This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

RPCA-KFE: Key Frame Extraction for Video using Robust Principal Component Analysis

Chinh Dang and Hayder Radha, *Fellow, IEEE*

***Abstract*—Key frame extraction algorithms consider the problem of selecting a subset of the most informative frames from a video to summarize its content. Several applications such as video summarization, search, indexing and prints from video can benefit from extracted key frames of the video under consideration. Most approaches in this class of algorithms work directly with the input video dataset, without considering the underlying low rank structure of the dataset. Other algorithms exploit the low rank component only; ignoring the other key information in the video. In this paper, a novel key frame extraction framework based on Robust Principal Component Analysis (RPCA) is proposed. Furthermore, we target the challenging application of extracting key frames from unstructured consumer videos. The proposed framework is motivated by the observation that RPCA decomposes an input data into (a) a low rank component that reveals the systematic information across the elements of the dataset and (b) a set of sparse components each of which containing distinct information about each element in the same dataset. The two information types are combined into a single ℓ_1-norm based non-convex optimization problem to extract the desired number of key frames. Moreover, we develop a novel iterative algorithm to solve this optimization problem. The proposed RPCA-based framework does not require shot(s) detection, segmentation, or semantic understanding of the underlying video. Finally, experiments are performed on a variety of consumer and other types of videos. A comparison of the results obtained by our method with the ground truth and with related state-of-the-art algorithms clearly illustrates the viability of the proposed RPCA-based framework.**

***Index Terms*—Video summarization, robust principal component analysis, consumer video.**

I. Introduction

GIVEN a huge dataset, the problem of finding a subset of important data points, also known as representatives or exemplars, which have the ability of describing the whole input dataset (at least to some extent) is emerging as a key approach for dealing with the massive growth of data, in general, and for video in particular.

Digital signals naturally belong to a high dimensional ambient space. Working directly in the high dimensional space generally involves much more complex algorithms. Meanwhile, signals are assumed to reside in an underlying low dimensional space, e.g. a set of high dimensional data points could be modeled by a union of multiple low dimensional subspaces. This modeling leads to the challenging problem of subspace clustering [1], which aims at clustering data points into multiple linear/affine subspaces. On a different low dimensional model, manifolds with a few degrees of freedom have been used successfully for the class of non-parametric signals, e.g. image of human faces and handwritten digits [2]. Numerous methods aiming at dimensionality reduction have been developed that could be classified into two main categories. On one hand, several well-known (linear and non-linear dimensionality reduction) methods, e.g. Principal Component Analysis (PCA), multidimensional scaling, isomap, multilayer auto-encoders [3], belong to the first category that mainly focus on preserving some particular desired properties. The algorithms in the other category aim at reconstructing the original dataset from the lower dimensional space measurement, e.g. compressive sensing and related random linear projections on a low dimensional manifold [4].

Beside the dimensionality reduction approach, finding a subset of important data points plays an important role in handling massive growth of digital data in many applications in machine learning, computer vision and clustering to name a few [5]. In video, the problem of finding a small subset of frames (called key frames or representative frames) is known as key frame extraction. A set of key frames in a video sequence plays an essential role in intelligent video management systems such as video retrieval and browsing, navigation, indexing, and prints from video. It helps to reduce computational complexity since the system could process a smaller set of representative frames instead of the whole video sequence. Key frames capture both the temporal and spatial information of the video sequence, and hence, they also enable the rapid viewing functionality [6].

In this paper, we adapt a dimensionality reduction technique for the problem of key frame extraction. The proposed approach has been originated from Robust Principal Component Analysis (RPCA) [17], which provides a stable tool for data analysis and dimensionality reduction. Under the RPCA framework, the input dataset is decomposed into a sum of low rank and sparse components. A majority of prior approaches works directly with the input high dimensional dataset, without considering the underlying low rank structure of input videos [6-7]. Other approaches focus on the low rank component only [8], ignoring the essential information from the other components. In this paper, we exploit both components into the problem of key frame extraction.

The main contributions of our work are:

1) A novel key frame extraction framework based on RPCA is proposed to automatically select a set of maximally

Manuscript received 2014. This work was supported in part by the National Science Foundation under Grants 1117709 and 1331852, and by an unrestricted gift from Google Inc., and by a Vietnam Education Foundation Fellowship.

Chinh Dang and Hayder Radha are with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, 48824-1226 USA; Corresponding e-mail: dangchin@msu.edu

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

informative frames from an input video. The framework is developed from a novel perspective of low rank and sparse components, in which the low rank component of a video frame reveals the relationship of that frame to the whole video sequence, referred to as *systematic* information, and the sparse component indicates the *distinct* information of particular frames.

2) A set of key frames are identified by solving an ℓ_1-norm based non-convex optimization problem where the solution minimizes the reconstruction error of the whole dataset for a given set of selected key frames and maximizes the sum of distinct information.
3) We propose a novel iterative algorithm to solve the aforementioned non-convex optimization problem. The algorithm allows to adapt to new observations and to update for the new set of key frames.

Our experiments employ two different datasets: consumer video, and traditional video datasets. It is important to note that consumer videos are different from professionally-produced videos in the following ways. First, consumer videos have rather diverse content with no predefined structure. Moreover, they may suffer from low quality due to poor lighting and camera shake. In addition, a majority of previous video summarization techniques are domain-dependent, in which they exploit specific properties of a video clip in a specific domain to generate a summary [9-10]. However, until now, there is little focus on solving challenges associated with consumer videos [11-13][33-34], not to mention that consumer videos have grown rapidly nowadays. Hence, this paper focuses on consumer videos, and the overall comparison of leading approaches with our RPCA-KFE method is presented to validate its effectiveness. On the other hand, we also show the experimental results on other types of videos taken from the Open Video Project [1], and validate the proposed method by comparing it with some recent leading approaches.

The outline of the rest of the paper is as follows. Section II briefly reviews some related works to our proposed approach. Section III formalizes the proposed RPCA-KFE method as well as the iterative algorithm to solve the optimization problem and how to deal with new observations. Experiments, results and comparison of the obtained results with other state-of-the-art methods are presented in Section IV. The last section outlines concluding remarks and discusses future works.

For easy reference, the following is a list of key symbols used in this paper: A capital notation will be used for a matrix:

$D = [d_1, d_2, \ldots, d_N] \in \mathbb{R}^{m \times N}$	Data points in matrix form
$L = [l_1, l_2, \ldots, l_N] \in \mathbb{R}^{m \times N}$	Low rank component of D
$S = [s_1, s_2, ..., s_N] \in \mathbb{R}^{m \times N}$	Sparse component of D
$D_r = [d_{t_1}, d_{t_2}, ..., d_{t_k}]$	The set of selected key frames
$L_r = [l_{t_1}, l_{t_2}, ..., l_{t_k}]$	Low rank component of D_r
$S_r = [s_{t_1}, \ldots, s_{t_k}]$	Sparse component of D_r
$C = [c_1, c_2, \ldots, c_N] \in \mathbb{R}^{k \times N}$	Coefficient matrix
$[C]_{ij}$	(i^{th} row, $j^{th} column$) element of C
$C_{i,:}$	i^{th} row of a matrix C
$C/C_{i,:}$	Matrix C without its i^{th} row
$C_{:,i}$	i^{th} column of a matrix C
$C/C_{:,i}$	Matrix C without its i^{th} column
$\|\|L\|\|_1 = \sum_{i=1}^{N} \|\|l_i\|\|_1 = \sum_{ij} \|L_{ij}\|_1$	The ℓ_1-norm of a matrix
$\|\|L\|\|_* := \sum_i \sigma_i(L)$	The nuclear norm of matrix L
$\#L_r$	Number of elements in the set L_r

II. Related Works

The general area of video summarization has been researched for years due to its important role in many video related applications. Comprehensive reviews of previous approaches could be found in [6][27-30]. Here, we briefly outline some related ideas and tools that have been exploited in our proposed approach.

