IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 16, 2017, accepted June 1, 2017, date of publication June 21, 2017, date of current version July 24, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2717998

Multi-Modal Visual Features-Based Video Shot Boundary Detection

SAWITCHAYA TIPPAYA[1,3], SUCHADA SITJONGSATAPORN[2], TELE TAN[3], MASOOD MEHMOOD KHAN[3], (Member, IEEE), AND KOSIN CHAMNONGTHAI[1], (Senior Member, IEEE)

[1]Department of Electronics and Telecommunication Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand
[2]Department of Electronic Engineering, Mahanakorn University of Technology, Bangkok 10530, Thailand
[3]Department of Mechanical Engineering, Faculty of Science and Engineering, Curtin University, Bentley Campus, Perth, WA 6102, Australia

Corresponding author: Kosin Chamnongthai (kosin.cha@kmutt.ac.th)

ABSTRACT One of the essential pre-processing steps of semantic video analysis is the video shot boundary detection (SBD). It is the primary step to segment the sequence of video frames into shots. Many SBD systems using supervised learning have been proposed for years; however, the training process still remains its principal limitation. In this paper, a multi-modal visual features-based SBD framework is employed that aims to analyze the behaviors of visual representation in terms of the discontinuity signal. We adopt a candidate segment selection that performs without the threshold calculation but uses the cumulative moving average of the discontinuity signal to identify the position of shot boundaries and neglect the non-boundary video frames. The transition detection is structurally performed to distinguish candidate segment into a cut transition and a gradual transition, including fade in/out and logo occurrence. Experimental results are evaluated using the golf video clips and the TREC2001 documentary video data set. Results show that the proposed SBD framework can achieve good accuracy in both types of video data set compared with other proposed SBD methods.

INDEX TERMS Cut transition detection, gradual transition detection, golf video analysis, logo transition detection, transition pattern analysis, video shot boundary detection.

I. INTRODUCTION

With the advancement and the popularity of social media technology, usage of digital video uploading has increased at a phenomenal rate resulting in a significant number of video databases. In general, it is difficult for users to manage this video data, especially to search for some specific video events from a large video database. Manual searching consumes more time and takes more effort from users to retrieve the desired event; for example, a long play video consists of multiple events that maybe of interest to only a small number of activities. Alternatively, it has a great benefit for users performing a semantic search that directly accesses some desired specific content, instead of querying the video from the massive video database, using the name of each video clip. Therefore, the research into video indexing, browsing and retrieving, known as Content Based Video Retrieval (CBVR), has been widely studied to help users achieve semantic searching [1], [2].

Semantic video analysis aims to relate language texts to abstract the visual representation of the video content. The challenging task here is to reduce the gap between the humans and the automatic search engine, represented as high-level semantic meanings and low-level information. The bottom-up searching approach [3] is useful when the user knows what to look for, thus increasing the demand of an efficient semantic indexing. In general, the structure of video content can be arranged as scenes, shots and frames [4]. Frames are the smallest unit of the video, whereas many frames constitute shots. Consequently, video shots (segmented video sequences) are considered to be the basic units to illustrate the video content, and the first pathway, to the high-level semantic indexing (annotation) and retrieval tasks. Therefore, it is essential to analyse the structure of video data from the shot level, to obtain an accurate video indexing system (i.e. event detection), represented in the higher semantic unit. The essential process here is the video shot boundary detection (SBD), also called video temporal segmentation. SBD aims to segment video sequences into many shots where these video shots are composed of frames conducted by a single camera operation. The tool also locates the position of the shot boundary and separates that shot boundary into various transition types (i.e. cut and gradual transition). It has received significant attention, which resulted in many approaches being proposed over the years [1]–[4].

Considering the SBD framework that has been proposed, it can be comprised of three modules, i.e.: the visual representation (frame-level feature extraction), the dissimilarity/similarity measures of visual content represented in the form of discontinuity/continuity signals and the classification (shot identification) of signals to transition types [1], [5]. Video frame features can be adopted by using various low-level feature extraction methods (i.e. color histogram, edge, and motion). The content signal showing the frame-pair difference can also be measured by various distance methods (i.e. Euclidean, Cosine, and Pearson's correlation). The shot identification or shot classification can be achieved by implementing the statistical machine learning methods (supervised or unsupervised), or the procedure based identification (PBI) method (rule-based classifiers). Most early SBD works that adopted a supervised learning approach employed a state-of-the-art discriminative classifier, such as K-Nearest Neighbour (KNN) and support vector machines (SVMs) to perform the transition detection process. The SBD result is very promising when using this approach, however, the speed and size of the training and testing data, and kernel selection, is one of the disadvantages. Moreover, the balance of the transition type sample, in the real implementation, is a significant limitation, especially when the special transition effect (i.e. logo) is added into the video content. In [6], the authors employ Singular Value Decomposition (SVD), with Hue Saturation Value (HSV) histogram, to propose a low computational complexity SBD scheme. The candidate segment selection using adaptive threshold is implemented, which can speed up the detection because it can eliminate the non-boundary video sequences. This method provides a high-speed transition detection. However, the result shows that it still needs improvement in the detection accuracy.

The overall performance of SBD systems can also rely on the effect of color variation, rapid changes of object movement, brightness, and special effects such as logo, and camera recording techniques in filmmaking. The SBD scheme using Walsh-Hadamard Transformation (WHT) is proposed in [7] to reduce these influences. This system has been focused on feature extraction, and feature fusion, to obtain a significant similarity of visual content. The transition detection is then simply performed by the PBI method using one global threshold selection. The results show that the detection accuracy can be improved by extracting more features from different domains. However, the frame feature is extracted using block calculation on every video frames. Therefore, it may increase the overall computational time. Another feature that can represent the object movement effect, while still tolerating the color variation, is the Speeded Up Robust Feature (SURF). The matching score can also measure the similarity between two video frames [8], [9]. The combination of the global color histogram and SURF has resulted in excellent detection performance as proposed in [9]–[11].

As an improvement on previous SBD systems, in this paper, we propose an SBD framework that captivates both cut and gradual transition. It is suggested in [6] that it is not necessary to process the whole video sequence to locate shot boundaries. However, shot boundaries usually present in a non-structured pattern. Hence, segmenting video frames using the method in [6] further requires frame adjustment to determine the candidate location. Unlike the previous method, we overcame this drawback by proposing the new method to collect video frames, without using threshold units, which results in potentially increasing the detection accuracy. We separate the process into three main stages: candidate segment selection, discontinuity signal representation, and the transition type detection. The objective of the first step is to find the group of consecutive frames, which represent any changes of video content as much as possible. Therefore, the discontinuity signal is calculated based on the SURF matching score and RGB histogram cosine distance value. By increasing the inter-frame distance, the signal can be enhanced resulting in the collection of most of the potential SBD segments. Cut and gradual (dissolve, fade in/out, logo) transitions are finally detected by using two particular PBI that directly analyse the inverse of similarity value of the SURF features and color histogram. Experiments on sport video data and TRECVID dataset show that the proposed SBD framework can provide high accuracy for detecting shot transitions.

