a distributed setting and best captures the overall essence of a video. The improved performance is particularly evident for the SFU Skating video (Figure 2(c)), where a diverse variety of actions is performed by the skater. The images selected in the final summary by all the methods for this dataset are shown in Figure 3. Our method captures samples from a wide range of actions, thereby generating a more informative summary in a distributed setup, which accounts for its superior performance.

Experiment 2: Proposed vs. Centralized: We also compare the solution obtained by the proposed algorithm against the centralized solution that we would obtain if we had the resources to store the entire video in a single computer. The centralized solution is obtained by running the greedy algorithm (used in the first step of our framework) once on the entire video stream. Figure 4 reports the results on the VIVID dataset. We note that the performance of our algorithm is very close to the optimum centralized solution. This corroborates the fact that the proposed framework produces high quality solutions, even when the data is partitioned across multiple machines.



(a) UT Egocentric Dataset

(b) VIVID Surveillance Dataset

(c) SFU Skating Dataset

Figure 2: Comparison of the Proposed Method against the Baselines. Best viewed in color.

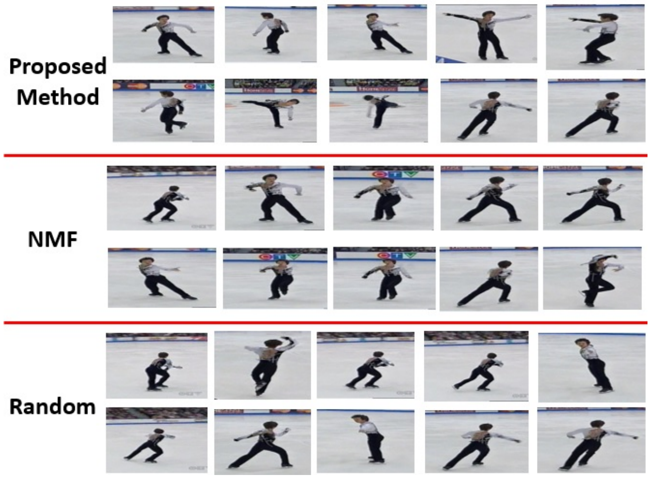


Figure 3: Images Selected in the Final Summary by all the Methods: SFU Skating Dataset

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel algorithm for distributed video summarization. We exploited distributed submodular maximization techniques to identify a set of exemplar frames from a video distributed across multiple machines. Our empirical results demonstrate tremendous promise in identifying the useful and relevant information from a video in a distributed environment. As part of our future work, we plan to conduct extensive large scale experiments to validate the merit of our framework.

5. REFERENCES

[1] S. Almeida, E. Cahuina, A. Araujo, G. Chavez, and D. Menotti. GPUs and multicore CPUs implementations of a static video summarization. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 2014.

[2] S. Almeida, A. Nazare, A. Araujo, G. Chavez, and D. Menotti. Speeding up a video summarization approach using GPUs and multicore CPUs. In *International Conference on Computational Science*, 2014.

[3] S. Avila, A. Lopes, A. Luz, and A. Araujo. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. In *Pattern Recognition Letters*, 2011.

[4] R. Collins, X. Zhou, and S. Teh. An open source tracking testbed and evaluation web site. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2005.

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

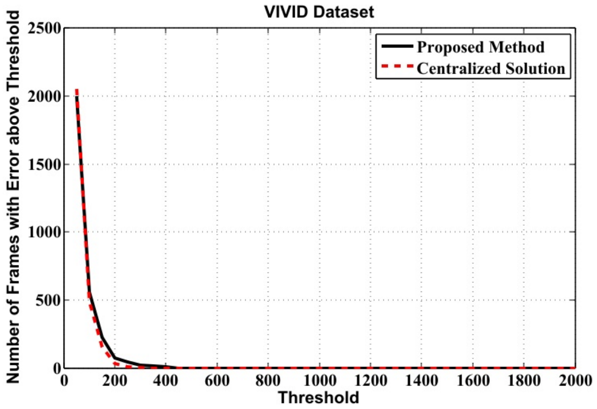


Figure 4: Comparison of the Proposed Framework against the Centralized Solution: VIVID Dataset. Best viewed in color.

[6] S. Fujishige. Submodular functions and optimization. In *Elsevier Science*, 2005.

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

[8] K. Ioannis, S. Tsevas, I. Maglogiannis, and D. Iakovidis. Enabling distributed summarization of wireless capsule endoscopy video. In *International Conference on Imaging Systems and Techniques*, 2010.

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

[10] B. Mirzasoleiman, A. Karbasi, R. Sarkar, and A. Krause. Distributed submodular maximization: Identifying representative elements in massive data. In *NIPS*, 2013.

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

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

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

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

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

[16] Y. Wang, H. Jiang, M. Drew, Z. Li, and G. Mori. Unsupervised discovery of action classes. In *CVPR*, 2006.

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