IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 4, 2018, accepted January 1, 2019, date of publication January 10, 2019, date of current version February 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2891834

Video Summarization via Nonlinear Sparse Dictionary Selection

MINGYANG MA[1], SHAOHUI MEI[1], SHUAI WAN[1], ZHIYONG WANG[2], AND DAGAN FENG[2], (Fellow, IEEE)

[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China
[2]School of Information Technologies, The University of Sydney, Sydney, NSW 2006, Australia

Corresponding author: Shaohui Mei (meish@nwpu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 3102018AX001 and in part by the National Natural Science Foundation of China under Grant 61671383.

ABSTRACT Video summarization (VS) is to identify important content from a given video, which can help users quickly comprehend video content. Recently, sparse dictionary selection (SDS) has demonstrated to be an effective solution for VS problems, which generally assumes a linear relationship between keyframes and non-keyframes. However, this assumption is not always true for video frames which possess intrinsic nonlinear structures and properties. In this paper, by exploiting the nonlinearity between video frames, a nonlinear SDS model is formulated for VS, in which the nonlinearity is transformed to linearity by projecting a video to a high-dimensional feature space induced by a kernel function. We also propose two greedy optimization algorithms to solve the resulting model, namely the standard kernel SDS (KSDS) greedy algorithm and the robust KSDS greedy algorithm with a backtracking strategy. In order to achieve an intuitive and flexible configuration of the VS process, an adaptive criterion, namely energy ratio, is devised to produce video summaries with different lengths for different video contents. Experimental results on two different benchmark video datasets demonstrate that the proposed algorithm outperforms several state-of-the-art VS algorithms.

INDEX TERMS Nonlinear representation, dictionary selection, sparse representation, video summarization.

I. INTRODUCTION

Nowadays, a vast amount of video data is produced and consumed every day with the rapid development of multimedia technology, Internet and intelligent terminals. For example, according to the survey of Smart insights – ''What Happens Online in 60 Seconds'', about 500 hours of video content is uploaded to YouTube every minute, which means 82.2 years' worth of content is uploaded every day. Such a huge amount of data results in lots of difficulties in video data management and analysis, such as video browsing and video retrieval. Meanwhile, video summarization (VS) aims to generate a compact and informative version, which is of great importance in the era of big video data [1].

VS has been extensively studied and there exist a large number of methods. Existing VS approaches can be categorized into two categories in terms of the forms of video summaries: keyframe extraction approaches and video skim approaches [2]. Keyframe extraction approaches select individual and salient frames as the summary of a given video, while video skim approaches produce a video summary by concatenating a number of important video segments. Although keyframes and video skims are often generated in different ways, these two types of video summarization can be easily converted into each other. Generally, keyframe extraction approaches can produce a smaller number of individual keyframes compared to video skim approaches, which will help human beings and computers comprehend and analyze video content more efficiently, particularly in the applications with huge video capacities and quantities. On the other hand, the keyframe set is not restricted by any timing or synchronization issues, therefore, can provide much more flexibility and adaptability [3]. In this paper, we focus on the keyframe extraction approach.

In the past decades, many keyframe based VS methods have been proposed, which can be classified into the following five categories:

1) Shot Boundary Based Methods: Here, a shot is defined as the longest sequence of frames between two cuts. This kind of methods firstly detect shot boundaries and the first frame, the last frame or the middle frame in a shot can be simply

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

selected as a keyframe. The shot boundary can be detected by identifying changes between of successive frames which implies the boundary. Many methods have been proposed to identify the changes, such as color histogram difference [4], object tracking [5], event analysis [6], dynamic mode decomposition [7] and multi-modal features-based detection [8]. Shot boundary based methods are suitable for the videos whose shots are of little change in content, and their performances heavily depend on the result of boundary detection.

2) Clustering Based Methods: The basic idea for clustering based methods is clustering similar frames/shots together and then selecting a limited number of frames from each cluster to generate the keyframes. For this kind of approaches, there are two key factors that influence the result of clustering: the selected feature upon which video frames are characterized (e.g., color, texture, motion), and the criterion employed to measure the similarity. Munder *et al.* [9] represented video frames as multi-dimensional data points and then used Delaunay Triangulation to cluster them. The frame that was nearest to its center was selected as the keyframe for each cluster. Furini *et al.* [10] proposed the STIMO based on a fast clustering algorithm that selected the most representative frames using HSV frame color distribution. Avila *et al.* [3] proposed VSUMM algorithm, where the color feature in HSV space was firstly extracted to represent video frames, then k-means clustering algorithm was adopted to group the frames, and one frame per cluster was selected. However, it is usually difficult to extract all clusters due to large intraclass and low interclass visual variance [11].

3) Motion Based Methods: Motion feature is considered to be a more prominent feature representing actions or events in a video. Therefore, keyframes can be selected by analyzing the motion information in a video. For example, Wolf [12] first utilized the optical flow analysis to measure the motion and selected keyframes at the local minimum of motion. Mendi *et al.* [13] also used optical flow computations to identify global extrema and local minimum which indicated the keyframe between two maximums in the motion. However, using traditional optical flow analysis to identify the change of motion state is extremely time consuming [14]. Recently, Zhang *et al.* [14] proposed to replace the optical flow with spatiotemporal motion trajectories to detect the change of motion, and the computational cost was reduced efficiently.

4) Sparse Representation Based Methods: In recent years, sparse representation [15], [16] has been widely used in computer vision tasks [17]–[19], and has achieved excellent performance. In sparse representation, a sub-dictionary is selected from an overcomplete dictionary, and the observed signal can be represented as a linear combination of the sub-dictionary. This shares the similar objective of keyframe extraction problems where a frame subset is selected as the keyframes from the original video, and each frame can be represented as a linear combination of the keyframes. As a result, the VS problem can be formulated as a sparse dictionary selection problem such that keyframes are selected according to the sparse representation coefficients. In recent years, sparse representation has been widely used for keyframe extraction based VS methods, and has demonstrated to be an effective and efficient solution to VS [11], [20]–[26] (see Section II for details).

5) Deep Learning Based Methods: Recently, some deep learning based methods have been proposed [27], [28], which aims at learning the summarizing capability using neural networks. A supervised learning based methods was represented [27], in which VS was considered as a structured prediction problem on sequential data, and a bidirectional LSTM was used to model the variable-range dependency. In [28], an unsupervised generative adversarial learning model called SUM-GAN was presented, in which the generator and discriminator were both LSTM networks. Its supervised version, namely SUM-GANsup, was proposed by adding a sparse regularization. In addition to summarizing a single video, some works generated summaries from large collections of videos. Gao *et al.* [29] proposed event video mashup to generate a short video from some related videos to describe an event, which could identify important frames/shots and temporally align them by simultaneously considering multiple videos.

In this paper, we focus on the sparse representation based method, which is a popular and effective technique used in unsupervised VS [30]. Although many sparse representation based VS methods have been proposed, existing works mainly focus on linear sparse representation which assumes a linear relation among video frames. However, many descriptors used for characterizing video frames have intrinsic nonlinear properties, and the relationships between extracted features are almost nonlinear. As a result, the summarization performance/quality of linear sparse representation models could be compromised. Recall that kernel trick [31] can project nonlinearly separable features into a high dimensional feature space where the projected features are linearly separable [32]. As a result, the higher order structure of the original data can be implicitly exploited with kernel trick under linear models. The linear sparse representation can also be extended into nonlinear cases with such kernel trick to handle signals with nonlinear structures. Therefore, in this paper, the nonlinearity among video frames is taken into account and a $\ell_{2,0}$ norm based nonlinear sparse dictionary selection (NSDS) model is formulated for VS. Specifically, two greedy algorithms, namely standard KSDS algorithm and robust KSDS algorithm, are proposed for model optimization. Experiments on two benchmark video datasets are then carried out to demonstrate the effectiveness of our proposed nonlinear model and methods for VS. Compared with our previous work [26], this work proposes a new robust algorithm which greatly improves the performance. In addition, various influencing factors are analyzed.

In summary, the key contributions of this paper are as follows:

1) The nonlinearity among video frames is taken into account and transformed into the linearity, and

a $\ell_{2,0}$ norm based nonlinear sparse dictionary selection model (NSDS) is proposed to formulate the keyframe extraction based VS problem.