A variety of pre-sampling techniques have been considered in prior works [25-26][32]. Such approach is acceptable for other types of videos due to the inherent redundancy in video. However, in the case of consumer video where the content tends to change abruptly and unpredictably, sampling at a predetermined rate [25] cannot guarantee the extraction of the best representative frames. Subsampling by selecting only I-frames [26] cannot ensure a viable set of representative frames either. This is due to the fact that, in general, no particular humanperception rules or video-summarization driven strategy are followed when coding video pictures as I-frame or B/P-frames. The video summarization based on compressed domain has a strong point of producing a video summary in a short time, which is potential for on-line applications [26]. However, creating a set of key frames while only a part of video available (for on-line application) cannot summarize the whole video content with a minimum number of key frames.

Zhang *et al.* [14] considered the problem of hybrid linear modeling (HLM), approximating a dataset with outliers by a mixture of d-dimensional linear subspaces. The paper concludes that replacing the ℓ_2-norm by the ℓ_1-norm improves significantly the robustness against outliers and noise. Yang *et al.* [15] considers the problem of sequential HLM that is of sequential recovery of multiple subspaces hidden in outliers. It leads to the problem of searching for the best ℓ_0 subspace (i.e. the subspace with largest number of data points) among multiple subspaces. G. Lerman and T. Zhang [16] studied the problem by minimizing the ℓ_p-averaged distance of data points from d-dimensional subspaces in high ambient dimensional space. The paper has an important conclusion that if $0 < p \leq 1$, then with overwhelming probability (i.e. the probability is at least $1 - u \times e^{-\frac{N}{u}}$, N is the size of dataset and u is a constant independent of N) the best ℓ_0 subspace can be recovered tractably. Even if some typical types of noise are added around the underlying subspaces, still the space can be recovered with overwhelming probability and the error will be proportional to the noise level. However, if $p > 1$, then the best ℓ_0 subspace cannot be recovered with overwhelming probability. The problem and results have been generalized into simultaneous recovery of multiple subspaces. In summary, the geometric properties of ℓ_p norm for $0 < p \leq 1$ lead

[1]The dataset could be download at http://www.npdi.dcc.ufmg.br/VSUMM/downloads/database.zip

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

to the ability of recovering the underlying subspaces with overwhelming probability, while the result is negative if $p > 1$.

RPCA [17] aims to recover a low rank matrix L_0 and its corresponding sparse component S_0 from the corrupted measurement D under a rather weak assumption, which can be solved by the following tractable convex optimization problem:

$$\begin{aligned} \{L_0, S_0\} &= \arg\min_{\{L,S\}} \|L\|_* + \lambda \|S\|_1 \\ s.t. D &= L + S \end{aligned} \tag{1}$$

The method has been exploited successfully in background and foreground separation [40], removing shadows from human face images, robust image alignment, robust image/video denoising [19][41], and image analysis [18]. As an application of background modeling in video sequences [17], the extracted sparse component from a frame implies additional information that does not belong to other video frames. This kind of information, although being sparse, becomes rather important under key frame extraction perspective, since we prefer a smallest number of key frames covering the maximum amount of information in the input video.

The key frame extraction problem has inherently a strong connection with clustering techniques, where a key frame can be considered as a medoid of each cluster [6][25]. k-means clustering is one of the most popular one, in which each data point will be assigned uniquely to one and only one of the clusters (hard assignment). The performance of clustering has been improved by adopting a probabilistic approach with soft assignment of each data point to these clusters. This naturally leads to linear combination of clustering centers (in our case they are key frames). However, there are still some challenges when applying linear-combination analysis directly on video due to camera motion, moving objects, etc. However, many prior works [12][20][21][25] successfully assumed that each data point can be expressed as a linear combination of a set of representative data points. In our proposed method, we formalize the optimization using a linear function for the low rank components only after performing RPCA. The question is why our approach is better? Recent works have confirmed that RPCA provides a robust estimation of the underlying subspace by decomposing observations into a low rank matrix and a sparse matrix [17][19]. In a recent application of RPCA into the problem of action recognition in video with moving camera [31], the basis of the low-rank matrix successfully captured both camera motion and body motion, and thus, illustrating that RPCA can be used as a sound framework for video analysis. We believe that an even better underlying framework, which we exploit in our paper, is that the low-rank component data point can be expressed as a linear combination of the representatives.

III. Key Frame Extraction based on Robust Principal Component Analysis

A. Problem Formalation

A given input video could be represented by a data matrix D, where each video frame is a column vector of that matrix in a high dimensional ambient space. Then, D is decomposed into a low rank component L and a sparse component S via a RPCA framework. Using the notations that we mentioned earlier in Section I, we have $D = L + S$ and $D_r = L_r + S_r$, where D_r is the data matrix of the selected key frames, L_r and S_r are the corresponding low rank and sparse components from D_r. Fig.1 shows an example of these two components for some videos.

Under the proposed RPCA-KFE framework, D_r will be analyzed jointly with systematic and distinct information corresponding to L_r and S_r, respectively. First, L_r will be evaluated quantitatively by considering accumulatively the reconstruction error of each data point $l_i \in L$, in a general form of $||l_i - f(L_r)||_q$, where $f(.)$ is chosen as a linear function in this work. This error indicates how well the set of key frames covers the content of data point l_i. Hence, the reconstruction error using L_r as a set of key frames to represent l_i will be computed by $||l_i - L_r c_i||_q$, in which

$$c_i = \arg\min_{c \in \mathbb{R}^{k \times 1}} ||l_i - L_r c||_q \tag{2}$$

where q is a constant (to be defined below). Then, the overall reconstruction error for a given set of key frames L_r becomes:

$$||L - L_r C||_q \triangleq \sum_{i=1}^{N} ||l_i - L_r c_i||_q \tag{3}$$

Second, the distinct information associated with each video frame, $s_i \in S_r$, can be by measured using its ℓ_1 norm:$\|s_i\|_1$. Hence, the total distinct information of the set of key frames is $\sum_{j=1}^{k} \|s_{t_j}\|_1$, which should be maximum for a good selection of key frames (with a fixed cardinality). Combining these two terms leads to an overall non-convex optimization problem:

$$\begin{aligned} &\{l_{t_1}, l_{t_2}, \ldots, l_{t_k}\} = \arg\min_{L_r} ||L - L_r C||_q - \gamma \sum_{j=1}^{k} \|s_{t_j}\|_1 \\ &s.t. \quad L_r \subseteq L \; and \; \#L_r = k \end{aligned} \tag{4}$$

Here, γ>0 is a regularization constant parameter that indicates the relative importance of two components. In the experiment, these two components are considered equally important, so we select γ=1.

As we mentioned earlier in Section II about the problem of HLM [14-16], we expect to search for a subspace that contains the largest number of data points. Hence, $0 < q \leq 1$ leads to the ability of recovering the underlying subspaces with an overwhelming probability. Therefore, in our work, we select $q = 1$, and the problem (4) can be considered as a specific case of recovering the best ℓ_0 subspace with an additional condition that the subspace must be spanned by elements from the input dataset (key frames).

$$\begin{aligned} &\{l_{t_1}, l_{t_2}, \ldots, l_{t_k}\} = \arg\min_{L_r} ||L - L_r C||_1 - \gamma \sum_{j=1}^{k} \|s_{t_j}\|_1 \\ &s.t. \quad L_r \subseteq L \; and \; \#L_r = k \end{aligned} \tag{5}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

a) "*SoloSurfer*" video b) "*SkylinefromOverlook*" video

Fig. 1: Examples of low rank and sparse components from several frames extracted from two video clips.

This selection distinguishes our ℓ_1-norm based optimization from other ℓ_2-norm based optimization methods in image collection/video summarization [21-22]. More interestingly, our result is also consistent with other results from the compressive sensing theory area [23]. In particular, let us denote $X \triangleq \left[\|l_1 - L_r C_{1,:}\|_1, \ldots, \|l_N - L_r C_{N,:}\|_1\right]^T \in \mathbb{R}^{N\times 1}$. Then $\|X\|_1 = ||L - L_r C||_1$. In this case, X is a vector of distances from a data point to the linear subspace spanned by the selected key frames L_r. Since the ℓ_1 norm-based minimization problem tends to encourage solutions to be sparse [23], the linear space spanned by L_r contains the maximum number of elements from the input dataset (or the best ℓ_0 subspace). Despite the merits of using ℓ_1-norm, the solution obtained from ℓ_1 norm based problem might not be unique. However, under this circumstance, the additional constraint of maximizing the total distinct information leads to the unique solution for (5). In addition, we take advantages of using ℓ_2 norm by considering the least square solution as an initial solution in an iterative process when solving (5). The detailed algorithm and corresponding solution is presented in the next subsection.