The paper is constructed as follows. The recent developed SBD schemes over the past years are reported in Section II. An overview of the proposed SBD framework is described in Section III. In Section IV, we introduce dataset and evaluation criteria with our proposed method. Section V demonstrates the experimental results using sports and documentary video datasets. The discussion of the parameter selection and the experimental results are described in VI. Finally, the conclusion is presented in Section VII.

II. RELATED WORKS

Numerous SBD methods have been proposed with various concepts. However, the overall framework usually starts by extracting the visual feature, constructing the continuity signal or discontinuity signal, and classifying the variation of content difference to the shot transition, regardless the type of video dataset, to evaluate the superiority of the system. In this section, some advanced approaches will be categorised, according to the formal structure of SBD, to analyse the pros and cons of various techniques in each stage.

A. VISUAL FEATURE EXTRACTION OF VIDEO SEQUENCES

The use of visual feature extraction in an SBD scheme, can be categorised into different groups such as: a pixel-based [7], [12]–[14], histogram based, edge based, motion vector, compressed domain feature [15], [16], descriptor based, multiple features based [17], [18] and combined features based [9], [19]–[22]. One common approach to representing the visual information of video frames is to extract the low-level feature of each video frame. This feature describes the image without shape information (spatial relationship) or distinct regions. It can be extracted in two main types: global feature and local feature.

For an SBD system, the global feature difference extracted from color histogram can be used. This method includes RGB color histogram [6], [12], [18], [23]–[25], intensity histogram [19], HSV histogram [6], [21], Illumination-invariant chromaticity histogram [26], and so on. color histograms are less sensitive to a small camera or object movement, due to the property that does not incorporate the spatial information of color variation. However, their drawback is the inability to distinguish two shots within the same scene, and they are more sensitive in a large camera or object movement, due to rapidly changed color information.

On the contrary, the local feature aims to measure specific properties of the image, for instance: edge-based features [19], [27], [28], texture energy [17], motion estimation [17], [19], entropy [8], scale invariant feature transform (SIFT) [18], [29]–[34], and SURF features [8], [9], [35]. Some approaches use color coherence [36] and luminance [14] to obtain the spatial arrangement of colors in the image. These local features are more tolerant of illumination changes and small movements than global features, but they also have higher computational complexity. The features, as mentioned earlier, mainly rely on the properties of images (frames).

The middle-level features can be constructed by the low-level features such as dominant color and motion vector [37] for use in a sports video. Linear image transformation using WHT is proposed in [7] to extract a basis image for formulating a feature vector, which can represent the color and motion change. Some dimensionality reduction techniques were also proposed to map the raw feature onto the smaller dimensional vector, while maintaining the temporal characteristic of video content. These methods include QR-Decomposition [23], SVD [6], [38], and Independent Component Analysis (ICA) [26], [39], and Adaptive Locality Preserving Projections (ALPP) [20].

B. CONSTRUCTING THE VISUAL CONTENT FLOW SIGNAL OF VIDEO SEQUENCES

The objective of SBD is to find the location of consecutive frames that identify the transition between shots. From this perspective, we can extract the temporal characteristic of video contents regarding the similarity (continuity) signal or dissimilarity (discontinuity) signal. These signals can be constructed from the difference value between either adjacent features or two features within l inter-frame distance [25]. To measure a dissimilarity value between frames, some of the novel distance methods include pixel-wise distance [14], Euclidean distance [26], Chi-square distance [25], City block distance [7], the Bhattacharya distance [9], and histogram intersection [17]. In contrast, the similarity measure between video frames includes Pearson's correlation coefficient [11], [18], mutual information [40], normalised correlation [9], and Cosine similarity [6]. For local descriptors, such as SIFT and SURF, the matching score can be used to measure the similarity [8], [9], [33], [34].

Some of the SBD approaches show that the SBD framework may not rely on only one feature to discriminate the frame difference due to its trade-off of several features. Therefore, feature weighting based on machine learning such as Fuzzy logic [15], Adaboost [16] and Naive Bayes [7] are proposed to find the weighted similarity/dissimilarity value. However, these methods require a learning process, which becomes one of their limitations. Another simple approach to combine two features is also proposed in [9], which still provides good SBD accuracy.

C. SHOT BOUNDARY IDENTIFICATION SCHEME

The 1-D signal representing similarity/dissimilarity between frames can be used to detect the boundary (transition) in video sequences. The objective of the SBD process is to classify all similarity/dissimilarity values into each transition types (cut, gradual, and others). Because SBD needs to treat all values to find where the boundary occurs exactly between two video shots, it can be considered time-consuming. To overcome this problem, some researchers suggest pre-processing the long video sequences, and then finding the smaller segment of video frames that potentially contain shot boundaries. In [6] and [14], the video sequence is partitioned into n frame segments and then calculates the distance value between the first and end frames. If the distance value is higher than the predefined threshold, the segment is considered as the candidate for the respective SBD. These methods show that they can reduce the computational time necessary. However, the accuracy of transition detection still needs improvement for both cut and gradual transition detection.

According to several approaches during the past decades, the shot identification scheme can be classified into the statistical machine learning-based approach, rules-based approach. The statistical machine learning-based classification for the SBD scheme includes supervised learning and unsupervised learning. SVMs is also one of the most commonly used for SBD [15], [21], [22], [25], [32], [36], [37], [41] due to its effectiveness in dealing with a broad range of features, and giving a good detection result. Classifiers such as Multilayer Perceptron (MLP) network [19] and another extension of Naive Bayes classification in [39] are also used in SBD systems. The main advantage of the supervised approach is that it does not involve the threshold selection to justify whether the value is a shot or non-shot. Furthermore, the detection performance may be improved by combining multiple features. However, it needs a training dataset that contains a good balance between positive and negative samples (shot or no shot). Unsupervised learning approaches are also proposed, such as methods in [26], [29], and [40] to avoid the training process. However, it does not perform as well in SBD compared to the