2) Two efficient and effective nonlinear sparse representation based algorithms using kernel tricks and matching pursuit are proposed for VS problems: a standard KSDS algorithm and a robust KSDS algorithm with a backtracking strategy.

3) An energy ratio (ER) of residual to video in the kernel-mapped high dimensional space is devised to achieve an intuitive and flexible configuration for VS, which makes the proposed algorithms adaptive to different kinds of videos.

The rest of this paper is organized as follows: we review the related work of sparse representation based video summarization in Section II. In Section III, we present the formulation of general sparse dictionary selection and its expansion for VS. In Section IV, we present the formulation of our nonlinear sparse dictionary selection model and optimize it using a standard KSDS algorithm and a robust KSDS algorithm. In Section V, we conduct experiments on two benchmark video datasets and compare the proposed algorithms with several state-of-the-art works. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

Sparse representation based video summarization methods can be grouped into two categories: *sparse representation with ℓ_0 or $\ell_{2,0}$ norm minimization* and *sparse representation with ℓ_1 or $\ell_{2,1}$ norm minimization.*

1) Sparse Representation With ℓ_0 or $\ell_{2,0}$ Norm Minimization: The ℓ_0 norm calculates the number of non-zero elements in a vector, and the $\ell_{2,0}$ norm calculates the number of the non-zero rows in a matrix, so the $\ell_{2,0}$ norm can be viewed as a special case with multi-vectors of ℓ_0 norm. Both of them are used to ensure the sparsity of the sparse representation coefficient, which is also the number of selected keyframes in VS, and the representation with ℓ_0 or $\ell_{2,0}$ norm minimization is usually solved by a greedy pursuit algorithm. For example, Mei *et al.* [24] formulated the VS as a sparse dictionary selection problem, and the $\ell_{2,0}$ norm was adopted to guarantee the simultaneous sparsity. The traditional SOMP greedy algorithm was extended to solve the $\ell_{2,0}$ norm sparse dictionary selection problem. Afterwards, Mei *et al.* [11] reformulated VS as a minimum sparse reconstruction (MSR) problem with the true sparsity constraint ℓ_0 norm. The algorithm used the selected keyframes to reconstruct unselected frame each iteration and the frame with maximum reconstruction error was selected as a keyframe. Recently, Cong *et al.* [25] designed a $\ell_{2,0}$ norm dictionary selection model with a forward-backward greedy optimization procedure. For model optimization, two methods in the forward step were also proposed: the standard forward greedy algorithm and the gradient cue based algorithm for speeding up. In addition, the nonlinear case has been explored preliminarily in our previous work [26].

2) Sparse Representation With ℓ_1 or $\ell_{2,1}$ Norm Minimization: The ℓ_1 norm calculates the sum of the absolute values of the elements in a vector, and the $\ell_{2,1}$ norm calculates the sum of the ℓ_2 norm of all rows in a matrix, so the $\ell_{2,1}$ norm can also be viewed as a special case with multi-vectors of ℓ_1 norm. The sparse representation with ℓ_0 or $\ell_{2,0}$ norm minimization is NP-hard, however, ℓ_1 or $\ell_{2,1}$ norm minimization can be solved by convex programming and ensure the sparsity of the representation coefficient as ℓ_0 or $\ell_{2,0}$ norm does. Therefore, the sparse representation with ℓ_1 or $\ell_{2,1}$ norm minimization has also been widely studied for VS. For example, Kumar and Loui [20] presented a ℓ_1 norm sparse representation based method to extract keyframes from unstructured consumer videos, and the sparse coefficient was solved by an interior-point method [33]. For $\ell_{2,1}$ norm constrained VS methods, various convex optimization algorithms have been proposed to solve this problem. Cong *et al.* [21] formulated VS as a dictionary selection problem using sparsity consistency with $\ell_{2,1}$ imposed to ensure sparsity, and the convex but nonsmooth optimization problem was solved by Nesterov's optimization method [34]. Liu *et al.* [22] proposed a $\ell_{2,1}$ structured optimization model for VS, and the nonconvex model was transformed into a weighted optimization problem which was solved by minimizing an equivalent objective function iteratively. Etezadifar and Farsi [23] also formulated VS as a optimization problem with a $\ell_{2,1}$ norm to guarantee the sparsity, and the problem was solved by performing dictionary learning and selection simultaneously in each iteration.

However, almost all the existing sparse representation based methods are based on the linear sparse representation without considering the nonlinearity between video frames, which limits the performance of keyframe extraction. Therefore, we proposed the nonlinear sparse dictionary selection (NSDS) model to formulate the VS problems to take the nonlinearity into consideration, in which the $\ell_{2,0}$ norm directly guarantees the sparsity of keyframes, and keyframes generating the minimum representation error are selected.

III. PRELIMINARY: SDS MODEL AND VS

A. SDS MODEL

SDS problem aims to extract a subset dictionary from an over-complete dictionary $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{n_1}] \in \mathbb{R}^{d\times n_1}$ such that the observed signal $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{n_2}] \in \mathbb{R}^{d\times n_2}$ can be well represented by the subset dictionary, where d is the observed dimensionality, n_1 and n_2 are the numbers of atoms in the over-complete dictionary and the observed signal, respectively. Generally, such SDS problem can be solved by optimizing the following sparse representation model:

$$\hat{\mathbf{X}} = \arg\min \|\mathbf{A} - \mathbf{BX}\|_F^2 \quad s.t.\ \mathcal{C}(\mathbf{X}) \le K_0, \tag{1}$$

in which $\mathbf{X} \in \mathbb{R}^{n_1\times n_2}$ represents the sparse representation coefficient matrix, $\|\cdot\|_F$ is the Frobenius norm, $\mathcal{C}(\cdot)$ is a constraint to induce sparsity, and K_0 is a given upper boundary on the sparsity level. As a result, the aiming subset dictionary, denoted as $\mathbf{B}_{sub} = [\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \cdots, \mathbf{b}_{i_m}] \in \mathbb{R}^{d\times m} \subseteq \mathbf{B}$ where $i_1, i_2, \cdots, i_m \in \{1, 2, \cdots, n_1\}$, can be selected by searching the nonzero rows of $\hat{\mathbf{X}}$. That is, if the i-th row of $\hat{\mathbf{X}}$ is a nonzero vector, the corresponding i-th column of $\mathbf{B}$ is selected for representation, and vice versa.

B. SDS MODEL BASED VS

In the VS problem, keyframes should represent all the important information in the video without significant information loss. Let $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n] \in \mathbb{R}^{d \times n}$ represent candidate video frames in a video, whose columns $\{\mathbf{f}_i\}_{i=1,2,\cdots,n}$ are d-dimensional feature vectors denoting n frames. The aim of VS is to find a video frame subset $\mathbf{F}_K = [\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \cdots, \mathbf{f}_{i_m}] \in \mathbb{R}^{d \times m}$ where $i_1, i_2, \cdots, i_m \in \{1, 2, \cdots, n\}$, m is the number of the selected keyframes, and $m \ll n$, such that it can represent all the information of interest among video frames. Such VS problem can be viewed as a special case of the sparse representation model defined in (1) that the observed signal is used as the over-complete dictionary, i.e. , $\mathbf{A} = \mathbf{B}$. As a result, the following SDS based model can be formulated for VS:

$$\hat{\mathbf{X}} = \arg\min \|\mathbf{F} - \mathbf{F}\mathbf{X}\|_F^2 \quad s.t. \ \|\mathbf{X}\|_{row,0} < K_0, \tag{2}$$

where $\|\cdot\|_{row,0}$ denotes the number of nonzero rows in the input. Equivalently, the VS problem defined in (2) can also be solved by by optimizing sparsity level within the approximation error:

$$\hat{\mathbf{X}} = \arg\min \|\mathbf{X}\|_{row,0} \quad s.t. \ \|\mathbf{F} - \mathbf{F}\mathbf{X}\|_F^2 < \sigma, \tag{3}$$

where σ is the error tolerance.

The sparse recovery problems for VS defined in (2) and (3) are NP-hard, which can be solved approximately by greedy pursuit algorithms [35], [36], or relaxed to convex programming in polynomial time [37], [38]. Correspondingly, many dictionary representation algorithms have been proposed as reviewed in Section II.

IV. PROPOSED NONLINEAR DICTIONARY REPRESENTATION FOR VS

In this section, we first formulate our kernel based framework for nonlinear dictionary dictionary selection (NSDS), then introduce two optimization algorithms, standard KSDS algorithm and a robust KSDS algorithm with a backtracking strategy.