B. Overall solution

The optimization problem (5) has a form that is close to dictionary learning for sparse representation [24]. However, there are some key differences between these two problems. Dictionary learning aims at finding good bases/frames for a given set of input data for sparse representation (minimizing ℓ_0 norm of coefficients). Hence, the number of learned elements in that basis is huge. Moreover, these learned bases may not contain exact elements in the dataset but sparse combination of atoms in the basis/frame, so they cannot be used as representatives of input dataset. As a result, most existing algorithms in dictionary learning and sparse coding cannot be directly applied into our optimization problem. In this paper, we propose a novel iterative algorithm to solve the problem (5) with some distinguished properties. Conventional iterative algorithms update all elements simultaneously at each step that leads to some main drawbacks of slow convergence and difficulty of solving sub-optimization problem inside a single step. We propose an algorithm that divides each main update step into smaller sub-steps, so that elements will be updated sequentially in a single sub-step. In addition, the updated formula guarantees to decrease the objective function in (5) after a single step.

Recall that the objective function is to find a set of indices $\{t_1, t_2, \ldots, t_k\}$, for a given number of k, that minimize the objective function:

$$||L - L_r C||_1 - \gamma \sum_{j=1}^{k} \left\| s_{t_j} \right\|_1 \tag{6}$$

Here, $L_r = [l_{t_1}, \ldots, l_{t_k}]$ is the corresponding low rank data matrix for the set of indices. Define ${L_r}^{i,\xi}$ as a matrix for the current set of key frames of i^{th} sub-step in the ξ^{th} main step: ${L_r}^{i,\xi} = \left[l_{t_1{}^{(\xi)}}, l_{t_2{}^{(\xi)}}, \ldots, l_{t_i{}^{(\xi)}}, l_{t_{i+1}{}^{(\xi-1)}}, \ldots, l_{t_k{}^{(\xi-1)}}\right]$ where the algorithm already update i elements $\left\{l_{t_1{}^{(\xi)}}, l_{t_2{}^{(\xi)}}, \ldots, l_{t_i{}^{(\xi)}}\right\}$. In the initial set of indices $\left\{{t_1}^{(0)}, {t_2}^{(0)}, \ldots, {t_k}^{(0)}\right\}$, the algorithm fixes ${L_r}^{0,1} = \left[l_{t_1{}^{(0)}}, l_{t_2{}^{(0)}}, \ldots, l_{t_k{}^{(0)}}\right]$ and computes the coefficient matrix:

$$C^{0,1} = \arg \min_{C\in\mathbb{R}^{k\times N}} ||L - {L_r}^{0,1} C||_1 \tag{7}$$

Since the solution of (7) becomes the input of the iterative process in the RPCA-KFE algorithm, the exact solution is not strictly demanded. Therefore, we convert the problem into the least square problem for fast and easy computation of the unique solution:

$$C^{0,1} = \arg \min_{C\in\mathbb{R}^{k\times N}} ||L - {L_r}^{0,1} C||_2 \tag{8}$$

Therefore,

$$C^{0,1} = \left(({L_r}^{0;1})^T {L_r}^{0;1}\right)^{-1} ({L_r}^{0;1})^T L \tag{9}$$

Let us consider the low rank component matrix of the current set of key frames ${L_r}^{i,\xi}$ and the corresponding coefficient matrix $C^{i,\xi} = \left[C_1^{(\xi)}, C_2^{(\xi)}, \ldots, C_i^{(\xi)}, C_{i+1}^{(\xi-1)}, \ldots, C_k^{(\xi-1)}\right]^T$ at the i^{th} sub-step of the ξ^{th} main step of the algorithm. In this sub-step, to update $l_{t_{i+1}{}^{(\xi-1)}}$ into $l_{t_{i+1}{}^{(\xi)}}$, the RPCA-KFE algorithm assumes that ${L_r}^{i,\xi} / \left\{l_{t_{i+1}{}^{(\xi-1)}}\right\}$ and $C^{i,\xi} / \left\{C_{i+1}^{(\xi-1)}\right\}^T$ are constants, and then the optimization problem focuses only on $l_{t_{i+1}{}^{(\xi-1)}}$ and its corresponding coefficient row.

Using the property of decomposition of a matrix product as a sum of rank one matrices, ${L_r}^{i,\xi} C^{i,\xi}$ will be decomposed into the sum of two matrices:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

Robust Principal Component Analysis based Key Frame Extraction (RPCA-KFE) Algorithm

Task: Finding the set of key frames to represent the video data samples $D = \{d_i\}_{i=1}^{N}$ by solving:

$$\min_{\substack{L_r \subseteq L \\ \#L_r = k}} ||L - L_r C||_1 - \gamma \sum_{j=1}^{k} \left\| s_{t_j} \right\|_1$$

Given the number of desired selected elements k and the constant γ.

1. % Initialization:

1) Find the low rank and sparse component L, S from input data D by using Robust PCA.

2) Initialize: the set of frame indices $\{{t_1}^{(0)}, {t_2}^{(0)}, \dots, {t_k}^{(0)}\}$, and set the current loop index $\xi = 0$ and ξ_{max}; find the initial coefficient matrix C by solving (using (9)): $C = \arg\min_C ||L - {L_r}^{0;\xi} C||_1$ where ${L_r}^{0;1} = [l_{t_1{}^{(0)}}, l_{t_2{}^{(0)}}, \dots, l_{t_k{}^{(0)}}]$

2. % Repeat*(until $\xi = \xi_{max}$)*

1) For each element $i = 1,2, \dots, k$, update the i^{th} element $l_{t_i{}^{(\xi)}}$ into $l_{t_i{}^{(\xi+1)}}$ by the following steps:

- Compute the constant component: $L^{i;\xi} = L - {L_r}^{i,\xi} / \left\{ l_{t_{i+1}{}^{(\xi-1)}} \right\} C^{i,\xi} / \left\{ C_{i+1}^{(\xi-1)} \right\}^t$
- Solve the optimization problem: $\left\{ l_{t_i{}^{(\xi)}}, C_{i+1}^{(\xi)} \right\} = \arg\min_{\substack{l_i \in L/{L_r}^{i,\xi} \\ c_i \in \mathbb{R}^{1\times N}}} \left\| L^{i;\xi} - l_i c_i \right\|_1 - \gamma \|s_i\|_1$
- Update: $l_{t_i{}^{(\xi)}} \to l_{t_i{}^{(\xi+1)}}$ and $C = C / C_{i,:} \cup C_{i+1}^{(\xi)}$

2) Set $\xi = \xi + 1$

$$\begin{aligned} {L_r}^{i,\xi} C^{i,\xi} = {L_r}^{i,\xi} / \left\{ l_{t_{i+1}{}^{(\xi-1)}} \right\} C^{i,\xi} / \left\{ C_{i+1}^{(\xi-1)} \right\}^T \\ + l_{t_{i+1}{}^{(\xi-1)}} \left\{ C_{i+1}^{(\xi-1)} \right\}^T \end{aligned} \tag{10}$$

Denote $L^{i;\xi} = L - {L_r}^{i,\xi} / \left\{ l_{t_{i+1}{}^{(\xi-1)}} \right\} C^{i,\xi} / \left\{ C_{i+1}^{(\xi-1)} \right\}^T$, then the sub-step optimization has the following form:

$$\begin{aligned} \left\{ l_{t_i{}^{(\xi)}}, C_{i+1}^{(\xi)} \right\} = \arg \min_{\{l_i, c_i\}} \left\| L^{i;\xi} - l_i c_i \right\|_1 - \gamma \|s_i\|_1 \\ s.t. \;\; l_i \in L/{L_r}^{\mathrm{i},\xi};\; c_i \in \mathbb{R}^{1\times N} \end{aligned} \tag{11}$$