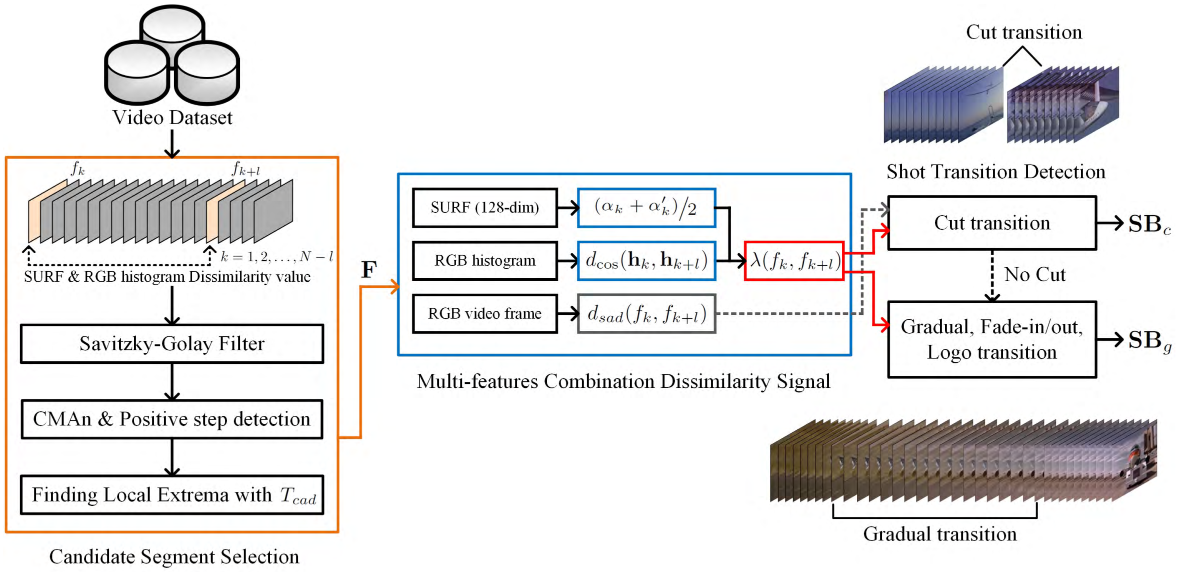
Video Dataset
f_k
f_{k+l}
$k = 1, 2, \ldots, N-l$
SURF & RGB histogram Dissimilarity value
Savitzky-Golay Filter
CMAn & Positive step detection
Finding Local Extrema with T_{cad}
Candidate Segment Selection
F
SURF (128-dim)
$(\alpha_k + \alpha'_k)/2$
RGB histogram
$d_{\cos}(\mathbf{h}_k, \mathbf{h}_{k+l})$
RGB video frame
$d_{sad}(f_k, f_{k+l})$
$\lambda(f_k, f_{k+l})$
Multi-features Combination Dissimilarity Signal
Cut transition
Shot Transition Detection
Cut transition
SB_c
No Cut
Gradual, Fade-in/out, Logo transition
SB_g
Gradual transition

FIGURE 1. Overview of the proposed multi-modal visual features-based video shot boundary detection.

supervised learning with regard to recognising transition patterns.

A rules-based approach is specifically developed to detect video shot boundaries by incorporating the characteristics of transition behaviours. This method can be designed using decision rules [16], a fuzzy system [17], [24], or PBI [27]. It can be designed to use the threshold mechanism to detect the occurrence of the shot transition. This threshold mechanism includes global [7], [8], [12], [13], [23], adaptive [26], [28], [31], or global and adaptive combined [6], [11], [14]. Global threshold-based systems select the same value over the video sequences, to identify the transition. However, one threshold value may not be suitable for various categories of video. Therefore, the adaptive threshold is proposed to eliminate the drawback by finding an appropriate threshold value, to conform to the temporal characteristic of video content locally. It is suggested that global threshold method may not perform well in the detection and the adaptive threshold mechanism may be difficult to determine [1], [5]. However, according to recent approaches (i.e. the SBD methods in [6] and [7]), these methods have shown that the overall performance may not rely on the threshold mechanism selection, but rather the chosen features and the shot identification process.

From the above reviewed SBD methods, great techniques have been proposed with various concepts, resulting in good detection accuracy, especially the cut transition. However, it is still a challenging task to improve the detection accuracy of SBD systems in both transition types, due to some disturbances caused by rapid movements, and advances in video editing technologies. Since the visual features have been developed for years, we found that various features can lead to good detection performance for each transition types. Moreover, the SBD system is considered as an essential pre-processing step, but to obtain the transition, it needs to process whole video sequences, and the detection result may still not provide a high percentage of detection accuracy at the same time. Hence, instead of looking for a transition from a long video sequence, in this paper, we aim to develop an SBD system that can improve the result, by methodically focusing the transition behaviours. Similarity and dissimilarity signal based system on both global and local feature extraction are calculated to detect the shot transition by our multi-modal visual features based approach. The proposed SBD framework is also evaluated with other proposed SBD methods, to illustrate the effectiveness of shot transition detection.

III. MULTI-MODAL VISUAL FEATURES BASED VIDEO SHOT BOUNDARY DETECTION

In this section, we introduce the concept of our SBD system using video frame features. Our proposed SBD framework is illustrated in Fig.1. At the first stage, we can query the potential video segment that contains the shot boundary to overcome the miss detection problem. The second step is to perform a quantitative feature extraction and construct the dissimilarity signal. Finally, the shot transition detection is conducted to detect a shot boundary and categorise it into a cut transition or a gradual transition. The following subsections describe our SBD system framework which comprises of three steps as aforementioned: candidate segment selection, multi-features combination dissimilarity signal and shot transition detection.

A. CANDIDATE SEGMENT SELECTION

The primary purpose of candidate segment selection is to reduce the processing time by eliminating many non-boundary frames from the video sequences. This concept was previously proposed in [6], [11], and [14] with an adaptive threshold mechanism to obtain a segment of video frames. However, it requires a parameter to formulate the threshold value and subsequent steps to find the group of video frames.

In our scheme, we aim to propose the candidate segment selection that requires less parameter. It is designed with the simple concept of analysing the behaviour of the temporal characteristic extracted from visual features, based on SURF descriptors and RGB histogram. It can be separated into two stages: feature extraction and local extrema calculation. The first stage is to perform a feature extraction to video frames. SURF with 128-dimensional feature vector (descriptor), is chosen to reduce the effects of rotation, illumination, and color variation in N video frame sequences. The RGB histogram is also extracted to obtain the histogram difference value. We can describe the similarity value based on the combined features between two frames f_k and f_{k+l} by multiplying the matching SURF score [8], [9] and cosine distance of RGB histogram as

$$\psi(f_k, f_{k+l}) = \left((\frac{\alpha_k + \alpha'_k}{2}) \cdot d_{\cos}(\mathbf{h}_k, \mathbf{h}_{k+l}) \right), \tag{1}$$

where $d_{\cos}(\mathbf{h}_k, \mathbf{h}_{k+l}) = \cos(\mathbf{h}_k, \mathbf{h}_{k+l})$ is the cosine similarity between two RGB color histograms, $\mathbf{h}_k$ and $\mathbf{h}_{k+l}$. $\alpha_k = M_k/|Q_k|$ and $\alpha'_k = M_k/|Q_{k+l}|$ denote the matching score. M_k is the number of matched keypoints (index pairs) between two frames. $|Q_k|$ and $|Q_{k+l}|$ are the number of descriptors (features) in the f_k and f_{k+l} frames, respectively. l denotes an inter-frame distance.

Small changes usually occur during the gradual transition, so setting a small l may not provide a good result. We can enhance the signal level by increasing l to retrieve all potential difference values. In the candidate segment selection, the inverse of $\psi(f_k, f_{k+l})$ value is used instead of (1) to cope with our scheme, which is given by $\lambda(f_k, f_{k+l}) = 1 - \psi(f_k, f_{k+l})$. Let λ be the discontinuity signal representing the temporal characteristic in the video. In this work, we also apply the Savitzky-Golay (SG) polynomial smoothing filter [42] with window length $w = 5$, to generate the smoothed signal $\widetilde{\lambda}$. SG filtering can smooth the noise from the signal, while preserving the original properties of the signal, in contrast to a moving-average filter.

