Nyström Approximated Temporally Constrained Multisimilarity Spectral Clustering Approach for Movie Scene Detection

Rameswar Panda, *Student Member, IEEE*, Sanjay K. Kuanar, and Ananda S. Chowdhury, *Senior Member, IEEE*

***Abstract*—Movie scene detection has emerged as an important problem in present day multimedia applications. Since a movie typically consists of huge amount of video data with widespread content variations, detecting a movie scene has become extremely challenging. In this paper, we propose a fast yet accurate solution for movie scene detection using Nyström approximated multisimilarity spectral clustering with a temporal integrity constraint. We use multiple similarity matrices to model the wide content variations typically present in any movie dataset. Nyström approximation is employed to reduce the high computational cost of constructing multiple similarity measures. The temporal integrity constraint captures the inherent temporal cohesion of the movie shots. Experiments on five movie datasets from different genres clearly demonstrate the superiority of the proposed solution over the state-of-the-art methods.**

***Index Terms*—Movie scene detection, Nyström approximation, similarity matrices, spectral clustering.**

I. Introduction

With the recent development of inexpensive digital multimedia technologies along with lower cost of publishing and wide potential reach, there has been a tremendous increase in the number of videos over the Internet [1], [2]. For example, one of the most prevalent social media services and Web sites like YouTube reported that over 1 billion unique users visit YouTube each month and 300 h of video, including movies, are uploaded every minute, amounting to nearly 1.5 billion hours of video every year. The task of managing this large amount of video information is an enormously challenging task. Computationally efficient methods are necessary to process, organize, summarize, and index this information in a semantically meaningful manner [3]–[5].

Manuscript received May 9, 2016; revised September 23, 2016 and November 29, 2016; accepted January 8, 2017. Date of publication February 7, 2017; date of current version February 14, 2018. This paper was recommended by Associate Editor H. Yin. *(Corresponding author: Ananda S. Chowdhury.)*

R. Panda was with Jadavpur University, Kolkata 700032, India. He is now with the Department of Electrical and Computer Engineering, University of California at Riverside, Riverside, CA 92521 USA (e-mail: rpand002@ucr.edu).

S. K. Kuanar is with the Department of Computer Science and Engineering, Gandhi Institute of Engineering and Technology, Gunpur 765022, India (e-mail: sanjay.kuanar@gmail.com).

A. S. Chowdhury is with the Department of Electronics and Telecommunications Engineering, Jadavpur University, Kolkata 700032, India (e-mail: aschowdhury@etce.jdvu.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2657692

Movie scene detection is an important problem in the area of multimedia content management. Scene(s), composed of groups of shots, usually emphasize specific concepts (e.g., a fixed setting or the same action) and are hence found to be semantically meaningful. Shot level video segmentation and video summarization, the two existing techniques for handling large amount of video data prevalent in the Internet are found to be inadequate to handle the current problem. This is largely because movies are of long durations and have widely varying contents. Shot level segmentation methods are inefficient to organize the chapters of a movie which correspond to various themes. On the other hand, browsing a movie can be often more convenient and meaningful using less number of scenes (say, 100) compared to a large number of key frames (say, 10 000) which can be obtained from typical video summarization methods [6]–[12]. The scene detection problem can be modeled as a shot clustering problem, where each cluster should be semantically distinct [13]–[15]. In this paper, we propose a novel fast yet accurate solution for detecting movie scenes using Nyström approximated multisimilarity spectral clustering with a temporal integrity constraint.

A. Related Work

This paper spans following areas of interest—video summarization, video scene detection, and multiview learning. We will review some representative works from these areas.

1) Video Summarization: There exists several methods for video summarization. For example, de Avila *et al.* [6] used color feature and an improved k-means algorithm to choose the key frames. Kuanar *et al.* [7] applied both color and texture features with a dynamic Delaunay clustering for the same purpose. Almeida *et al.* [8] utilized the notion of color similarity between successive frames to extract the key frames in the MPEG compressed domain. Han *et al.* [9] examined color in combination with human in the loop guidance for personalized video summarization. Recently, a work on scalable video summarization using skeleton graph and random walk is reported [10]. For most recent reported results on key frame summarization in YouTube videos (please see [16], [17]). A more comprehensive review of video summarization approach can be seen [11], [12].

2) Video Scene Detection: Since, we propose a graph-theoretic solution, we first discuss some prominent graph-based approaches for video scene detection. Yeung *et al.* [18]

represented video as a scene transition graph, where shots are clustered and then each cluster is represented by a node in the graph. The complete link method is used to split the graph into several subgraphs (i.e., scenes). In [19], a weighted shot similarity graph (SSG) is constructed, where each node represents a shot and the edges between shots are weighted by color and motion similarity information. Then normalized cut is used for recursive bi-partitioning of SSG, to maximize intrasubgraph similarities while minimizing intersubgraph similarities. These partitions depict individual scenes present in the video. A similar approach is presented in [20], where shot clustering was achieved using Ncuts algorithm in the first step and the resulting clusters are represented by a temporal graph. In another graph partition-based method [21], a 1-D signal is constructed for each feature. Chasanis *et al.* [15] proposed a spectral graph-based approach using visual similarity of individual shots. A label is assigned to each shot depending on the cluster it belongs to. Then, a global sequence alignment algorithm is applied to detect the change in shot label pattern. Odobez *et al.* [22] proposed a spectral method with automatic model selection for video scene detection. However, the approach is only restricted to scene detection in home videos. Another spectral clustering method for scene segmentation is presented in [23], where Zhang *et al.* used the concept of JSEG measure to capture the local information embedded in video shots. Although the method presented in [23] is independent of video genres but the incorporation of temporal information in form of a sliding window while computing the shot similarity greatly influences the detection results. Sakarya *et al.* [24] introduced a movie scene detection method based on finding dominant sets in SSG. In this paper, two graph partitioning approaches, i.e., tree-based approach (TBM) and order-based approach (OBM), using dominant set clustering are applied for movie scene detection. However, inaccurate estimation of control variables, such as temporal distance decay factor and outlier detection factor may adversely affect the scene detection result.

Other than graph-theoretic methods, statistical approaches are also applied for the scene detection problem. A Markov Chain Monte Carlo method is presented in [13]. The authors use three types of updates, i.e., diffusion, merge, and split to determine the scene boundaries. However, this method is highly sensitive to model prior and the number of shots. Tan and Lu [25] proposed a Gaussian mixture model for scene segmentation, where each scene is modeled as a Gaussian density assuming similar visual features for the shots constituting a scene. This method is able to discover scene-level semantics for sports videos. However, for more general video genres, such as movies, only using the features of individual shot is not sufficient. The impact from the neighboring shots (i.e., temporal integrity) should also be considered. Sundaram and Chang [26] proposed a computational scene model to achieve video scene segmentation. In this paper, video and audio scenes were detected separately and these two were used with some cinematic rules in order to construct scenes.

Apart from the above unsupervised approaches, there has been a growing interest in developing supervised or semi-supervised algorithms for detecting video scene boundaries [27]–[29]. A generic framework based on semi-supervised learning for video annotation can be seen in [30]. Zhang *et al.* [30] used combination of multimodal information by developing a graph-based multiple instance learning framework for video annotation. It jointly explores small-scale expert labeled videos which are obtained from analysis and alignment of well-structured video related text (e.g., movie script and captions) and large-scale unlabeled videos which are obtained by querying related events from the video search engine (e.g., YouTube and Google) to train a discriminative model for video annotation. A novel metric for evaluating scene segmentation methods is presented in [31]. Recently, deep learning-based approaches have been developed for scene detection in broadcast videos [31], [32]. For most recent reported results on scene detection in YouTube videos (please see [5], [31]).

3) Multiview Learning: In recent years, many methods of clustering from multiview data by considering different views have been proposed. These views may be obtained from multiple sources or different feature subsets [33]–[42]. This paper is closely related to different multiview learning methods since we use different sets of features of a movie for efficient detection of scene boundaries. In contrast to the prior works, our proposed approach is different in two significant ways. First, we consider Nyström approximation in constructing the feature similarity matrices which are then used in the spectral clustering. The use of Nyström approximation reduces the computation burden to a large extent with only a marginal compromise in the detection performance. Second, we explicitly use a temporal constraint in our formulation to enforce the temporal cohesion between video shots which is crucial in movie scene detection.

By reviewing the related works on scene detection, we found that the following three key problems in movie scene detection still remain unaddressed to a considerable extent.