A. NONLINEAR DICTIONARY REPRESENTATION MODEL FOR VS

Suppose there exists a nonlinear function $\Phi(\cdot) : \mathbb{R}^d \rightarrow \mathcal{F} \subset \mathbb{R}^{\tilde{d}}$ representing nonlinear mapping from the original video feature space $\mathbb{R}^d$ into a new high-dimensional feature space $\mathcal{F}$ ($\tilde{d} \gg d$, may be *infinite* dimensional), such that the nonlinear relationship in $\mathbb{R}^d$ can be converted into linear one in $\mathbb{R}^{\tilde{d}}$. Based on such mapping, new feature vectors $\Phi(\mathbf{F}) = [\Phi(\mathbf{f}_1), \Phi(\mathbf{f}_2), \cdots, \Phi(\mathbf{f}_n)] \in \mathbb{R}^{\tilde{d} \times n}$ can be generated to represent video frames, and thus, the nonlinearity among video frames vectors in $\mathbf{F}$ can be transformed into linearity among frame feature vectors in $\Phi(\mathbf{F})$. As a consequence, the keyframes $\Phi(\mathbf{f}_i)_K, i = 1, 2, \ldots, m$ can be viewed as a dictionary to linearly represent all the video frames $\Phi(\mathbf{f}_i), i = 1, 2, \ldots, n$, such that nonlinear sparse dictionary selection (NSDS) models are proposed for VS according to the SDS models defined in (2) and (3):

$$\hat{\mathbf{X}^{\phi}} = \arg\min \|\Phi(\mathbf{F}) - \Phi(\mathbf{F})\mathbf{X}^{\phi}\|_F^2, \quad s.t. \ \|\mathbf{X}^{\phi}\|_{row,0} < K_0, \tag{4}$$

or,

$$\hat{\mathbf{X}^{\phi}} = \arg\min \|\mathbf{X}^{\phi}\|_{row,0}, \quad s.t. \ \|\Phi(\mathbf{F}) - \Phi(\mathbf{F})\mathbf{X}^{\phi}\|_F^2 < \sigma, \tag{5}$$

where $\mathbf{X}^{\phi}$ represents the nonlinear sparse representation coefficient matrix. Similarly, $\mathbf{X}^{\phi}$ is a *row-sparse* matrix to ensure $\Phi(\mathbf{F}_K) \triangleq [\Phi(\mathbf{f}_i)_K]$ is a subset of $\Phi(\mathbf{F})$. Compared to the SDS models in (2) and (3), the proposed NSDS models in (4) and (5) linearly represent the corresponding high dimensional features that are induced by a nonlinear mapping function, so that the nonlinearity among frames can be considered. Besides, the model can not only guarantee minimum approximation error within error tolerance in the least-squares sense, but also satisfy the requirement of fewer keyframes with the constraint of sparse level upper boundary. The NSDS for VS seeks the sparse representation $\hat{\mathbf{X}^{\phi}}$ of mapped frames in the high dimensional space. Similar to the linear SDS model, keyframes can be further extracted by searching for rows of $\hat{\mathbf{X}^{\phi}}$ with nonzero elements in the NSDS model. Note that, in order to guarantee row-sparsity of $\mathbf{X}^{\phi}$, $\|\mathbf{X}^{\phi}\|_{row,0}$ can be flexibly replaced by arbitrary $\ell_{p,0}$ norm where ($p \geq 1$), i.e., $\|\mathbf{X}^{\phi}\|_{p,0} = \|\{y | y = \|{\mathbf{X}^{\phi}}_{i\cdot}\|_p\}\|_0$, where ${\mathbf{X}^{\phi}}_{i\cdot}$ is the i-th row of $\mathbf{X}^{\phi}$.

One of the key problems in the NSDS model defined in (4) and (5) is to find the nonlinear mapping $\Phi(\cdot)$. However, such mapping maybe exist by mapping the original feature $\mathbf{F}$ to a extremely high-dimensional space $\mathbb{R}^{\tilde{d}}$, resulting in *the curse of dimensionality* in solving the NSDS. Therefore, in this paper, the mapping is restricted to the Hilbert spaces so that the *Mercer* kernel based method can be used to execute computations implicitly without venturing into the high-dimensional feature space.

A Mercer kernel is a function $\mathcal{K}(\mathbf{x}, \mathbf{y})$ that satisfies the Mercer's condition: for all data $\{\mathbf{y}_i\}_{i=1}^n$, the function gives rise to a positive semi-definite matrix $[\mathbf{K}_{ij}] = [\mathcal{K}(\mathbf{y}_i, \mathbf{y}_j)]$. If $\mathcal{K}$ satisfies the Mercer's condition, it can be shown that $\mathcal{K}$ corresponds to some mapping Φ in the Hilbert feature space $\mathcal{F}$. That is, a kernel function is defined as $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$. Mercer kernels are often used to implicitly impose the mapping Φ. Although the kernel $\mathcal{K}(\mathbf{x}, \mathbf{y})$ is in the form of inner product, it usually does not explicitly calculate the inner product of two mapped vectors in the high-dimensional space. The kernel function cleverly solves the above problem, and the inner product of two mapped vectors in the high-dimensional space can be calculated by the kernel function of a low-dimensional space, which helps to avoid the expensive computation when mapping the data into the high-dimensional feature space. This technique is also called "*kernel trick*". The commonly-used kernels include the q-th order polynomial kernel $\mathcal{K}(\mathbf{f}_i, \mathbf{f}_j) = (\langle \mathbf{f}_i, \mathbf{f}_j \rangle + c)^q$ and radial basis function (RBF) kernel $\mathcal{K}(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\frac{1}{2\sigma^2}\|\mathbf{f}_i - \mathbf{f}_j\|_2^2)$ with σ controlling the width of the RBF.

B. STANDARD KSDS ALGORITHM

Orthogonal matching pursuit (OMP) [39] is an iteratively greedy algorithm for sparse reconstruction, where the atom most correlated with the residuals is selected. Moreover, the residuals and the chosen atoms are always orthogonal, which means the atoms similar to previously selected atoms will not be selected again. Tropp and Gilbert [40] theoretically showed its remarkable efficiency, and extended it to solve the simultaneous sparse approximation, namely simultaneous OMP (SOMP) [35]. SOMP-based algorithms are also very effective to solve the linear SDS problem for VS [24], [25]. Therefore, based on SOMP and kernel trick, a kernel SDS (KSDS) algorithm is proposed to solve the NSDS problem for VS defined in (4) and (5).

Let $\Lambda = [i_1, i_2, \ldots, i_m]$ represents the index of selected frames as keyframes. Therefore, $\mathbf{F}_K = \mathbf{F}_{:,\,\Lambda}$. Under the nonlinear representation assumption, all the video keyframes can be represented by the keyframes:

$$\Phi(\mathbf{F}) = \Phi(\mathbf{F})\mathbf{X}^{\phi} = \Phi(\mathbf{F})_{:,\,\Lambda}\mathbf{X}^{\phi}_{\Lambda,\,:}, \tag{6}$$

where $\mathbf{X}^{\phi}$ represents the nonlinear representation coefficients of all the video frames by the keyframes that correspond to the non-zero rows in $\mathbf{X}^{\phi}$. If all the frames can be well reconstructed by the selected keyframes, the nonlinear representation error $\mathbf{R} = \Phi(\mathbf{F}) - \Phi(\mathbf{F})_{:,\,\Lambda}\mathbf{X}^{\phi}_{\Lambda,\,:}$ will be negligible. On the contrary, if $\mathbf{R}$ is not trivial, more keyframes should be selected. In the SOMP algorithm, the frames simultaneously yielding best approximation to the residuals of all frames are selected as keyframes. Therefore, in the proposed KSDS algorithm, the nonlinear case for such approximation is considered by using the correlation as a criterion to select new keyframes:

$$\lambda = \arg\max_{1 \le i \le n} \left\| \Phi(\mathbf{f}_i)^T \mathbf{R} \right\|_p^2, \tag{7}$$

where λ indicates index for a new keyframe and $\|\cdot\|_p$ represents the ℓ_p norm of vectors. By introducing kernel trick, such nonlinear-mapping-based optimization problem can be solved using the following kernel-based optimization:

$$\lambda = \arg\max_{1 \le i \le n} \left\| \mathbf{K}_{(\mathbf{F},\mathbf{R})i,:} \right\|_p^2, \tag{8}$$

where $\mathbf{K}_{(\mathbf{F},\mathbf{R})} \in \mathbb{R}^{n \times n}$ is a kernel matrix whose (i,j)-th element is $\mathcal{K}(\mathbf{f}_i, \mathbf{r}_j)$ and $\mathbf{r}_j (j = 1, 2, \ldots, n)$ denotes the j-th column vector of $\mathbf{R}$ representing the nonlinear representation error of j-th video frame. According to such kernel based optimization, all the keyframes can be selected iteratively.