Calculating the mean value over time can detect any changed content of the video. It is known as a running average or a cumulative moving average (CMAn). The smoothed discontinuity value between f_k and f_{k+l} frames based on SURF descriptors are denoted by $\widetilde{\lambda}(k)$ where $k = 1, 2, \ldots, N - l$. The running operation of the average values is to compute the mean value of each sample in the $\widetilde{\lambda}(k)$ concerning all the previous samples of $\widetilde{\lambda}(k)$, up until the current time. Let $\overline{\lambda}(k)$ be the cumulative average of $\widetilde{\lambda}(k)$; therefore, two cumulative average values are defined as

$$\overline{\lambda}(k) = \frac{\widetilde{\lambda}(1) + \ldots + \widetilde{\lambda}(k)}{k}, \tag{2}$$

$$\overline{\lambda}(k+1) = \frac{\widetilde{\lambda}(k+1) + k \cdot \overline{\lambda}(k)}{k+1}, \tag{3}$$

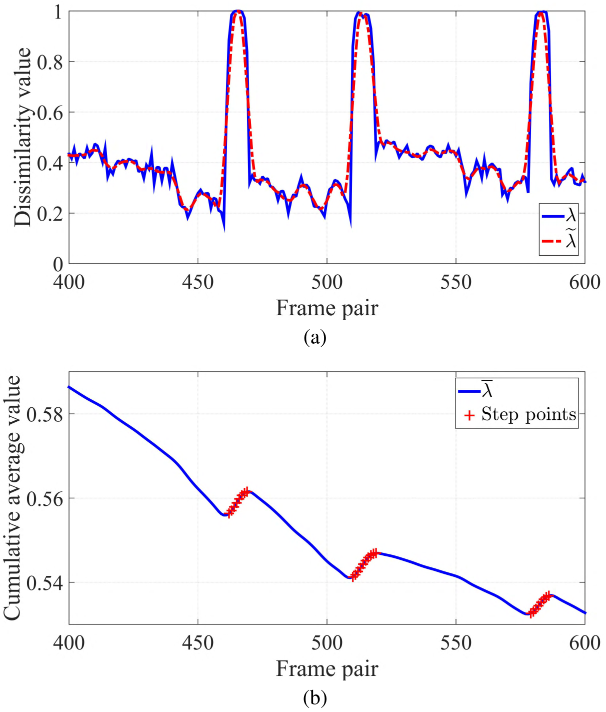

FIGURE 2. Example of qualified candidate frames using positive step detection. (a) The dissimilarity value λ and smoothed signal $\widetilde{\lambda}$ between f_k and f_{k+l} frames with inter-frame $l = 5$. (b) Cumulative moving average signal ($\overline{\lambda}$) and the positive step points. .

where $\overline{\lambda}(k)$ is the current cumulative average and $\overline{\lambda}(k+1)$ is the updated cumulative value as the new $\widetilde{\lambda}(k+1)$ value arrives. $\overline{\lambda}$ is the cumulative moving average signal.

Finally, we can use the advantage of the output characteristic to construct one signal that represents a changed video frame pair as a time-series data, $\overline{\lambda}$. The cumulative moving average signal, $\overline{\lambda}$ is supplied to the positive step detection. Let $\overline{\lambda}$ be a noisy time series data, and we can locate positive transient steps by calculating the first derivative of $\overline{\lambda}$, where the step height is the difference between $\overline{\lambda}(k)$ and $\overline{\lambda}(k+1)$ values, over a specified number of a data points denoted as k^{th} frame pair. Each positive step point denotes a pair of video frames within inter-frame distance l as p_k.

A result of our proposed candidate segment selection scheme can be illustrated in Fig.2. It shows that the positive step positions Fig.2(b) correspond to the dissimilarity value in Fig.2(a). Now, we can construct the start, and end frame of each group of positive step points producing $P_{j,l}$ where $j = 1, 2, \ldots, n_{cad}$ respectively. n_{cad} represents the total number of candidate segments. By using the implementation as above, the length of a candidate segment varies to the duration of the shot boundary. We also add l frames before and after the segment to ensure that the similarity/dissimilarity value does not fall at the edge of the segment. The candidate segment $P_{j,l}$ is defined as

$$P_{j,l} = \left\{ p_{k-l}, p_{k-l+1}, \ldots, p_{k+l+\hat{n}-1} \right\}, \tag{4}$$

where p_k is the k^{th} video frame pair and $\hat{n}$ is a number of positive points in the segment. The segment, which has more than

Algorithm 1 Candidate sub-segment selection.

Notation: $P = \{P_{j,l}, P_{(j+1),l}, \ldots, P_{n_{cad}}\}$: candidate segment of video frame pair, $\hat{n}$: number of positive points, $\psi^i_{j,l}$: raw similarity signal, $\widetilde{\psi}^i_{j,l}$: smoothed similarity signal, $\overline{\psi^i_{j,l}}$: mean value of $\psi^i_{j,l}$, T_{cad}: candidate threshold, w: filter window size, l: candidate inter-frame distance, l_c: adjacent inter-frame distance, i_{np}: number of video frame pairs, Y_{max} and Y_{min}: matrix store local maxima and local minima to compute $f_{\max L}, f_{\max R}, f_{min}$.

Require: $P_{j,l}, \psi^i_{j,l}, T_{cad}, w, l, l_c$.

Ensure: $F_{j,l}$: sub-segment output.

```
for $j = 1, 2, \ldots, n_{cad}$ do
    $i = [p_{k-l}\ p_{k-l+1}\ \ldots\ p_{k+l+\hat{n}-1}]^T$;
    $\psi^i_{j,l} = [\psi(i)\ \psi(i+1)\ \ldots\ \psi(i_{np})]^T$;
    if min $\psi^i_{j,l} \leq T_{cad}$ then
        Obtain $\widetilde{\psi}^i_{j,l}$ by performing SG-filter with window
    size $w$ to $\psi^i_{j,l}$.
        Find $Y_{max}$ and $Y_{min}$ matrix of size $(c_{max} \times 2)$ and
    $(c_{min} \times 2)$ by local extrema calculation with $T_{cad}$
        if $c_{min} \neq 0$ then
            for $jj = 1, 2, \ldots, c_{min}$ do
                $f^i_{min} = y_{min}(jj, 1)$;
                $f^i_{maxL}(jj) = \min_{m<f^i_{min}, m \in i} |f^i_{min} - f^i_{max}(m)|$;
                $f^i_{maxR}(jj) = \min_{m>f^i_{min}, m \in i} |f^i_{min} - f^i_{max}(m)|$;
            end for
            $f_{maxL} = [f^i_{maxL}(1)\, f^i_{maxL}(2)\ \ldots\ f^i_{maxL}(c_{min})]^T$;
            $f_{maxR} = [f^i_{maxR}(1)\, f^i_{maxR}(2)\ \ldots\ f^i_{maxR}(c_{min})]^T$;
            $F_{j,l} = \{f_{maxL}, f_{maxR} + l\}$;
        end if
    else if min $\psi^i_{j,l} > T_{cad}$ && min $\psi^i_{j,l_c} < 2T_{cad}$
  && $\overline{\psi^i_{j,l_c}} > 4T_{cad}$ then
        $F_{j,l} = \{p_{k-l}, p_{k+l+\hat{n}-1} + l\}$;
    end if
end for
```

w frame pair will be analysed. Otherwise, the segment will be discarded and considered as the non-boundary segment. $P_{j,l}$ can be a long video frame segment that may contain more than one shot boundary. Therefore, it is essential to find the sub-segment, which includes accurate start frame and end frames.