1) *Proper Selection of Features:* It has been observed that different features and homogeneity criteria generally lead to different segmentations of the same video. This problem is even more prominent in case of movie scene detection as different types of content variations (due to variations in shooting and editing effects at various stages of the video life cycle) exist across the shots. So, selection of multiple features and an optimal weight assignment policy for their combination is highly necessary, a task missed by most of the prior works.
2) *Use of Computationally Efficient Technique for Shot Similarity Calculation:* Several recently proposed scene detection techniques compute pair-wise similarities for the clustering purpose [14], [15]. Such computation has to be carried out for all possible shot pairs in a video [43]. A case in the point is the movie gone in 60 s with nearly 2900 shots. Processing this movie for five similarity measures can easily involve 42 million pair-wise comparisons! Hence, an efficient approximation technique is required for clustering of large data like movie to substantially reduce the computational costs.

3) *Representation of Temporal Cohesion of Video Shots:* It can be easily noticed that a video scene has temporal integrity. So, temporal cohesion of movie shots is required to achieve accurate scene detection [14]. Most of the previous shot clustering approaches use temporal distance as a solution for this problem [14], [24]. Due to the absence of prior knowledge about the video content and the duration of scenes, it is difficult to determine an appropriate weight parameter that will account for the contribution of the temporal distance in the computation of the overall similarity between shots.

In this paper, we address the above three key problems in our solution pipeline. First, a combination of multiple shot similarity matrices involving color, texture, motion, and semantics is proposed to capture the diverse characteristics of different types of movie scenes. For example, color feature can effectively model a stationary scene whereas motion features are essential for action scenes. Second, Nyström approximation is employed to reduce the high computational cost of constructing multiple similarity measures. As a third contribution, we have directly incorporated temporal integrity constraints in the multisimilarity spectral clustering thereby obviating incorporation of temporal distance in form of a similarity matrix. Note that in the later case, prior knowledge is essential which is always difficult to obtain for movie scenes [15]. In addition, from pure theoretical perspective, the proposed Nyström approximated temporally constrained multisimilarity spectral clustering approach has not been applied in the field of pattern clustering to the best of our knowledge. Note that, the convergence is guaranteed for the Nyström approximated eigenvectors but not for the generalized eigenvectors [43].

B. Theoretical Foundations

Our movie scene detection approach is primarily based on spectral clustering and application of Nyström extension for similarity matrix completion. Some useful theories pertaining to our approach are briefly reviewed in this section.

1) Nyström Extension: The Nyström method is a technique for finding numerical approximations to eigenfunction of integral equations of the form [43]

$$\int W(x, y)\phi(y)p(y)dy = \lambda\phi(x) \tag{1}$$

where $p(y)$ represents the underlying probability density function, $\phi(y)$ indicates the eigenfunction and $W(x, y)$ denotes the similarity between x and y. For finding numerical approximation to (1), one need to choose n_s number of landmark points $Z = \{Z_1, Z_2 \ldots Z_{n_s}\}$ from the given dataset $X = \{X_1, X_2 \ldots X_n\}$ with $n_s << n$. For any given point x in X, using Nyström approximation we can write

$$\frac{1}{n_s}\sum_{i=1}^{n_s} W(x, Z_i)\widehat{\phi}(Z_i) = \lambda\widehat{\phi}(x) \tag{2}$$

where $\widehat{\phi}(x)$ is an approximation to the true $\phi(x)$. Equation (2) cannot be solved directly as λ and $\widehat{\phi}(x)$ are both unknown. In order to solve (2), one needs to substitute x with Z_i, and write it in matrix form $A\widehat{\Phi} = P\widehat{\Phi}\wedge$, where A denotes the similarity matrix between landmark points and $\widehat{\Phi}$ represent the eigenvectors of A. $\wedge = \text{diag}\{\widehat{\lambda_1}, \widehat{\lambda_2}, \ldots, \widehat{\lambda_{n_s}}\}$ is a diagonal matrix. For an unsampled point, the jth eigenfunction at x can be approximated as

$$\Phi_j(x) = \frac{1}{n_s\widehat{\lambda_j}}\sum_{i=1}^{n_s} W(x, Z_i)\widehat{\phi}_j(Z_i). \tag{3}$$

With the above equation, the eigenvector for any arbitrary point x can be approximated by the eigenvectors of the landmark similarity matrix.

2) Nyström Extension to Spectral Clustering: Let $A = U_A \wedge_A U_A^T$ be the eigen-decomposition of A. Further, let B denotes the similarity matrix between sample points and the remaining points, with $B \in R^{m\times(n-n_s)}$. From (3), the matrix form of the Nyström extension is then $B^T U_A \wedge_A^{-1}$, where B^T corresponds to $W(Z_i, \cdot)$, the columns of U_A correspond to the $\widehat{\phi}_j(Z_i)$s, and $\wedge_A^{-1}$ corresponds to the $1/\widehat{\lambda}_j$s. Let $W \in R^{n\times n}$ be the similarity matrix between all data points. For simplicity in notation, let us rearrange the points such that n_s number of randomly sampled points come first and remaining samples come next. Now, partition the similarity matrix W as

$$W = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \tag{4}$$

where $C \in R^{(n-n_s)\times(n-n_s)}$ is the similarity matrix between unsampled points. Using the approximated eigenvectors $\widehat{U} = [U_A; B^T U_A \wedge_A^{-1}]$, W can be estimated as

$$\widehat{W} = \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix} = \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-1} \begin{bmatrix} A & B \end{bmatrix}. \tag{5}$$

For spectral clustering, the similarity matrix is required to be normalized, i.e., one has to calculate row sums of W to acquire D. Depending on definiteness of A and D can be estimated through the row sums of $\widehat{W}$ in two different ways.

Case 1 (A Is Positive Definite): When matrix A is positive definite, all the eigenvalues of matrix A are positive and $A^{-1/2}$ is defined. The orthogonalized approximate eigenvectors are obtained by

$$\widehat{V} = \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-1/2} U_S \wedge_S^{-1/2} \tag{6}$$

where $S = A + A^{-1/2} B B^T A^{-1/2}$ with eigen decomposition $U_S \wedge_S U_S^T$.

Case 2 (A Is Indefinite): When A is indefinite, then two steps are required to get the normalized solution. Let $\overline{U}_S^T = [U_S^T \wedge_S^{-1} U_S^T B]$ and $Z = \widehat{U}\wedge^{1/2}$ so that $\widehat{W} = ZZ^T$. $F \sum F^T$ denote the diagonalization of $Z^T Z$. Then matrix $V = ZF \sum^{-1/2}$ contains the leading eigenvectors of $\widehat{W}$. More details about the Nyström approximation and its extension to spectral clustering can be seen at [43].

II. Proposed Framework

Our proposed method consists of four major steps, namely: 1) shot detection and representation; 2) shot similarities computation; 3) spectral grouping of shots; and 4) cluster sequence analysis. A block diagram of the proposed method is shown in Fig. 1. Now, we describe these four steps in details under four sections.

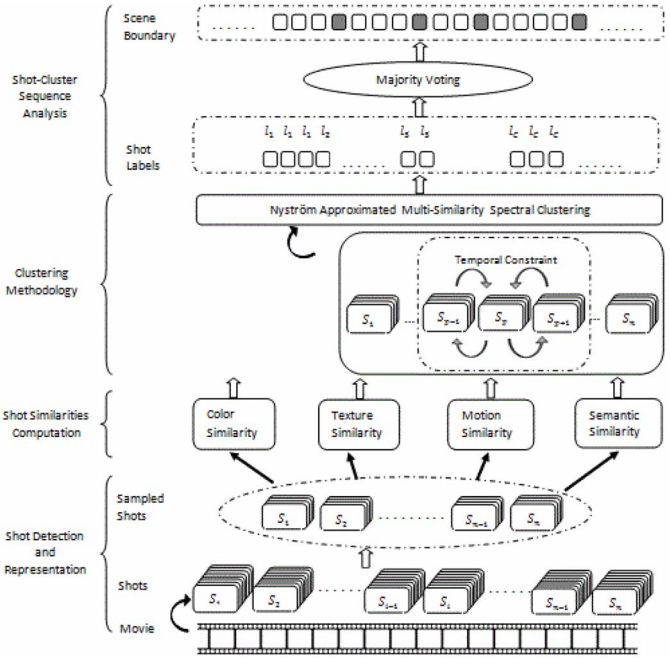

Fig. 1. Flowchart of the proposed method. We first divide the movie into shots using an information theory-based shot detection method and represent each shot by its middle frame. We employ Nyström sampling to select a group of shots and compute different shot similarities (color, texture, motion, and semantic). Nyström approximation is employed to reduce the high computational cost of constructing multiple similarity measures. We then apply multisimilarity spectral clustering with a temporal integrity constraint to cluster the shots. Finally, shot-cluster sequence analysis is used to detect the precise scene boundaries.