In order to select keyframes according to (8), the nonlinear representation coefficients $\mathbf{X}^{\phi}_{\Lambda,\,:}$ must be determined to calculate the corresponding representation residuals $\mathbf{R}$. According to (6), the following optimization is formulated to obtain $\mathbf{X}^{\phi}_{\Lambda,\,:}$:

$$\hat{\mathbf{X}}^{\phi}_{\Lambda,:} = \arg\min \left\| \Phi(\mathbf{F})_{:,\Lambda}\mathbf{X}^{\phi} - \Phi(\mathbf{F}) \right\|_F^2. \tag{9}$$

We adopt the least square method to estimate the reconstruction coefficient:

$$\begin{aligned} \hat{\mathbf{X}}^{\phi}_{\Lambda,:} &= \left(\Phi(\mathbf{F})^T_{:,\Lambda}\Phi(\mathbf{F})_{:,\Lambda}\right)^{-1}\Phi(\mathbf{F})^T_{:,\Lambda}\Phi(\mathbf{F}) \\ &= \mathbf{K}^{-1}_{\Lambda,\Lambda}\mathbf{K}_{\Lambda,:}, \end{aligned} \tag{10}$$

in which $\mathbf{K} \in \mathbb{R}^{n \times n}$, whose (i,j)-th element is $\mathcal{K}(\mathbf{f}_i, \mathbf{f}_j)$.

Once the nonlinear reconstruction coefficient is obtained, the residual of the video in the high-dimensional feature space can be calculated as follows:

$$\begin{aligned} \mathbf{R} &= \Phi(\mathbf{F}) - \Phi(\mathbf{F})_{:,\Lambda}\hat{\mathbf{X}}^{\phi}_{\Lambda,\,:} \\ &= \Phi(\mathbf{F}) - \Phi(\mathbf{F})_{:,\Lambda}\mathbf{K}^{-1}_{\Lambda,\Lambda}\mathbf{K}_{\Lambda,:}. \end{aligned} \tag{11}$$

It is observed from (11) that the representation residuals $\mathbf{R}$ cannot be explicitly calculated, because the nonlinear mapping is implemented implicitly without using an explicit mapping. However, the inner product used for selecting keyframes in (7) and (8) can be calculated as follows:

$$\begin{aligned} \mathbf{K}_{(\mathbf{F},\mathbf{R})i,:} &= \Phi(\mathbf{f}_i)^T\mathbf{R} \\ &= \Phi(\mathbf{f}_i)^T\left(\Phi(\mathbf{F}) - \Phi(\mathbf{F})_{:,\Lambda}\mathbf{K}^{-1}_{\Lambda,\Lambda}\mathbf{K}_{\Lambda,:}\right) \\ &= \Phi(\mathbf{f}_i)^T\Phi(\mathbf{F}) - \Phi(\mathbf{f}_i)^T\Phi(\mathbf{F})_{:,\Lambda}\mathbf{K}^{-1}_{\Lambda,\Lambda}\mathbf{K}_{\Lambda,:} \\ &= \mathbf{K}_{i,:} - \mathbf{K}_{i,\Lambda}\mathbf{K}^{-1}_{\Lambda,\Lambda}\mathbf{K}_{\Lambda,\,:} \end{aligned} \tag{12}$$

Generally, it is difficult to determine the number of keyframes exactly for VS problem [41]. In the sparse representation based VS applications, in addition to the sparsity, an effective criterion which can adaptively adjust the number of selected keyframes for different kinds of videos should be designed. Thus, the energy ratio (ER) of residual to the video in the high dimensional feature space is devised as a stopping criterion:

$$ER = \frac{\mathcal{E}(\mathbf{R})}{\mathcal{E}(\Phi(\mathbf{F}))}, \tag{13}$$

where $\mathcal{E}(\cdot)$ represents the energy in a matrix and the F-norm is adopted. Thus, the energy in video frames matrix $\Phi(\mathbf{F})$ and their corresponding residuals matrix in nonlinear representation $\mathbf{R}$ are as follows:

$$\mathcal{E}(\Phi(\mathbf{F})) = \|\Phi(\mathbf{F})\|_F^2 = Tr\left(\Phi(\mathbf{F})^T\Phi(\mathbf{F})\right) = Tr(\mathbf{K}), \tag{14}$$

$$\begin{aligned} \mathcal{E}(\mathbf{R}) &= \|\mathbf{R}\|_F^2 = Tr\left(\mathbf{R}^T\mathbf{R}\right) \\ &= Tr\left(\left[\Phi(\mathbf{F}) - \Phi(\mathbf{F})_{:,\Lambda}\hat{\mathbf{X}}^{\phi}_{\Lambda,:}\right]^T\left[\Phi(\mathbf{F}) - \Phi(\mathbf{F})_{:,\Lambda}\hat{\mathbf{X}}^{\phi}_{\Lambda,:}\right]\right) \\ &= Tr\left(\mathbf{K} - 2\mathbf{K}_{:,\Lambda}\hat{\mathbf{X}}^{\phi} + (\hat{\mathbf{X}}^{\phi})^T\mathbf{K}_{\Lambda,\Lambda}\hat{\mathbf{X}}^{\phi}\right), \end{aligned} \tag{15}$$

where $Tr(\cdot)$ represents the trace of a matrix, and especially when RBF is used as the kernel function, $Tr(\mathbf{K}) = n$. As a result, the energy ratio (ER) is determined as follows:

$$ER = \frac{Tr\left(\mathbf{K} - 2\mathbf{K}_{:,\Lambda}\hat{\mathbf{X}}^{\phi} + (\hat{\mathbf{X}}^{\phi})^T\mathbf{K}_{\Lambda,\Lambda}\hat{\mathbf{X}}^{\phi}\right)}{Tr(\mathbf{K})}. \tag{16}$$

The proposed adaptive stopping criterion presents a decreasing trend with the decreasing of the residual energy. Therefore, when ER decreases below a predefined threshold ER_{thr}, the keyframe selection iteration will stop and all the keyframes are identified.

C. ROBUST KSDS ALGORITHM

According to (7), the selection of keyframes in each iteration is determined by the correlation between video frames and representation residuals. However, several frames may possess very close correlations, such as temporal adjacent frames which own very similar content. Under such circumstance, the selection of the frame possessing maximum correlation is not robust. As a result, in each iteration of the proposed KSDS, the best keyframe is selected using a local backtracking mechanism, in which several candidate keyframes are firstly selected according to (7) and the best one is then refined. Therefore, n_c most related frames to the current residuals are selected first:

$$\lambda_j = \underset{1\leq i\leq n, i\neq\{\lambda_1,\lambda_2,\ldots,\lambda_{j-1}\}}{\arg\max} \left\| \Phi(\mathbf{f}_i)^T \mathbf{R} \right\|_p^2, \quad j = 1, 2, \cdots, n_c. \tag{17}$$

Such greedy selection only focuses on selecting the appropriate frames that have high correlation with the current residuals. However, the selected keyframes are expected to well represent the original video frames, instead of the current residuals. Therefore, all the candidate keyframes should be used to reconstruct all the video frames and the significance of each candidate keyframe is evaluated by the reconstruction coefficients. The smaller the value of the significance, the less important the corresponding keyframe. As a result, the candidate keyframe corresponding to the largest significance value is selected as a keyframe:

$$\lambda_s = \underset{i=1,2,\ldots,n_c}{\arg\max} \; \|(\mathbf{X}_{\mathcal{L}})_{i,:}\|_p^2, \tag{18}$$

where $\mathbf{X}_{\mathcal{L}}$ represents the reconstruction coefficient using candidate keyframes to reconstruct all the video frames, which can be obtained using least-square solution as follows:

$$\begin{aligned} \mathbf{X}_{\mathcal{L}} &= \left(\Phi(\mathbf{F})_{:,\mathcal{L}}^T \Phi(\mathbf{F})_{:,\mathcal{L}}\right)^{-1} \Phi(\mathbf{F})_{:,\mathcal{L}}^T \Phi(\mathbf{F}) \\ &= \mathbf{K}_{\mathcal{L},\mathcal{L}}^{-1} \mathbf{K}_{\mathcal{L},:}, \end{aligned} \tag{19}$$

where $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_j\}$ is the index set of selected candidate keyframes, and each row in $\mathbf{X}_{\mathcal{L}}$ corresponds to one candidate keyframe. According to the proposed refinement strategy, the best keyframe is selected in each iteration and the selection is robust to the frames which have similar correlations, such as temporally adjacent frames. Note that KSDS is a special case of R-KSDS when $n_c = 1$. The proposed standard KSDS and robust KSDS (R-KSDS) are summarized in Algorithm 1.