Let $\psi^i_{j,l}$ be the similarity signal of $P_{j,l}$ segment where $i = (p_{k-l}, p_{k-l+1}, \ldots, p_{k+l+\hat{n}-1})$. Local extrema calculation is performed to select two maxima points ($f_{\max L}$ and $f_{\max R}$), and at least one minima point that are lower than the threshold T_{cad} to ensure that a shot boundary presents in the current segment. The summary of the candidate sub-segment selection process is described in Algorithm 1. From the proposed candidate segment selection step, it should be denoted that more than one shot is collected in the same segment, depending on the signal representation. Also, there is no constraint on the length of transition because the local maxima and local minima are corresponding to the transition occurrence itself. Finally, the final candidate segment indexing the list of video frame pair is defined by

$$\mathbf{F} = \begin{pmatrix} F_{j,l} \\ F_{j+1,l} \\ \vdots \\ F_{N_{cad},l} \end{pmatrix}; \quad F_{j,l} = [f_{\max L}\ (f_{\max R} + l)]. \qquad (5)$$

According to (5), we also add l frame length to the segment to ensure that there are a sufficient number of video frames over the transition period. N_{cad} denotes the total number of video frame segment. By using the list of candidate segments $\mathbf{F}$, the only potential segments of shot boundary frames are collected for the subsequent SBD process, regardless of the length of the transition. The next step is to construct the signal for each video segment. At this stage, a different l value is obtained for each transition detected.

The cut (abrupt) characteristic is a sudden change between two adjacent frames, so the system requires a dominant similarity/dissimilarity value that effectively identifies the cut frames. Meanwhile, it has been found that SURF is a powerful feature to discriminate between two images, by matching their respective interest points. However, it can be insensitive to a broad movement of camera or object when using high l value results in the miss matched descriptors. In sports videos, the disappearance/appearance of the small effects such as logo, scoreboard, or texts, can affect the matching corresponding points between two frames. Moreover, the SBD framework has a parameter constraint when extracting features from a large number of video databases. This limitation may lead to a false matching score value. To conclude, we need the quantitative analysis on the segment that may contain these effects.

B. MULTI-FEATURES COMBINATION DISSIMILARITY SIGNAL

Visual feature extraction is the primary process leading to effective shot detection. In this paper, we propose a frame-based approach. Therefore, the boundary will be identified when its similarity value is below the predefined threshold during the candidate segment selection process. From the review of SBD methods in Section II, SURF and RGB color histograms perform well in the shot detection.

For our proposed system, shot boundaries are detected by using the discontinuity signal between two video frames, $\lambda(f_k, f_{k+l})$. The similarity signal used in shot transition detection is initially calculated based on SURF matching score, and RGB histogram, based on Cosine similarity. However, the limitation is that when calculating the similarity between two adjacent frames, it may produce an undefined value, or an outlier value resulting in misleading shot classification. Hence, it is important to extract additional features, which can discriminate two frames regardless of their having very similar color information. Consequently, the dissimilarity signal denoted as λ, can be effectively implemented in the shot transition detection.

The cut transition detection also obtains the sum of absolute difference (SAD) value between two video frames, to define the shot boundary. However, the normalised SAD within the candidate segment is used instead of the un-normalised value, because we need the value in the range [0, 1]. The frame difference based SAD can be sensitive to some disturbances caused by illumination or flashlight. In this case, we make use of these weaknesses as the false shot detection when performing the cut detection process. Let $d_{sad}(f_k, f_{k+l})$ be the normalised SAD between f_k and f_{k+l} frames, which is calculated by

$$d_{sad}(f_k, f_{k+l}) = \frac{1}{|\mathbf{M}|} \sum_{\mathbf{u} \in \mathbf{M}} |f_k(\mathbf{u}) - f_{k+l}(\mathbf{u})|. \tag{6}$$

where $f_k(\mathbf{u})$ and $f_{k+l}(\mathbf{u})$ denote the pixel value at coordinates $\mathbf{u}$ of f_k and f_{k+l} frames, respectively. $\mathbf{M}$ is the pixels of the overall video frame, and $|\mathbf{M}|$ represents the total number of pixels in a video frame. From (6) it should be denoted that we calculate the sum of difference for all R, G, and B values.

C. SHOT TRANSITION DETECTION

In this step, it should be noted that the dissimilarity signal λ of each segment in $\mathbf{F}$ obtained from (5) will be used instead of similarity, to cope with our proposed SBD criteria. In general, the cut (abrupt) transition is a significant change in a triangle shape in a candidate, when choosing an inter-frame distance as $l_c = 1$. From the purpose of a multi-modal features based approach, cut transition is the point where a high value presents. Our assumption is that if there is a significant change, $\lambda(f_k, f_{k+l_c})$ should be as high as possible, caused by the multiplication result. Now, let $\lambda_{l_c}^{j1}$ be the dissimilarity signal, $\hat{d}_{sad}^{j1}$ be the normalised SAD signal for each $(j1)^{th}$ segment where $j1 = 1, 2, \ldots, N_{cad}$, and N_{cad} is the total number of video frame segment in $\mathbf{F}$.

For each segment, the cut transition occurs when the dissimilarity value is high. In contrast, values during other transitions such as gradual, or no transition can be low. Hence, we can use the peak detection to identify a segment of video frames that contains a cut transition when its local maxima are above the cut transition threshold, T_c. The cut transition will be modelled into a criterion depending on the E_α; the total number of the first derivative of SURF matching score that equals to 1 or ∞, and non-cut effect; $E_{sad} = \sum(\hat{d}_{sad}^{j1} > T_c - 0.1)$. We define that E_{sad} is no more than 2 for each cut transition segment, which limits the number of an expected cut boundary.

The conditions are made to prevent the false detection caused by the failure of the SURF feature extraction and non-cut effect. If any candidate segment does not meet the E_α criteria, the segment will be discarded from the cut detection. Consequently, the cut transition is declared when one of the following conditions is satisfied:

$$\mathcal{CT} \begin{cases} \left|\Delta\hat{d}_{sad,-}^{j1}\right| > \beta T_c \;\&\&\; \left|\Delta\hat{d}_{sad,+}^{j1}\right| > \beta T_c \\ \left|\Delta\hat{d}_{sad,-}^{j1}\right| < \beta T_c \;\&\&\; \left|\Delta\hat{d}_{sad,+}^{j1}\right| > T_c \\ \left|\Delta\hat{d}_{sad,-}^{j1}\right| > T_c \;\&\&\; \left|\Delta\hat{d}_{sad,+}^{j1}\right| < \beta T_c \end{cases}, \tag{7}$$

where $\left|\Delta\hat{d}_{sad,-}^{j1}\right| = \left|\hat{d}_{sad}^{j1}\left(\Gamma_\lambda(p)\right) - \hat{d}_{sad}^{j1}\left(\Gamma_\lambda(p) - 1\right)\right|$ and $\left|\Delta\hat{d}_{sad,+}^{j1}\right| = \left|\hat{d}_{sad}^{j1}\left(\Gamma_\lambda(p)\right) - \hat{d}_{sad}^{j1}\left(\Gamma_\lambda(p) + 1\right)\right|$ are the difference of $\hat{d}_{sad}$ value between each video frame pair with the peaks of $\lambda_{l_c}^{j1}$ signal. Γ_λ and Γ_{sad} are the peak vector of λ_{l_c} and $\hat{d}_{sad}$, respectively. N_p is a total number of detected peak where $p = 1, 2, \ldots, N_p$. However, if E_α is greater than 1, the cut transition will be declared using the similar criteria in (7) but refer to $\Gamma_{sad}(\hat{p})$ instead of $\Gamma_\lambda(p)$. It should be noted that the performance of cut transition of our proposed system relies on parameters β and T_c. More discussion on the parameter selection will be discussed in Section VI.