Fig. 2. Shot representative frames using different methods from the movie "Kingdom of Heaven." Representative frames detected by different methods for two shots are arranged in rows. First row: representative frames selected by the method [24], second row: frames selected by the method [15], third row: frames selected by the key frame extraction method [8], fourth row: key frames selected by the method [7], and fifth row: middle frame. As can be seen from the figure, there is a little difference among the different methods in representing a shot. Thus, the middle frame is a good choice for representing the movie shots as it can capture the general view of the overall content by avoiding considerable computational overhead in selecting the key frames.

A. Shot Detection and Representation

We first divide the movie into shots using information theory-based shot detection method by Černeková *et al*. [44]. This method is shown to yield high detection accuracy on the TRECVID 2003 video test set. Various schemes exist to represent the video shots using a single key frame or a set of key frames [12], [18], [21]. Fig. 2 shows a comparative analysis of shot representation methods. In general, on analyzing Fig. 2, we can conclude that the middle frame is a good choice for representing the movie shots as it can capture the general view of the overall content. Hence, we have taken the middle frame of a shot as its representative thereby avoiding considerable computations in selecting the key frames of a shot [14].

B. Shot Similarities Computation

To properly capture diverse characteristics of different types of movie scenes, we apply multiple feature similarity matrices. Color, texture, and motion similarity functions between two shots (i.e., representative key frames) are calculated. Color histogram is obtained using the HSV color space, as it is found to be more resilient to noise [45]. We use 16 ranges of H, 4 ranges of S, and 4 ranges of V to form a 256-D color feature vector and an edge histogram descriptor [46] to form a 80-D texture feature vector. Histogram-based visual features are found to work well for stationary scene boundary detection [14], [24], [47], [48]. For calculating the color and texture similarity, we adopt the method of [24]. However, for action scenes, these visual features cannot work when scene changes are encountered very frequently and it may result in oversegmentation. Hence, histogram-based motion activity analysis is required for action scene detection. The above visual similarities do not take into account the shot semantics. So, we also consider semantic similarity between shots in this paper. Semantic similarity between documents is addressed using the Bag of words model [49]. Bag of visual words model for a video captures semantic meaning which can improve clustering of shots [50], [51]. We compute visual words using K-means clustering on SIFT features [52] extracted from all shot representative frames of a movie. Each visual word is represented by a cluster. A visual word w_j appears in a shot s_i if there exists some SIFT feature points in the shot representative frame within the jth cluster. Finally, a shot is represented by

$$S = \nu_1, \nu_2, \ldots, \nu_j, \ldots, \nu_k. \tag{7}$$

In (7), ν_j represents the normalized frequency of the jth visual word and k is the total number of visual words/clusters. We consider 1 00 000 SIFT features of a movie and group them into 1000 clusters. Similar number of visual words is also used by Kumar *et al.* [53], where the authors clustered the sift features into 1500 visual words. SemanticSim function captures the semantic similarity between two shots i and j and is given by

$$\text{SemanticSim}(i,j) = 1 - \sum_{l=1}^{k} \min\left(\nu_{il}, \nu_{jl}\right) \tag{8}$$

where ν_{il} is the normalized frequency of the lth visual word in shot i and k is the total number of visual words. A general

shot similarity matrix is represented by

$$W(i,j) = e^{-a*\mathrm{Sim}(i,j)}. \tag{9}$$

In (9), $\mathrm{Sim}(i,j)$ represents the similarity function between any two shots i and j and a is a control parameter [54].

C. Spectral Grouping of Shots

Given n movie shots $s_1, s_2, \ldots, s_n$, m similarity matrices $W_k, (k = 1, \ldots, m)$, $w_{i,j;k}$ denotes the similarity between s_i and s_j for the kth feature. Let $V = [v_1, v_2, \ldots, v_m]$ be a weight vector acting as selectors for the similarities. The objective of multisimilarity spectral clustering [55] is to divide these movie shots into c clusters by finding n indicators which minimizes

$$\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} v_k^2 w_{i,j;k}\|f_i - f_j\|^2 \tag{10}$$

where $f_i \in \mathbb{R}^c$ represent the cluster indicator variable for the ith movie shot. We now incorporate a temporal continuity constraint within the above objective function to address temporal cohesion of the shots. The constrained objective function is given by

$$\begin{aligned} J = &\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} v_k^2 w_{i,j;k}\|f_i - f_j\|^2 \\ &+ \delta\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} v_k^2 |w_{i,j;k} - w_{i,l;k}|\varphi_{jl}\|f_i - f_j\|^2 \end{aligned} \tag{11}$$

where δ is a weight parameter for the second term of the objective function and φ_{jl} accounts for temporal integrity between movie shots j and l. The temporal integrity function must satisfy the following properties while grouping the movie shots.

1) $\varphi_{jl} = 1$, if $|j-l| = 0$, indicates that the same shot must be in one cluster.
2) $\varphi_{jl} \to 0$, if $|j-l| \to \infty$, means that if two shots are very far in time order, then the effect of one shot on the other is negligible. In other words, grouping the first and the last shot of a movie into one cluster is highly unacceptable.
3) φ_{jl} increases when $|j-l|$ becomes smaller, means that neighboring shots that are close in time order to a specific shot, have more effect on clustering as compared to further shots. We choose the following temporal integrity function which satisfies the above properties:

$$\varphi_{jl} = e^{-|j-l|}. \tag{12}$$

This objective function is minimized under the constraint that the weighted sum of v_k's p-norm is normalized, that is

$$\sum_{k=1}^{m} v_k^p = 1;\ 1 \le p \le 2, v_k \ge 0. \tag{13}$$

In addition, for satisfying normalized spectral clustering, we require $f^T Df = 1$, where D is the diagonal matrix with its ith diagonal element being the sum of ith row of W_k. Mathematically, that is expressed as

$$f^T Df = \sum_{k=1}^{m} \alpha_k v_k^2 = 1 \tag{14}$$

where $\alpha_k = f^T Df$. The goal is to minimize (11) subject to constraints (13) and (14). We construct the corresponding unconstrained objective function by applying Lagrange multipliers

$$\begin{aligned} J_{\lambda_1,\lambda_2} = &\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} v_k^2 w_{i,j;k}\|f_i - f_j\|^2 \\ &+ \delta\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} v_k^2 |w_{i,j;k} - w_{i,l;k}|\varphi_{jl}\|f_i - f_j\|^2 \\ &- \lambda_1\left(\sum_{k=1}^{m} v_k^p - 1\right) - 2\lambda_2\left(\sum_{k=1}^{m}\alpha_k v_k^2 - 1\right). \end{aligned} \tag{15}$$

Note that in (15), there are two sets of variables, indicators f_i and weights v_k. A good strategy is to solve one set of variables at a time while fixing the other set of variables [55].

Case 1: Weights v_k are given; the goal is to determine indicators f_i. If the weights v_k are given, the problem becomes a standard spectral clustering problem and the similarities are set as $w(i,j) = \sum_k v_k^2 w_{i,j;k}$. Thus, the indicators f_i can be determined from the eigenvectors of the Laplacian matrix [56].

Case 2: Indicators f_i are known; the goal is to find weights v_k. Let us first assume that indicators f_i are given and fixed. By taking partial derivative of (15) with respect to v_k^p and setting them to zero, we have

$$\begin{aligned} \frac{\partial J_\lambda}{\partial v_K^p} = &\frac{2}{p} v_k^{2-p} \\ &\times \begin{pmatrix} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j;k}\|f_i - f_j\|^2 \\ +\ \delta\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|w_{i,j;k} - w_{i,l;k}| \\ *\ \varphi_{jl}\|f_i - f_j\|^2 - 2\lambda_2\alpha_k \end{pmatrix} - \lambda_1 = 0. \end{aligned} \tag{16}$$

For simplification, let

$$\begin{aligned} \beta_k = &\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j;k}\|f_i - f_j\|^2 \\ &+ \delta\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|w_{i,j;k} - w_{i,l;k}|\varphi_{jl}\|f_i - f_j\|^2. \end{aligned}$$

The solution becomes

$$v_k = \left(\frac{\lambda_1}{2p^{-1}}\right)^{\frac{1}{2-p}} (\beta_k - 2\lambda_2\alpha_k)^{\frac{1}{2-p}}. \tag{17}$$

It is difficult to solve (17) as v_k is dependent on two variables λ_1 and λ_2. So, we first solve for λ_1 and λ_2. Substituting the above in (13), we obtain

$$\lambda_1 = \left(\frac{p}{2}\right)^{-1}\left[\sum_{k=1}^{m}(\beta_k - 2\lambda_2\alpha_k)^{\frac{-p}{2-p}}\right]^{-\left(\frac{2-p}{p}\right)}. \tag{18}$$

Furthermore, from (14) and (17), we have

$$(\lambda_1)^2 = \left(\frac{p}{2}\right)^{-2}\left[\sum_{k=1}^{m}\alpha_k(\beta_k - 2\lambda_2\alpha_k)^{\frac{-2}{2-p}}\right]^{-(2-p)}. \quad (19)$$

Replacing λ_1 in the above equation from (18), we have

$$\left[\sum_{k=1}^{m}(\beta_k - 2\lambda_2\alpha_k)^{\frac{-p}{2-p}}\right]^{-2\left(\frac{2-p}{p}\right)} = \left[\sum_{k=1}^{m}\alpha_k(\beta_k - 2\lambda_2\alpha_k)^{\frac{-2}{2-p}}\right]^{-(2-p)}. \quad (20)$$

Note that (20) contains only one variable λ_2. Thus, we now have a 1-D search problem which can be solved by Newton–Raphson method [57]. After finding λ_2, we obtain λ_1 from (18). Finally, v_k can be determined from (17). In our current work, we set $p = 1$ [55] and $m = 4$ (four similarity matrices). We have experimentally chosen $\delta = 0.5$ for all the video segments. This alternating process of determining indicators f_i and weights v_k is repeated till the convergence is reached.