Algorithm 1 Proposed KSDS and R-KSDS for VS

```
Input: video frames $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n]$, ER threshold
    $ER_{thr}$, sparsity $K_0$, the number of candidate frames $n_c$
    (only for R-KSDS).
Output: index set of keyframes $\Lambda$
    Initialization: the kernel matrix $\mathbf{K}$, the representation
    residual $\mathbf{R}_0 = \Phi(\mathbf{F})$, the index set $\Lambda_0 = \emptyset$, and the
    iteration counter $t = 1$.
 1: while $t \leq K_0$ and $ER_t \geq ER_{thr}$ do
 2:     //For Standard KSDS: using Step 3.
 3:     Find the index of the selected frame according to (7).
 4:     //For Robust KSDS: using Steps 5, 6, and 7.
 5:     Find the indices of $n_c$ candidate keyframes according
        to (17).
 6:     Calculate the significance for the candidate keyframes
        according to (19).
 7:     Extract the index of best candidate keyframes according to (18).
 8:     Augment the index set: $\Lambda_t \leftarrow \Lambda_{t-1} \cup \{\lambda_t\}$.
 9:     Update the reconstruction coefficient by
        solving (9) or using (10).
10:     Calculate the energy ratio (ER) of residual to video
        according to (16).
11:     Increment iteration: $t \leftarrow t + 1$.
12: end while
```

TABLE 1. Computation complexity analysis of the t-th iteration in the proposed KSDS algorithm.

Calculation	Complexity
initialization: $\mathbf{K}$	$O(n^2 d)$
$\mathbf{K}_{(\mathbf{F},\mathbf{R})i,:}, i = 1, \cdots, n$	$max\{O(N_{\Lambda_t}^3, nN_{\Lambda_t}^2, n^2 N_{\Lambda_t})\} = O(n^2 N_{\Lambda_t})$
$\hat{\mathbf{X}}_{\Lambda,:}^{\phi}$	$max\{O(N_{\Lambda_t}^3, nN_{\Lambda_t}^2)\} = O(nN_{\Lambda_t}^2)$
$\mathcal{E}(\mathbf{R})$	$max\{O(nN_{\Lambda_t}^2, n^2 N_{\Lambda_t})\} = O(n^2 N_{\Lambda_t})$
whole iteration	$\sum_{t=1}^{K} O\left(n^2 N_{\Lambda_t}\right) = O(n^2 K^2)$
whole algorithm	$max\{O(n^2 K^2, n^2 d)\}$

The computational complexity analysis is shown in Table 1, in which N_{Λ_t} is the number of the keyframes Λ_t in the t-th iteration. Because $N_{\Lambda_t} \ll n$, the computational complexity of the whole iteration is n^2K^2, which is determined by the number of video frames n and the sparsity K. Before the iteration, the kernel matrix $\mathbf{K}$ should be initialized, whose computational complexity is n^2d.

V. EXPERIMENTS AND DISCUSSION

In this section, we present various experiments and comparisons to validate the effectiveness of our proposed method. Firstly, we describe the datasets utilized for the evaluations and introduce the quantitative evaluation metrics. Secondly, the influences of some settings are also investigated. Finally, our proposed algorithms are compared quantitatively with several state-of-the-art algorithms.

A. EXPERIMENTAL DATASETS

In order to evaluate the proposed method, two video datasets have been used as follows:

1) VSUMM

The VSUMM dataset [3] contains 50 videos from the Open Video Project [42] across several genres (e.g., documentary, educational, ephemeral, historical, and, lecture) with durations varying from 1 to 4 minutes (approximately 75 minutes in total), and the frame rate is about 30 frames per second. The ground truth summaries of each video clip in VSUMM were obtained by 5 different users, so there are 5 summarized sets for each video clip, which are available for all 50 video clips in the dataset.

2) TVSum

The TVSum dataset [43] consists of 50 videos collected from YouTube, the videos vary across 10 categories (5 videos per category) from the TRECVid Multimedia Event Detection task [44], such as news, how-toąŕs, documentaries, and user-generated content. Their durations vary from 1.6 to 10.8 min, and the average is 4.2 min. The ground truth summaries of the dataset were obtained by 1000 responses collected from Amazon Mechanical Turk, so each video was summarized by 20 responses.

Like most video summarization works [3], [11], [25], each video of both datasets is down-sampled at 5 frames per second in our experiments. The frame size of the videos in VSUMM is 352 × 240, and the frame size of most videos in TVSum is 640 × 360 (2 videos with the size of 540 × 360, 7 videos with the size of 480 × 360).

B. EVALUATION METRICS

Since we adopt two datasets provided by different groups, the forms of the ground truth summaries are different. More specifically, the ground truth summaries of VSUMM are keyframes, however the TVSum provides an importance score from 1 (not important) to 5 (very important) to each shot that has a uniform length of two seconds. Therefore, two kind of evaluation metrics are adopted in responding to different ground truth summaries.

1) VSUMM

In order to quantitatively evaluate the performance of VS, automatic summaries (AS) generated by different algorithms are compared with the user summaries (US). Three evaluation metrics, including Precision, Recall, and F-score, are used to measure summarization quality of each algorithm:

$$Precision = \frac{n_{matched}}{n_{AS}}, \tag{20}$$

$$Recall = \frac{n_{matched}}{n_{US}}, \tag{21}$$

$$F\text{-}score = \frac{(\beta^2+1)Precision \cdot Recall}{\beta^2 Precision + Recall}, \tag{22}$$

where $n_{matched}$ is the number of matched keyframes from an automatic summary, n_{AS} is the number of keyframes in the automatic summary, n_{US} is the number of keyframes in a user summary, and β controls the balance between precision and recall. In our evaluation setup, we set $\beta = 1$ and F-score is equivalent to the harmonic mean of precision and recall.

2) TVSum

The ground truth of the TVSum dataset is the importance score of the uniform two seconds length shots. As the provider has stated, a shot length of two seconds can capture local context with good visual coherence, and after watching these videos, we also find that video content in the uniform-length shot does not change much. In other words, the video content of a shot can be represented by any of its video frames. Though our summarization results are a set of keyframes rather than video shots, a keyframe can obtain the same importance score of the two-second shot which contains the keyframe.

Through this way, we convert the score of a shot to that of a keyframe. In addition, if more than one frames belong to the same shot, the scores of these keyframes should be penalized due to their redundancy. Specifically, the keyframes belonging to the same shot will share the shot's importance score equally. Therefore, we propose the mean score of selected keyframes as the metrics to evaluate the performance.

$$Score = \frac{\sum_{i=1}^{m} f_i}{m} = \frac{\sum_{i=1}^{m} \frac{Score_i}{r_i}}{m}, \tag{23}$$

where m is the number of the selected keyframes, f_i is the score of the i-th keyframes, $Score_i$ is the importance score of the shot that the i-th keyframe belongs to, and r_i is the number of keyframes belonging to the same shot.

C. IMPACT OF SETTINGS

In this part, the influences of three settings are explored by implementing the proposed KSDS algorithm over VSUMM, including the type of kernels, different $\ell_{p,0}$, and the number of candidate keyframes n_c in R-KSDS algorithm.