From the cut transition detection, any segment without a cut transition will be listed as the gradual candidate segment. According to Algorithm 1, the small segment, which belongs to the gradual segment may be collected separately. These overlapped segments will be merged before performing the gradual detection. Moreover, our candidate segment selection does not depend on the length of the transition; therefore the segment may contain more than one shot. We apply the peak detection to $\tilde{\lambda}_{l_c}^{j1}$ again. If a minimum value between two detected peaks is less than 0.5, the candidate segment is divided into smaller multiple segments. Finally, the segment must contain more than $2w$ video frame pairs to proceed to the gradual detection process.

Gradual transition detection is subsequently applied to the rest of the candidate segments that do not meet the cut transition condition. In this step, we define the SG-filter input dissimilarity signal as $\tilde{\lambda}_{l_g}^{j2}$. l_g is the inter-frame distance parameter that directly influences the performance of gradual detection and overall shot detection. More discussion on this parameter is described in Section VI. Peak detection is applied to detect the maximum value of $\tilde{\lambda}_{l_g}^{j2}$ that is greater than T_g. Let $\Gamma_{\tilde{\lambda}_g}$ be the peak location (video frame pair) and $\phi_{\tilde{\lambda}_g}$ denotes the highest value of $\tilde{\lambda}_{l_g}^{j2}$. Consequently, the gradual transition is revealed when one of the following criteria is satisfied:

$$\mathcal{GT} \begin{cases} \left|\phi_{\tilde{\lambda}_g} - \min \tilde{\lambda}_{l_g}^{j2,LT}\right| > T_g \\ \quad \&\& \left|\phi_{\tilde{\lambda}_g} - \min \tilde{\lambda}_{l_g}^{j2,RT}\right| > T_g \\ \left|\phi_{\tilde{\lambda}_g} - \min \tilde{\lambda}_{l_g}^{j2,LT}\right| < T_g \\ \quad \&\& \left|\phi_{\tilde{\lambda}_g} - \min \tilde{\lambda}_{l_g}^{j2,RT}\right| > 2T_g \\ \left|\phi_{\tilde{\lambda}_g} - \min \tilde{\lambda}_{l_g}^{j2,LT}\right| > 2T_g \\ \quad \&\& \left|\phi_{\tilde{\lambda}_g} - \min \tilde{\lambda}_{l_g}^{j2,RT}\right| < T_g \end{cases}, \tag{8}$$

where $\min \tilde{\lambda}_{l_g}^{j2,LT}$ and $\min \tilde{\lambda}_{l_g}^{j2,RT}$ denote the minimum value of the first and second part of the gradual segment at $\Gamma_{\tilde{\lambda}_{g,LT}}$ and $\Gamma_{\tilde{\lambda}_{g,RT}}$ video frame pair, respectively. To eliminate the false detection, λ_{LT} and λ_{RT} are calculated to ensure that a shot boundary exists in the current gradual segment. The summary of our proposed shot boundary detection is illustrated in Algorithm 2.

IV. DATASET AND EVALUATION CRITERIA

To evaluate the performance of the proposed SBD algorithm, we have experimented with our algorithm, over the set of professional golf video frame sequences. The video dataset contains various transition effects including cut, gradual (fade-in, fade-out), and logo transition.

A. VIDEO DATASET CHARACTERISTICS

Six video clips (140 minutes) of the golf video frame sequences [43] were chosen as our sport video data, which contain cut, gradual and other transition (e.g. logo and scoreboard changes). However, the logo and scoreboard occurrence have been considered a part of the shot boundary in our experiment. Therefore, we have assigned these effects as our gradual transition ground-truth. The golf video sequences contain 432 cut transitions and 316 gradual transitions in total.

Our experiments also aim to ensure that the proposed algorithm is evaluated in comparison to the recently-proposed methods. Hence, we have selected four video sequences from the video dataset provided by the US National Institute of Standards (NIST) benchmark dataset [44] for benchmarking the proposed SBD system. The results can be described as both cut and gradual transition detections obtained from the reported results. The dataset is TREC2001 which includes: NASA 25th Anniversary, Airline Safety and Economy, Perseus Global Watcher. These video sources are publicly available on the "Open-Video Project" website [45]. TRECVID collection website [45] provides the ground-truth of the video dataset for all transition types. There are 414 shot transitions in total, 219 cut transitions, and 195 gradual transitions that consist of fade-in, fade-out, wipe and dissolve types. The detail of the video sequences can be illustrated in Table 1.

B. PERFORMANCE EVALUATION CRITERIA

To illustrate the efficiency of the proposed SBD scheme, we also adopt the similar measurement to other frameworks using the following criteria:

$$\text{Recall } (R) = \frac{N_c}{N_c + N_m} \times 100, \tag{9}$$

$$\text{Precision } (P) = \frac{N_c}{N_c + N_f} \times 100, \tag{10}$$

$$\text{F-measure } (F1) = \frac{2RP}{R + P} \times 100, \tag{11}$$

where N_c is the number of correctly detected transitions, N_m is the number of missed detected transitions, and N_f is the number of false transitions detection. Recall (R) is the rate of

Algorithm 2 Shot transition detection.

