A Salient Dictionary Learning Framework for Activity Video Summarization Via Key-Frame Extraction

Ioannis Mademlis, Anastasios Tefas, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract

Recently, dictionary learning methods for unsupervised video summarization have surpassed traditional video frame clustering approaches. This paper addresses static summarization of videos depicting activities, which possess certain recurrent properties. In this context, a flexible definition of an activity video summary is proposed, as the set of key-frames that can both reconstruct the original, full-length video and simultaneously represent its most salient parts. Both objectives can be jointly optimized across several information modalities. The two criteria are merged into a "salient dictionary" learning task that is proposed as a strict definition of the video summarization problem, encapsulating many existing algorithms. Three specific, novel video summarization methods are derived from this definition: the Numerical, the Greedy and the Genetic Algorithm. In all formulations, the reconstruction term is modeled algebraically as a Column Subset Selection Problem (CSSP), while the saliency term is modeled as an outlier detection problem, a low-rank approximation problem, or a summary dispersion maximization problem. In quantitative evaluation, the Greedy Algorithm seems to provide the best balance between speed and overall performance, with the faster Numerical Algorithm a close second. All the proposed methods outperform a baseline

clustering approach and two competing state-of-the-art static video summarization algorithms.
Keywords: Video Summarization, Key-Frame Extraction, Column Subset Selection Problem, Video Saliency, Genetic Algorithm.

---

1. Introduction

Massive amounts of digital visual media data are publicly available nowadays, accelerating the transformation of global culture into a vision-dominated one [38]. Thus, the need for compact and succinct visual data presentation has arisen. It is a problem of broad interest in domains where large-scale video footage must be stored, archived, analysed or visualized, that typically demands tedious human intervention and manual effort.

Automated *video summarization* offers one solution to the video presentation problem, by generating concise versions of a video stream that only retain its most informative and representative content. Relevant algorithms are expected to meticulously strike a balance between summary compactness, conciseness, enjoyability and content coverage. *Static summarization* typically extracts a set of represenative video frames, i.e., *key-frames*, that in a sense summarize the entire video content. When editing cuts between clearly separated video shots are discernible (e.g., in television or film content [29]), a shot cut/boundary detector [4] is typically employed before key-frame extraction, to facilitate the summarization process by operating independently at each shot [16].

Several other possibilities exist for video summarization, such as *dynamic summarization* (also called *skimming*) [8], *video synopsis* [37] or *temporal video segmentation* exploiting semantic activity cues [39]. Despite their advantages,

they may only be suitable for specific applications (for instance, the use of synopses is limited to cases where the video frames are not visually crowded and retaining the original content is not a requirement), or even require key-frame extraction as a pre-processing / post-processing step. Thus, this paper focuses on key-frame extraction, used hereafter synonymously with video summarization.

In most of the relevant literature, video summarization is implicitly defined as a video frame sampling problem, constrained by an attempt to simultaneously satisfy several intuitive heuristic criteria, such as representativeness (extraction of key-frames that are jointly indicative of the original video content), compactness (lack of redundancy in the selected key-frames), outlier inclusion (selection of atypical key-frames) and content coverage (representation of the entire original video in the produced summary) [25]. Additionally, the summary should be as concise (i.e., short in length) as possible, or as desired by the end-user.

The traditional summarization method based on the constrained video frame sampling philosophy is video frame clustering, where frames closest to the estimated cluster centroids, or medoids, are selected as key-frames [44]. Thus, the video summarization problem is simply cast as a distance-based data partitioning task, with all semantic content description offloaded solely to the underlying video frame description/representation algorithm.

The modern alternative route is extracting a key-frame set as a dictionary of representative video frames that can linearly reconstruct the entire original video stream. This is an effective approach, supported by a sound theoretical background, that does not depend on shot cut/boundary detection or temporal video segmentation and formalizes the representativeness, compactness and content coverage criteria. Under a reasonable linear representatives assumption [12],

i.e., all original video frames can be approximately reconstructed as linear combinations of a representative subset of them, it can be argued that such methods, when supported by appropriate underlying video frame description/representation schemes, are able to incorporate scene semantics into the summarization algorithm itself. The reason is that the extracted key-frames will inherently tend to depict disjoint subsets of visually important scene objects, spatial segments, activities etc. In contrast, with a distance-based clustering approach, such a semantically meaningful partitioning of the key-frames will only be a serendipitous outcome.

However, dictionary-of-representatives approaches do not guarantee outlier inclusion. A related issue is that the reconstructive advantage conveyed by a video frame (i.e., the sole factor typically considered) cannot be the only criterion for its inclusion in the extracted key-frame set. The reason is that this leads to the extraction of unimportant video frames that happen to depict common but uninteresting visual building blocks of the entire video (e.g., the background), at the expense of engaging video frames, containing atypical visual elements which do not contribute much to the reconstruction.

Since most human activities can easily be decomposed into specific combinations of elementary actions [1], activity videos tend to satisfy the linear representatives assumption. However, the problems of dictionary-of-representatives approaches are especially pronounced when summarizing activity videos, due to their characteristic properties: static camera, static background, heavy inter-frame visual redundancy and lack of editing cuts. This paper, which integrates and extends preliminary work [30] [31], introduces a framework for activity video summarization that attempts to overcome the above issues. Its contributions are

four-fold.

First, video summarization is explicitly formalized under a flexible, multi-modal framework that can accommodate several existing algorithms as special cases. The proposed framework generalizes most current dictionary methods, which only consider the reconstructive ability, and conceptually places video summarization at the crossroad between video saliency estimation and video dictionary learning, thus defining it as a "salient dictionary" learning task. Second, three key-frame extraction algorithms are formulated in this context, where the video reconstruction term is modeled as a Column Subset Selection Problem (CSSP) [2]. This guarantees summary conciseness, favors summary compactness and had not been utilized for key-frame extraction before the preliminary work that this paper extends ([30]). Third, the saliency terms for the Numerical and for the Greedy algorithms are novel, video-oriented modifications of state-of-the-art image saliency algorithms ([26] and [10], respectively). Fourth, a novel metric, called Independence Ratio (IR), is proposed as an objective performance indicator of activity video key-frame extraction.

The three presented video summarization methods are compared against a baseline clustering approach [8] and two competing, dictionary-of-representatives state-of-the-art static video summarization algorithms [32] [7]. Each of the proposed algorithms is evaluated in five separate variants, characterized by a different balance between the reconstruction and the saliency term. All of the above algorithms, including both the proposed and the competing ones, are different, specific formulations of the presented salient dictionary learning framework for static video summarization.

2. Related Work

Current static video summarization methods can be broadly classified into supervised and unsupervised learning methods. Supervised approaches have surfaced lately [42] [43], in the wake of the success of deep learning. Such methods do not rely on heuristic summarization criteria, but attempt to implicitly learn them from human-created manual video summaries. However, due to the subjectiveness inherent in the problem (different persons may produce widely differing summaries from the same video source) and the lack of manual activity video summaries readily available for training in most use-cases, this paper is focused solely on unsupervised algorithms.

Established unsupervised approaches can be partitioned into the traditional video frame clustering methods [44] [8] [33] and the more modern dictionary-of-representatives algorithms [12]. General dictionary learning is an established methodology consisting in learning a dictionary of atoms from training data, such that the data themselves can be linearly re-expressed in terms of the dictionary (usually, as sparse codes) [35]. It has been applied on several tasks and extended in various ways. For instance, in [20] the method of Low-Rank Representation (LRR) [23], which segments data drawn from a union of originally unknown affine subspaces, is employed in an algorithm for human pose recovery from video footage. In [19], a multi-view extension of Laplacian Sparse Coding (LCR) [14], which preserves locality and similarity properties of data during coding, is also employed for human pose recovery.