Here, s_i is the sparse component that corresponds to the low rank component $l_i \in L/{L_r}^{i,\xi}$. The optimization problem (11) can be solved by scanning all possible value of $l_i \in L/{L_r}^{i,\xi}$, and for a fixed value of l_i, the coefficient vector $c_i \in \mathbb{R}^{1\times N}$ of the problem could be computed based on the following results:

Lemma 1. *Given two positive vectors $\boldsymbol{u} = [u_i]_{m\times 1}$ and $\boldsymbol{v} = [v_i]_{m\times 1}$, $(\boldsymbol{u}, \boldsymbol{v} \in (\mathbb{R}^+)^m)$ then a scalar parameter of the solution for $\min_{\alpha\in\mathbb{R}} | \, |\boldsymbol{u} - \alpha \boldsymbol{v}| \, |_1$ belongs to a particular set:*

$$\alpha_0 = \arg\min_{\alpha\in\mathbb{R}} | \, |\boldsymbol{u} - \alpha \boldsymbol{v}| \, |_1 \in \left\{ \frac{u_i}{v_i} | 1 \le i \le m \right\} \tag{12}$$

This lemma allows seeking an optimal value for each single element in the coefficient vector $c_i \in \mathbb{R}^{1\times N}$ which belongs to that particular set. To avoid considering a single element in all m possible values of the set $\left\{ \frac{u_i}{v_i} | 1 \le i \le m \right\}$, the following simple result helps to determine the exact solution:

Lemma 2. *Without loss of generality, assuming that the sequence $\left\{ \frac{u_i}{v_i} | 1 \le i \le m \right\}$ is a non-decreasing sequence. Then, the unique solution for (12) is in the form of $\frac{u_{t_0}}{v_{t_0}}$ where:*

$$t_0 = \min_{1\le t\le m} t \; s.t. \; \sum_{i=1}^{t} v_i \ge \sum_{i=t+1}^{m} v_i \tag{13}$$

The detail proof for Lemma 1 and 2 are given in the APPENDIX. Lemma 2 helps to determine the exact solution without scanning all m possible solutions. In the experiment, m is the dimension of high dimensional data points that is the number of image pixels in a visual dataset. Therefore, not scanning all m possible solutions significantly improves the speed of convergence of the algorithm.

C. RPCA-KFE with New Observations

In this section, we show how the proposed RPCA-KFE algorithm could be adapted to deal with new observations. Matrices $D^{(0)}, L^{(0)}, S^{(0)}$ are respectively the current set of data points, low rank, and sparse components as before. $D_r^{(0)} = L_r^{(0)} + S_r^{(0)}$ is the set of selected key frames for the current dataset. Let us use $D^{(0)}$ to denote the set of new observations and $D_r^{(1)} = L_r^{(1)} + S_r^{(1)}$ where $L^{(1)}$ and $S^{(1)}$ for the low rank and sparse components, founding using Robust PCA. The overall problem (6) could be rewritten as:

$$\begin{aligned} &\arg\min_{L_r} \left\| \left[L^{(0)} L^{(1)} \right] - L_r \left[C^{(0)} C^{(1)} \right] \right\|_1 - \gamma \sum_{j=1}^{k} \left\| s_{t_j} \right\|_1 \\ &s.t. \quad L_r \subseteq \left[L^{(0)} L^{(1)} \right] \; and \; \#L_r = k \end{aligned} \tag{14}$$

Here, $\left\{ s_{t_j} \right\}_{j=1}^{k}$ are sparse components corresponding to L_r. Instead of starting solve the problem from the beginning as in Section III.B, the algorithm will be adapted as follows. Since $L_r^{(0)}$ is the set of selected key frames for $L^{(0)}$ (the low

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

TABLE 1. VIDEO CLIP DESCRIPTION USED FOR EVALUATION [11]

Video Name	*# KF*	*# Frames*	*Camera Motion*	*Persp. Changes*	*Bright. Changes*
HappyDog	4	376	Yes	Yes	Yes
MuseumExhibit	4	250	Yes	No	No
SoloSurfer	6	618	Yes	Yes	Yes
SkylinefromOverlook	6	559	Yes	Yes	Yes
FireworkAndBoat	4	656	Yes	No	No
BusTour	5	541	Yes	Yes	Yes
LiquidChocolate	6	397	Yes	Yes	yes

rank component), it becomes the initial set of key frames. Hence, the initial coefficient matrix for the new dataset will be computed by:

$$C^{(1)} = \left(L_r^{(0)^T} L_r^{(0)}\right)^{-1} L_r^{(0)^T} L^{(1)} \tag{15}$$

In the iterative process, the search space for each element is restricted among elements from the new observations $L^{(1)}$ only, not the whole dataset $L = \left[L^{(0)} L^{(1)}\right]$. In particular, the algorithm considers the cost of changing from a current key frame in $L_r^{(0)}$ into a new frame in $L^{(1)}$. The new frame will be selected as a key frame if it leads to a smaller cost than the current one. In particular, we consider the algorithm at the i^{th} sub-step of the ξ^{th} main step, similar to the previous section. To update the current key frame $l_{t_{i+1}(\xi-1)}$ into $l_{t_{i+1}(\xi)}$, the adapted RPCA-KFE algorithm consider $l_{t_{i+1}(\xi)} \in L^{(1)}$ only.

IV. EXPERIMENTAL RESULTS

Our experiments are performed onto two different types of videos: (1) a consumer video dataset; and (2) videos from the Open Video Project.

A. *Consumer Video Dataset*

While most prior efforts were applied to structured videos and used certain publically available datasets, here, we worked on a dataset of consumer videos. In particular, our simulations were run on the Kodak Home Video Database [11]. These clips were captured using KodakEasyShare C360 and V550 zoom digital cameras, with a VGA resolution (frame size of [640x480]). We showed seven clips for evaluation and comparison. The detailed description of these clips is provided in Table I. They vary in duration from 250 frames to 656 frames, approximately 485 frames per clip on average. The average number of key frames is five per clip, and it depends on the number of key frames in the ground truth (explained below). The proposed algorithm does not perform any pre-sampling as in previous approaches, such as at a predetermined rate [24] or by selecting only I-frames [25]. Therefore, it is rather straightforward to extend the proposed algorithm for longer video clips in conjunction with simple sub-sampling (e.g. 15 minutes if a pre-sampling rate at one frame/sec is employed). We would mention that point in our next experiment for longer video dataset. We will revisit this point when showing the results for more traditional longer videos.

1. Parameter Selection:

For a given input video, each frame was first converted into YCbCr format, and down-sampled into a resolution of 80×60. The algorithm works with the luminance channel only. A frame of size 80×60 is converted to a column vector of dimension 4800×1. The input video becomes a dataset of high (normally full) rank matrix, dimension of [4800, number of frames]. Robust PCA method has been exploited to decompose the input data matrix into the low rank and sparse components. We use the augmented Lagrange multiplier method for this kind of decomposition because of its high accuracy in a small number of iterations. Some other parameters for this decomposition include: the maximum number of iterations is set to 100, and the tolerance of stopping criterion equals to $1e-5$, and the constant parameter balancing two components is $\lambda_0 = 1/\sqrt{\max(4800, \text{numberframes})}$ as suggested by Candes *et al.* [17]. Algorithm 1 has been performed for the two obtained components. In the experiment, the initial set of key frames is sampled uniformly from the video sequence. The parameter γ is selected as a rule of thumb, γ=1. That means we consider these two types of information (distinct and systematic) to be equally important. We test the obtained result with some different values of maximum iteration (stopping rule), $\xi_{\max}$, and see that the algorithm converges quickly to the stable results in many cases. There is only two videos (SoloSurfer and SkylinefromOverlook), where the obtained set of selected key frames in second iteration ($\xi_{\max}$=3) is slightly different from the set of selected key frames from the first iteration ($\xi_{\max}$=2). Therefore, in our experiments, we select the maximum number of iterations $\xi_{\max}$=2 to minimize the computation burden. This implies that the algorithm requires only one iteration with k sub-steps to stop.

2. Baseline Algorithms

Sparse modeling finding representative (SMFR) [22] method was developed based on an assumption that each data point can be expressed as a linear combination of the representatives. The self-expressiveness property has been studied along with additional constraints for the representative selection problem. (The software implementation of this method is available online).

