Multimed Tools Appl (2011) 51:1127–1145
DOI 10.1007/s11042-010-0462-x

A unified scheme of shot boundary detection and anchor shot detection in news video story parsing

Hansung Lee · Jaehak Yu · Younghee Im · Joon-Min Gil · Daihee Park

Published online: 19 January 2010

Abstract In this paper, we propose an efficient one-pass algorithm for shot boundary detection and a cost-effective anchor shot detection method with search space reduction, which are unified scheme in news video story parsing. First, we present the desired requirements for shot boundary detection from the perspective of news video story parsing, and propose a new shot boundary detection method, based on singular value decomposition, and a newly developed algorithm, viz., *Kernel-ART*, which meets all of these requirements. Second, we propose a new anchor shot detection system, viz., *MASD*, which is able to detect anchor person cost-effectively by reducing the search space. It consists of skin color detector, face detector, and support vector data descriptions with non-negative matrix factorization sequentially. The experimental results with the qualitative analysis illustrate the efficiency of the proposed method.

Keywords News video parsing · Shot boundary detection · Anchor shot detection · Adaptive resonance theory · Support vector data description

H. Lee
Electronics and Telecommunications Research Institute,
138, Gajeongno, Yueseong-gu, Daejeon, Republic of Korea
e-mail: mohan@etri.re.kr

J. Yu · Y. Im · D. Park (✉)
Department of Computer and Information Science, Korea University,
208, Seochang-Ri, Chochiwon, Chungnam, 339-700, South Korea
e-mail: dhpark@korea.ac.kr

J.-M. Gil
Department of Computer Science Education, Catholic University of Daegu,
330 Geumnak, Hayang-eup, Gyeongsan-si, Gyeongbuk 712-702, South Korea

1 Introduction

At present, video materials and video services are more available than ever. In particular, with the growing popularity of digital news video, the numbers of collections of news video databases have recently exploded. Consequently, news videos are valuable to data analysts, information providers, and TV consumers, because of their information richness [12]. Therefore, news video databases were the subject of extensive research over the past decade, to develop effective and efficient tools for manipulation and analysis of news videos. An important step towards effective news video indexing and retrieval is news video story parsing, which partitions a news video into stories. This process generally involves three steps: news shot boundary detection, anchor shot detection, and news story segmentation [4, 5, 9, 16].

News shot boundary detection is the process of dividing a news video into shots by detecting transition boundaries between shots. The transitions between shots can be mainly classified into two types: abrupt cuts and gradual transitions. Gradual transitions can be further subdivided into fade-ins, fade-outs, and dissolves according to the characteristics of the different editing effects [26]. According to the recent literature, a large number of methods have been proposed for detecting shot boundaries [2, 3, 6, 10, 11, 19, 25]. Z. Cernekova et al. [3] proposed a shot boundary detection method based on mutual information (MI) and joint entropy (JE). They used MI and JE for detecting abrupt cuts and gradual transitions, respectively. X. Ling et al. [19] suggested a three-step shot boundary detection method based on SVM, which consists of three modules and operates sequentially. It firstly abstracts the reordered frame sequence (RFS) and detects abrupt cuts using SVM, and then detects gradual transitions by temporal multi-resolution analysis of RFS. M. Cooper et al. [6] presented a shot boundary detection method using supervised classification. It combines binary kNNs for detecting abrupt cuts and gradual transitions. In the first step, abrupt cut boundaries are detected and then, the non-abrupt cut frames are classified as either gradual transition frames or non-transition (normal) frames. In summary, most recently developed methods have focused on general purpose shot boundary detection [2, 3, 6, 10, 19], and a few of them were designed for news video story parsing [11]. Besides, most existing studies are detecting abrupt cut and gradual transition separately and they are made up of two-pass structure [3, 11] that requires scanning dataset twice for shot boundary detection or multi-step structure [6, 19].

Anchor shot detection is the process of finding shots which contain the anchor person. Concerning anchor shot detection, in general, there are two predominant research paradigms [21]: 1) The model (template) matching method and 2) The unsupervised (clustering) method. The former defines a set of predefined models for an anchor shot, then, matches them against all shots in a news video, in order to detect potential anchor shots. A. Hanjalic et al. [15] proposed an approach for template-based anchor shot detection by using sequence-own video shots as the template for detecting anchor shots of that sequence. However, it is infeasible to construct a generic model which can represent all types of news, because of the large variety of existing news programs. On the other hand, the latter constructs an unsupervised anchor shot model, or aggregates shots with similar visual content that frequently occur throughout the news video, using a clustering algorithm. X. Luan et al. [20] proposed an anchor shot detection method based on clustering, viz., AnchorClu. After fast clustering of a shot's key frames, some clusters

including the anchorperson cluster can be obtained. Then via some proper rules, the anchorperson cluster is selected. M. Santo et al. [21] also proposed an anchor shot detection method based on clustering. This algorithm firstly uses a clustering method for candidate anchor shots and then employs a two-stage pruning technique for reducing the number of falsely detected anchor shots. Both clustering and pruning are performed in an unsupervised manner. However, according to the assumptions used, there are inevitably corresponding drawbacks, for example a shot with a reporter (not an anchorperson) can be erroneously recognized as an anchor shot, an interview shot can result in a falsely detected anchor shot, etc.

Thus far, we have addressed problems associated with shot boundary detection and anchor shot detection in terms of news video story parsing. In general, shot boundary detection methods and anchor shot detection methods have been developed independently, although these are tightly related in news video story parsing. In this paper, we propose an efficient one-pass shot boundary detection algorithm and a cost-effective anchor shot detection method, which are unified scheme in news video story parsing.