As discussed in the theoretical foundations section, Nyström extension can be applied to find the approximated eigenvectors in spectral grouping. In the following section, we show the application of Nyström extension for approximated eigenvector computation in multisimilarity spectral clustering of movie shots. Let W_T denote the combined matrix which can be represented as addition of individual scalar multiplied-similarity matrices. Then, we can write

$$W_T = v_1 W_{\mathrm{CS}} + v_2 W_{\mathrm{MS}} + v_3 W_{\mathrm{TS}} + v_4 W_{\mathrm{SS}} \quad (21)$$

where $(v_1, v_2, v_3, v_4 > 0)$

$$W_T = v_1\begin{bmatrix} A_{\mathrm{CS}} & B_{\mathrm{CS}} \\ B_{\mathrm{CS}}^T & C_{\mathrm{CS}}\end{bmatrix} + v_2\begin{bmatrix} A_{\mathrm{MS}} & B_{\mathrm{MS}} \\ B_{\mathrm{MS}}^T & C_{\mathrm{MS}}\end{bmatrix} + v_3\begin{bmatrix} A_{\mathrm{TS}} & B_{\mathrm{TS}} \\ B_{\mathrm{TS}}^T & C_{\mathrm{TS}}\end{bmatrix} + v_4\begin{bmatrix} A_{\mathrm{SS}} & B_{\mathrm{SS}} \\ B_{\mathrm{SS}}^T & C_{\mathrm{SS}}\end{bmatrix} \quad (22)$$

where A_{CS} and B_{CS} represent the pair-wise similarity matrix between sampled movie shots and similarity matrix between sampled shots and remaining movie shots. Using the properties of matrix algebra [3]

$$(v_1 B_{\mathrm{CS}} + v_2 B_{\mathrm{MS}} + v_3 B_{\mathrm{TS}} + v_4 B_{\mathrm{SS}})^T = v_1 B_{\mathrm{CS}}^T + v_2 B_{\mathrm{MS}}^T + v_3 B_{\mathrm{TS}}^T + v_4 B_{\mathrm{SS}}^T. \quad (23)$$

The above equation can be represented as follows:

$$W_T = \begin{bmatrix} \widetilde{A} & \widetilde{B} \\ \widetilde{B}^T & \widetilde{C}\end{bmatrix} \quad (24)$$

where $\widetilde{A} = v_1 A_{\mathrm{CS}} + v_2 A_{\mathrm{MS}} + v_3 A_{\mathrm{TS}} + v_4 A_{\mathrm{SS}}$, $\widetilde{B} = v_1 B_{\mathrm{CS}} + v_2 B_{\mathrm{MS}} + v_3 B_{\mathrm{TS}} + v_4 B_{\mathrm{SS}}$, and $\widetilde{C} = v_1 C_{\mathrm{CS}} + v_2 C_{\mathrm{MS}} + v_3 C_{\mathrm{TS}} + v_4 C_{\mathrm{SS}}$. Using Nyström extension, we approximate W_T by $\widehat{W}_T$ in the following manner:

$$\widehat{W}_T = \begin{bmatrix} \widetilde{A} \\ \widetilde{B}^T\end{bmatrix}\widetilde{A}^{-1}\begin{bmatrix} \widetilde{A} & \widetilde{B}\end{bmatrix}. \quad (25)$$

As discussed earlier, the eigenvectors in Nyström approximated spectral clustering can be computed in two different ways depending on the definiteness of $\widetilde{A}$. In our movie scene detection problem, the matrix $\widetilde{A}$ is positive definite. Hence, the orthogonal approximate eigenvectors are obtained by

$$\widehat{V} = \begin{bmatrix} \widetilde{A} \\ \widetilde{B}^T\end{bmatrix}\widetilde{A}^{-1/2} U_S \wedge_S^{-1/2} \quad (26)$$

where $S = \widetilde{A} + \widetilde{A}^{-1/2}\widetilde{B}\widetilde{B}^T\widetilde{A}^{-1/2}$ with eigen decomposition $S = U_S \wedge_S U_S$. But, one of the most important aspects of Nyström extension is sampling of shots to extrapolate the complete grouping solution. For this purpose, we adopt random sampling-based cross-validation approach to select the landmark shots that will give low clusterability difference of eigenvectors [43]. Now, we show all the steps of our spectral clustering approach in Algorithm 1.

Algorithm 1 Spectral Grouping of Shots

Given n number of shots s_i, m similarity matrices between sampled shots and all other shots ($|AB| = m = 4$), group the shots into c clusters (number of scenes).

Procedure (s_i, m, c)

1: Initialize the weights $v_k = 1/m$
2: **Repeat**
3: ▷ Fix weights and find indicators f_i ▷
4: Assume $W_T = \sum_{k=1}^{m} v_k^2 w_{i,j;k}$
5: Find approximated eigenvectors $V_2 \ldots V_{c+1}$ using Nyström extension.
6: Indicator f_i = the i^{th} row of $[V_2 \ldots V_{c+1}]$
7: ▷ Fix indicators f_i and find weights v_k ▷
8: Solve a 1-D search problem of λ_2 in (20)
9: Obtain λ_1 by substituting λ_2 in (18)
10: Weight $v_k = \left(\frac{\lambda_1 p}{2}\right)^{\frac{1}{2-p}}(\beta - 2\lambda_2\alpha_k)^{\frac{-1}{2-p}}$
11: **Until** Convergence
12: Run K-means on $f_1, f_2, \ldots, f_n$ to group the shots into c clusters.

End Procedure

D. Cluster Sequence Analysis

Once the clustering algorithm has grouped the shots into c clusters, a label is assigned to each shot according to the cluster it belongs to. This shot cluster sequence is then analyzed to detect the scene boundaries. A scene boundary exists when two adjacent shot labels are different. The optimal number of scenes (c^*) required for spectral grouping of shots is obtained using MDL principle [58]. However, under-segmentation can happen in the case where $c < c^*$, and over-segmentation in the case $c > c^*$. Hence, to select more robust boundaries, we perform the clustering repeatedly with number of clusters c around the optimal number (c^*) and follow a majority voting procedure as follows:

$$\mathrm{Vote}(i) = \frac{1}{(2 \times Tw) + 1}\sum_{c=c^*-Tw}^{c=c^*+Tw}\mathrm{Boundary}^{(c)}(i) \quad (27)$$

where Tw is the temporal window size and $\mathrm{Boundary}(i) = 1$ if the ith shot is selected as scene boundary; and set to zero otherwise. Please see Algorithm 2 in this connection.

TABLE I
EVALUATION MOVIE DATASETS (SOURCE: INTERNET MOVIE DATABASE). THE TOTAL DURATION OF THE TEST SET IS 11 H 33 MIN 17 S, CONTAINING TOTAL 10 656 VIDEO SHOTS

Video Name	Video ID	Year	Length	Detected Shots	Genre
1492:Conquest of Paradise	1	1992	02:26:52	1975	Adventure/Biography/Drama/History
Gone in Sixty Seconds	2	2000	01:48:13	2881	Action/Crime/Thriller
A Beautiful Mind	3	2001	02:06:17	1564	Biography/Drama
Kingdom of Heaven	4	2005	02:16:53	2711	Action/Adventure/Drama
The Message	5	1976	02:55:02	1525	Adventure/Biography/Drama

Algorithm 2 Cluster Sequence Analysis

Input: Temporal window size, Tw
Number of movie shots, n
Output: Final scene boundaries
1: Compute the optimal number of scenes (c^*) using MDL principle.
2: **For** $c = c^* - Tw : c^* + Tw$
3: Cluster all the movie shots into c groups using Algo.1.
4: **For** $i = 1 : n$
5: Set $Boundary^{(c)}(i) = 1$ if i-th and $(i+1)$-th shot are assigned to different cluster; set to zero otherwise.
6: **End For**
7: **End For**
8: **For** $i = 1 : n$
9: Compute $Vote(i)$ using (27).
10: **End For**
11: Assign final scene boundaries at shot i if $Vote(i) \geq \frac{Tw+1}{(2\times Tw)+1}$.