Dictionary-of-representatives is an offshoot of general dictionary learning, well-suited to the key-frame extraction task, where the dictionary consists of unaltered training data points. In [12] such an approach is introduced for video sum-

marization, called sparse modeling representative selection, as a follow-up work to sparse subspace clustering [11]. An initial reconstruction objective function is defined, where the dictionary is given by the original data and the sparse codes to be estimated essentially select the representatives. A convex relaxation of the original objective is solved, while outliers are purposefully disregarded.

In [6] [32] video summarization is also framed under a dictionary-of-representatives perspective. In [6] convex relaxation using the $\mathcal{L}_{2,1}$ norm is employed, while in [32] an iterative algorithm called Minimum Sparse Reconstruction (MSR) attempts to indirectly minimize the actual $\mathcal{L}_0$ sparsity norm. In both cases, the outliers are entirely disregarded.

In [7] RPCA-KFE is presented, a key-frame extraction algorithm that takes into account both the contribution to video reconstruction and the distinctness of each video frame. The idea is to select as a summary the subset of video frames that simultaneously minimizes the aggregate reconstruction error and maximizes the total distinctness. RPCA-KFE is executed after performing Robust PCA [3], an operation entailing joint nuclear norm-based and $\mathcal{L}_1$ norm-based minimization that decomposes a matrix into a low-rank and a sparse component.

To the best of our knowledge, RPCA-KFE is the only relevant dictionary-of-representatives method that explicitly considers a saliency measure, thus easily fitting into the presented framework. However, its reliance on Robust PCA makes it inflexible, with the saliency and the reconstruction measures tightly coupled (as they are complementary in nature) and, thus, essentially fixed. Moreover, RPCA-KFE is designed to operate on spatially sub-sampled versions of the raw luminance video frame channels, thus ignoring semantic content elicited by feature-based video description/representation methods.

3. Key-Frame Extraction as a Salient Dictionary Learning Task

We assume that an input video consists in a temporally ordered set of N video frames. The desired output summary is an order-preserving subset of the input video, containing C of its elements/video frames ($N, C \in \mathbb{N}$, $0 < C < N$). Moreover, we assume that a preliminary video frame description and representation process, across K different video modalities, has resulted in the following video representation: a set of K matrices $\mathbf{D}_j \in \mathbb{R}^{V_j \times N}$, $0 \leq j < K$, where V_j is the representation vector size for the j-th modality. Column vector $\mathbf{d}^j_{:i}, 0 \leq i < N$ is the representation vector of the j-th modality of the i-th video frame. The desired output summary can either be expressed as a set of K matrices $\mathbf{S}_j \in \mathbb{R}^{V_j \times C}$, each one containing an ordered set of video key-frame representation vectors in the corresponding modality, or as a single binary-valued frame selection vector $\mathbf{s} \in \{0, 1\}^N$ indicating which frames of the original video are contained in the summary (equivalently, which columns of $\mathbf{D}_j$ are contained in $\mathbf{S}_j$). These two expressions are interchangeable.

The proposed framework can best be expressed by the simultaneous optimization of two terms: the first one (the "reconstruction" term) represents the ability of the extracted key-frame set to reconstruct the original full-length video, while the second one (the "saliency" term) represents the saliency of the extracted key-frame set, i.e., the degree to which its elements are distinct with regard to the complete video frame set, thus more likely to attract viewer attention. Thus, the desired video summary is defined as a salient dictionary of unaltered atoms/video frames, in a generalization of the dictionary-of-representatives tradition.

Both the reconstruction and the saliency terms are computed separately across all the K available video modalities, in order to jointly optimize different aspects

of the desired summary. The K modalities may be derived from K different description/representation processes pre-applied to the video frames. For instance, one such process may capture mid-level semantic information (e.g., activity, scene objects) and another one may convey low-level image properties (e.g., color, luminance or stereoscopic disparity distribution). If sound or text is available, they may also be described and represented as separate modalities in a per-frame fashion. In the simplest scenario, i.e., $K = 1$, all representation vectors for the i-th video frame (derived from different modalities) are concatenated into a single V-dimensional vector $\mathbf{d}_{:i}, 0 \leq i < N$, i.e., the i-th column of the single video matrix $\mathbf{D} \in \mathbb{R}^{V \times N}$.

Table 1 contains the symbols employed in the proposed framework. Given these definitions, the summarization process can be expressed in the following manner:

$$\begin{aligned} \min_{\mathbf{s}} :(1-\alpha) \sum_{j=0}^{K-1} (\|\mathbf{D}_j - \mathbf{S}_j \mathbf{A}_j\|_n) + \lambda R(\mathbf{s}) \quad - \\ - \alpha \sum_{j=0}^{K-1} (\mathbf{s}^T \mathbf{p}_j + L(\mathbf{S}_j)), \end{aligned} \tag{1}$$

where $\|\cdot\|_n$ is a matrix norm.

In the above expression, the goal is to find the binary-valued frame selection vector $\mathbf{s}$ that minimizes the objective. Since each matrix $\mathbf{S}_j$ consists in C of the columns of $\mathbf{D}_j$, selected based on the entries of $\mathbf{s}$, minimizing with respect to $\mathbf{s}$ is equivalent to finding the most suitable summary matrices $\mathbf{S}_j$.

Equation (1) is a general framework that can accommodate many different video summarization algorithms, given specific choices for the reconstruction,

Table 1: Summarization Nomenclature.

$\mathbf{D}_j \in \mathbb{R}^{V_j \times N}, 0 < j < K$	The j-th modality of the original video, with its i-th column $\mathbf{d}^j_{:i}, 0 \leq i < N$ representing the i-th video frame
$\mathbf{S}_j \in \mathbb{R}^{V_j \times C}, 0 < C < N$	The j-th modality of the desired video summary, with all its columns $\mathbf{s}^j_{:i}, 0 \leq i < C$ having been sampled without replacement from the set of all the columns in $\mathbf{D}_j$
$\mathbf{s} \in \{0,1\}^N$	A binary-valued frame selection vector, common for all modalities, according to which each $\mathbf{S}_j$ is constructed
$\mathbf{A}_j \in \mathbb{R}^{C \times N}$	A matrix containing reconstruction coefficients for the j-th video modality
$R(\mathbf{s}), \{0,1\}^N \rightarrow \mathbb{R}$	An application-specific regularization function
$\lambda \in [0, \infty)$	Regularizer weight
$\alpha \in [0,1]$	User-provided parameter regulating the contribution of the saliency terms
$\mathbf{p}_j \in \mathbb{R}^N$	A vector containing precomputed, constant per-frame saliency values for the j-th video modality
$L(\mathbf{S}_j), \mathbb{R}^{V_j \times C} \rightarrow \mathbb{R}$	A function assigning a global saliency value to a summary modality

saliency and regularization terms. The remainder of this Section provides concrete examples for these choices and describes more thoroughly the nomenclature defined in Table 1.

The reconstruction term is meant to bias key-frame selection towards video frames which constitute representative building blocks of the entire video content. Using the above definition, coefficients matrix $\mathbf{A}_j$ may be implicitly derived from any suitable video reconstruction method among several alternatives, which can be embedded in the proposed framework. Such methods are Principal Component Analysis (PCA), Orthogonal Subspace Projection (OSP) [17] or the Column

Subset Selection Problem (CSSP) [2].