1) IMPACT OF KERNELS

The q-th order polynomial kernel $\mathcal{K}(\mathbf{f}_i, \mathbf{f}_j) = (\langle \mathbf{f}_i, \mathbf{f}_j \rangle + 1)^q$ and radial basis function kernel $\mathcal{K}(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\frac{1}{2\sigma^2}\|\mathbf{f}_i - \mathbf{f}_j\|_2^2)$ are adopted to explore the influence in the proposed KSDS algorithm. The parameters in each kernel vary within a certain range respectively, and the values which achieve the best performances are adopted to make a comparison. Specifically, $q = 13$ is for the polynomial kernel, and $\sigma = 0.22$ is for the RBF kernel. The experimental results of the proposed KSDS algorithm with different kernels are shown in Fig. 1, in which the ER threshold varies from 32% to 12%. It can be observed that the RBF kernel generally has a better performance than the polynomial kernel, except when the ER threshold is high, i.e., the number of keyframes is very small. Thus, the RBF kernel is adopted in our experiments.

2) IMPACT OF $\ell_{p,0}$

Empirically, different $\ell_{p,0}$ constraints on sparse representation coefficient will have an impact on the selection of keyframes. Therefore, the influence of $\ell_{p,0}$ is investigated

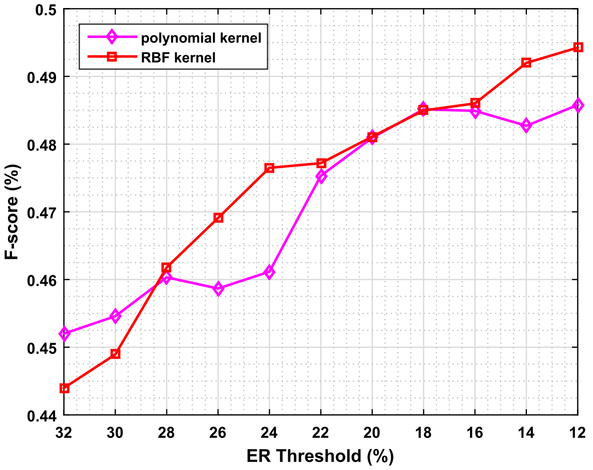

FIGURE 1. Performance comparison of the polynomial kernel and RBF kernel on VSUMM.

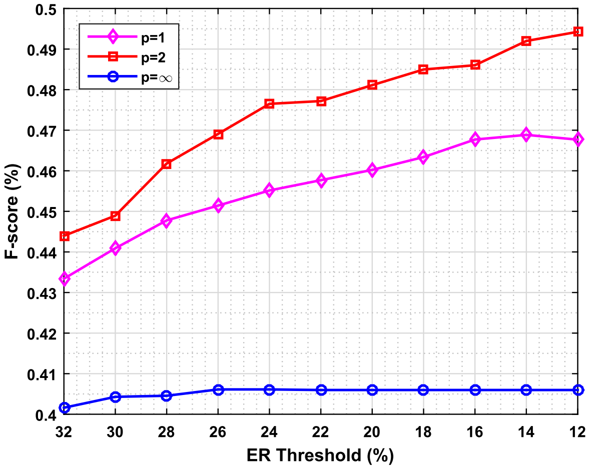

FIGURE 2. Performance comparison of different $\ell_{p,0}$ on VSUMM.

in this part, and p is set to 1, 2 and ∞, respectively. The performance comparison of different $\ell_{p,0}$ is shown in Fig. 2. It can be obviously observed that $\ell_{2,0}$ and $\ell_{\infty,0}$ achieve the best and worst performance separately, and the performance of $\ell_{1,0}$ is moderate. When $\ell_{\infty,0}$ is adopted, the frame which is important for representing a few frames will be selected. However, $\ell_{2,0}$ and $\ell_{1,0}$ select the frame which has a contribution to the representation of all frames. In this sense, $\ell_{2,0}$ and $\ell_{1,0}$ can achieve better summarization performance than $\ell_{\infty,0}$. Therefore, we use $\ell_{2,0}$ in this paper.

3) IMPACT OF n_c

The performances of R-KSDS with different numbers of candidate keyframes (n_c) and KSDS are shown in Fig. 3, in which n_c varies from 2 to 30. It can be observed that when n_c is less than 20, the performance of R-KSDS is better than that of KSDS, and vice verse. This demonstrates that the backtracking strategy is effective and necessary. When n_c equals 4, R-KSDS achieves the best performance. In addition, there are local best performances when n_c equals 8, 14 and 24 respectively. Thus, four candidate keyframes are adopted in our experiments.

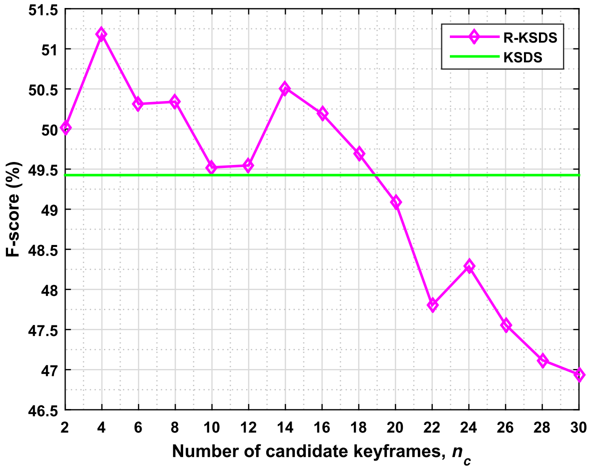

FIGURE 3. Performance comparison of different number of candidate keyframes in R-KSDS on VSUMM.

D. EVALUATIONS ON BENCHMARK DATASETS

1) EVALUATIONS ON VSUMM

The human-selected ground-truth keyframes of VSUMM are available on the VSUMM official website. When calculating the quantitative metrics, the average of the results among 5 ground-truth sets of keyframes is adopted. Our proposed KSDS algorithms including KSDS and R-KSDS are compared with various kinds of methods.

(i) *The Open Video Project storyboard* (OVP) [42] from service provider;

(ii) *Clustering based*: including Delaunay Triangulation clustering (DT) [9], STIMO [10] and VSUMM [3];

(iii) *Shot segmentation based*: including Multidimensional Time Series Analysis (MTSA) [45] and Fast Shot Segmentation (FSS) [46];

(iv) *Sparse representation based*: including Sparse Modeling Representative Selection (SMRS) [47], SOMP [24], Minimum Sparse Reconstruction (MSR) [11], Adaptive Greedy Dictionary Selection (AGDS) [25], Structured Sparse Dictionary Selection (SSDS) [48], and $\ell_{2,1}$ norm based Nonlinear Sparse Modeling ($\ell_{2,1}$ NSM) [49].

In all the sparse representation based methods, the 360 dimensional feature designed in [21] is adopted, which includes both structure and color information. In our KSDS and R-KSDS, the sparsity upper boundary K_0 is set to 13, and ER_{thr} is set to 12%.

According to the definition of the three quantitative metrics, *Precision* reveals the ability to select matched keyframes over all the selected keyframes, while *Recall* reflects the ability to select matched keyframes over all ground truth keyframes, and *F-score* balances these two metrics and evaluates overall performance of summarizing videos. The comparison on VSUMM is shown in Table 2. It is obviously observed that our proposed KSDS based algorithms achieve

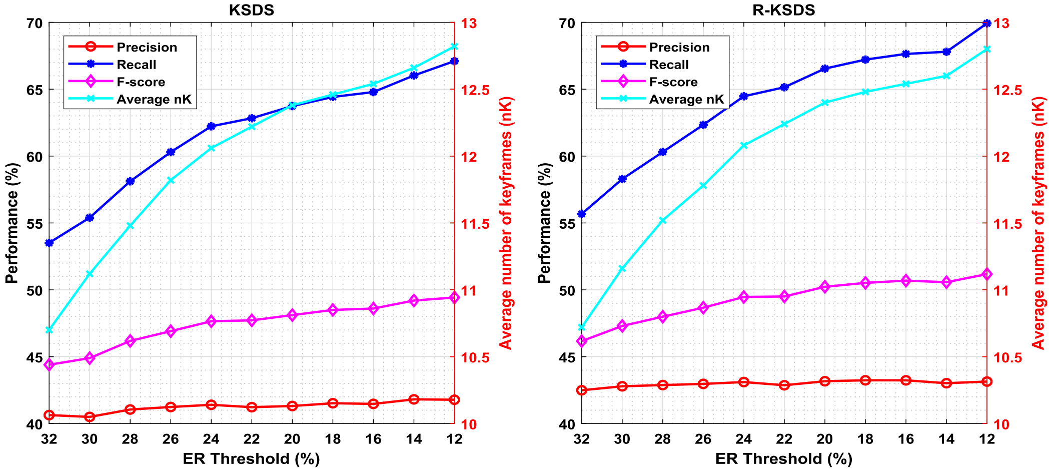

FIGURE 4. The quantitative performance of proposed KSDS based algorithms with different ER thresholds on VSUMM dataset, the left is KSDS and the right is R-KSDS.