In experiments, the value of Tw is set as 4 [58]. The true (i.e., final) scene boundary is detected at shot i if Vote(i) is above the threshold of 0.55 (i.e., 5 out of 9).

E. Computational Complexity

Now, we show the advantage of the proposed method by analyzing its computational complexity. Let n be the total number of shots and n_s be the number of Nyström sampled shots (where $n_s << n$). Following [43], we can write the Nyström approximation takes $O(n_s^3) + O(nn_s)$ operations to build one similarity matrix. The time-complexity of spectral clustering with n shots is $O(n^3)$. So, the overall complexity of our spectral clustering method with Nyström sampling being used for approximating the similarity matrices is: $(O(n_s^3) + O(nn_s)) + O(n^3) = O(n^3)$. Considering the post-processing part (cluster sequence analysis), the total computational complexity of our proposed method is $O((2 * Tw + 1) * n^3) = O(n^3)$, where Tw is the temporal window size and $Tw << n$. In this paper, we have set Tw to 4 throughout the experiments. Please note that if the Nyström approximation was not used, the complexity of generating a single similarity matrix would have been: $O(n^2) >> O(n_s^3) + O(nn_s)$, the complexity of the same with the Nyström approximation. Usually computation of more than one similarity matrices is necessary to achieve better clustering. For example, in this paper, we have used four similarity matrices, namely, color, texture, motion, and semantic. With the necessity to compute more similarity matrices for large datasets like the movies, it is imperative that the computational benefit with the Nyström approximation is even more pronounced.

III. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the proposed method is analyzed and compared with two recent graph-theoretic approaches [15], [24]. Five Hollywood movies (without commercials) from the Internet movie database [59] (three of which are also used in [24]) are chosen for performance evaluation (see Table I). Chasanis *et al.* [15] and Sakarya *et al.* [24] have created their own ground-truths. Since movie scene is somewhat a subjective concept and is a problem of general interest, we invite five people from different backgrounds (two film study experts and three graduate students) to create the ground-truths for us. We compare each such ground truth with the algorithmically detected result using F_1 measure [13]–[15].

To determine if a detected scene is correct or not, the best match method [14] is adopted with a sliding window of τ shots as the tolerance factor. The detected scene boundary is regarded as true positive if the offset is less than the tolerance factor τ. We report here mean F_1 value for each movie dataset.

A. Performance Evaluation of Different Similarity Matrices

We first evaluate the efficiency of the proposed combination of multiple similarities over a single similarity. The results are shown in Table II. All the reported F_1 measures are calculated with tolerance factor $\tau = 4$. An interesting observation is that combination of color and edge similarity provides better scene detection performance for Biographic/Drama movies (Video ID #1, 3, 5) whereas the color and motion combination provides better detection accuracy for action movies (Video ID #2, 4). It can be seen from Table II that the proposed combination of similarity measures (shown in bold) easily provides best scene detection performance for all movie datasets independent of their genre. We have used the value of the parameter a [in (3)] as 0.1 [54]. We also show the effect of changing the value of this control parameter a on different shot similarities. From Fig. 3(a)–(d), it can be noticed that the F_1 measure changes only marginally with change in a, which indicates the adopted form of the similarity measures are quite robust.

B. Performance Evaluation of Nyström Approximation

In this section, we make a comparative performance analysis of our method with and without Nyström approximation based

TABLE II
PERFORMANCE COMPARISON WITH DIFFERENT SIMILARITY MEASURES. ALL THE REPORTED F_1 MEASURES ARE CALCULATED WITH TOLERANCE FACTOR $\tau = 4$. BEST PERFORMANCES ARE SHOWN IN BOLD

Methods	Video ID #1	Video ID #2	Video ID #3	Video ID #4	Video ID #5	Mean
Color (C)	0.6714	0.6589	0.7053	0.6865	0.7241	0.6892
Color + Edge (C+E)	0.6828	0.6775	0.7096	0.7099	0.7294	0.7018
Color + Motion (C+M)	0.6441	0.7051	0.7046	0.7340	0.7261	0.7027
Semantic (S)	0.6745	0.6853	0.7059	0.7045	0.7272	0.6994
C+E+M+S	**0.7176**	**0.7283**	**0.7486**	**0.7984**	**0.8069**	**0.7599**

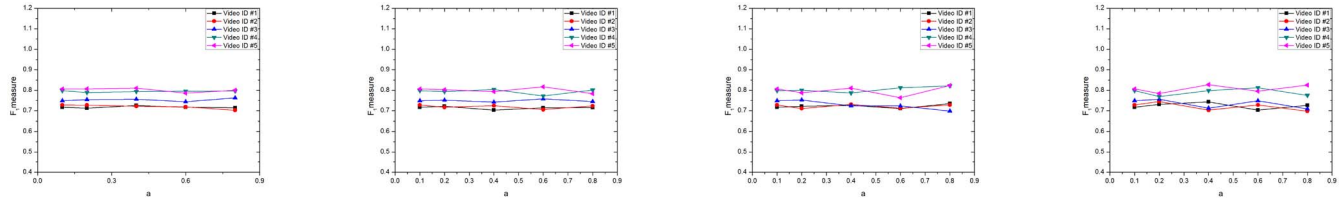

Fig. 3. Effect of varying control parameter a in similarity matrices. As can be seen, F_1 measure changes only marginally with change in a, which indicates the adopted form of the similarity measures Are quite robust. (a) Visual similarity. (b) Texture similarity. (c) Motion similarity. (d) Semantic similarity.

TABLE III
PERFORMANCE COMPARISON WITH/WITHOUT NYSTRÖM APPROXIMATION. ALL THE REPORTED F_1 MEASURES ARE CALCULATED WITH TOLERANCE FACTOR $\tau = 4$

Methods	Video ID #1	Video ID #2	Video ID #3	Video ID #4	Video ID #5	Mean
F_1 Measure Comparison						
With Nyström Approximation	0.7176	0.7283	0.7486	0.7984	0.8069	0.7599
Without Nyström Approximation	0.7352	0.7334	0.7502	0.8153	0.8321	0.7732
Execution Time Ratio Comparison						
With Nyström Approximation	3.2	6.1	2.3	6.7	1.7	4.0
Without Nyström Approximation	9.4	16.7	9.6	14.2	7.4	11.46

on both execution time ratio and F_1 measure. Our proposed graph clustering-based movie scene segmentation method has following computational steps.

1) Shot detection and representation.
2) Shot similarities computation.
3) Clustering methodology.
4) Scene boundaries determination step.

Steps 1) and 4) are independent of Nyström approximation whereas steps 2) and 3) are dependent on the approximation. Motivated by [24], we show a comparative performance analysis of this paper with and without Nyström approximation based on the execution time ratio over the minimum value indicated by 1.0 [only steps 1) and 4) are included] in Table III. The results given in Table III show that our method with the Nyström approximation significantly reduces the execution time (mean value of 4.0 versus 11.46). This is due to the fact that the complete grouping solution is efficiently approximated using Nyström extension. The same table also presents a comparative performance analysis with and without Nyström approximation based on F_1 measure. All the reported F_1 measures are calculated with tolerance factor $\tau = 4$. From Table III, we can conclude that the Nyström approximation substantially improves the execution time (mean value decreases from 11.46 to 4.0 if we use it) and only marginally affects the F_1 value (mean value increases from 0.7599 to 0.7732 if we use it). This type of speedup (about threefold in the present experiments) is highly important, where huge amount of data like that of a movie needs to be processed. We next examine the quality of the Nyström approximation by measuring their approximation errors in terms of Frobenius norm of the difference similarity matrix with and without Nyström approximation and report the average value in Table IV. From the results, it can be observed that the approximation errors are significantly low, which once again validates the use of Nyström extension for the present problem. The proposed method on an average takes about 15 min to detect the scene changes in the movie datasets in Table I on a desktop PC with Intel core i5-2400 processor and 8GB of DDR2 memory.