The color histogram based method of UCF [27] exploits the color histogram intersection similarity measure to extract key frames. The algorithm depends on how to select the first key frame (normally the first frame of video is selected as key frame), and then next frames are selected based on how they differ from the current selected key frames.

Motion based key frame extraction (MKFE) [11] method was developed based on analysis of spatio-temporal changes over time for capturing semantically meaningful information from the scene and camera motions. In particular, a video clip is segmented into homegenerous parts based on camera motion types, e.g. pan, zoom, pause, and steady. A set of candidate key frames has been extracted based on the motion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

type corresponding to each segment. Finally, the set of key frames is extracted from the initial candidate key frame set based on confidence measure of each frame.

Sparse representation based method (SR) [12] is a typical sparse representation based clustering method in which each video frame is projected onto a low dimensional random feature space. A symmetric matrix is created from the projected dataset based on sparse representation, and then normalized clustering algorithm is applied for clustering the dataset. Middle frame in term of temporal order from each cluster will be selected as a key frame.

Bi-layer group sparsity (BGS) [13] approach segments each input video frame into visually homogenous patches, and proposed the bi-layer group sparsity approach by combining the traditional group sparse Lasso and Moreau-Yosida regulation. The spatial correlation is considered among frames for the first layer and the temporal correlation among frames is then considered for the final frame quality score.

The two first methods (SMFR and UCF) are designed for a general video dataset, while the last three methods (MKFE, SR, and BGS) particularly target consumer videos. Other methods (e.g. [10], [28], [29]) were devoted for news and sport videos, which require shot detection and segmentation. In particular, TRECVID [28] extracts one key frame per shot, which is not suitable for the consumer video dataset, since each video clip in our experiment can be considered as one shot in the TRECVID video [11]. Here, shots are defined as the longest continuous sequence that originates from a single camera take [35].

Fig. 2: *HappyDog* video. The visual comparison includes different methods: a) SR [12], b) BGS [13], c) MKFE [11], d) our proposed RPCA-KFE, and e) the ground truth. Solid red border implies good match: 1 points, and dashed red border implies fair match:0.5 point.

3. Evaluation and the ground truth

Types of Evaluation: In general, methods for video summarization evaluation can be classified into three different groups: (i) result description, (ii) objective metrics, and (iii) subjective metrics or user-based studies. Some authors may prefer combining two or more of these methods to provide additional information within the summary results.

In the proposed RPCA-KFE framework, we exploit the result description (visual comparison) and subjective metric (quantitative comparison) approaches. In particular, our results are compared with the ground truth agreed by multiple human judges. The goals of creating the ground truth are to: (1) create a reference database of video clips, particularly for consumer video space; (2) identify a foundation by which automated algorithms can be used for comparison; (3) uncover the criteria used by human evaluation so they may influence algorithm design [11]. To establish the ground truth, three human judges were asked to independently browse the video clips and provide the key frames. Photographers who actually captured the videos were not selected as the judges. The key frames estimated by the three judges were reviewed in a ground session with a fourth judge (arbitrator) to derive final key frames for each of the video clips. Furthermore, the judges also need to keep the purpose of the frame selection task as a summarization of input video when making their decision [11]. The number of key frames was determined by the human judges based on the representatives and quality of the corresponding video clips.

Quantitative Comparison: In order to quantitatively evaluate the performance of an automated algorithm in selecting key frames relative to the key frames in the ground truth, we examine both image content and time differences as have been done in prior efforts [11-13]. In particular, if the selected key frame by an automated algorithm has (a) similar content and (b) is within one second (30 frames) of the corresponding key frame in the ground truth, then the algorithm receives one full point. Otherwise, if the predicted key frame is only similar to the frame in the ground truth, but the time difference is larger than the threshold (30 frames), then the algorithm gets 0.5 point. In the latter case, if the selected key frame does not have similar content to the frame in the ground truth, then the algorithm receives no points. This happens in some videos, where a selected key frame is close to a key frame in the ground truth, however the content is not similar.

We compare the proposed RPCA-KFE algorithm with five other key frame extraction methods. The results are summarized in Table 2. The score here could be understood as the number of good key frames selected by each method. The difference between the number of key frames in the ground truth and the obtained score could be considered as the missing frames. Since in all of the algorithms being compared, the number of desired key frames selected by the automatic algorithms are set to equal to the number of frames from the ground truth, the two factors of precision and recall (and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

TABLE 2. SUMMARY OF EXPERIMENTAL RESULTS UNDER KEY FRAME EXTRACTION FOR SEVEN VIDEO CLIPS

Video Name	SMFR [22]	UCF [27]	SR [12]	BGS [13]	MKFE [11]	RPCA-KFE	#Key Frame
HappyDog	1	2	2	3	3	**3.5**	4
MuseumExhibit	3	2	3	3	3	**3.5**	4
SoloSurfer	3.5	2	4	5.5	4.5	4	6
SkylinefromOverlook	4	4	3.5	4	3	**5**	6
FireworkAndBoat	1	0	0	1	3	1	4
BusTour	1	3	3	1	2	**3**	5
LiquidChocolate	3	3	3.5	5	4	4	6
Summary	16.5 47.1%	16 45.7%	19 54.2%	22.5 64.3%	22.5 64.3%	**24** 68.6%	35

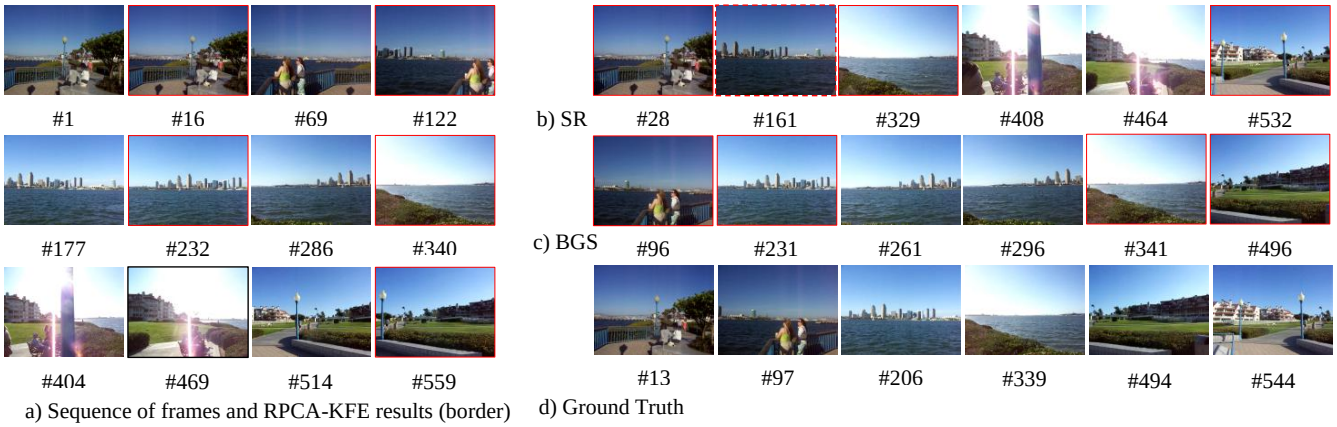


Fig. 3: *SkylinefromOverlook* video. a) Set of frames (we show 12 frames including our selected results to capture the video content). Solid red border implies good match: 1 point, dashed red border implies fair match:0.5 point). The comparison is shown in the right figure, includes the SR(1st row) [12], the BGS (2nd row)[13], and the ground truth (the last row).

F measure [32]) are not used in this work (since in this case precision = recall). The other evaluation method, called accuracy rate and error rate [25], are not used as well, since when the number of desired key frame is set to equal the number of frames from the ground truth, sum of accuracy rate and error rate is exactly one, and the accuracy rate is equivalent to the given score by our method. From Table 2, the proposed approach achieves the best results among the evaluated methods. More importantly, the RPCA-KFE algorithm does not require shot detection, segmentation, or semantic understanding.