TABLE 2. Performance comparison on VSUMM.

Algorithms	Precision(%)	Recall(%)	F-score(%)	Average nK
OVP [42]	44.58	50.72	44.81	9.7
DT [9]	38.06	28.73	31.64	6.2
STIMO [10]	36.65	42.88	37.95	10.0
VSUMM [3]	**47.44**	42.63	43.75	7.7
MTSA [45]	40.46	55.54	44.26	11.8
FSS [46]	40.57	54.76	44.41	12.4
SMRS [47]	38.10	55.83	43.82	12.3
MSR [11]	41.42	53.82	44.98	11.0
SOMP [24]	40.92	56.58	45.46	12.0
AGDS [25]	38.06	63.48	45.54	13.0
SSDS [48]	40.97	63.51	47.71	12.0
$\ell_{2,1}$ NSM [49]	39.88	64.86	47.33	13.0
KSDS	41.78	67.11	49.43	12.8
R-KSDS	43.14	**69.93**	**51.18**	12.8

the two best performances among all the compared algorithms. The performance of our R-KSDS is better than that of standard version KSDS, so it is necessary and useful to adopt the backtracking mechanism.

Generally, the more keyframes are selected, the more ground truth summaries are matched, thus, Recall can be higher. According to the average number of keyframes (nK) shown in Table 2, shot segmentation and sparse representation based algorithms select more keyframes than others, so they may have more matched keyframes and achieve a higher Recall. Though VSUMM has the highest precision, it could miss a certain number of keyframes due to the limitation of the number of selected frames, which leads to its lower Recall and F-score. Especially, our KSDS and R-KSDS have achieved the two best performances on Recall, which demonstrates that most of the keyframes in the ground truth summaries have been selected by our algorithms. Meanwhile, the Precisions of our algorithms are only lower than VSUMM and OVP, so only a few keyframes are not matched with the ground truthes.

The results of our proposed KSDS based VS algorithms with different ER thresholds on VSUMM are shown in Fig. 4. It can be observed that when the ER threshold decreases, more keyframes are selected. Recall increases gradually, reflecting more keyframes are matched with the ground truth summaries, and Precision basically maintain a stable level, both of them indicate the stable efficiency and capacity of our proposed KSDS based algorithms. The stable Precision and gradually increasing Recall result in F-score increasing gradually. As demonstrated in Fig. 4, when $ER_{thr} = 32\%$, the number of selected keyframes is smaller than most algorithms and a little greater than OVP and STIMO, however our proposed R-KSDS algorithm still outperforms all compared algorithms, which indicates that our proposed KSDS based VS algorithms can provide very robust and effective summarization results.

2) EVALUATIONS ON TVSum

The human-selected ground-truth importance scores of TVSum are available as part of Yahoo! WebScope program [50], and the average importance score of 20 users is adopted for evaluation. In the representative skim based VS [43], [51], 15% of the video length is adopted to limit the length of summarized skims, therefore, we also use 15% of the average number of uniform shots (approx. 18) as the limitation to the number of keyframes. Specifically, the limitation varies from 5 to 25. In TVSum dataset, our algorithms are only compared with sparse representation based methods.

The experimental results of our proposed KSDS based algorithms, together with other sparse representation based algorithms, are shown in Table 3. It can be observed that our proposed KSDS and R-KSDS achieve the two best performances. All the scores vary between 1.7 and 2.1, and the reason is that the distribution of importance scores is skewed toward low values. Specifically, target ranges were defined before making ground truth for each score assignment [43]: more than 65% of the shots got the low importance scores of 1 or 2, and only a few shots got high scores.

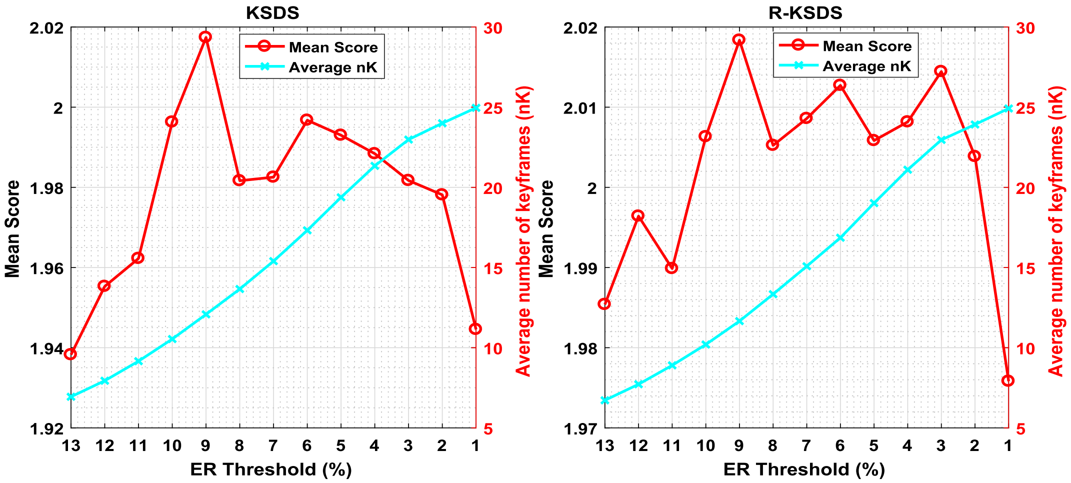

FIGURE 5. The quantitative performance of the proposed KSDS based algorithms with different ER thresholds on TVSum dataset, the left is KSDS and the right is R-KSDS.

TABLE 3. Performance comparison on TVSum.

Algorithms	nK=5	nK=10	nK=15	nK=20	nK=25
MSR [11]	1.892	1.895	1.837	1.784	1.748
SOMP [24]	1.951	1.955	1.952	1.930	1.917
AGDS [25]	1.910	1.935	1.961	1.943	1.910
SMRS [47]	2.009	1.995	1.965	1.956	—
SSDS [48]	1.887	1.870	1.799	1.734	1.699
$\ell_{2,1}$ NSM [49]	1.757	1.787	1.836	1.832	1.835
KSDS	1.995	2.004	2.003	1.970	1.945
R-KSDS	**2.007**	**2.034**	**2.038**	**2.004**	**1.975**

According to [43], the shots which are most representative of video content were most likely to get high importance scores. Our proposed algorithms perform best no matter how many keyframes are selected, which indicates the keyframes summarized by our algorithms are more representative of video contents.

According to the definition of mean score in (23), if more than one keyframes belong to the same shot, the scores of the keyframes will be penalized. Coupled with the skewed low scores of shots, so there could exist a threshold on the number of selected keyframes. When the number of keyframes is more than this threshold, the performance will drop due to the redundancy in the same shot and the difficulty of achieving high scores. It can also be verified by the experimental results in Table 3, when the number of key frames increases excessively, the performances of almost all algorithms have degraded.

The results of our proposed KSDS based VS algorithms with different ER thresholds on TVSum are shown in Fig. 5. In this experiment, K_0 is set to 25 and ER_{thr} varies from 13% to 1%. It can also be observed that when the ER threshold decreases, the number of selected keyframes increases monotonically. However, the change of the performance in term of the mean score is irregular, because there is no clear relationship between the importance score and the number of selected keyframes, which means that the importance score of the m-th keyframe may be higher or lower than the mean score of the previous $(m-1)$ keyframes. As explained earlier, there exists a threshold on the number of keyframes. It can be verified from Fig. 5 that when the number of keyframes is greater than 17 (ER threshold equals 6%), the performance of KSDS degrades gradually. For R-KSDS, the threshold is about 22 (ER threshold equals 3%), and the performance degrades more quickly than the standard KSDS.

VI. CONCLUSION

In this paper, we focus on the SDS based video summarization, and present a $\ell_{2,0}$ norm based nonlinear SDS model to take the nonlinearity among video frames into account. Moreover, two optimization algorithms, i.e., a standard KSDS algorithm and a robust KSDS algorithm are developed. In addition, the energy ratio (ER) of residual to video in the projected space is devised to adaptively produce summaries with different lengths to achieve an intuitive and flexible configuration of the VS process. Experimental results on the VSUMM dataset and the TVSum dataset demonstrate quantitatively and qualitatively that our nonlinear KSDS based VS approaches outperform the state-of-the-art VS algorithms and are flexible to generate keyframe based summaries with different lengths.