TABLE IV
AVERAGE NYSTRÖM APPROXIMATION ERROR

Video ID	Error
1	$(2.15 \pm 0.72) \times 10^{-2}$
2	$(1.09 \pm 0.12) \times 10^{-1}$
3	$(1.67 \pm 0.78) \times 10^{0}$
4	$(2.64 \pm 0.08) \times 10^{-2}$
5	$(7.15 \pm 2.43) \times 10^{-2}$

C. Performance Evaluation of Cluster Sequence Analysis

In this section, we present the effectiveness of our cluster scene analysis step using Table V. The results in Table V show that our proposed method performs better in presence

TABLE V
IMPACT OF CLUSTER SCENE ANALYSIS. MEAN F_1 VALUE INCREASES FROM 0.7047 TO 0.7599

Methods	Video ID #1	Video ID #2	Video ID #3	Video ID #4	Video ID #5	Mean
Number of Scenes/Clusters (c^*)	92	76	98	112	64	-
Without Cluster Scene Analysis	0.6814	0.6975	0.7169	0.7177	0.7103	0.7047
With Cluster Scene Analysis	0.7176	0.7283	0.7486	0.7984	0.8069	0.7599

TABLE VI
MEAN F_1 PERFORMANCE COMPARISON WITH NN CLUSTERING AND A SPECTRAL FACTORIZATION-BASED GRAPH PARTITIONING ALGORITHM. ALL THE REPORTED VALUES ARE COMPUTED USING ONLY COLOR FEATURE AND THE CLUSTER SCENE ANALYSIS STEP IS KEPT SAME FOR ALL THE RESULTS. THE RESULTS CLEARLY INDICATE THE SUPERIORITY OF OUR CLUSTERING COMPARED TO BOTH NN CLUSTERING AND SPECTRAL FACTORIZATION-BASED GRAPH PARTITIONING. BEST VALUES ARE SHOWN IN BOLD

Methods	Video ID #1	Video ID #2	Video ID #3	Video ID #4	Video ID #5	Mean
Proposed Method	**0.6714**	**0.6589**	**0.7053**	**0.6865**	**0.7241**	**0.6892**
Graph Partitioning	0.4321	0.5672	0.5863	0.6037	0.6170	0.5612
NN Clustering	0.3771	0.3684	0.4255	0.3720	0.4962	0.4078

of cluster scene analysis with a mean F_1 value of 0.7599 as compared to a mean F_1 value of 0.7047 when no such step is taken. The difference in result is due to the fact that the number of scenes obtained using the MDL principle [58] can result in over-segmentation or under-segmentation. Moreover, obtaining the accurate number of scenes *a priori* using any principle for a movie is a difficult task.

D. Performance Comparison With Other Methods

In this section, we make a comparative performance analysis to evaluate the results of the proposed method. For that reason we have implemented two state-of-the-art methods presented in the literature. Both the works are based on graph-theoretic approaches for video scene detection. The first work is presented in [15]. This method computes visual similarity between shots as the maximum color similarity among all possible pairs of their key-frames. The key-frames are extracted using an improved version of spectral clustering algorithm, where fast global k-means algorithm is used. This comprehensive shot similarity calculation is not feasible for movie data set. Moreover, shots are grouped into clusters using the same spectral clustering, where the number of clusters is estimated based on the magnitude of the eigenvalues of the similarity matrix. Finally, a sequence alignment procedure was applied over shot sequence labels to detect the scene boundaries. However, determination of window size, threshold for global minimum selection in scoring function profile, and the weight parameter α is a tedious task. We have implemented and tested this method using the same movie data set for different values of the above parameters. In our comparisons, we found distinct values for each movie that provide the best performance. It can be noticed that F_1 measures reported in [15] are in the range of 0.85–0.90 as the method is tested for small duration video clips. But, it is not possible to get this range of values for large movie datasets as there exist wide content variation across any movie due to various dynamics of alternating sequences in addition to the movie editing effects.

The second method has been proposed in [24]. This method clusters shots into groups taking into account both color and motion similarity of video shots. Moreover, the temporal shot similarity function is also used along with visual similarity to cluster the shots into groups. This method is based on the idea of finding dominant movie scene boundary using dominant sets framework [24]. After determining the most probable movie scene in the first round, two partitioning strategies are examined to obtain the boundaries of the remaining scenes: 1) a TBM and 2) an OBM. As reported by Sakarya *et al.* [24], TBM is preferred over OBM from a tradeoff between F-measure performance and computational complexity. Hence, we have implemented and tested TBM method using the same movie data set and human generated ground truths. We set the parameters as $r = 2.24, c = 7$ and different values of d for different movies [24]. It must be noted that the F_1 measures reported in [24] are different to the values presented in our performance comparison as former F_1 measures are according to the frame level comparison of clusters and their ground truth results are not available to make a fair comparison.

Table VII presents a comparative performance analysis of our proposed method, methods in [15] and [24] with varying tolerance factor (i.e., τ varied from 1 to 10). The reported F_1 measure values represent the average of F_1 measures obtained from comparing the results of different methods with the ground-truths of each movie dataset. From Table VII, it is observed that our proposed method outperforms both [15] and [24] even with low τ which indicates that the proposed method is able to achieve more precise boundaries. In sharp contrast to our results, the outputs of [15] and [24] suffer from both over/under segmentation problems due to inaccurate scene detections. From Table VII, we can conclude by considering all values of the tolerance factor that the proposed method clearly outperforms [15] and [24] in as many as 46 out of a total of 50 cases.

In order to further demonstrate the superiority of our clustering methodology, we have included comparisons with two popular clustering approaches, namely, a graph partitioning algorithm based on spectral factorization [60] and nearest neighborhood (NN) [61] clustering. Only color feature is used and the cluster scene analysis step is kept the same. Table VI presents the mean F_1 measure obtained by applying both the approaches for five movies. The results clearly indicate that the

TABLE VII
MEAN F_1 MEASURE COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT METHODS WITH VARYING TOLERANCE FACTOR. PROPOSED METHOD OUTPERFORMS [15] AND [24] IN 46 OUT OF 50 CASES. BEST VALUES ARE SHOWN IN BOLD

Tolerance Factor τ	Methods	Video ID #1	Video ID #2	Video ID #3	Video ID #4	Video ID #5	Mean
1	Proposed Method	**0.5529**	**0.5714**	0.5123	**0.4924**	**0.6726**	**0.5603**
	Chasenis *et al.* [15]	0.5089	0.5217	**0.5225**	0.4712	0.5252	0.5099
	Sakarya *et at.* [24]	0.4915	0.5396	0.4815	0.4356	0.5138	0.4924
2	Proposed Method	**0.6235**	**0.5934**	**0.5775**	**0.5517**	**0.7563**	**0.6204**
	Chasenis *et al.* [15]	0.5799	0.5548	0.5340	0.5489	0.5788	0.5592
	Sakarya *et at.* [24]	0.5586	0.5819	0.5025	0.4785	0.5765	0.5396
3	Proposed Method	**0.6823**	0.6493	**0.6737**	**0.6924**	**0.7825**	**0.6960**
	Chasenis *et al.* [15]	0.6153	0.6087	0.6387	0.6327	0.6021	0.6195
	Sakarya *et at.* [24]	0.6033	**0.6543**	0.6153	0.6144	0.6359	0.6246
4	Proposed Method	**0.7176**	**0.7283**	**0.7486**	**0.7984**	**0.8069**	**0.7599**
	Chasenis *et al.* [15]	0.6745	0.6412	0.7120	0.6690	0.6850	0.6763
	Sakarya *et at.* [24]	0.6368	0.6582	0.7076	0.6877	0.6971	0.6774
5	Proposed Method	**0.7529**	**0.7472**	**0.7486**	**0.8007**	**0.8251**	**0.7749**
	Chasenis *et al.* [15]	0.7100	0.7064	0.7120	0.7153	0.7058	0.7059
	Sakarya *et at.* [24]	0.6854	0.7195	0.7076	0.7348	0.7380	0.7170
6	Proposed Method	**0.7882**	**0.7802**	**0.7701**	**0.8249**	**0.8394**	**0.8005**
	Chasenis *et al.* [15]	0.7692	0.7333	0.7434	0.7356	0.7265	0.7416
	Sakarya *et at.* [24]	0.7374	0.7407	0.7384	0.7459	0.7328	0.7390
7	Proposed Method	**0.8235**	**0.8241**	**0.8021**	**0.8327**	**0.8480**	**0.8260**
	Chasenis *et al.* [15]	0.8047	0.7934	0.7853	0.7544	0.7549	0.7785
	Sakarya *et at.* [24]	0.8044	0.7936	0.7794	0.7521	0.7526	0.7764
8	Proposed Method	**0.8705**	**0.8571**	**0.8449**	**0.8568**	**0.8766**	**0.8611**
	Chasenis *et al.* [15]	0.8047	0.8260	0.8272	0.8369	0.7867	0.8163
	Sakarya *et at.* [24]	0.8156	0.8464	0.8102	0.8085	0.8092	0.8179
9	Proposed Method	**0.8941**	0.8901	**0.8770**	**0.8767**	**0.9056**	**0.8887**
	Chasenis *et al.* [15]	0.8282	**0.8913**	0.8691	0.8415	0.8129	0.8486
	Sakarya *et at.* [24]	0.8379	0.8878	0.8205	0.8309	0.8153	0.8384
10	Proposed Method	**0.9058**	0.9120	**0.8983**	**0.9215**	**0.9287**	**0.9132**
	Chasenis *et al.* [15]	0.8520	0.9021	0.8795	0.8971	0.8450	0.8751
	Sakarya *et at.* [24]	0.8714	**0.9235**	0.8404	0.8634	0.8746	0.8746