Visual Comparison: Fig.2 shows the results of and *Happy-Dog* video. The video includes four key frames in the ground truth (row e) that focus on capturing different positions of the dog. This video includes different challenging visual effects, such as camera motion, pan, zom, and moving objects. Our RPCA-KFE method obtains the best result (quantitatively 3.5 points) in comparison with other methods.

Fig.3 shows the result of *SkylinefromOverlook* video. The video contains six key frames in the ground truth (the last row on the right figure), which was captured outdoors with a significant amount of change in perspective and brightness. In this video, the SR-based method [12] obtains 3.5 points. There are three frames (#28, 329, and 532) that get full one points due to the similarity of content as well as the within the threshold time difference. The second frame (#161) gets 0.5 points since it has similar content to the key frame #206; however, the time difference is beyond the threshold. The BGS [13] method performs slightly better with full 4 points for this video. However, there are two redundant frames of similar content. Our proposed RPCA-KFE method extracts successfully five key frames in this video, missing only the last key frame from the ground truth. As before, the ground truth is shown in the last row for comparison.

Computational Complexity: since the source codes of the other methods being compared here are not available, and the time required for producing a set of key frames or a video skimming excerpt depends on a particular hardware, it is almost impossible to produce a fair comparison in term of complexity among these methods. In this paper, we evaluate the average processing time per frame, as appeared in [26] to evaluate the complexity. According to those experiments, our HIP-based technique takes 1.469 second on average to process a single frame, including 0.233 second per frame for the RPCA decomposition of the input signal into low rank and sparse components, and then solving the optimization problem (on average 1.236 second per frame). This particular number depends on the computational power of the underlying

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing

hardware. In our work, we used an Intel Core E7500 2.93Ghz platform. The average processing time per frame could be reduced by a factor that depends on a pre-sampling rate (if used), and the image size relative to the size of 80x60, which we used in our experiments. For example, using a pre-sampling rate of 1frame/sec, the average time per a single frame could be reduced to 0.0612 sec/frame.

B. *The Open Video Database*

In this part, we focus on evaluating our algorithm for different types of videos. We select 50 videos taken from the Open Video Project. All of those videos are in the MPEG-1 format (30 frames per second, and frame size of [352x240], and length from 1 to 4 min), which are distributed among several genres (documentary, educational, ephemeral, lecture, and historical).

Pre-processing: For many traditional types of videos, scenes normally change slowly; therefore, it allows for pre-sampling the input video without significantly impacting the quality of the summary. In our work, the sampling rate is fixed to one frame per 25 frames of the input video. Each frame was converted into YCbCr format, and down-sampled into a resolution of [88x60] (to be consistent with the aspect ratio of the input frame resolution of [352x240]). The proposed RPCA-KFE algorithm is applied to the luminance channel only.

Initial number of key frames: After the pre-processing stage, we select uniformly one frame per 200 frames for the initial set of key frames. The number of key frames from the initial set is slightly higher than the average number of key frames selected subjectively from consumer video since scenes in consumer videos tend to change more abruptly than other types of videos. Since the initial number of key frames is proportional to the input video length, the obtained set of key frames may contain redundant frames. Hence, we add one final step to discard redundant frames simply using color histogram correlation. Under the final step, if two frames that have color histogram correlation greater than or equal a predetermined threshold (we use threshold equals 0.9), then only one frame, the one with higher color histogram variance, would be selected as a key frame; the other one would be discarded. We note that this final step is not employed in the consumer video database case. In that case, the input number of key frames for the RPCA-KFE algorithm is the same as the number of key frames from the ground truth.

The ground truth and evaluation metrics: Our results are compared with the ground truth agreed by multiple human judges. Different from the consumer video case, here we have five different ground truths selected by a number of users. In our work, we exploit a prior evaluation method, called Comparison of User Summaries (CUS) [25], which compares each user summary directly with the automatic summaries, thus keeping the original opinion of every user. We denote n_{US}, n_{AS} as the number of key frames from one ground truth and from an automatic summary. Comparing the set of selected key frames to the ground truth, we obtain n_{mAS} key frames that represent a good match with the ground-truth key frames; and hence $n_{\bar{m}AS}$ key frames do not represent a good match ($n_{mAS} + n_{\bar{m}AS} = n_{AS}$). The quality of an automatic video summary algorithm is then evaluated based on two metrics:

- The accuracy rate $\mathrm{CUS}_A = \frac{n_{mAS}}{n_{US}}$
- The error rate $\mathrm{CUS}_E = \frac{n_{\bar{m}AS}}{n_{US}}$

Here, we note that the accuracy rate is smaller than or equal to one, but the error rate could be greater than one.

TABLE III. COMPARISON OF USER SUMMARIES AND NUMBER OF SELECTED FRAMES

	KFE – DT	*STIMO*	*RPCA – KFE*
# selected KF	311	496	383
Avg # KF	434	434	434
CUS_A	0.48	**0.66**	**0.64**
CUS_E	**0.32**	0.62	**0.30**

TABLE IV. DIFFERENCE BETWEEN OUR RPCA-BASED TECHNIQUES AND OTHER TYPICAL METHODS AT A CONFIDENCE OF 95%

Measure	RPCA – DT		RPCA-STIMO	
	min.	max.	min.	max.
CUS_A	0.157	0.15	−0.0204	−0.0189
CUS_E	−0.0147	−0.0084	−0.2989	−0.3412

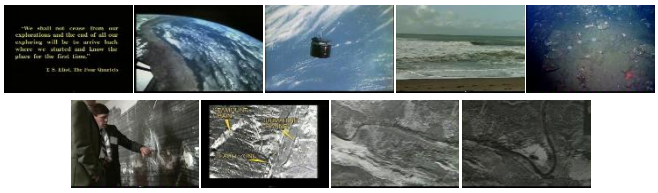
Quantitative Comparison: For the consumer video dataset, we examine both image content and time differences as have been done in prior efforts [11-13][34] to quantitatively evaluate the performance of an automated algorithm in selecting key frames relative to the key frames in the ground truth. Here, the tested videos from the Open Video Project are longer, and scenes tend to change slower. We do not select a predetermined threshold to give score for a selected key frame; only evaluate a key frame based on similar image content to be such that it is consistent with a human observer.

Compared methods: We compare the proposed RPCA-KFE algorithm with two different automatic video summarization methods including Key Frame Extraction based on Delaunay Triangulation (KFE-DT) [39], and the STIll and Moving storyboard (STIMO) algorithm [38]. The overall evaluation scores and comparison of our proposed RPCA-KFE algorithm with the other approaches are summarized in Table III.

Statistical Test: We employed the technique that has been used in [26] to verify the statistical significance of our method in comparison with the other methods being evaluated. Table IV shows this comparison using confidence interval of 95%. In the table, if the confidence interval includes zero, the difference is not significant at that confidence level; otherwise, the sign of the difference indicates which alternative is better [26]. The min. (max.) values in the table indicate the difference between the minimum (maximum) values between two compared methods. The statistical analysis shows that the proposed RPCA-KFE algorithm leads to a high accuracy rate (better than KFE-DT, and almost equal to STIMO) while maintaining a very small error rate.

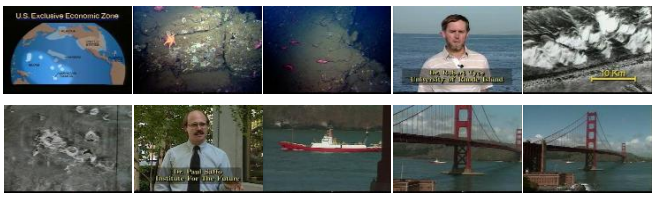
We can make some important conclusions based on the results from Table III and Table IV. First, the number of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2015.2445572, IEEE Transactions on Image Processing



(a) RPCA-KFE (video 46)

(b) User summary #3, $CUS_A = 8/9$, $CUS_E = 1/9$



(a) RPCA-KFE (video 50)

(b) User summary #3, $CUS_A = 1$, $CUS_E = 1/9$

Fig. 4: RPCA-KFE summary for two videos and user summaries

selected key frames in our RPCA-KFE method (383 frames) is the closest number to the average number of key frames from the set of ground truths. It is not as small as in KFE-DT method (311 frames) or as large as in STIMO method (496 frames). Second, despite the fact that the number of selected key frames under RPCA-KFE is larger than the one in KFE-DT, the average error rate ($\mathrm{CUS_E}$) is still lower (0.32 and 0.30) while the accuracy rate ($\mathrm{CUS_A}$) of the proposed RPCA-KFE method is much higher than the KFE-DT algorithm. On the other hand, the proposed RPCA-KFE method leads to approximately equal accuracy rate in comparison with the STIMO approach using a smaller number of selected key frames (as a result, much smaller in terms of the error rate $\mathrm{CUS_E}$). Fig. 4 shows some experimental results and ground truths for two sample videos.