Function $L(\mathbf{S}_j)$ assigns an aggregate saliency value to an entire candidate solution/summary, while the vector $\mathbf{p}_j$ contains a precomputed saliency value per video frame. Both operate on each video modality separately. $L(\mathbf{S}_j)$ is meant to be estimated adaptively, during the optimization process, and contributes a global perspective on the summary saliency, while $\mathbf{p}_j$ is fixed and conveys static, temporally localized information regarding the video content. It may not be necessary for both of them to be included in a specific implementation of the proposed framework, with a choice between them allowing better adjustment of summarization to specific applications.

Function $R(\mathbf{s})$ adds flexibility to the proposed framework by biasing the solution towards desired summary properties, e.g., in the form of a sum of matrix norms and/or vector dot products. For instance, it can be the $\mathcal{L}_0$ norm of a key-frame selection matrix, in order to enforce summary conciseness through a sparsity constraint, as in [32], or the Frobenius norm of a key-frame similarity matrix, to ensure summary compactness.

4. Salient Dictionary Learning on Activity Videos

Given the proposed framework and notation (detailed in Section 3), as well as our stated goal of summarizing activity videos via key-frame extraction (due to their characteristic properties, detailed in Section 1), three specific algorithms have been developed and are presented in this Section.

Without loss of generality, we assume below that the various available video representations (one for each modality) have been merged into a single representation by vector concatenation ($K = 1$). Therefore, in the following, a single

matrix $\mathbf{D}$, a single matrix $\mathbf{S}$, a single representation vector size V and a single precomputed per-frame saliency vector $\mathbf{p}$ replace the multiple matrices $\mathbf{D}_j$, $\mathbf{S}_j$, representation vector sizes V_j and vectors $\mathbf{p}_j$ ($0 \leq j < K$), respectively.

Actually following this route during key-frame extraction, would suggest a compromise between reducing the computational cost of multimodal representation and jointly exploiting multiple video aspects/modalities.

4.1. Video Reconstruction Using Column Subset Selection

Since activity videos are mainly composed of elementary actions assembled in various combinations [1], the linear representatives assumption is especially applicable in their case. Thus, it is reasonable to expect $\mathbf{D} \in \mathbb{R}^{V \times N}$ to have a pronounced low-rank structure. In the context of Equation (1), the reconstruction term should bias towards obtaining a solution $\mathbf{S}$ that can serve as a dictionary of representatives. $\mathbf{S}$ will contain unaltered columns of $\mathbf{D}$ which, ideally, form a set of linearly independent basis vectors that approximately span all columns of $\mathbf{D}$.

The Column Subset Selection Problem (CSSP) [2] has been selected for modeling the reconstruction term, due to its definition as directly selecting the "best" subset of columns of a given matrix. The cardinality of this subset is pre-fixed. These properties of CSSP make it well-suited to the problem of extracting a dictionary of representatives that is as succinct as desired.

Given $\mathbf{D}$ and a parameter $C < N$, the CSSP consists in selecting a subset of exactly C columns of $\mathbf{D}$ which form a new $V \times C$ matrix $\mathbf{S}$. This is equivalent to obtaining the most suitable binary-valued column selection vector $\mathbf{s} \in \{0, 1\}^N$. Given these definitions, the CSSP objective is the following one:

$$\min_{\mathbf{S}} : \|\mathbf{D} - (\mathbf{S}\mathbf{S}^{+})\mathbf{D}\|_F. \tag{2}$$

$\|\cdot\|_F$ is the Frobenius matrix norm and $\mathbf{S}^+$ is the pseudoinverse of $\mathbf{S}$. $\mathbf{S}$ approximates $\mathbf{D}$ in a projection sense: $\mathbf{S}\mathbf{S}^+$ projects $\mathbf{D}$ onto the span of the C columns contained in $\mathbf{S}$. Thus, $\mathbf{S}\mathbf{S}^+\mathbf{D}$ is the rank-C approximation of $\mathbf{D}$ achieved with the column subset matrix $\mathbf{S}$. Minimizing the corresponding reconstruction error is equivalent to finding a matrix $\mathbf{S}$ that is as close to full-rank as possible.

The CSSP is a combinatorial optimization problem, believed to be NP-hard and typically employed in a feature selection setting [13]. As exhaustive search requires $\mathcal{O}(N^C)$ time [2], approximate algorithms with lower computational complexity have been proposed, with the goal of finding a suboptimal but acceptable solution. In the context of this paper, three such approximate algorithms have been adopted and adapted under the proposed framework. First, a landmark numerical algorithm based on the Singular Value Decomposition (SVD) [2]. Second, a greedy approach that picks one column for addition to the currently selected subset at each iteration, such that the reconstruction error for the new subset is minimal [13]. Third, a genetic approach that directly uses Equation (2) as a fitness function [22].

Due to the nature of the CSSP, there is no need for a regularizing function $R(\mathbf{s})$ enforcing sparsity on the selection vector $\mathbf{s}$. The degree of summary conciseness is directly regulated by a strict, user-provided parameter C.

4.2. Video Saliency Using Outlier Detection and Low-Rank Approximation

Intuitively, the reconstruction term alone will tend to favor video frames solely containing common, elementary visual building blocks of the entire video, which facilitate the reconstruction process. These include not only video frames that are representative of the depicted activities, but also uninteresting video frames which do not contribute to discrimination among activities (e.g., emphasizing re-

curring static background, or human body poses common to multiple activities). Additionally, outlier video frames that do not significantly contribute to the reconstruction may be excluded, which is undesirable for a video summarization task.

4.2.1. Local Outlier Detection

The first of the proposed saliency terms models saliency on temporally localized outlier detection, following the established center-surround retinal receptive field organization paradigm [36]. Thus, the degree of estimated video frame saliency depends on how different a video frame is from its temporal neigbours. The method is adapted from the spatial, intra-frame component of the image saliency estimation algorithm presented in [10]. In the context of this paper, a preliminary saliency value is assigned to each video frame representation, instead of each raw image block. Correspondingly, spatial distance between the image blocks is replaced by temporal distance between video frames.

We define the fully connected, undirected, weighted distance graph $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ derived from the matrix $\mathbf{D}$, where $\mathbf{x}_i \in \mathcal{X}, 0 \leq i < N$, $\mathbf{y}_j \in \mathcal{Y}, 0 \leq j < N(N-1)/2$. Each vertex $\mathbf{x}_i$ corresponds to a column $\mathbf{d}_{:i}$ in $\mathbf{D}$, i.e., a video frame representation, and each edge $\mathbf{y}_j$ is weighted by the Euclidean distance between its two incident vertices/video frames, namely, the distance between the corresponding columns in $\mathbf{D}$, normalized by the temporal distance between the two vertices. The degree $deg(\mathbf{x}_i)$ of each vertex, i.e., the sum of the weights of all its incident edges, is employed as an initial measure of video frame saliency.

Thus, a preliminary saliency value for each column $\mathbf{d}_{:i}$ is computed. That is,

the i-th entry of an initial per-frame saliency vector $\hat{\mathbf{p}}$ is given by:

$$\hat{\mathbf{p}}_i = deg(\mathbf{x}_i) = \sum_{j=0}^{N-1} \left(\frac{\|\mathbf{d}_{:i} - \mathbf{d}_{:j}\|_2}{1 + |i - j|} \right). \tag{3}$$

The above saliency measure obviously favors local outlier inclusion. However, additionally, the most salient video frames should be temporally distant, similarly to how salient image regions are typically selected so as to be spatially distant, with less salient regions suppressed, in image saliency map estimation algorithms [26]. Such a consideration also fits well with the demands of video summarization, where the heuristic of maximum content coverage requires the extracted key-frames to be temporally dispersed, as long as this does not contradict the remaining summarization heuristic criteria (representativeness, compactness, conciseness and outlier inclusion).