TABLE VIII
PERFORMANCE COMPARISON WITH [15] AND [24] USING SAME SET OF FEATURES. OURS PERFORMS BEST IN TERMS F_1 VALUES AND AT THE SAME TIME IS ALSO FASTER AS CAN BE SEEN FROM THE EXECUTION TIME RATIOS

Methods	Video ID #1	Video ID #2	Video ID #3	Video ID #4	Video ID #5	Mean
			F_1 **Measure Comparison**			
Proposed Method (C)	0.6714	0.6589	0.7053	0.6865	0.7241	0.6892
Chasenis *et al.* [15]	0.6745	0.6412	0.7120	0.6690	0.6850	0.6763
Proposed Method (C+M)	0.6441	0.7051	0.7046	0.7340	0.7261	0.7027
Sakarya *et al.* [24]	0.6368	0.6582	0.7076	0.6877	0.6971	0.6774
			Execution Time Ratio Comparison			
Proposed Method (C)	1.1	1.6	1.3	1.7	1.4	1.4
Chasenis *et al.* [15]	3.2	4.1	6.7	7.9	5.2	5.3
Proposed Method (C+M)	3.1	4.5	2.7	5.8	1.9	3.6
Sakarya *et al.* [24]	11.2	13.1	14.2	16.3	17.1	14.3

proposed method performs significantly better than the graph partitioning algorithm and NN clustering for all the movie segments.

E. Performance Comparison of Methods With Same Similarity Measures

Boundaries between scenes in the previous sections are determined by the Nyström approximated multisimilarity spectral clustering. In order to explicitly verify the superiority of our clustering methodology, we include a comparison of our method with that of [15] and [24] using the same shot similarity matrix. Table VIII shows a performance analysis of our method using same shot similarity measures as in [15] and [24]. Table VIII also represents a comparison of these methods in terms of execution time ratio over the minimum value indicated by 1.0 [only step 4) of Section III-B is included]. The reason for only inclusion of step 4) instead of both steps 1) and 4) as in Section III-B, is due to the fact that only step 1) is common to all of the three compared methods. Proposed method (C) uses only color similarity as in [15], whereas proposed method (C + M) uses both color and motion similarity as in [24]. The same shot detection and representation (i.e., the middle frame) are used in the preprocessing steps for all the implementations. On considering the mean values of F_1 measures in Table VIII, the performance of proposed method (C) and [15] are close to each other. However, proposed method (C + M) clearly outperforms [24]. The reason for the small performance difference between proposed method (C) and Chasenis *et al.* [15] is

due to the fact that both the methods are based on spectral clustering. The superiority of our method is due to the formulation of spectral clustering with temporal integrity constraint. On the other hand, the reason for the superior performance over the method by Sakarya *et al.* [24] is due the robustness of our clustering strategy with temporal information. At the same time, on considering the execution time ratio, our method performs quite well as compared to both of the methods. Specifically, our clustering strategy uses Nyström approximated eigenvectors that substantially reduces the computational burden in detecting precise scene boundaries from long duration movies.

IV. Conclusion

In this paper, we presented a novel method for high-level segmentation of movies into scenes using Nyström approximated multisimilarity spectral clustering with a temporal integrity constraint. Multiple shot similarity matrices are used to model the diverse characteristics of different types of movie scenes. Comprehensive experimentations clearly indicate that the superiority of the proposed method over some recently published works. In future, we will focus on integration of more extensive set of video features to further improve the scene detection results. Another direction of future research will be to assign semantic labeling to the detected movie scenes for more effective movie navigation.

Appendix

In the following section, we will show the positive definiteness of combined matrix $\widetilde{A}$.

1) Let X be a non empty set. A function (likewise for matrix) $K : X \times X \rightarrow R$ is called positive definite if and only if it is symmetric, i.e., $K(x, x^{'}) = K(x^{'}, x)$ for all $x, x^{'} \in X$ and if for an arbitrary finite non-zero vector c

$$C^T K_{ij} C = \sum_{i,j=1}^{n} C_i C_j K\big(x_i, x_j\big) > 0 \tag{28}$$

for $x_1, \ldots, x_n \subseteq X$; $C_1, \ldots, C_n \subseteq R$.

2) Multiplication of a finite positive constant with a positive definite function or matrix is also positive definite. In other words, if K is positive definite then $v * K$ is also positive definite ($v > 0$).
3) Exponential of a positive definite function is also positive definite, i.e., if K is positive definite, then $\exp(K)$ is also positive definite.

Lemma 1: Individual similarity matrices $W_{\mathrm{CS}}, W_{\mathrm{MS}}, W_{\mathrm{TS}}$, and W_{SS} are positive definite [please see (9)].

Proof of Lemma 1:

$$\begin{aligned} W_{\mathrm{CS}}(i,j) &= e^{-a*\mathrm{ColorSim}(i,j)}, a > 0 \\ &= e^{-a\left[1-\sum_{h=1}^{m}\min\left(H_i(h),H_j(h)\right)\right]} \\ &= e^{-a+a\sum_{h=1}^{m}\min\left(H_i(h),H_j(h)\right)} \\ &= e^{-a} \cdot e^{a\sum_{h=1}^{m}\min\left(H_i(h),H_j(h)\right)} \\ &= v \cdot e^{a\sum_{h=1}^{m}\min\left(H_i(h),H_j(h)\right)}; v > 0. \end{aligned}$$

The above equation is positive definite if $\sum_{h=1}^{m}\min(H_i(h), H_j(h))$ function is positive definite. The function $K(H_i, H_j) = \sum_{h=1}^{m}\min(H_i(h), H_j(h))$ is a positive definite function or Mercers kernel. Hence, the matrix W_{CS} is positive definite. Similarly, it can be shown that other similarity matrices are also positive definite.

Lemma 2: The combined similarity matrix $\widetilde{A}$ is positive definite [please see (21)].

Proof of Lemma 2:

$$\widetilde{A} = v_1 A_{\mathrm{CS}} + v_2 A_{\mathrm{MS}} + v_3 A_{\mathrm{TS}} + v_4 A_{\mathrm{SS}}.$$

$\widetilde{A}$ is positive definite iff $C^T \widetilde{A} C > 0$, that is

$$C^T(v_1 A_{\mathrm{CS}} + v_2 A_{\mathrm{MS}} + v_3 A_{\mathrm{TS}} + v_4 A_{\mathrm{SS}})C > 0.$$

LHS of the inequality can also be written as

$$C^T(v_1 A_{\mathrm{CS}})C + C^T(v_2 A_{\mathrm{MS}})C + C^T(v_3 A_{\mathrm{TS}})C + C^T(v_4 A_{\mathrm{SS}})C.$$

Each individual component, $C^T(v_1 A_{\mathrm{CS}})C > 0$ (from Lemma 1). Hence, $\widetilde{A}$ is positive definite.