V. Conclusion

We proposed an automatic algorithm to extract a set of key frames from a video using RPCA. Our work was based on the assumption that the low rank component contains systematic information along with the distinct information from the sparse component. We formulate the problem of key frame extraction from a video as an optimization problem and analyzed the advantages of using ℓ_1 norm based optimization. A greedy algorithm has been proposed to solve the non-convex optimization problem. Experiments were performed on a consumer video dataset and videos from the Open Video Project, and the obtained results were compared with state-of-the-art methods to validate the effectiveness of the proposed algorithm.

Appendix

Proof of Lemma 1 and 2.

Denote $f(\alpha) = ||u - \alpha \times v||_1 = \sum_{i=1}^{m} |u_i - \alpha \times v_i|$. Without loss of generality, we assume that:

$$\frac{u_0}{v_0} = -\infty < \frac{u_0}{v_0} \leq \frac{u_1}{v_1} \leq \ldots \leq \frac{u_m}{v_m} < \frac{u_{m+1}}{v_{m+1}} = +\infty$$

Then, denote $S_t = \left(\frac{u_{t-1}}{v_{t-1}}, \frac{u_t}{v_t}\right]$ for $1 \leq t \leq m+1$, we have the following properties:

$$\begin{cases} S_i \cap S_j = \emptyset (\forall 1 \leq i \neq j \leq m+1) \\ R = \bigcup_{t=1}^{m+1} S_t \end{cases}$$

Assuming that $\alpha \in S_t$, then

$$\begin{aligned} f(\alpha) &= \sum_{i=1}^{t-1} |u_i - \alpha \times v_i| + \sum_{i=t}^{m} |u_i - \alpha \times v_i| \\ &= \alpha \times \left(\sum_{i=1}^{t-1} v_i - \sum_{i=t}^{m} v_i\right) + \left(\sum_{i=t}^{m} u_i - \sum_{i=1}^{t-1} u_i\right) \end{aligned}$$

Take the derivative of $f(\alpha)$ with $\alpha \in \left(\frac{u_{t-1}}{v_{t-1}}, \frac{u_t}{v_t}\right)$, we obtain the following result:

$$\begin{cases} f(x) \leq f(y) \, \forall \frac{u_t}{v_t} \geq x \geq y > \frac{u_{t-1}}{v_{t-1}} \text{ if } \left(\sum_{i=1}^{t-1} v_i - \sum_{i=t}^{m} v_i\right) \leq 0 \\ f(x) \geq f(y) \, \forall \frac{u_t}{v_t} \geq x \geq y > \frac{u_{t-1}}{v_{t-1}} \text{ if } \left(\sum_{i=1}^{t-1} v_i - \sum_{i=t}^{m} v_i\right) > 0 \end{cases}$$

Denote $t_0 = \min_{1 \leq t \leq m} t$ s.t. $\sum_{i=1}^{t} v_i \geq \sum_{i=t+1}^{m} v_i$. Since $f(\alpha)$ is a continuous fucntion of α, the property holds for $\mathbb{R}$. In particular, we have:

$$\begin{cases} f(x) \leq f(y) \ \forall \ \frac{u_{t_0}}{v_{t_0}} \geq x \geq y > \frac{u_0}{v_0} \\ f(x) \geq f(y) \ \ \forall \ \frac{u_{m+1}}{v_{m+1}} > x \geq y \geq \frac{u_{t_0}}{v_{t_0}} \end{cases}$$

Since $\begin{cases} \left(\sum_{i=1}^{t_0-1} v_i - \sum_{i=t}^{m} v_i\right) \leq 0 \\ \left(\sum_{i=1}^{t_0} v_i - \sum_{i=t_0+1}^{m} v_i\right) > 0 \end{cases}$

The full set of experimental results and dataset could be found at http://www.egr.msu.edu/~dangchin/ExperimentalResultsRPCA.rar

As a result, $f\left(\frac{u_0}{v_0}\right) = \min_{\alpha \in R} f(\alpha)$.

ACKNOWLEDGMENT

The authors would like to thank Dr. M.Kumar from Eastman Kodak Company for the dataset, the experimental results of some compared methods, along with useful discussions. We also thank Dr. Yang Cong from the Chinese Academy of Science regarding some parts of the experimental results. We also thank Professor John R. Deller from Michigan State University, and Dr. Moghadam from Sony R & D Lab for their useful comments. We also thank Professor Anuj Srivastava, and anonymous reviewers for their valuable comments and constructive feedback to improve the quality of the paper.

REFERENCES

[1] Elhamifar, Ehsan, and Ren Vidal. "Sparse subspace clustering." in *Proc. IEEE Int. Conf. Computer Vision,* 2009, pp.2790-2797.

[2] Hinton, Geoffrey E., Peter Dayan, and Michael Revow. "Modeling the manifolds of images of handwritten digits." *Neural Networks*, IEEE Transactions on 8.1 (1997): 65-74.

[3] Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." *Science* 315.5814 (2007): 972-976.

[4] Baraniuk, Richard G., and Michael B. Wakin. "Random projections of smooth manifolds." *Foundations of Computational Mathematics* 9.1 (2009): 51-77.

[5] Elhamifar, Ehsan, Guillermo Sapiro, and Rene Vidal. "Finding Exemplars from Pairwise Dissimilarities via Simultaneous Sparse Recovery." *Advances in Neural Information Processing Systems* 25. 2012.

[6] B. T. Truong and S. Venkatesh,"Video abstraction: a systematic review and classification. " *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 1, pp. 3es, Feb. 2007.

[7] Tang, Lin-Xie, Tao Mei, and Xian-Sheng Hua. "Near-lossless video summarization." In *Proceedings of the 17th ACM international conference on Multimedia*, pp. 351-360. ACM, 2009.

[8] Gong, Yihong, and Xin Liu. "Video summarization using singular value decomposition." In *Proc. IEEE Int. Conf. Computer Vision,* vol. 2, pp. 174-180. IEEE, 2000.

[9] Ekin, Ahmet, A. Murat Tekalp, and Rajiv Mehrotra. "Automatic soccer video analysis and summarization. " *IEEE Trans. Image Processing* no. 7 (2003): 796-807.

[10] Wang, Meng, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua."Event driven web video summarization by tag localization and key-shot identification." *IEEE Trans. Multimedia* no. 4 (2012): 975-985.

[11] Luo, Jiebo, Christophe Papin, and Kathleen Costello. "Towards extracting semantically meaningful key frames from personal video clips: from humans to computers." *IEEE Trans. Circuits and Systemss for Video Technology* 19.2 (2009): 289-301.

[12] Kumar, Mrityunjay, and Alexander C. Loui. "Key frame extraction from consumer videos using sparse representation." In *Proc. IEEE Int. Conf. Image Processing* pp. 2437-2440, 2011.

[13] Wang, Zheshen, Mrityunjay Kumar, Jiebo Luo, and Baoxin Li. "Extracting key frames from consumer videos using bi-layer group sparsity." In *Proc. ACM Int. Conf. Multimedia*, pp. 1505-1508, 2011.

[14] Zhang, Teng, Arthur Szlam, and Gilad Lerman. "Median k-flats for hybrid linear modeling with many outliers." In *Proc. Int. Conf. Computer Vision* pp. 234-241, 2009.

[15] Yang, Allen Y., Shankar R. Rao, and Yi Ma. "Robust statistical estimation and segmentation of multiple subspaces." In *Computer Vision and Pattern Recognition Workshop* pp. 99-99. IEEE, 2006.