First, we present the desired requirements for shot boundary detection from the perspective of news video story parsing, and suggest a new shot boundary detection method which meets all of these requirements as follows: 1) Detecting abrupt cuts and gradual transitions using a single algorithm so as to divide a news video into shots with a single scan of the dataset; 2) Additionally determining shots into static (a shot with anchor(s)) or dynamic (a shot without anchor(s)), therefore, reducing the search space for the subsequent stage of anchor shot detection; 3) Minimizing the incorrect data in the dataset for anchor shot detection by emphasizing of the *recall* ratio. The proposed method, based on singular value decomposition (SVD) and the incremental clustering with mercer kernel, has additional desirable features. 4) By applying SVD, noises or trivial variations in the video sequence are removed; 5) The mercer kernel improves the probability of detection of shots which are not separable in input space by mapping data onto a high dimensional feature space.

Second, we propose a new anchor shot detection system, viz., Multi-phase Anchor Shot Detection (*MASD*), which consists of three modules and operates sequentially: 1) A skin color detection module for reducing candidate face regions; 2) A face detection module for finding key-frames with face data; 3) A One-class Support Vector Machine (OSVM) module for determining anchor shots using Support Vector Data Description (SVDD) and Non-negative Matrix Factorization (NMF). It achieves a high speed by greatly reducing the search space, and a high accuracy by SVDD using NMF.

The rest of this paper is organized as follows. In Section 2, we present an efficient one-pass algorithm for shot boundary detection. A cost-effective algorithm for anchor shot detection is proposed in Section 3. In Section 4, we perform simulations to check the possible validity of our approach. Finally, in Section 5, we conclude with a brief summary and suggest future research directions.

2 News video shot boundary detection

In this section, we introduce a one-pass news video shot boundary detection method, based on SVD, and a newly developed algorithm (*Kernel-ART*) with many desirable features. It satisfies all of the desired requirements for the shot boundary detection

method from the perspective of news video story parsing, as we mentioned in Section 1.

2.1 Theoretical background for news video shot boundary detection

In this paper, color histograms are chosen as the raw feature vector used to represent video frames. The three-dimensional histograms in the RGB color space with 16 respective bins for R, G, and B, are calculated. To incorporate the spatial information of the color distribution, we divide each image into 2×2 blocks. Thus, the dimensionality of the feature vector is $m = 4 \times 16^3 = 16{,}384$. With the feature vector of frame i as i-th column, the $m \times n$ feature-frame matrix A for the video sequence f_i; $i = 1, 2, \cdots, n$ is constructed.

The SVD is now applied to the aforementioned feature-frame matrix A. By choosing the largest k singular values, the m-dimensional input space is mapped onto the reduced k-dimensional refined feature space. Concerning SVD, we summarize the following two important properties that have been widely used for indexing, clustering, and retrieval in the context of text and image processing areas. These theorems are later used in the construction of our system. The complete proof can be found in [13] and [14].

Theorem 1 [13] *For $A \in R^{m \times n}$, if $k < r = rank(A)$ and*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T, \tag{1}$$

then

$$\min_{rank(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1} \tag{2}$$

Theorem 1 has significant implications. Discarding small singular values means removing the linearly semi-dependent or non-essential axes of the feature space. That is, the truncated SVD still captures most of the important underlying structure in the association of histograms and video frames, without losing important information. As a result, the removal of noises and trivial variations in video frames can be achieved.

Theorem 2 [14] *Let $A = [A_1 \cdots A_i \cdots A_n]$, $V^T = [\psi_1 \cdots \psi_i \cdots \psi_n]$. Define the distance of ψ_i to the origin of the refined feature space as:*

$$\|\psi_i\| = \sqrt{\sum_{j=1}^{rank(A)} v_{ij}^2} \tag{3}$$

If $rank(A) = n$, then, from the orthonormal property of matrix V, we have $\|\psi_i\|^2 = 1$, where $i = 1, 2, \cdots n$. Let $A' = [A_1 \cdots A_i^{(1)} \cdots A_i^{(k)} \cdots A_n]$ be the matrix obtained by duplicating column vector A_i in A k times ($A_i^{(1)} = \cdots = A_i^{(k)} = A_i$), and $V'^T = [\psi_1' \cdots \phi_1' \cdots \phi_k' \cdots \psi_n']$ be the corresponding right singular vector matrix obtained from SVD. Then, $\|\phi_j'\|^2 = 1/k$, where $j = 1, 2, \cdots, k$.

Theorem 2 provides the theoretical basis for search space reduction for the subsequent stage of anchor shot detection. A linearly-independent column vector A_i of matrix A will be projected onto the vector ψ_i whose distance defined by Eq. 3 is the one in the refined feature space. While, if A_i has some duplicates $A_i^{(j)}$, the distance of its projected vector ϕ_i' decreases. In the context of the news video domain, this can be interpreted as the fact that in the refined feature space, frames in a static shot will be projected onto points closer to the origin, while frames in a dynamic shot will be projected onto points farther from origin.

2.2 New shot boundary detection algorithm: *Kernel-ART*

The valuable findings stated in the previous subsection help us discover distinct features of low level patterns, and then use this knowledge to develop our targeting system. Since frames with similar histogram patterns will be mapped onto near points, a shot is represented as a cluster in the refined feature space. In this paper, we interpreted news video shot boundary detection problem as clustering one. In the case of abrupt cut, as independent cluster is formed on input space, clustering is easy. However, in the case of gradual transition, as cluster has large dispersion, it is not easy to separate the border. Therefore, we adopted the mercer kernel in our algorithm which can improve the probability of detection of shots in the case of gradual transition which are not separable in input space by mapping data onto a high dimensional feature space.