Notation: F: candidate segment, N_{cad}: number of candidate segments, w: filter window size, l_c: cut inter-frame distance, λ: dissimilarity signal, $\hat{d}^{j1}_{sad}$: normalised d_{sad}, Γ_λ: peak vector of λ_{l_c}, Γ_{sad}: peak vector of $\hat{d}_{sad}$, ϕ_λ: detected peak value of λ, ϕ_{sad}: detected peak value of $\hat{d}^{j1}_{sad}$, T_c: cut candidate threshold, β: threshold for normalised d_{sad} signal, N_p, $N_{\acute{p}}$: number of detected peaks, E_α: predefined parameter for SURF matching score, E_{sad}: predefined parameter for $\hat{d}^{j1}_{sad}$, $\mathcal{CT}$: cut transition condition, $\mathbf{SB}_c$: cut transition, G: gradual candidate segment, N_g: number of gradual candidate segments, l_g: gradual inter-frame distance, $\tilde{\lambda}^{j2}_{l_g}$: smoothed dissimilarity signal, $\Gamma_{\tilde{\lambda}_g}$: peak vector of $\tilde{\lambda}_g$, $\phi_{\tilde{\lambda}_g}$: detected peak value of $\tilde{\lambda}_g$, T_g: gradual transition threshold, λ_{LT}, λ_{RT}: calculated value to ensure a gradual transition exists, $\mathcal{GT}$: gradual transition condition, $\mathbf{SB}_g$: gradual transition.

```
Require: $F, w, \lambda_{l_c}, \lambda_{l_g}, d_{sad}, T_c, \beta, l_c, T_g, l_g$
Initialization: $c = g = 0$.
Ensure: $\mathbf{SB}_c$ and $\mathbf{SB}_g$
for $j1 = 1, 2, \ldots, N_{cad}$ do
    $c1 = 0; f_s = F(j1, 1); f_e = F(j1, 2);$
    if $|f_s - f_e| + 1 > 2w$ then
        Perform cut transition detection.
        Calculate $\hat{d}^{j1}_{sad}$.
        Find $\Gamma_\lambda(p)$ where $\phi_\lambda(p) > T_c, p = 1, 2, \ldots, N_p$
        Find $\Gamma_{sad}(\acute{p})$ where $\phi_{sad}(\acute{p}) > \beta T_c, \acute{p} = 1, 2, \ldots, N_{\acute{p}}$
        Calculate $E_\alpha$ and $E_{sad}$
        if $E_\alpha = 0$ && $E_{sad} \leq 2$ then
            for $p = 1, 2, \ldots, N_p$ do
                if one of $\mathcal{CT}$ condition in (7) is true then
                    $c1 = c1 + 1; c = c + 1;$
                    $\mathbf{SB}_c(c, :) = [\Gamma_\lambda(p)\ \Gamma_\lambda(p) + 1];$
                end if
            end for
        else if $E_\alpha \neq 0$ && $E_{sad} \leq 2$ then
            for $\acute{p} = 1, 2, \ldots, N_{\acute{p}}$ do
                if $\lambda(\Gamma_{sad}(\acute{p})) > T_c$ && one of $\mathcal{CT}$ is true then
                    $c1 = c1 + 1; c = c + 1;$
                    $\mathbf{SB}_c(c, :) = [\Gamma_{sad}(\acute{p})\ \Gamma_{sad}(\acute{p}) + 1];$
                end if
            end for
        end if
        if $c1 = 0$ then
            Declare $F(j1, :)$ as the gradual candidate segment.
        end if
    end if
end for
Combine the overlapped segment
Check if more than one gradual transition may exist in the segment.
Generate the final gradual candidate segment $G$.
for $j2 = 1, 2, \ldots, N_g$ do
    $f_s = G(j2, 1); f_e = G(j2, 2);$
    Compute $\tilde{\lambda}^{j2}_{l_g}$
    Find $\Gamma_{\tilde{\lambda}_g}$ where $\phi_{\tilde{\lambda}_g} > T_g$.
    if $\Gamma_{\tilde{\lambda}_g} \neq 0$ then
        Compute $E_\alpha$.
        if one of $\mathcal{GT}$ condition in (8) is true && $E_\alpha = 0$ then
            Calculate $\lambda_{LT} = \lambda(\Gamma_{\tilde{\lambda}_{g,LT}}, \Gamma_{\tilde{\lambda}_g})$.
            Calculate $\lambda_{RT} = \lambda(\Gamma_{\tilde{\lambda}_g}, \Gamma_{\tilde{\lambda}_{g,RT}})$.
            if $\lambda_{LT} > 0.9 \parallel \lambda_{RT} > 0.9$ then
                $g = g + 1;$
                $\mathbf{SB}_g(g, :) = [\Gamma_{\tilde{\lambda}_{g,LT}}\ \Gamma_{\tilde{\lambda}_g}\ \Gamma_{\tilde{\lambda}_{g,RT}}];$
            end if
        end if
    end if
end for
```

TABLE 1. **Video sequences used and their respective descriptions.**

Video	Frames	Minutes	Transition types			Sources
			CT	GT	Total	
G1	43030	23.91	78	68	146	Golf Video
G2	45169	25.09	64	31	95	
G3	41200	22.89	60	52	112	
G4	44045	24.47	85	50	135	
G5	41213	22.9	45	67	112	
G6	43194	24	100	48	148	
D2	16586	9.21	42	31	73	TREC2001
D3	12304	6.84	39	64	103	
D4	31389	17.44	98	55	153	
D6	13648	7.58	40	45	85	
Total	331778	184.33	651	511	1162	

CT: Cut (abrupt) transition, GT: Gradual transition.

missed detections and is high when missed detections are low. Precision (P), on the other hand, represents the false detection rate. The overall performance can be described by the F-measure ($F1$), which is a measure considering both recall and precision value. Our experimental results are performed by Matlab R2015a on a Windows 7 Professional with Intel Xeon 2.6GHz CPUs.

V. EXPERIMENTAL RESULTS

Our proposed SBD method is evaluated over the available dataset under two conditions to show the performance regarding the percentage of detection accuracy. Firstly, we performed our approach and SBD system in [9] with sports video dataset. The available online software was obtained from [46] and we carried out the test based on the provided default parameter. With the limitation on publicly available experiments on golf video, in this work, we can compare the detection performance with the method mentioned above. Secondly, the evaluation was performed over TREC2001 dataset and compares our results with the results previously reported in the literature.

A. SPORT VIDEO DATASET EVALUATION

The experiment evaluated over golf video is shown in Table 2. The results indicate that our cut detection scheme provides an excellent detection performance for both precision and recall. Our proposed cut detection, therefore, can provide very promising detection performance regardless of the type of video dataset. For gradual detection, the results include all dissolve transitions, logo transitions and scoreboard changes, because sports videos has large camera/object movements, and rapid background changes. The precision of gradual detection is, therefore, lower than in cut detection. In this paper, it should be noted that the flyball camera movement effect is not identified as a shot transition, because the frames before and after this effect should be recognised in the same semantic event as ‘‘tee-off shot’’ or ‘‘fairway approaching shot’’.

To compare the detection accuracy over the golf video dataset, we compared the result of our proposed scheme, and the global and local descriptors approach proposed in [9]. Based on our empirical experiments, the recommended T_g and l_g are 0.35 and 1, respectively. The available software from [46] does not provide the result by each transition type, therefore in this paper, the comparison of detection accuracy is listed by the performance of overall detection (detecting all transitions as the shot boundary).

From the results in Table 3, it is revealed that the proposed method performs better than the method in [9] on: precision, recall and $F1$-measure. Our proposed SBD also yields a better precision value, which concludes that the false shot elimination performs well by calculating λ and E_α over the gradual segment. The missed shot boundary, such as in G5, failed to detect because the video sequences between two shots are in the similar color histogram. We observe that in the golf video, some of the gradual transitions occurs before or after a large camera/object movement where the characteristic of $\lambda(f_k, f_{k+lg})$ may not effectively distinguish the dissolve change from these disturbances caused by the camera/object movements.

B. TREC2001 DATASET EVALUATION

To show the superiority of the proposed SBD scheme, the precision, recall and $F1$ measures, of four videos from the TREC2001 video collection, are compared with our proposed framework and methods in [6] and [7], respectively. Table 4 shows that our cut detection outperforms other methods with a similar parameter selection of golf video dataset. We also test the proposed scheme with two more video datasets from the TREC2001 collection [45]: anni005 and anni009, to compare the cut detection results with the method in [19], and [47], respectively. Our proposed SBD framework has the highest $F1$ performance with an average cut detection at 98.0%, while the method in [19] is 93.0% and 97.6% in [47].