References

[1] M. Naaman, "Social multimedia: Highlighting opportunities for search and mining of multimedia data in social media applications," *Multimedia Tools Appl.*, vol. 56, no. 1, pp. 9–34, 2012.
[2] W. Gao, Y. Tian, T. Huang, and Q. Yang, "Vlogging: A survey of videoblogging technology on the Web," *ACM Comput. Surveys*, vol. 42, no. 4, 2010, Art. no. 15.
[3] Y. Pang, H. Yan, Y. Yuan, and K. Wang, "Robust CoHOG feature extraction in human-centered image/video management system," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 458–468, Apr. 2012.
[4] Y. Yuan, Y. Feng, and X. Lu, "Statistical hypothesis detector for abnormal event detection in crowded scenes," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–12, Jun. 2016.
[5] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal, "Harnessing object and scene semantics for large-scale video understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2016, pp. 3112–3121.
[6] S. E. F. de Avila, A. P. B. Lopes, A. da Luz, 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.
[7] S. K. Kuanar, R. Panda, and A. S. Chowdhury, "Video key frame extraction through dynamic delaunay clustering with a structural constraint," *J. Vis. Commun. Image Represent.*, vol. 24, no. 7, pp. 1212–1227, 2013.
[8] J. Almeida, N. J. Leite, and R. da Silva Torres, "VISON: Video summarization for online applications," *Pattern Recognit. Lett.*, vol. 33, no. 4, pp. 397–409, 2012.
[9] B. Han, J. Hamm, and J. Sim, "Personalized video summarization with human in the loop," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Kailua-Kona, HI, USA, 2011, pp. 51–57.
[10] R. Panda, S. K. Kuanar, and A. S. Chowdhury, "Scalable video summarization using skeleton graph and random walk," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Stockholm, Sweden, 2014, pp. 3481–3486.
[11] 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.
[12] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *J. Vis. Commun. Image Represent.*, vol. 19, no. 2, pp. 121–143, 2008.
[13] Y. Zhai and M. Shah, "Video scene segmentation using Markov chain Monte carlo," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 686–697, Aug. 2006.
[14] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1097–1105, Dec. 2005.
[15] V. T. Chasanis, A. C. Likas, and N. P. Galatsanos, "Scene detection in videos using shot clustering and sequence alignment," *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 89–100, Jan. 2009.
[16] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 766–782.
[17] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarizatio," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2016, pp. 1059–1067.
[18] M. Yeung, B.-L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Comput. Vis. Image Understand.*, vol. 71, no. 1, pp. 94–109, 1998.

[19] Y. Zhao *et al.*, "Scene segmentation and categorization using Ncuts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, 2007, pp. 1–7.

[20] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 296–305, Feb. 2005.

[21] U. Sakarya and Z. Telatar, "Video scene detection using graph-based representations," *Signal Process. Image Commun.*, vol. 25, no. 10, pp. 774–783, 2010.

[22] J.-M. Odobez, D. Gatica-Perez, and M. Guillemot, "Spectral structuring of home videos," in *Int. Conf. on Image and Video Retrieval*. Heidelberg, Germany: Springer, 2003, pp. 310–320.

[23] Z. Zhang, B. Li, H. Lu, and X. Xue, "Scene segmentation based on video structure and spectral methods," in *Proc. 10th Int. Conf. Control Autom. Robot. Vis. (ICARCV)*, Hanoi, Vietnam, 2008, pp. 1093–1096.

[24] U. Sakarya, Z. Telatar, and A. A. Alatan, "Dominant sets based movie scene detection," *Signal Process.*, vol. 92, no. 1, pp. 107–119, 2012.

[25] Y.-P. Tan and H. Lu, "Model-based clustering and analysis of video scenes," in *Proc. Int. Conf. Image Process.*, vol. 1. Rochester, NY, USA, 2002, pp. I-617–I-620.

[26] H. Sundaram and S.-F. Chang, "Computable scenes and structures in films," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 482–491, Dec. 2002.

[27] A. Kowdle and T. Chen, "Learning to segment a video to clips based on scene and camera motion," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, 2012, pp. 272–286.

[28] B. Wu *et al.*, "A novel horror scene detection scheme on revised multiple instance learning model," in *Proc. Int. Conf. Multimedia Model.*, Taipei, Taiwan, 2011, pp. 359–370.

[29] C. Liu, D. Wang, J. Zhu, and B. Zhang, "Learning a contextual multi-thread model for movie/TV scene segmentation," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 884–897, Jun. 2013.

[30] T. Zhang, C. Xu, G. Zhu, S. Liu, and H. Lu, "A generic framework for video annotation via semi-supervised learning," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1206–1219, Aug. 2012.

[31] L. Baraldi, C. Grana, and R. Cucchiara, "A deep siamese network for scene detection in broadcast videos," in *Proc. 23rd ACM Int. Conf. Multimedia*, Brisbane, QLD, Australia, 2015, pp. 1199–1202.

[32] L. Baraldi, C. Grana, A. Messina, and R. Cucchiara, "A browsing and retrieval system for broadcast videos using scene detection and automatic annotation," in *Proc. ACM Multimedia Conf.*, Amsterdam, The Netherlands, 2016, pp. 733–734.

[33] J. Yu, Y. Rui, and D. Tao, "Click prediction for Web image reranking using multimodal sparse coding," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2019–2032, May 2014.

[34] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multiview stochastic learning in image classification," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2431–2442, Dec. 2014.

[35] J. Yu, M. Wang, and D. Tao, "Semisupervised multiview distance metric learning for cartoon synthesis," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4636–4648, Nov. 2012.

[36] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[37] C. Xu, D. Tao, and C. Xu, "Large-margin multi-viewinformation bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.

[38] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. IJCAI*, Beijing, China, 2013, pp. 2598–2604.

[39] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity," in *Proc. ICML*, Atlanta, GA, USA, 2013, pp. 352–360.

[40] S. Sun, X. Xie, and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3272–3284, Dec. 2016.

[41] J. Li, Y. Wu, J. Zhao, and K. Lu, "Low-rank discriminant embedding for multiview learning," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–14, May 2016.

[42] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1083–1094, May 2015.

[43] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nystrom method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, Feb. 2004.

[44] Z. Černeková, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 82–91, Jan. 2006.

[45] G. Paschos, "Perceptually uniform color spaces for color texture analysis: An empirical evaluation," *IEEE Trans. Image Process.*, vol. 10, no. 6, pp. 932–937, Jun. 2001.

[46] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.

[47] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[48] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "Biologically inspired features for scene classification in video surveillance," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 307–313, Feb. 2011.

[49] D. Bollegala, Y. Matsuo, and M. Ishizuka, "A Web search engine-based approach to measure semantic similarity between words," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 977–990, Jul. 2011.

[50] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Nice, France, 2003, pp. 1470–1477.

[51] N. Rasiwasia and N. Vasconcelos, "Scene classification with low-dimensional semantic spaces and weak supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Anchorage, AK, USA, 2008, pp. 1–6.

[52] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[53] N. Kumar, P. Rai, C. Pulla, and C. V. Jawahar, "Video scene segmentation with a semantic similarity," in *Proc. IICAI*, Tumkur, India, 2011, pp. 970–981.

[54] Y. Fu *et al.*, "Multi-view video summarization," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 717–729, Nov. 2010.

[55] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen, "Multi-affinity spectral clustering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Kyoto, Japan, 2012, pp. 2089–2092.

[56] U. Von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[57] R. Baldick, *Applied Optimization: Formulation and Algorithms for Engineering Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[58] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 580–588, Jun. 1999.

[59] *IMDB*. Accessed on Jan. 26, 2017. [Online]. Available: http://www.imdb.com/stats

[60] J. P. Hespanha, "An efficient MATLAB algorithm for graph partitioning," Dept. Elect. Comput. Eng., Univ. California, Santa Barbara, CA, USA, Tech. Rep., 2004. [Online]. Available: http://www.ece.ucsb.edu/~hespanha/techrep.html

[61] M. A. Wong and T. Lane, "A kth nearest neighbour clustering procedure," in *Proc. 13th Symp. Interface Comput. Sci. Stat.*, Pittsburgh, PA, USA, 1981, pp. 308–311.

Rameswar Panda (S'16) received the M.E. degree in electronics and telecommunication engineering from Jadavpur University, Kolkata, India, in 2013.

His current research interests include computer vision, multimedia analysis, and pattern recognition.

Sanjay K. Kuanar received the M.Tech. and Ph.D. degrees in electronics and telecommunication engineering from Jadavpur University, Kolkata, India, in 2007 and 2015, respectively.

His current research interests include pattern recognition, multimedia analysis, and computer vision.

Ananda S. Chowdhury (M'01–SM'16) received the Ph.D. degree in computer science from the University of Georgia, Athens, GA, USA, in 2007.

He is currently an Associate Professor with the Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India, where he leads the Imaging, Vision and Pattern Recognition Group. He was a Post-Doctoral Fellow with the Department of Radiology and Imaging Sciences, National Institutes of Health, Bethesda, MD, USA, from 2007 to 2008. He has authored or co-authored over 50 papers in leading international journals and conferences, in addition to a monograph in the Springer *Advances in Computer Vision and Pattern Recognition Series*. His current research interests include computer vision, pattern recognition, biomedical image processing, and multimedia analysis.

Dr. Chowdhury is a member of the IAPR TC on Graph-Based Representations. He currently serves as an Editorial Board Member of the *Pattern Recognition Letters*. His Erdős number is 2.