Therefore, $\hat{\mathbf{p}}$ is post-processed in the following manner. Initially, the preliminary saliency value $\hat{\mathbf{p}}_i$ of video frame $\mathbf{d}_{:i}$ is subtracted from the average saliency of its temporal neighborhood $[i-M, i+M]$. This is implemented by first performing moving average filtering on $\hat{\mathbf{p}}$, using a filtering window of length $2M + 1$. Subsequently, all negative per-frame saliency values (corresponding to video frames which, on average, are less salient than their neighbors) are set to zero, giving rise to the final precomputed, per-frame saliency vector $\mathbf{p}$.

4.2.2. Regularized SVD-based Low-Rank Approximation

The second of the proposed approaches, models precomputed per-frame saliency on a regularized SVD-based reconstruction of $\mathbf{D}$, adapted from the no-learning raw image saliency estimation method in [26]. First, the SVD decomposition $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is obtained. Then, the singular values of $\mathbf{D}$, lying ordered on the diagonal of $\mathbf{\Sigma}$, are clustered into three groups: large, intermediate and small. The

large ones and the small ones are set to zero and, thus, the regularized matrix $\tilde{\mathbf{\Sigma}}$ is derived. Subsequently, the video matrix is approximately reconstructed using $\tilde{\mathbf{\Sigma}}$:

$$\tilde{\mathbf{D}} = \mathbf{U}\tilde{\mathbf{\Sigma}}\mathbf{V}^T. \tag{4}$$

The underlying intuition, as presented in [26], is that large, intermediate and small singular values correspond to non-salient/visually dominating image regions (e.g., the background), salient/important image regions and noise/fine-grained visual details, respectively. In this paper, the video frame representation $\mathbf{D}$ (encoding spatiotemporally varying content) is employed in place of raw image data (directly conveying spatially varying content). The proposed saliency term relies on the hypothesis that the above-mentioned intuition applies to such a scenario also.

In [26], the three largest singular values of the image matrix are simply set to zero. Here, instead, the singular values are adaptively clustered into three discrete groups (large, intermediate, small) using a fast, dynamic programming-based variant [18] of the Jenk's Natural Breaks Optimization algorithm for one-dimensional clustering [21]. The latter operates by exploiting a scalar version of the Fisher ratio, typically employed in Linear Discriminant Analysis (LDA), thus by attempting to simultaneously minimize intra-cluster variance and maximize inter-cluster variance.

The resulting matrix $\tilde{\mathbf{D}}$ is, in essence, a two-dimensional spatiotemporal video saliency map. A preliminary saliency value for the i-th video frame can easily be extracted from $\tilde{\mathbf{D}}$ in the following manner:

$$\tilde{\mathbf{p}}_i = \|\tilde{\mathbf{d}}_{:i}\|_1, \tag{5}$$

where $\tilde{\mathbf{d}}_{:i}$ is the the i-th column of $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{p}}$ is a preliminary, per-frame saliency

vector.

The final, precomputed per-frame saliency vector $\mathbf{p}$ can then be derived by applying the post-processing saliency enhancement step, previously described in Section 4.2.1, on $\tilde{\mathbf{p}}$.

4.2.3. Maximum-Dispersion Global Summary Saliency

Finally, a global summary saliency function is also examined in the context of the proposed algorithms:

$$L(\mathbf{S}) = \operatorname{trace}\left(\mathbf{S}\mathbf{S}^T\right). \tag{6}$$

This global saliency term simply seeks to maximize the dispersion of the desired summary, in order to indirectly push towards greater compactness and outlier inclusion.

4.3. Solving the Optimization Problem

The desired solution is the binary-valued video frame selection vector $\mathbf{s} \in \{0,1\}^N$, or, equivalently, the actual video summary representation $\mathbf{S} \in \mathbb{R}^{V \times C}$, constructed based on $\mathbf{s}$ ($\|\mathbf{s}\|_0 = C$).

Three different algorithms were devised for extracting the solution, i.e., three concrete implementations of the proposed salient dictionary learning framework, by coupling one of the CSSP algorithms discussed in Subsection 4.1 with one of the saliency terms discussed in Subsection 4.2. The global saliency term could only be combined trivially with the Genetic CSSP method. Precomputed per-frame saliency terms were only employed for both the Numerical and the Greedy Algorithm.

4.3.1. Numerical Algorithm

In the context of the proposed framework, the objective function implicitly optimized in the Numerical Algorithm is the following:

$$\min_{\mathbf{s}} : (1-\alpha)\|\mathbf{D} - \mathbf{S}\mathbf{S}^{+}\mathbf{D}\|_F - \alpha \mathbf{s}^T \mathbf{p}, \tag{7}$$

where the entries of $\mathbf{p}$ are a priori given by the proposed Regularized SVD-based Low-Rank Approximation saliency estimation method. A solution to this optimization problem is obtained in the way described below.

The Numerical Algorithm relies on a landmark SVD-based method for solving the CSSP [2], based on the notion of *statistical leverage score sampling* [9]. The intuition behind it is that the SVD decomposition $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ can be exploited to quickly remove the less outlying columns from $\mathbf{D}$ in a preliminary, randomized step, before deterministically constructing the final matrix $\mathbf{S}$. Below, $\mathbf{V}_C \in \mathbb{R}^{N \times C}$ denotes the matrix whose columns are the top C right singular vectors of $\mathbf{D}$.

The method operates in two stages. First, approximately $C\log C$ columns are randomly sampled from matrix $\mathbf{D}$. Assuming that p_i is the probability of selecting the i-th column of $\mathbf{D}$, sampling follows a probability distribution computed based on information coming from the top-C right singular subspace of $\mathbf{D}$, which is spanned by the columns of $\mathbf{V}_C$:

$$p_i = \|(\mathbf{V}_C)_{i:}\|_2^2 / C, \tag{8}$$

where $(\mathbf{V}_C)_{i:}$ denotes the i-th row of $\mathbf{V}_C$.

In the second stage, exactly C columns are selected from the sample using any deterministic CSSP algorithm (e.g., based on Rank-Revealing QR decomposition [5]).

In order to adapt the method to the proposed framework, matrix $\mathbf{D}$ is modified in the following manner:

$$\hat{\mathbf{D}} = (1-\alpha)\mathbf{D} + \alpha\mathbf{D}\left(\mathbf{diag}(\mathbf{n})\mathbf{diag}(\mathbf{p})\right), \tag{9}$$

where $\mathbf{n} \in \mathbb{R}^N$ is a vector containing normalization coefficients, so as to map the precomputed saliency factors to the interval $[0, 1]$. In $\hat{\mathbf{D}}$, less salient columns (corresponding to less salient video frames) have been scaled down to a degree directly proportional to their saliency and to the provided saliency contribution parameter α. Subsequently, the algorithm in [2] is applied on $\hat{\mathbf{D}}$, in order to obtain the desired summary.

4.3.2. Greedy Algorithm

In the context of the proposed framework, the objective function implicitly optimized in the Greedy Algorithm is also given by Equation (7), but the entries of $\mathbf{p}$ are provided by the proposed Local Outlier Detection saliency estimation method, described in Eq. (3). A solution to this optimization problem is obtained in the way described below.

The Greedy Algorithm is based on an efficient iterative method for approximately solving the CSSP [13]. One video frame is added to the currently extracted key-frame set at each iteration, so that the reconstruction error is greedily minimized, until C key-frames have been extracted (equivalently, after C iterations). Therefore, for the t-th iteration, the following quantities are defined:

1. $\mathbf{s}^{t-1}$: the currently extracted key-frame set/summary binary selection vector, prescribing the current summary $\mathbf{S}^{t-1}$. It holds that $\|\mathbf{s}^{t-1}\|_0 = t-1$.
2. $\overline{\mathcal{R}}^{t-1}$: the set of the temporal indices of all video frames not contained in $\mathbf{S}^{t-1}$. It contains $N-(t-1)$ elements, all in the interval $[0, N-1]$.