To detect the shot boundary, the angle between row vectors ψ_i and ψ_j is used as a similarity measure:

$$S(\psi_i, \psi_j) = \cos(\psi_i, \psi_j) = \frac{< \psi_i, \psi_j >}{||\psi_i|| ||\psi_j||} \tag{4}$$

By Theorem 2, since the degree of visual changes in a shot is closely related to the location of its corresponding shot cluster in the refined feature space, we define the decision function classifying shots into static and dynamic, as follows:

$$D(S_i) = \frac{1}{n_i} \sum_{\psi_i \in S_i} ||\psi_i||^2 \tag{5}$$

where S_i is i-th video segment (shot) and n_i is the number of frames pertaining to S_i.

Here, we propose a *Kernel-ART* clustering algorithm, which combines Adaptive Resonance Theory (ART) and the mercer kernel, to detect a shot boundary effectively. Details of the proposed algorithm are as follows:

Initialization The number of clusters c is initially set to one. The first input pattern is assigned to its initial weight vector as follows:

$$w_1 = \psi_1 \tag{6}$$

This ensures that the first input pattern is assigned to the first shot cluster corresponding to any vigilance parameter $\rho \in [0, 1]$.

Activation Function The mercer kernel improves the linear separability of data which are not separable in input space, by mapping data onto high dimensional feature space [7]. By substituting the inner product in the similarity measure of Eq. 4

with the Radial Basis Function (RBF) kernel, we can obtain a similarity measure function in the feature space, and the activation function is defined as follows:

$$AF(\psi_i, w_j) = \exp\left\{-\frac{1}{c}||\psi_i - w_j||^2\right\} \tag{7}$$

where w_j is the mean vector of the cluster j.

Matching Function If the activation function $AF(\cdot)$ and the matching function $MF(\cdot)$ are chosen under condition (8), then the mismatch reset condition and the template matching process of the original ART is eliminated in the resonance domain [1].

$$MF(\psi_i, w_1) > MF(\psi_i, w_2) \Longleftrightarrow AF(\psi_i, w_1) > AF(\psi_i, w_2) \tag{8}$$

The simplest means to define the activation and matching functions under condition (8) is to set the activation function to equal the matching function.

Shot Boundary Condition According to the simple setting of the matching function, the shot boundary condition is selected as follows:

$$AF(\psi_i, w_j) \geq \rho \tag{9}$$

Regarding shot boundary detection, our concern is testing whether frame ψ_i pertains to the most recent cluster j formed prior to i-th time instant. If the shot boundary condition is not satisfied, the cluster j is classified as either static or dynamic using Eq. 5, then, a new cluster unit is created, $j = j + 1$, and the input pattern is assigned to it. Otherwise w_j is updated, subsequent to the inclusion of ψ_i.

The proposed algorithm reduces the search space for the subsequent stage of anchor shot detection, by using only static shots for inputs of anchor shot detection. It has two parameters for controlling the results. One is the vigilance parameter ρ affecting the support of clusters. The other is the RBF kernel-width parameter, c. So the proposed algorithm is flexible and we have the capability to control the results with a high *recall* ratio. It also detects the abrupt cuts and the gradual transitions using a single algorithm so as to divide news video into shots with a single scan of the dataset. By applying SVD, noises or trivial variations in the video sequence are removed. Therefore, the separability is improved. The mercer kernel improves the probability of detection of shots which are not separable in input space, by mapping data onto high dimensional feature space. Accordingly, the proposed algorithm meets all of the design requirements presented in the introduction. The proposed shot boundary detection algorithm is summarized in Table 1.

Complexity analysis of proposed *Kernel-ART* algorithm is as follows: Let's assume there are n frames, and s shot clusters. The first frame of input data is unconditionally assigned to the first shot cluster and in the case new shot cluster is generated as previous shot cluster is excluded from calculation, $n - 1$ activation function (Eq. 7) calculations and $n - 1$ shot boundary condition (Eq. 9) comparisons are required. As the first vector of all shot clusters does not need cluster mean calculation, $n - s$ cluster mean calculations are required. Therefore, for shot boundary detection, total of $O(n)$ activation function calculations, $O(n)$ shot boundary condition comparisons,

Table 1 The proposed shot boundary detection algorithm (*Kernel-ART*)

1. Initialize weights: $w_1 = \psi_1$

2. For each element of input data,

2.1. Compute activation function by Eq. 7:

$AF(\psi_i, w_j) = \exp\{-\frac{1}{c}||\psi_i - w_j||^2\}$

2.2. Test the shot boundary condition:

If $AF(\psi_i, w_j) \geq \rho$ then

2.2.1. Assign frame ψ_i to cluster j and update cluster median w_j

Else

2.2.2. Classify shot S_i as either static or dynamic by Eq. 5:

$D(S_i) = \frac{1}{n_i} \sum_{\psi_i \in S_i} ||\psi_i||^2$

2.2.3. Create a new cluster: $j = j + 1$

2.2.4. Initialize the new cluster: $w_j = \psi_i$

and $O(n)$ cluster mean calculations are required. Consequently, though n of frame increases, the amount of total calculation of algorithm increases with pseudo linear regarding n.

3 Anchor shot detection

The overall architecture of our newly proposed anchor shot detection system is given in Fig. 1, viz., Multi-phase Anchor Shot Detection (*MASD*), which consists of three modules and operates sequentially: 1) Skin color detection module for reducing candidate face regions; 2) Face detection module for finding key-frames with a facial data; 3) One-class SVM module for determining the anchor shots using a Support Vector Data Description (SVDD). Each module will be described in turn.