[16] Lerman, Gilad, and Teng Zhang. "ℓ_p-Recovery of the Most Significant Subspace among Multiple Subspaces with Outliers." *arXiv preprint*:1012.4116 (2010).

[17] Candes, Emmanuel J., et al. "Robust principal component analysis?" *Journal of the ACM* 58, no. 3 (2011): 11.

[18] Zhou, Xiaowei, Can Yang, Hongyu Zhao, and Weichuan Yu. "Low-Rank Modeling and Its Applications in Image Analysis," *arXiv* preprint:1401.3409 (2014). (To appear in ACM Computing Surveys)

[19] Peng, Yigang, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images." *IEEE Trans. Pattern Analysis and Machine Intelligence* no. 11 (2012): 2233-2246.

[20] Ji, Hui, Chaoqiang Liu, Zuowei Shen, and Yuhong Xu. "Robust video denoising using low rank matrix completion." In *IEEE Conf. on Computer Vision and Pattern Recognition* , pp. 1791-1798, 2010.

[21] Yang, Chunlei, Jialie Shen, Jinye Peng, and Jianping Fan. "Image collection summarization via dictionary learning for sparse representation." *Pattern Recognition* 46, no. 3 (2013): 948-961.

[22] Elhamifar, Ehsan, Guillermo Sapiro, and Rene Vidal. "See all by looking at a few: Sparse modeling for finding representative objects." In *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1600-1607, 2012.

[23] Donoho, David L. "For most large underdetermined systems of linear equations the minimal ℓ_1 norm solution is also the sparsest solution." *Communications on pure and applied mathematics* 59.6 (2006): 797-829.

[24] Aharon, Michal, Michael Elad, and Alfred Bruckstein. "k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. " *IEEE Trans. Signal Processing*, no. 11 (2006): 4311-4322.

[25] de Avila, Sandra Eliza Fontes, and Ana Paula Brando Lopes. "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method." *Pattern Recognition Letters* 32.1 (2011): 56-68.

[26] Almeida, Jurandy, Neucimar J. Leite, and Ricardo da S. Torres. "Online video summarization on compressed domain." *Journal of Visual Communication and Image Representation* 24, no. 6 (2013): 729-738.

[27] Rasheed, Zeeshan, and Mubarak Shah. "Detection and representation of scenes in videos." *IEEE Trans. Multimedia*, no. 6 (2005): 1097-1105.

[28] Over, Paul, Alan F. Smeaton, and Philip Kelly. "The TRECVID 2007 BBC rushes summarization evaluation pilot." In *Proc. Int. Workshop on TRECVID video summarization*, pp. 1-15. ACM, 2007.

[29] Over, Paul, Alan F. Smeaton, and George Awad. "The TRECVID 2008 BBC rushes summarization evaluation." In *Proc. Int. Workshop on TRECVID video summarization*, pp. 1-20. ACM, 2008.

[30] Li, Ying, Shih-Hung Lee, Chia-Hung Yeh, and C-CJ Kuo. "Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques." *IEEE Signal Processing Magazine*, no. 2 (2006): 79-89.

[31] Wu, Shandong, Omar Oreifej, and Mubarak Shah. "Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories." In *IEEE Int. Conf. Computer Vision*, pp. 1419-1426. IEEE, 2011.

[32] Almeida, Jurandy, Neucimar J. Leite, and Ricardo da S. Torres. "Vison: Video summarization for online applications. " *Pattern Recognition Letters* 33.4 (2012): 397-409.

[33] Dang, C. T., M. Kumar, and H. Radha. "Key frame extraction from consumer videos using epitome." In *Proc. IEEE Int. Conf. Image Processing*, pp. 93-96. IEEE, 2012.

[34] Dang, Chinh T., and Hayder Radha. "Heterogeneity Image Patch Index and Its Application to Consumer Video Summarization," *IEEE Transactions on Image Processing* 23, no. 6 (2014).

[35] Cotsaces, Costas, Nikos Nikolaidis, and Ioannis Pitas. "Video shot detection and condensed representation: a review," *Signal Processing Magazine, IEEE* 23, no. 2 (2006): 28-37.

[36] Drew, Mark S., and James Au. "Video keyframe production by efficient clustering of compressed chromaticity signatures (poster session) " In *Proceedings of the eighth ACM international conference on Multimedia*, pp. 365-367. ACM, 2000.

[37] Dirfaux, F. "Key frame selection to represent a video," In *International Conference on Image Processing*, vol. 2, pp. 275-278. IEEE, 2000.

[38] Furini, Marco, Filippo Geraci, Manuela Montangero, and Marco Pellegrini. "STIMO: STIll and MOving video storyboard for the web scenario," *Multimedia Tools and Applications* 46, no. 1 (2010): 47-69.

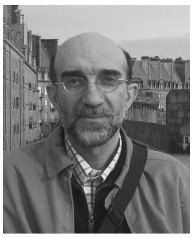
[39] Mundur, Padmavathi, Yong Rao, and Yelena Yesha. "Keyframe-based video summarization using Delaunay clustering," *International Journal on Digital Libraries* 6, no. 2 (2006): 219-232.

[40] T. Bouwmans, E. Zahzah, "Robust PCA via Principal Component Pursuit: A Review for a Comparative Evaluation in Video Surveillance," Special Issue on Background Models Challenge, Computer Vision and Image Understanding, CVIU 2014, Volume 122, pages 2234, May 2014.

[41] Gu, Shuhang, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. "Weighted Nuclear Norm Minimization with Application to Image Denoising " In *IEEE Conf. on Computer Vision and Pattern Recognition*. 2014.

Chinh Dang received B.E. degree (Hons.) in electronics and telecommunications engineering from Talented Engineer Programs, Hanoi University of Science and Technology, Vietnam, in 2010. He is a Ph.D student in electrical and computer engineering department, Michigan State University, East Lansing, MI, USA.

He is currently a research assistant with the Wireless and Video Communications (WAVES) Laboratory, Michigan State University, under the supervision of Prof. Hayder Radha (since 8/2010). His research interests include image/video quality enhancement, multimedia understanding, video coding, compressed sensing, and manifold/geometric topology based signal processing. He received a Gold Medal from 16th Vietnam National Olympiad of Mathematics for undergraduate students in 2008. He is a fellow of Vietnam Education Foundation.



Hayder Radha received the Ph.M. and Ph.D. degrees from Columbia University in 1991 and 1993, the M.S. degree from Purdue University in 1986, and the B.S. degree (with honors) from Michigan State University (MSU) in 1984 (all in electrical engineering). Currently, he is a Professor of Electrical and Computer Engineering (ECE) at MSU and the Director of the Wireless and Video Communications Laboratory. Professor Radha was with Philips Research (1996-2000), where he worked as a Principal Member of Research Staff and then as a Consulting Scientist in the Video Communications Research Department. He was a Member of Technical Staff at Bell Laboratories where he worked between 1986 and 1996 in the areas of digital communications, image processing, and broadband multimedia.

Professor Radha is a Fellow of the IEEE, and he was appointed as a Philips Research Fellow in 2000 and a Bell Labs Distinguished Member of Technical Staff in 1992. He was elected as a member of the IEEE Technical Committee on Image, Video, and Multidimensional Signal Processing (IVMSP) and the IEEE Technical Committee on Multimedia Signal Processing (MMSP). He served as Co-Chair and Editor of a Video Coding Experts Group of the International Telecommunications Union Telecommunications Section (ITUT) between 1994-1996. He is a Senior Editor of IEEE Signal Processing Letters, and he served on the Editorial Board of IEEE Transactions on Multimedia. He also served as a Guest Editor for the special issue on Network-Aware Multimedia Processing and Communications of the IEEE Journal on Selected Topics in Signal Processing. Professor Radha is a recipient of the Bell Labs Distinguished Member of Technical Staff Award, the AT&T Bell Labs Ambassador Award, AT&T Circle of Excellence Award, the MSU College of Engineering Withrow Distinguished Scholar Award for outstanding contributions to engineering, and the Microsoft Research Content and Curriculum Award. His current research areas include compressed sensing and signal sparsification, signal processing of network graphs, analysis of social networks, and visual processing, coding, and communications. He has more than 200 peer-reviewed papers and 30 patents.