3. l^t: the temporal index of the video frame $\mathbf{d}_{:l_t}$ that is actually selected for inclusion in $\mathbf{S}^t$ during iteration t. Obviously, $l^t \in \overline{\mathcal{R}}^{t-1}$, but $l^t \notin \overline{\mathcal{R}}^t$.

The method operates by recursively maintaining two vectors, $\mathbf{f}, \mathbf{g} \in \mathbb{R}^N$. Each one keeps track of a scalar score for each video frame $\mathbf{d}_{:i}, 0 \leq i < N$. At the start of the t-th iteration, the most suitable l^t is selected for addition to the extracted key-frame set/summary in the following manner:

$$l^t = \arg\max_i \frac{f_i^{t-1}}{g_i^{t-1}}, \quad i \in \overline{\mathcal{R}}^{t-1}, \tag{10}$$

where f_i^{t-1}, g_i^{t-1} is the i-th entry of current vector $\mathbf{f}, \mathbf{g}$, respectively. Subsequently, $\mathbf{f}^t$ and $\mathbf{g}^t$ are computed, by updating $\mathbf{f}^{t-1}$ and $\mathbf{g}^{t-1}$ based on the value of l^t. The specific formulas for initializing and updating $\mathbf{f}$ and $\mathbf{g}$, as well as their derivation, can be found in [13], in the form of the so-called "Memory-Efficient Criterion".

In order to adapt the method to the proposed framework, $\tilde{\mathbf{p}} \in \mathbb{R}^N$ is initially precomputed once. It is a modified version of $\mathbf{p}$, with its entries (the per-frame saliency factors) normalized into the interval $[0, 1]$. Subsequently, the Greedy Algorithm is iteratively executed as described above, but Equation (10) is modified in the following manner:

$$l^t = \arg\max_i \left((1-\alpha)\frac{f_i^{t-1}}{g_i^{t-1}} + \alpha\tilde{p}_i \frac{f_i^{t-1}}{g_i^{t-1}} \right), \quad i \in \overline{\mathcal{R}}^{t-1}. \tag{11}$$

where $\tilde{p}_i$ is the i-th entry of $\tilde{\mathbf{p}}$. Thus, at each iteration, vectors $\mathbf{f}$ and $\mathbf{g}$ are updated based on the reconstructive advantage currently conveyed by each video frame, but the actual selection of a candidate video frame for inclusion in the summary also depends on its precomputed saliency and the provided saliency contribution parameter α. The algorithm is completed after C iterations.

4.3.3. Genetic Algorithm

The Genetic Algorithm is the third proposed implementation of the key-frame extraction framework defined in Equation (1). It approximates an optimal solution to the problem prescribed by the following objective function:

$$\max_{\mathbf{S}} : -(1-\alpha)\|\mathbf{D} - \mathbf{S}\mathbf{S}^{+}\mathbf{D}\|_F + \alpha \operatorname{trace}\left(\mathbf{S}\mathbf{S}^{T}\right). \tag{12}$$

Therefore, it is a straightforward combination of the CSSP definition and the Maximum-Dispersion global saliency term from Section 4.2.3, using a typical genetic approach. Equation (12) is directly employed as a fitness function. Each candidate is encoded in the form of a sequence of integer column indices sorted in increasing order, with every such candidate/chromosome having length C. Deterministic tournament selection at each iteration is adopted as the mating pool formation strategy. Order-preserving variants of 1-point crossover and of mutation are adopted from [22], where a genetic approach was first introduced for solving the CSSP. They were also employed in [29] [28], using a slightly different problem formulation and fitness function, for multimodal shot selection in movie summarization.

4.4. Computational Complexity Analysis

The computational complexity of the proposed methods is shown in Table 2, where competing algorithms are also included for reference purposes. **NUM**, **GRE** and **GEN** denote the Numerical, the Greedy and the Genetic Algorithm, respectively. In the Genetic Algorithm, P and G refer to the population size and the number of generations, respectively. Details are provided in the Supplementary Material.

Table 2: Computational Complexity of Activity Video Summarization Algorithms.

Method	Complexity Class
NUM	$\mathcal{O}(min\{VN^2, V^2N\})$
GRE	$\mathcal{O}(VN^2)$
GEN	$\mathcal{O}(PGV^2N)$, if $V < C$ $\mathcal{O}(max\{PGVCN, PGV^2C\})$, if $V > C$
[8]	$\mathcal{O}(VCN)$
[32]	$\mathcal{O}(CNV^2)$
[7]	$\mathcal{O}(VCN^2)$

5. Quantitative Evaluation

5.1. Video Description and Representation

Provenly effective hand-engineered video description and representation methods were employed for deriving the original video representation $\mathbf{D}$. Three different feature descriptors were applied per video frame, using their default parameters: LMoD [27] (low-level, local, spatial descriptor, computed on four different video frame channels, i.e., luminance, color hue, optical flow magnitude and edge map), SIFT [24] (mid-level, local, spatial, semantic descriptor) and Improved Dense Trajectories (IDT) [40] (mid-level, local, spatiotemporal, semantic descriptor). The resulting three descriptions per video frame can be seen as different modalities, since they capture different aspects of the underlying data. The Improved Fisher Vector (IFV) approach [34] was selected for feature aggregation per video frame (and, thus, video frame representation).

5.2. Evaluation Datasets

Single-view subsets of three publicly available, annotated, multi-view activity video datasets were employed. The datasets were slightly processed to better suit an activity video summarization task (e.g., several videos, each one depicting a single activity, were temporally concatenated, so as to form a long video composed of multiple consecutive activities). In each case, a specific camera angle was chosen from the original multi-view dataset for all activity sessions. The processed versions are briefly described below:

1. The IMPART video dataset [39], depicting 3 actors in 2 different settings: an outdoor one and a living-room. A total of 116 indoor and 214 outdoor activity sessions with static camera are included, where the actors perform a series of activities one after another, moving along approximately fixed trajectories via predefined waypoints. 4 different activity types were performed, namely "Walk", "Hand-wave", "Run" and "Other". The dataset consists of 6 video files with a resolution of 720 $\times$ 540 pixels and mean duration of about 4542 video frames.
2. The IXMAS dataset [41], depicting 10 actors in an indoor setting. A total of 467 activity sessions with static camera are included, where the actors perform a series of activities one after another, with varying/unconstrained body poses. In total, 11 different activities were performed. The dataset consists of 4 video files with a resolution of 390 $\times$ 290 pixels and mean duration of about 9055 video frames. This is the most challenging dataset, due to the low video resolution, the relatively high number of video frames and activity segments, as well as the very high visual similarity between video frames belonging to different activity segments.

3. The i3DPOST dataset [15], depicting 8 actors in a blue-screen backdrop. A total of 104 activity sessions with static camera are included, where either the actors perform a series of activities one after another, moving along approximately fixed trajectories, or two actors interact. In total, 12 different activities were performed. The dataset consists of 3 video files with a resolution of 640 $\times$ 480 pixels and mean duration of about 5358 video frames.