3.1 Skin color detection

To increase the probability of anchor shots in a candidate set, and reduce the search space for the face detection module in the next stage, we simplified the fast skin color detection algorithm [22] which eliminates key-frames of each shot without a skin color region and/or a region that is too large. The simple but cost-effective heuristic

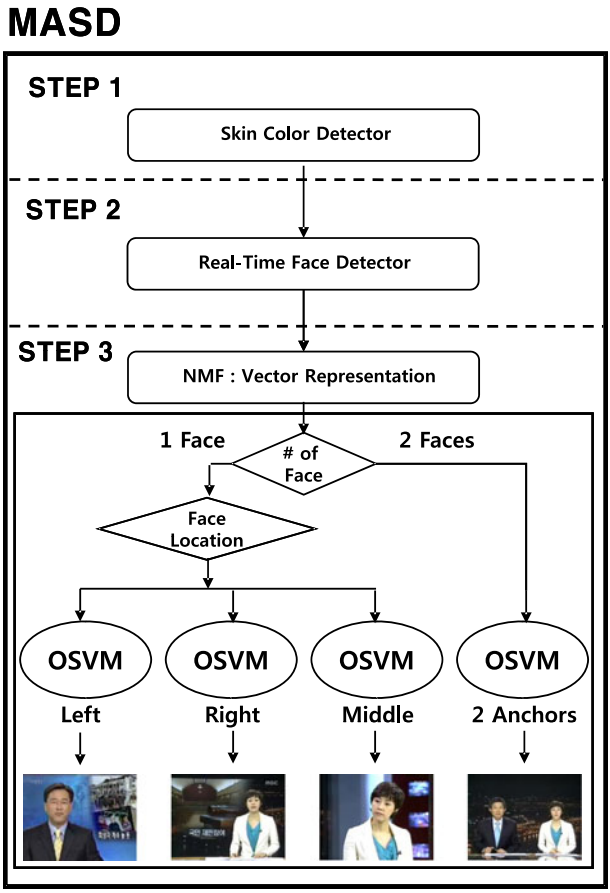

Fig. 1 The overall architecture of proposed anchor shot detection method

rule defined in Eq. 10 is highly appropriate here. Consequently, the reduced images with a skin color region are rapidly obtained.

$$\begin{aligned} &(R, G, B) \quad is \quad classified \quad as \quad skin \quad if: \\ &R > 95 \quad and \quad G > 40 \quad and \quad B > 20 \quad and \\ &R - G > 15 \quad and \quad R > B \end{aligned} \tag{10}$$

With regard to our skin color detection rule, six comparisons per pixel are conducted. If we assume that there exist n frames whose size is k pixels, it is necessary that $6k$ comparisons should be conducted in order to detect skin color regarding a frame. Finally, to detect skin color regarding all frames, $6kn$ comparisons are necessary. Consequently, to detect skin color, $O(n)$ simple relational comparison operations are required.

3.2 Fast face detection

With the resulting reduced key-frames obtained from the prior step, the main job of a real-time face detection module is to classify them into frames with facial data and frames with non-facial data. A key-frame without a face or with over three-faces is automatically discarded. Accordingly, the search space for the next stage will be considerably reduced. To meet our design requirements (i.e., speed and precision), we choose a real-time face detection model [24] which consists of several weak classifiers (Haar-like features) in a chain of cascading structure and a SVM in the last row. A series of weak classifiers incrementally generates candidate faces at a high speed but at the cost of a low precision. Since the false-negative ratio is high but the false-positive ratio is close to zero, they definitely have the capability to generate candidate faces with a low computational cost. On the other hand, a SVM in the last row does finally detect a face with a high precision. The overall architecture of the real-time face detection module is shown in Fig. 2.

3.3 One-class SVM

In the last stage of *MASD*, SVM is used to detect the anchor shot with high precision along with a vector representation based on NMF. The NMF provides part-based representations of dataset while SVD and Principal Component Analysis (PCA) provide holistic representations [18]. Therefore, the feature vectors with NMF have not only color histogram information in key-frames, but also spatial information of objects in key-frames. Accordingly, it is feasible to select NMF for representation of feature vectors.

The anchor shot detection may have data size differences according to type of anchor shot (the anchor shot has one of four spatial types as shown in Fig. 1). As a result, because of the unbalanced size of the training data, the learning result of one anchor shot type might be influenced by a different anchor shot type. Moreover, it is hard to conclude that the current training data represents the entire class, since anchor person and studio condition can sometimes be changed. Therefore, it is probable that the binary classifier SVM results in misclassification of the new training data by creating a decision boundary including an unobserved area. Accordingly, it is preferable to select a decision boundary function that uses a One-class SVM (OSVM) (one of the most well-known types of OSVM is the Support Vector Data Description (SVDD [23])) that independently expresses the corresponding class.

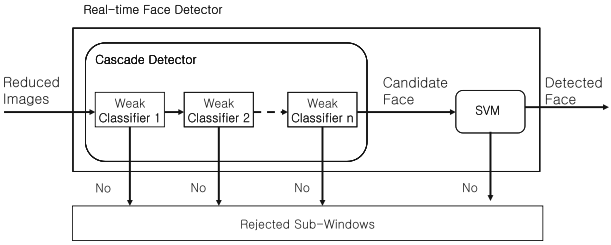

Fig. 2 A cascading face detector

In this paper, we trained total four classifiers that are applicable to each anchor shot type by using SVDDs (see Fig. 1), and final anchor shot is determined by selecting one of four SVDDs through the number of faces and center point value of faces that have been detected in face detection stage with regard to key-frame that represents shot.