REFERENCES

[1] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury, "Context-aware surveillance video summarization," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5469–5478, Nov. 2016.

[2] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 3, no. 1, 2007, Art. no. 3.

[3] S. E. F. de Avila, A. P. B. Lopes, A. da Luz, Jr., and A. de Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 56–68, 2011.

[4] K. Fujimura, K. Honda, and K. Uehara, "Automatic video summarization by using color and utterance information," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2002, pp. 49–52.

[5] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2178–2190, Dec. 2010.

[6] B. Li and M. I. Sezan, "Event detection and summarization in sports video," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries*, Dec. 2001, pp. 132–138.

[7] C. Bi *et al.*, "Dynamic mode decomposition based video shot detection," *IEEE Access*, vol. 6, pp. 21397–21407, 2018.

[8] S. Tippaya, S. Sitjongsataporn, T. Tan, M. M. Khan, and K. Chamnongthai, "Multi-modal visual features-based video shot boundary detection," *IEEE Access*, vol. 5, pp. 12563–12575, 2017.

[9] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *Int. J. Digit. Libraries*, vol. 6, no. 2, pp. 219–232, 2006.

[10] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: Still and moving video storyboard for the Web scenario," *Multimedia Tools Appl.*, vol. 46, no. 1, pp. 47–69, 2010.

[11] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video summarization via minimum sparse reconstruction," *Pattern Recognit.*, vol. 48, no. 2, pp. 522–533, Feb. 2015.

[12] W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1996, pp. 1228–1231.

[13] E. Mendi, H. B. Clemente, and C. Bayrak, "Sports video summarization based on motion analysis," *Comput. Elect. Eng.*, vol. 39, no. 3, pp. 790–796, 2013.

[14] Y. Zhang, R. Tao, and Y. Wang, "Motion-state-adaptive video summarization via spatiotemporal analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1340–1352, Jun. 2017.

[15] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[16] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[17] L. Jia *et al.*, "Image denoising via sparse representation over grouped dictionaries with adaptive atom size," *IEEE Access*, vol. 5, pp. 22514–22529, 2017.

[18] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face and person recognition from unconstrained video," *IEEE Access*, vol. 3, pp. 1783–1798, 2015.

[19] L. Zhao, Q. Sun, and Z. Zhang, "Single image super-resolution based on deep learning features and dictionary model," *IEEE Access*, vol. 5, pp. 17126–17135, 2017.

[20] M. Kumar and A. C. Loui, "Key frame extraction from consumer videos using sparse representation," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2437–2440.

[21] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66–75, Feb. 2012.

[22] H. Liu, Y. Liu, Y. Yu, and F. Sun, "Diversified key-frame selection using structured $L_{2,0}$ optimization," *IEEE Trans Ind. Informat.*, vol. 10, no. 3, pp. 1736–1745, Aug. 2014.

[23] P. Etezadifar and H. Farsi, "Scalable video summarization via sparse dictionary learning and selection simultaneously," *Multimedia Tools Appl.*, vol. 76, no. 6, pp. 7947–7971, 2017.

[24] S. Mei, G. Guan, Z. Wang, M. He, X.-S. Hua, and D. D. Feng, "$L_{2,0}$ constrained sparse dictionary selection for video summarization," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.

[25] Y. Cong, J. Liu, G. Sun, Q. You, Y. Li, and J. Luo, "Adaptive greedy dictionary selection for Web media summarization," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 185–195, Jan. 2017.

[26] M. Ma, S. Mei, J. Hou, S. Wan, Z. Wang, and D. Feng, "Nonlinear kernel sparse dictionary selection for video summarization," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2017, pp. 637–642.

[27] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 766–782.

[28] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2982–2991.

[29] L. Gao, P. Wang, J. Song, Z. Huang, J. Shao, and H. T. Shen, "Event video mashup: From hundreds of videos to minutes of skeleton," in *Proc. AAAI*, 2017, pp. 1323–1330.

[30] Z. Ji, K. Xiong, Y. Pang, and X. Li. (2017). "Video summarization with attention-based encoder-decoder networks." [Online]. Available: https://arxiv.org/abs/1708.09545

[31] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artif. Neural Netw.*, 1997, pp. 583–588.

[32] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Kernel sparse representation for image classification and face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

[33] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[34] Y. Nesterov, "Gradient methods for minimizing composite functions," *Math. Program.*, vol. 140, no. 1, pp. 125–161, 2013.

[35] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[36] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

[37] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, Mar. 2006.

[38] E. van den Berg and M. P. Friedlander, "Theoretical and empirical results for recovery from multiple measurements," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2516–2527, May 2010.

[39] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst. Comput.*, 1993, pp. 40–44.

[40] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[41] G. Guan, Z. Wang, S. Lu, J. D. Deng, and D. D. Feng, "Keypoint-based keyframe selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 729–734, Apr. 2013.

[42] (2011). *Open Video Project*. [Online]. Available: http://www.open-video.org/

[43] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing Web videos using titles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5179–5187.

[44] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. ACM Int. Workshop Multimedia Inf. Retr.*, 2006, pp. 321–330.

[45] Z. Gao, G. Lu, and P. Yan, "Key-frame selection for video summarization: An approach of multidimensional time series analysis," *Multidimensional Syst. Signal Process.*, vol. 29, no. 4, pp. 1–21, 2017.

[46] E. Apostolidis and V. Mezaris, "Fast shot segmentation combining global and local visual descriptors," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 6583–6587.

[47] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1600–1607.

[48] H. Wang, Y. Kawahara, C. Weng, and J. Yuan, "Representative selection with structured sparsity," *Pattern Recognit.*, vol. 63, pp. 268–278, Mar. 2017.

[49] F. Dornaika and I. K. Aldine, "Instance selection using nonlinear sparse modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1457–1461, Jun. 2018.

[50] *Yahoo! Webscope Program*. Accessed: 2015. [Online]. Available: http://webscope.sandbox.yahoo.com/

[51] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 505–520.

MINGYANG MA received the B.S. degree in communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2015, where he is currently pursuing the Ph.D. degree in information and communication engineering. His main research interests include image processing and video summarization.

SHAOHUI MEI received the B.S. degree in electronics and information engineering and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2005 and 2011, respectively, where he is currently an Associate Professor with the School of Electronics and Information. He was a Visiting Student with The University of Sydney, from 2007 to 2008. His research interests include hyperspectral remote sensing image processing, deep learning, video processing, and pattern recognition.

SHUAI WAN received the B.E. degree in telecommunication engineering and the M.E. degree in communication and information system from Xidian University, Xi'an, China, in 2001 and 2004, respectively, and the Ph.D. degree in electronics engineering from the Queen Mary University of London, in 2007. She is currently a Professor with Northwestern Polytechnical University, Xi'an, and an Adjunct Professor with RMIT University, Australia. Her research interests include advanced video coding, video quality assessment, and hyperspectral image compression.

ZHIYONG WANG received the B.Eng. and M.Eng. degrees in electronic engineering from the South China University of Technology, Guangzhou, China, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong. He is currently an Associate Professor with the School of Information Technologies, The University of Sydney, Australia. His research interests focus on multimedia computing, including multimedia information processing, retrieval and management, Internet-based multimedia data mining, human-centred multimedia computing, and pattern recognition.

DAGAN FENG received the M.E. degree in electrical engineering and computing science from Shanghai JiaoTong University, in 1982, and the M.Sc. degree in biocybernetics and the Ph.D. degree in computer science from the University of California at Los Angeles, Los Angeles, in 1985 and 1988, respectively. He was an Assistant Professor with the University of California at Riverside, Riverside. He joined The University of Sydney as a Lecturer, in 1988, where he is currently a Professor with the Department of Computer Science and the Head of the School of Information Technologies. He is also an Honorary Research Consultant with Royal Prince Alfred Hospital, Sydney, the Chair Professor of information technology with The Hong Kong Polytechnic University, an Advisory Professor with Shanghai JiaoTong University, and a Guest Professor with Northwestern Polytechnic University, Northeastern University, and Tsinghua University.