5.3. Evaluation Metric

Video summarization methods are typically evaluated either subjectively, where a group of end-users rates the *informativeness* and the *enjoyability* of the extracted summary (e.g., in [29]), or in a semi-objective manner, where several video frames have been manually pre-selected by end-users, according to their judgment, as "ground-truth" key-frames. In the latter case, the automatically extracted key-frames can be compared to this "ground truth" using various evaluation metrics, including the F-Score (as in [32]) or the summarization-specific "Comparison of User Summaries" (CUS) metric, introduced in [8]. However, a significant degree of subjectivity is unavoidable due to reliance on manually prepared summaries.

During empirical evaluation of this work, the special characteristics of activity videos were exploited so as to avoid the subjectivity outlined above. Temporal video segmentation ground-truth annotation data, describing obvious temporal boundaries between consecutive activity video segments, were employed for evaluating the proposed methods as objectively as possible, similarly to [30]. Given a summary $\mathbf{s}$ of an input video $\mathbf{D}$, the number I_s of extracted key-frames derived from actually different activity segments (hereafter called *independent key-frames*) is used as an indirect indication of summarization success. Obviously, I_s equals the number of different activity segments represented in the summary $\mathbf{s}$.

Thus, the *Independence Ratio* (IR) score is defined:

$$IR(\mathbf{s}) = \frac{I_s}{C}, \tag{13}$$

where C is the total number of requested key-frames. That is, the percentage of independent key-frames in the set of all extracted key-frames serves as a practically objective evaluation metric, automatically adjustable to the desired degree of summary conciseness (regulated by the user through C).

This way, any two video frames belonging to the same activity segment are treated as interchangeable and, from the aspect of its empirical evaluation, video summarization is reduced to a variant of temporal video segmentation, so as to bypass any subjective judgement regarding which video frames are more representative of a specific activity segment.

5.4. Algorithm Parameters

A crucial, user-provided parameter controlling the grain of summarization is the desired number of requested key-frames per video (C, i.e., the length of the summary). It corresponds to the number of clusters, in clustering, and of columns of $\mathbf{S}$, in the proposed methods. In video summarization literature, C is typically set to a fixed or adaptable percentage of the original video duration (e.g., [8], [33], [29]), unless the algorithm itself converges to its own estimation of C (e.g., [32]). In this work, in order to most effectively compare the competing algorithms, the actual number of different activity segments (known from the ground truth) was employed as C for each video. This was also used in our implementation of [32], so as to achieve a fair comparison. However, in principle, it is perfectly possible to rely on any pre-existing method for estimating a proper C.

Regarding the proposed algorithms, saliency contribution parameter α was set to five different values: 0.00, 0.25, 0.50, 0.75, 1.00, ordered from less to more contribution of the saliency term at the expense of the reconstruction term. Therefore, in total, 15 different variants of the proposed methods were evaluated, including reconstruction-only ($\alpha = 0.00$) and saliency-only ($\alpha = 1.00$) forms. Note that when $\alpha = 0.00$, the proposed algorithms reduce to traditional CSSP applied for key-frame extraction.

Other implementation issues are detailed in the Supplementary Material.

5.5. Evaluation Setup

The 15 presented video summarization method variants were compared against a baseline clustering approach [8] and two competing, state-of-the-art static video summarization algorithms: OffMSR from [32] and RPCA-KFE from [7]. OffMSR was selected as a relatively recent dictionary-of-representatives video summarization method. RPCA-KFE was selected not only because it fits perfectly within the proposed framework, being a dictionary-of-representatives approach that also considers a form of video frame saliency, but also because it is a method operating on raw pixel luminance video frame representations, unlike the feature-based proposed approaches. To achieve a fair comparison, the video representation scheme described in Section 5.1 was employed for all competing feature-based methods, but the proposed deterministic Greedy Algorithm was additionally tested with the 4800-dimensional raw pixel luminance video frame representations employed by RPCA-KFE. Results of trivial key-frame extraction via random video frame sampling are also presented for comparison purposes.

In order to achieve a reasonable execution time, videos were pre-sampled in the case of RPCA-KFE (every tenth video frame was retained), as suggested in

[7]. This was not necessary for the other methods. All other algorithm parameters were tuned according to the original papers, in the cases of OffMSR and RPCA-KFE.

The algorithm underpinning key-frame extraction in [8] is simple K-Means clustering, similarly to many other video summarization methods which only differ with respect to the employed video description/representation scheme and are based on K-Means or K-Medoids variants (e.g., [33]). Since our own choice for video representation was used in all methods, comparing the proposed algorithms with [8] is (to a degree) equivalent to comparing them with any of the above-mentioned clustering approaches.

5.6. Evaluation Results

Below, **NUM**, **GRE** and **GEN** denote the Numerical, the Greedy and the Genetic Algorithm, respectively, while **GRE-RAW** refers to the Greedy Algorithm using the raw pixel luminance video frame representations employed by RPCA-KFE. Tables 3, 4 and 5 present the mean IR scores obtained by all competing methods, for the IMPART, i3DPOST and IXMAS datasets, respectively. The presented IR scores for random sampling have been averaged over 1000000 executions per video. The presented IR scores for all randomized algorithms (both from the proposed and the competing ones) have been averaged over 5 executions per video. In all tables, the best IR performance for each value of α is highlighted in bold.

Tables 6, 7 and 8 present the mean required execution time per-frame (in milliseconds) for all competing methods, for each dataset. These measurements do not include the computation time necessary for initial video description/representation. In all tables, the fastest algorithm runtime for each value of α is highlighted in

Table 3: Mean IR scores for all competing methods in the IMPART dataset (higher is better).

	Random	**NUM**	**GRE**	**GEN**	[8]	[32]	[7]	**GRE-RAW**
$\alpha = 0.00$	58.86%	72.16%	75.21%	**75.85%**	72.94%	68.03%	47.15%	68.97%
$\alpha = 0.25$	-	69.86%	**74.74%**	74.72%	-	-	50.17%	70.71%
$\alpha = 0.50$	-	70.40%	**75.05%**	75.03%	-	-	48.70%	65.04%
$\alpha = 0.75$	-	68.80%	**77.13%**	76.01%	-	-	49.13%	63.57%
$\alpha = 1.00$	-	56.09%	**63.63%**	58.49%	-	-	43.17%	52.00%

Table 4: Mean IR scores for all competing methods in the i3DPOST dataset (higher is better).

	Random	**NUM**	**GRE**	**GEN**	[8]	[32]	[7]	**GRE-RAW**
$\alpha = 0.00$	59.01%	70.94%	67.95%	72.56%	**72.65%**	65.81%	37.18%	67.52%
$\alpha = 0.25$	-	71.62%	**74.79%**	72.05%	-	-	44.87%	67.52%
$\alpha = 0.50$	-	**75.64%**	70.94%	70.68%	-	-	40.60%	71.79%
$\alpha = 0.75$	-	**73.93%**	73.50%	71.37%	-	-	32.91%	68.80%
$\alpha = 1.00$	-	62.65%	62.39%	56.75%	-	-	20.09%	**67.09%**

Table 5: Mean IR scores for all competing methods in the IXMAS dataset (higher is better).

	Random	**NUM**	**GRE**	**GEN**	[8]	[32]	[7]	**GRE-RAW**
$\alpha = 0.00$	59.40%	**66.33%**	62.94%	62.00%	65.29%	66.16%	46.66%	61.02%
$\alpha = 0.25$	-	**66.38%**	61.22%	60.07%	-	-	44.73%	61.88%
$\alpha = 0.50$	-	**65.08%**	62.07%	62.90%	-	-	45.38%	61.24%
$\alpha = 0.75$	-	**63.80%**	60.58%	60.07%	-	-	46.66%	61.24%
$\alpha = 1.00$	-	61.87%	**67.66%**	59.22%	-	-	37.91%	59.32%

Table 6: Mean execution time per video frame (in milliseconds) for all competing methods on the IMPART dataset.