4 Experiments

To evaluate the performance of our proposed system, we collected a dataset from news programs of the two leading TV stations in Korea; Korean Broadcasting System (KBS) and Munwha Broadcasting Corporation (MBC). The ground truth is manually pre-labeled. The experiment of this paper has been conducted in PC of Intel(R) Core(TM)2 Duo CPU E8400 @ 3.00GHZ specification and all algorithms have been realized by using Matlab. We use the standard *precision* and *recall* criteria as the evaluation measure. As a single figure of merit for comparing different algorithms, the so-called F-measure [4] combining *precision* and *recall* is also used:

$$F = \frac{2 \times recall \times precision}{(recall + precision)} \tag{11}$$

4.1 Performance evaluations of the proposed shot boundary detection

The rank parameter k of SVD is set to 50, and we constructed a 50-dimensional refined induced feature space. Here, the width parameter of mercer kernel c is fixed to 0.005, and parameter ρ under the shot boundary condition is optimized experimentally in its dataset. The total dataset consists of ten new video datasets including static shots and dynamic shots. In addition, each dataset is also divided into abrupt cut and gradual transition, which include fade-in, fade-out, and dissolve. The dataset is described in detail in Table 2.

We apply SVD to the first dataset consisting of 150 shots, and plot the portion of the dataset in the refined subspace, as shown in Fig. 3. The data distributions of static and dynamic shots in the refined subspace of the first and second dimension are given. As shown in Fig. 3, it is clear that static shots are projected onto data points with a small dispersion, while dynamic shots are projected onto data points

Table 2 The dataset for performance evaluation of the proposed shot boundary detection method

Datasets	Total shots	Shots with abrupt cuts	Shots with gradual transitions	Static shots	Dynamic shots
D1	150	105	45	63	87
D2	145	117	28	69	76
D3	148	132	16	67	81
D4	147	115	32	63	84
D5	152	124	28	75	77
D6	153	130	23	59	94
D7	155	127	28	81	74
D8	144	115	29	58	86
D9	143	121	22	72	71
D10	154	132	22	96	58

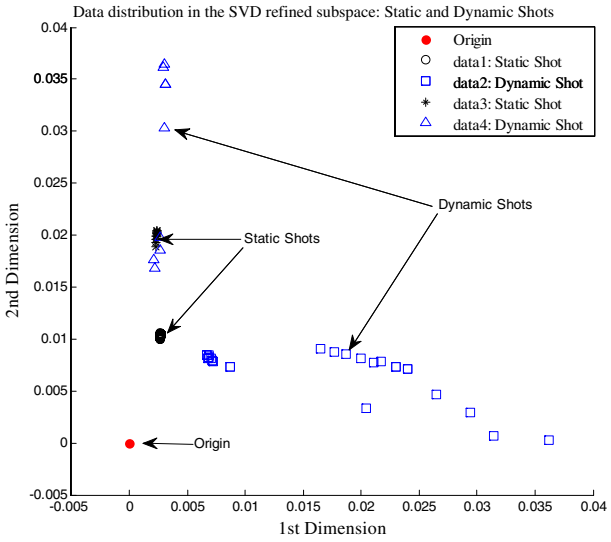

Fig. 3 Data distributions of static and dynamic shots in the refined subspace derived by SVD

(a) A typical static shot (anchor shot)

(b) A dynamic shot which involves a large camera motion

(c) A static shot with a small change of object motions

(d) A dynamic shot with a large change of object motions

Fig. 4 The examples of the frame sequences used in Fig. 3

Table 3 The experimental evaluation of the proposed shot boundary detection algorithm (*Kernel-ART*)

Datasets	Abrupt cuts			Gradual transitions			Overall performance		
	P	R	F	P	R	F	P	R	F
D1	91.9	97.1	94.4	95.7	97.8	96.7	93.0	97.3	95.1
D2	95.1	100	97.5	96.2	89.3	92.6	95.3	97.9	96.6
D3	92.2	98.5	95.2	93.8	93.8	93.8	92.4	98.0	95.1
D4	95.0	98.3	96.6	100	93.8	96.8	96.0	97.3	96.6
D5	90.4	99.2	94.6	96.6	100	98.3	91.5	99.3	95.2
D6	89.6	99.2	94.2	100	100	100	91.0	99.3	95.0
D7	91.9	98.4	95.0	100	96.4	98.2	93.3	98.1	95.6
D8	96.6	99.1	97.8	100	96.6	98.3	97.3	98.6	97.9
D9	84.5	99.2	91.3	100	95.5	97.7	86.5	98.6	92.2
D10	96.4	100	98.2	100	100	100	96.9	100	98.4
Average	92.36	98.90	95.48	98.23	96.32	97.24	93.32	98.44	95.78

P precision, *R* recall, F *F*-measure

with a large dispersion. It is also clear that frames in a static shot are projected onto data points closer to the origin, whereas, frames in a dynamic shot are projected onto data points farther from the origin. As a result, static shots and dynamic shots can be easily classified by Theorem 2 and Eq. 5 of *Kernel-ART*. By using only static shots as inputs in anchor shot detection, the search space in anchor shot detection can be

Table 4 Summary of quantitative/qualitative analysis for the shot boundary detection

	Gao et al. [11]	Cernekova et al. [3]	Fang et al. [8]	Proposed method
Data size	Total: 3,893 CUTs: NA GTs: NA	Total: 3,315 CUTs: 3,045 GTs: 270	Total: 284 CUTs: NA GTs: NA	Total: 1491 CUTs: 1218 GTs: 273
Precision	Overall: 98.07 CUTs: NA GTs: NA	Overall: NA CUTs: 98 GTs: 93.75	Over all: 96.07 CUTs: NA GTs: NA	Over all: 93.32 CUTs: 92.36 GTs: 98.23
Recall	Overall: 96.48 CUTs: NA GTs: NA	Overall: NA CUTs: 97.38 GTs: 98.12	Over all: 98.97 CUTs: NA GTs: NA	Over all: 98.44 CUTs: 98.90 GTs: 96.32
Used data	Private	Private & TRECVID2003	Charleton Univ. data	Private
Application domain	News video	General	General	News video
Methods	Fuzzy classifier	Mutual information & joint entropy	Fuzzy logic & C4.5	Incremental clustering
Strategy for CUT & GT	Multi-step	Multi-step	Multi-step	Single algorithm
Unified scheme of SBD & ASD	No	No	No	Yes
Search space reduction for ASD	No	No	No	Yes

CUT abrupt cut, *GT* gradual transition, *SBD* shot boundary detection, *ASD* anchor shot detection

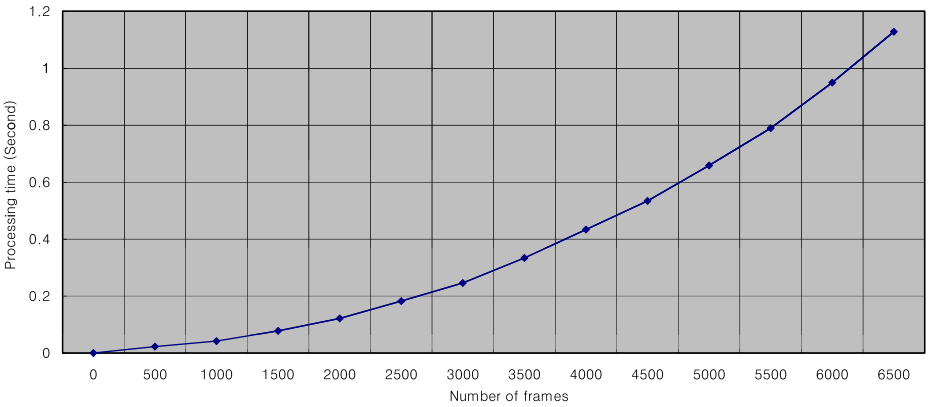

Fig. 5 A number of input frames versus processing time of *Kernel-ART* clustering

greatly reduced. This satisfies the second of the design requirements presented in the introduction. The example of the frame sequences used in Fig. 3 is presented in Fig. 4. Figure 4a is the anchor shot, which is a typical static shot, and Fig. 4b is a dynamic shot, which involves a large camera motion. A static shot with a small change of object motions and a dynamic shot with a large change of object motions are shown in Fig. 4c and d, respectively.

In Table 3, we summarize the experimental results of our proposed shot boundary detection algorithm, *Kernel-ART*. This indicates that our system detects almost all shot boundaries (the average *recall* is 98.44%), although it sometimes misclassifies a shot as two shots (the average *precision* is 93.32%). In this paper, we designed the shot boundary detection algorithm as part of a news video story parsing system. Since missing shots have a negative impact on later anchor shot detection, news video story parsing may ultimately have unfortunate consequences. Therefore, the high *recall* ratio is an important factor of overall system performance, and is very valuable. This high *recall* ratio of our proposed shot boundary detection algorithm satisfies the third requirement of the aforementioned design requirements. Equation 5, which is based on theorem 2, has the capability to greatly reduce the search space in the anchor shot detection step. We obtain 63.89% on the average as inputs into the anchor shot detection step, and eliminate an average 36.11% of the shots which is classified as dynamic shot. The *recall* rate of static shots is 100%, which means that our method does not miss any static shots and all of the anchor shots will be used as inputs in the anchor shot detection step. Also, we summarize the quantitative/qualitative analysis with existing methodologies in Table 4.

Table 5 The dataset for performance evaluation of the proposed anchor shot detection method

Datasets	Total shots	Anchor shots	Reporter shots	Interview shots
D1	624	25	13	42
D2	602	30	30	28
D3	897	42	12	90
D4	718	41	18	62
D5	517	43	10	46

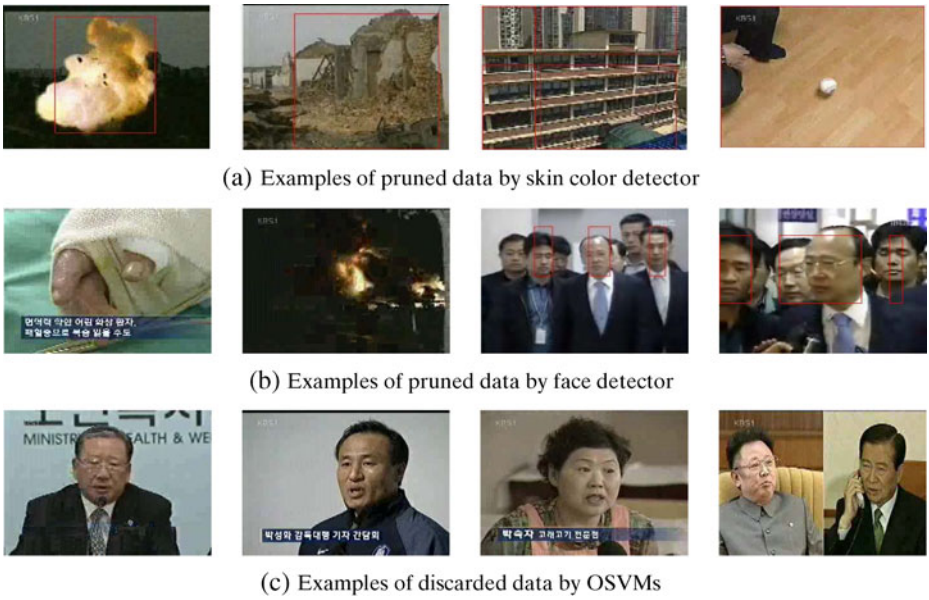

Fig. 6 A snapshot of the proposed anchor shot detection system, *MASD*, operation with a typical example