	NUM	**GRE**	**GEN**	[8]	[32]	[7]	**GRE-RAW**
$\alpha = 0.00$	17.90	1.26	552.92	76.85	4043.82	425.34	$\mathbf{0.46}$
$\alpha = 0.25$	$\mathbf{45.96}$	216.24	551.07	-	-	427.84	61.07
$\alpha = 0.50$	$\mathbf{45.28}$	216.26	553.59	-	-	431.45	61.06
$\alpha = 0.75$	$\mathbf{44.98}$	216.27	553.58	-	-	420.84	61.07
$\alpha = 1.00$	36.14	216.27	1.89	-	-	$\mathbf{0.401}$	61.05

Table 7: Mean execution time per video frame (in milliseconds) for all competing methods on the i3DPOST dataset.

	NUM	**GRE**	**GEN**	[8]	[32]	[7]	**GRE-RAW**
$\alpha = 0.00$	11.28	1.26	517.80	70.01	2544.20	400.37	$\mathbf{0.43}$
$\alpha = 0.25$	$\mathbf{41.49}$	251.26	517.85	-	-	385.35	75.71
$\alpha = 0.50$	$\mathbf{42.05}$	251.29	517.13	-	-	410.03	75.70
$\alpha = 0.75$	$\mathbf{43.10}$	251.27	519.45	-	-	380.41	75.68
$\alpha = 1.00$	36.49	251.26	0.99	-	-	$\mathbf{0.76}$	75.67

bold.

One should notice the similarity between the no-saliency variant of the proposed GRE (when $\alpha = 0.00$) and the OffMSR method in [32]: they are both iterative procedures attempting to greedily optimize almost identical objectives (the CSSP and the OSP definitions, respectively) and, thus, incrementally construct the desired summary. However, GRE is not only entire orders of magnitude faster (due to the intelligent way it keeps track of the reconstructive advantage each video frame conveys per iteration, using the $\mathbf{f}$ and $\mathbf{g}$ vectors), but also leads

Table 8: Mean execution time per video frame (in milliseconds) for all competing methods on the IXMAS dataset.

	NUM	**GRE**	**GEN**	[8]	[32]	[7]	**GRE-RAW**
$\alpha = 0.00$	33.75	2.42	734.34	225.45	8594.31	897.57	$\mathbf{0.96}$
$\alpha = 0.25$	$\mathbf{80.82}$	427.44	748.63	-	-	905.70	130.49
$\alpha = 0.50$	$\mathbf{81.41}$	427.44	760.21	-	-	900.88	130.46
$\alpha = 0.75$	$\mathbf{82.01}$	427.42	749.12	-	-	891.55	130.48
$\alpha = 1.00$	65.92	427.41	2.52	-	-	$\mathbf{1.28}$	130.48

to much improved IR scores.

Secondly, RPCA-KFE completely fails to handle the desired task, performing far below even random sampling. The IR performance of GRE-RAW (significantly higher than random sampling, lower than GRE), indicates that this cannot be solely attributed to the limitations of raw pixel luminance video frame representations (when compared to feature-based representations). Therefore, it is reasonable to conclude that RPCA-KFE is not suitable for processing activity videos having the property of heavy visual inter-frame redundancy. The behaviors of RPCA-KFE and GRE-RAW on a specific video from the i3DPOST dataset are visualized and contrasted in Figure 1. As it can be seen, RPCA-KFE mainly favors specific activity segments, assessed as more salient by the algorithm, producing multiple key-frames for them, at the expense of other activity segments that remain under-represented in the produced summary.

In general, GRE and NUM provide the best IR performance, while NUM is the fastest method (in fact, near-real-time). When both reconstruction and saliency terms are considered in the experiments (i.e., $\alpha = 0.25$, $\alpha = 0.50$ or

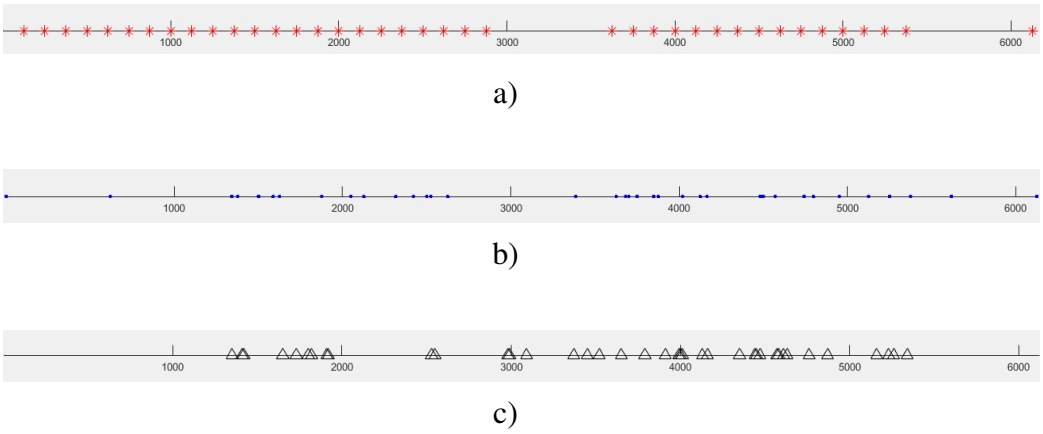

Figure 1: Behavior of RPCA-KFE and GRE-RAW algorithms in an example video composed of 6125 video frames: a) The ground truth activity segment boundaries are marked as red stars, b) The key-frames selected by GRE-RAW are marked as blue dots, c) The key-frames selected by RPCA-KFE are marked as black triangles.

$\alpha = 0.75$), one of the three proposed methods is the best performer in all three employed datasets: GRE in IMPART, NUM in i3DPOST and IXMAS. When only the reconstruction term is considered ($\alpha = 0.00$), as in traditional dictionary-of-representatives methods, one of the proposed methods is the best performer in two out of three datasets (GEN in IMPART, NUM in IXMAS), while clustering-based [8] performs best in i3DPOST. The latter result highlights the contribution of the saliency term in good key-frame extraction performance when employing dictionary-of-representatives methods, which is the main idea behind the proposed framework.

When only the saliency term is considered ($\alpha = 1.00$, i.e., no reconstruction term), one of the three proposed methods is the best performer in all three employed datasets: GRE in IMPART and IXMAS, GRE-RAW in i3DPOST. Local Outlier Detection alone, i.e., the saliency term of GRE, leads to acceptable results,

while Maximum-Dispersion Global Summary Saliency alone, i.e., the saliency term of GEN, fails to surpass even random video frame sampling.

Absolute best performance, over all tested values of α, is obtained by one of the proposed methods, employing both reconstruction and saliency terms, in two out of three datasets (GRE with $\alpha = 0.75$ in IMPART, NUM with $\alpha = 0.50$ in i3DPOST). In the third dataset IXMAS, absolute best performance is also achieved by one of the proposed methods (GRE), when only considering the saliency term of Local Outlier Detection ($\alpha = 1.00$). This is most likely due to the very challenging nature of this dataset (low resolution, very high visual similarity between video frames belonging to different activity segments), which results in enhanced reconstructive advantage conveyed by the background and by human body poses common across multiple different activity segments, at the expense of summarization performance. This outcome also highlights the contribution of the saliency term in good key-frame extraction.