Kernel-ART is based on ART algorithm that scalability has already been demonstrated in Artificial Neural Networks [27]. It means that though the size of data increases, performance above certain level can be shown. We identified it through experiment. The proposed *Kernel-ART* algorithm requires satisfactory processing time of 1.2983 s on the average in clustering total 6,876 frames under the environment of this experiment. To identify the change of processing speed according to the increase in data, we measured the number of frames up to 6,500 while increasing by 500. As we have already identified in complexity analysis, we confirmed that processing speed of proposed *Kernel-ART* clustering algorithm increases close to the linear according to the number of input frames through experiment (see Fig. 5).

4.2 Experiment for anchor shot detection

We conduct performance evaluation of our anchor shot detection (*MASD*) using five test datasets which include real anchor shots, reporter shots, and interview shots. The dataset is summarized in Table 5.

Table 6 Experimental evaluation of the proposed anchor shot detection method

	Datasets					
	D1	D2	D3	D4	D5	Average
Precision	100	96.6	97.6	95.2	97.6	97.40
Recall	100	93.3	95.2	97.6	95.3	96.28
F-measure	100	94.9	96.4	96.4	96.4	96.83

Table 7 Summary of quantitative/qualitative analysis for the anchor shot detection

	Gao et al. [11]	Luan et al. [20]	Lan et al. [17]	Proposed method
Data size	Total: 3,830 Anchor shots: 255	Total: NA Anchor shots: 44	Total: 3,470 Anchor shots: 305	Total: 3,358 Anchor shots: 181
Precision	97.64	98	90.49	97.40
Recall	97.25	93	92.64	96.28
Used data	Private	Private	Private	Private
Methods	Graph-theoretical clustering	AnchorClu clustering	Multimodal associated clustering	One-class SVMs
Postprocessing	Yes	Yes	Yes	No

Figure 6 illustrates a snapshot of our system operation, using a typical example. Figure 6a is the examples of pruned data by skin color detector which is the first step of *MASD*, and Fig. 6b is the examples of pruned data by face detector, the second step of *MASD*, which involves no face or more than three faces. The examples of discarded data by OSVMs, the last step of *MASD*, are shown in Fig. 6c.

Our simulation indicates that an average 31.79% of all candidate sets is obtained without any loss of anchor shots, after the skin color detection phase. Furthermore, an average 13.27% of all candidate sets is determined, subsequent to the face detection phase including all anchor shots. Ultimately, an average 86.73% of non-candidate anchor shots is successfully eliminated by skin color detection and face detection, as we expected. With the reduced search space obtained from prior steps, the anchor shot detection module classifies face data as anchor shots and non-anchor shots.

In Table 6, we summarize the experimental results of our proposed anchor shot detection system, *MASD*. According to the simulation, our system shows an average 97.40% for *precision*, 96.28% for *recall* and 96.83% for F-measure, respectively. As previously stated, our system methodology achieves not only a high speed by means of greatly reducing the search space, but also a high accuracy by means of a SVDD using NMF. Also, we summarize the quantitative/qualitative analysis with existing methodologies in Table 7.

5 Conclusion

In this paper, we introduced an efficient one-pass shot boundary detection algorithm and a cost-effective anchor shot detection method for news video story parsing, which are tightly coupled. First, we proposed a new shot boundary detection method, based on SVD, and a newly developed algorithm, viz., *Kernel-ART*, which meets all of the design requirements for shot boundary detection in terms of news video story parsing. It can greatly reduce the search space in anchor shot detection by using only static shots as inputs in anchor shot detection and showed a very high *recall* ratio. It also detects the abrupt cuts and the gradual transitions using a single algorithm with a single scan of the dataset. By applying SVD, noises or trivial variations in the video sequence are removed, and the mercer kernel improves the probability of detection of shots which are not separable in input space. Second, we proposed

a new anchor shot detection system, viz., *MASD*, which consists of three modules and operates sequentially: 1) A skin color detection module for reducing candidate face regions; 2) A face detection module for finding key-frames with face data; 3) An OSVM module for determining the anchor shots using a SVDD. The proposed system achieves not only a high speed by means of greatly reducing the search space, but also a high accuracy by means of a SVDD using NMF. We checked the validity of our approach with simulations.

We are very interested in video data mining. So, we will analyze the inter-structure of a news video from the perspective of video data mining in future work.

Acknowledgements This research was supported by a Korea University Grant; This research was financially supported by the Ministry of Education, Science Technology (MEST) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation.

References

1. Baraldi EC (1998) Simplified ART: a new class of ART algorithms. International Computer Science Institute, TR 98-004
2. Cernekova Z, Kotropoulos C, Pitas I (2003) Video shot segmentation using singular value decomposition. In: Procs of international conference on acoustics, speech, and signal processing, vol 3, pp 181–184
3. Cernekova Z, Pitas I, Nikou C (2006) Information theory-based shot cut/fade detection and video summarization. IEEE Trans Circuits Syst Video Technol 16(1):82–91
4. Chaisorn L, Chua T, Lee C (2003) A multi-modal approach to story segmentation for news video. In: World Wide Web: internet and web information systems, vol 6, pp 187–208
5. Colace F, Foggia P, Percannella G (2005) A probabilistic framework for TV-news stories detection and classification. In: Procs of international conference on multimedia and expo, pp 1350–1353
6. Cooper M, Liu T, Rieffel E (2007) Video segmentation via temporal pattern classification. IEEE Trans Multimedia 9(3):610–618
7. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, UK
8. Fang H, Jiang J, Feng Y (2006) A fuzzy logic approach for detection of video shot boundaries. Pattern Recogn 39:2092–2100
9. Fang Y, Zhai X, Fan J (2006) News video story segmentation. In: Procs of the international conference on multi-media modeling, pp 397–400
10. Feng H, Fang W, Liu S, Fang Y (2005) A new general framework for shot boundary detection based on SVM. In: Procs of international conference on neural networks and brain, vol 2, pp 1112–1117
11. Gao X, Tang X (2002) Unsupervised video shot segmentation and model free anchor person detection for news video story parsing. IEEE Trans Circuits Syst Video Technol 12(9): 765–776
12. Gao X, Li J, Yang B (2003) A graph-theoretical clustering based anchorperson shot detection for news video indexing. In: Procs of international conference on computational intelligence and multimedia applications, Washington, DC, USA, pp 108–113
13. Golub G, Van Loan C (1996) Matrix computations, 3rd edn. The Johns Hopkins University Press, USA
14. Gong Y, Liu X (2000) Video summarization using singular value decomposition. In: Procs of international conference on computer vision and pattern recognition, vol 2, pp 174–180
15. Hanjalic A, Lagendijk R, Biemond J (1998) Template-based detection of anchorperson shots in news programs. In: Procs of IEEE international conference on image processing, pp 148–152
16. Ko C, Xie W (2008) News video segmentation and categorization techniques for content-demand browsing. In: Procs of congress on image and signal processing, vol 2, pp 530–534
17. Lan D, Ma Y, Zhang H (2004) Multi-level anchorperson detection using multimodal association. In: Procs of 17th international conference on pattern recognition, vol 3, pp 890–893

18. Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
19. Ling X, Yuanxin Q, Huan L, Zhang X (2008) A method for fast shot boundary detection based on SVM. In: Procs of congress on image and signal processing, vol 2, pp 445–449
20. Luan X, Xie Y, Wu L, Wen J, Lao S (2005) AnchorClu: an anchorperson shot detection method based on clustering. In: Procs of 6th international conference on parallel and distributed computing, applications and technologies, pp 840–844
21. Santo M, Foggia P, Sansone C, Percannella G, Vento M (2006) An unsupervised algorithm for anchor shot detection. In: Procs of 18th international conference on pattern recognition, vol 2, pp 1238–1241
22. Solina F, Peer P, Batagelj B, Juvan S, Kovac J (2003) Color-based face detection in the 15 seconds of fame art installation. In: Procs of mirage 2003. INRIA Rocquencourt, France, pp 10–11
23. Tax D, Duin R (2004) Support vector data description. Mach Learn 54(1):45–66
24. Viola P, Jones M (2004) Robust real-time face detection. Int J Comput Vis 7(2):137–154
25. Yoo HW, Cho SB (2007) Video scene retrieval with interactive genetic algorithm. Multimed Tools Appl 34:317–336. doi:10.1007/s11042-007-0109-8
26. Yuan J, Wang H, Xiao L, Zheng W, Li J, Lin F, Zhang B (2007) A formal study of shot boundary detection. IEEE Trans Circuits Syst Video Technol 17(2):168–186
27. Zurada JM (1992) Introduction to artificial neural systems. Info Access Distribution, Singapore

Hansung Lee received his B.S., M.S., and Ph.D. degrees in computer science from Korea University, Korea, in 1996, 2002, and 2008, respectively. He is currently a senior member of Electronics and Telecommunications Research Institute, Korea. From July 1996 to July 1999, he worked for DAEWOO Engineering Company. His recent research interests include Data Mining, Network Mining, Multimedia Mining, Intelligent Data Base, Machine Learning and Soft Computing.

Jaehak Yu received the B.S degree from Dept. of Computer Science of Konkuk University in 2001. He received M.S. degree from Department of Computer Science of Korea University in 2003, and he is currently a Ph.D. candidate at the DB & Data Mining laboratory, Dept. of Computer Science at Korea University, Korea. He is broadly interested in data and information analysis with a focus on data mining and machine learning. In particular, his research interests include intelligent network management, image mining, home network security, intrusion detection.

Younghee Im received her BS degree in computer science from Korea University, Korea, in 1994, and her PhD degree in computer science from Korea University, Korea, in 2001. She joined Korea University in 2005, where she is currently an Invitational Professor in the Dept. of Computer and Information Science. Her research interests include machine learning, context awareness, and intelligent database.

Joon-Min Gil received his B.S. and M.S. degrees in computer science from Korea University, Chochiwon, Korea in 1994 and 1996, respectively. He received his Ph.D. degree in computer science and engineering from Korea University, Seoul, Korea in 2000. From 2001 to 2002, he was a Visiting Research Associate in Dept. of Computer Science at University of Illinois at Chicago, USA. From 2002 to 2005, he was a Senior Research Engineer in Supercomputing Center at Korea Institute of Science and Technology Information, Daejeon, Korea. He is currently an assistant professor in Dept. of Computer Science Education at Catholic University of Daegu, Korea. His recent research interests include grid computing, Internet computing, and data mining.

Daihee Park received his BS degree in mathematics from Korea University, Korea, in 1982, and his PhD degree in computer science from the Florida State University, USA, in 1992. He joined Korea University in 1993, where he is currently a Professor in the Dept. of Computer and Information Science. His research interests include data mining and intelligent database.