The execution time deviates significantly for the extreme values of saliency contribution parameter α, i.e., for $\alpha = 0.00$ and $\alpha = 1.00$. This is because for $\alpha = 0.00$ the saliency term is not calculated at all, while for $\alpha = 1.00$ less computations have to be performed, since there is no need to fuse the degree of reconstruction advantage and the degree of saliency corresponding to each video frame (only saliency is considered). In the cases of GEN and RPCA-KFE, the reconstruction term is not calculated at all for $\alpha = 1.00$, resulting in extremely fast execution since their saliency terms are much faster to compute than their reconstruction terms. In contrast, GRE runs extremely fast for $\alpha = 0.00$ (no saliency term), since its reconstruction term is much faster to compute than Local Outlier Detection. Overall, Regularized SVD-based Low-Rank Approximation,

i.e., the saliency term of NUM, balances very well speed and IR performance, while GRE implements the fastest reconstruction term.

As a final note, one should notice the relatively good performance of simple K-Means clustering in activity video summarization, which explains its continuing dominance in key-frame extraction literature.

6. Conclusions and Discussion

Static activity video summarization was explicitly formalized under a flexible, multimodal framework that follows the unsupervised dictionary-of-representatives approach, thus integrating video semantics into the summarization process itself and guaranteeing summary representativeness. The proposed framework can accommodate several existing algorithms as special cases and firmly locates video summarization at the intersection of video saliency estimation and video dictionary learning. Thus, unlike most dictionary-of-representatives methods, key-frame extraction is defined as a salient dictionary learning task that attempts to balance a reconstruction and a saliency term, guaranteeing outlier inclusion.

In this context, three key-frame extraction algorithms were formulated as concrete instances of the presented framework. The video reconstruction term was modeled algebraically as a Column Subset Selection Problem (CSSP), thus favoring summary conciseness and compactness, while the saliency term was modeled as an outlier detection, a low-rank approximation, or a summary dispersion maximization problem, thus favoring outlier inclusion and/or content coverage in the summary. A numerical, SVD-based approach (Numerical Algorithm), a recursive, greedy approach (Greedy Algorithm) and an evolutionary, genetic approach (Genetic Algorithm) were detailed. In all cases, pre-existing solutions to

the CSSP have been modified and adapted to the proposed framework. Moreover, the saliency terms for the Numerical and for the Greedy algorithms are novel, video-oriented modifications of state-of-the-art image saliency algorithms. The first one exploits regularized SVD-based low-rank approximation of the video data matrix, while the second one models video frame saliency estimation upon local outlier detection.

Objective quantitative evaluation results on three public, activity video datasets, using a combination of provenly effective video description/representation methods, indicated the superiority of the proposed framework/algorithms in comparison to traditional data partitioning/clustering, as well as to competing dictionary-of-representatives approaches. The contribution of the saliency term to this effect is especially highlighted. In general, the Greedy Algorithm seems to provide the best balance between speed and overall performance, with the faster Numerical Algorithm a close second. Additionally, the importance of good, semantically meaningful video frame representations was showcased by comparing the performance of the Greedy Algorithm when using a rich, feature-based representation and a raw pixel luminance value representation.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 287674 (3DTVS) and 316564 (IMPART).

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16:1–16:43, 2011.

[2] C. Boutsidis, M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the Column Subset Selection Problem. In *Proceedings of the Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2009.

[3] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[4] Z. Cernekova, I. Pitas, and C. Nikou. Information theory-based shot cut/fade detection and video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(1):82–91, 2006.

[5] T. F. Chan and P. C. Hansen. Low-rank revealing QR factorizations. *Numerical Linear Algebra with Applications*, 1(1):33–44, 1994.

[6] Y. Cong, J. Yuan, and J. Luo. Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Transactions on Multimedia*, 14(1):66–75, 2012.

[7] C. Dang and H. Radha. RPCA-KFE: Key frame extraction for video using robust principal component analysis. *IEEE Transactions on Image Processing*, 24(11):3742–3753, 2015.

[8] S. E. F. De Avilla, A. P. B. Lopes, A. L. Jr. Luz, and A. A. Araujo. VSUMM:

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

[9] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

[10] L. Duan, T. Xi, S. Cui, H. Qi, and A. C. Bovik. A spatiotemporal weighted dissimilarity-based method for video saliency detection. *Signal Processing: Image Communicatoin*, 38(C):45–56, 2015.

[11] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[12] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] A. K Farahat, A. Ghodsi, and M. S. Kamel. Efficient greedy feature selection for unsupervised learning. *Knowledge and Information Systems*, 35(2):285–310, 2013.

[14] S. Gao, I.W. Tsang, and L.T. Chia. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(1):92–104, 2013.

[15] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas. The i3DPOST multi-view and 3D human action/interaction database. In *Proceedings of*

the IEEE Conference for Visual Media Production (CVMP), pages 159–168, 2009.

[16] G. Guan, Z. Wang, K. Yu, S. Mei, M. He, and D. Feng. Video summarization with global and local features. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2012.

[17] J. C. Harsanyi and C. I. Chang. Hyperspectral image classification and dimensionality reduction: an Orthogonal Subspace Projection approach. *IEEE Transactions on Geoscience and Remote Sensing*, 32(4):779–785, 1994.

[18] M. Hilferink. Fisher's Natural Breaks classification. http://wiki.objectvision.nl/index.php/Fisher2013.

[19] C. Hong, J. Yu, D. Tao, and M. Wang. Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval. *IEEE Transactions on Industrial Electronics (TIE)*, 62(6):3742–3751, 2015.

[20] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang. Multimodal deep autoencoder for human pose recovery. *IEEE Transactions on Image Processing (TIP)*, 24 (12):5659–5670, 2015.

[21] G. F. Jenks. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7(1):186–190, 1967.

[22] P. Kromer, J. Platos, and V. Snasel. Genetic algorithm for the column subset selection problem. In *Proceedings of the IEEE Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, 2014.

[23] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by Low-Rank Representation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[24] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1999.

[25] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, 2013.

[26] X. Ma, X. Xie, K.-M. Lam, J. Hu, and Y. Zhong. Saliency detection based on Singular Value Decomposition. *Journal of Visual Communication and Image Representation*, 32:95–106, 2015.

[27] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas. Compact video description and representation for automated summarization of human activities. In *Proceedings of the INNS Conference on Big Data*. Springer, 2016.

[28] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas. Movie shot selection preserving narrative properties. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2016.

[29] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas. Multimodal stereoscopic movie summarization conforming to narrative characteristics. *IEEE Transactions on Image Processing*, 25(12):5828–5840, 2016.

[30] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas. Summarization of human activity videos via low-rank approximation. In *Proceedings of the*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[31] I. Mademlis, A. Tefas, and I. Pitas. Summarization of human activity videos using a salient dictionary. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017.

[32] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng. Video summarization via minimum sparse reconstruction. *Pattern Recognition*, 48(2): 522–533, 2015.

[33] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya. Video summarization using deep semantic features. *arXiv preprint arXiv:1609.08758*, 2016.

[34] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher Kernel for large-scale image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 143–156. Springer, 2010.

[35] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[36] D. Sen and M. Kankanhalli. A bio-inspired center-surround model for salience computation in images. *Journal of Visual Communication and Image Representation*, 30(C):277–288, 2015.

[37] X. Song, L. Sun, J. Lei, D. Tao, G. Yuan, and M. Song. Event-based large scale surveillance video summarization. *Neurocomputing*, 187:66–74, 2016.

[38] M. Sturken and L. Cartwright. *Practices of looking: an introduction to visual culture*. Oxford University Press, 2nd edition, 2009.

[39] T. Theodoridis, A. Tefas, and I. Pitas. Multi-view semantic temporal video segmentation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2016.

[40] H. Wang and C. Schmid. Action recognition with Improved Trajectories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[41] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, 2006.

[42] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with Long Short-Term Memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

[44] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 1998.