From the gradual detection results in Table 4, our scheme does not perform well in the gradual transition as opposed to the cut detection. We observe that most of the false detection is caused by the effect of significant changes in background and object movement. Video frame features cannot tolerate these disturbances. Missed gradual shot boundaries in the video frame pair can occur on two occasions: missed collected in the candidate segment selection step and failed detection during the gradual transition detection. However, from the result in Table 5, our SBD method can yield the highest accuracy considering the overall detection performance when choosing the $l_g = 3$ and $T_g = 0.15$, respectively.

The overall performance generated by the software from [46] shows that the global and local descriptors do not perform well for documentary type videos, in contrast to the result evaluated using golf videos. The result supports our observation that the length of gradual transition has a high impact on the performance of overall gradual detection. Furthermore, the false elimination process, which is proposed in [9] and performed using [46], is calculated based on the behaviours of similarity signals within a certain number of

TABLE 2. Performance of the proposed method evaluated over golf video sequences ($l = 5, T_{cad} = 0.1, l_c = 1, T_c = 0.9, \beta = 0.6, l_g = 1, T_g = 0.35$).

Video	Cut transition			Gradual transition			Overall transition		
	P	R	F1	P	R	F1	P	R	F1
G1	100.0	98.7	99.4	91.9	83.8	87.7	96.4	91.8	94.0
G2	98.4	98.4	98.4	95.7	71.0	81.5	98.9	90.5	94.5
G3	100.0	98.3	99.2	91.7	84.6	88.0	96.3	92.0	94.1
G4	100.0	100.0	100.0	84.3	86.0	85.1	94.1	94.8	94.5
G5	100.0	95.6	97.7	93.3	83.6	88.2	96.1	88.4	92.1
G6	100.0	96.0	98.0	77.6	79.2	78.4	92.4	90.5	91.5
Average	99.7	97.8	98.8	89.1	81.4	84.8	95.7	91.3	93.4

TABLE 3. Comparison of overall transition detection of the proposed method with the system proposed in [9] ($T_c = 0.9, \beta = 0.6, l_c = 1, T_g = 0.35, l_g = 1$).

Video	SBD system in [9]			Proposed SBD system		
	P	R	F1	P	R	F1
G1	79.0	84.9	81.8	96.4	91.8	94.0
G2	83.2	93.7	88.1	98.9	90.5	94.5
G3	83.8	87.5	85.6	96.3	92.0	94.1
G4	84.7	90.4	87.5	94.1	94.8	94.5
G5	94.0	83.9	88.7	96.1	88.4	92.1
G6	87.4	93.9	90.6	92.4	90.5	91.5
Average	85.3	89.1	87.0	95.7	91.3	93.4

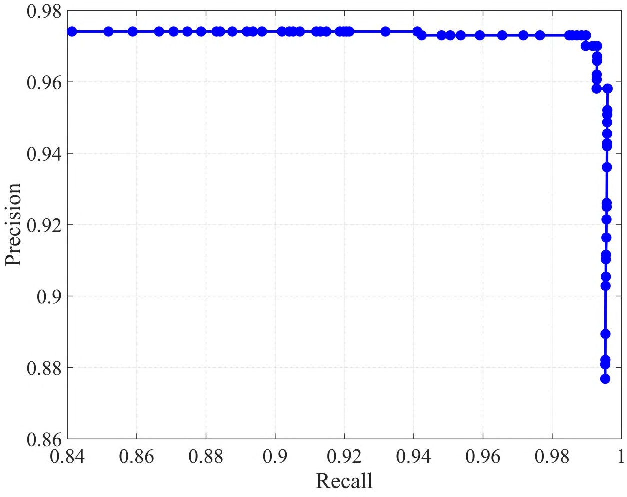

FIGURE 3. The average recall and precision graph of the cut transition detection of two video sources by varying threshold $\beta \in [0.01, 1]$.

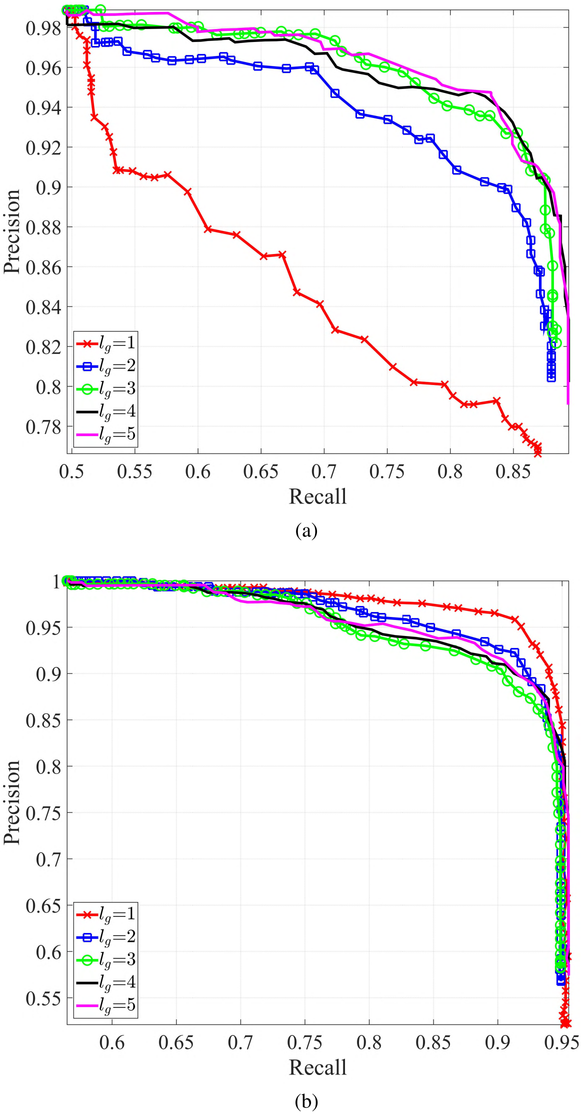

FIGURE 4. The recall and precision graph of the overall transition detection of two video sources by varying threshold $T_g \in [0.01, 1]$ and $l_g \in [1, 5]$. (a) TREC2001 video dataset. (b) Golf video sequences.

video frames using SURF and global histogram. It may cause missed shot boundaries where most are gradual transitions. The comparison between the proposed SBD scheme and method in [9] has shown that the similarity based on SURF and global features between two adjacent frames do not perform well in gradual detection; especially longer transition lengths experienced in the documentary video dataset. We can conclude that the trade-off of using SURF matching score is its disadvantage in representing the dissolve changes, but it still provides a good representation of the cut change.

VI. DISCUSSION

In our proposed shot boundary detection scheme, inter-frame distance l and threshold T_{cad} have a significant effect on the candidate segment selection. The purpose of this process is to select the potential frame pairs that contain boundaries. Therefore, larger l significantly collect all boundaries espe-

TABLE 4. Comparison of the proposed method with other systems evaluated over TREC2001 video Dataset ($l_c = 1, T_c = 0.9, \beta = 0.6, T_g = 0.15, l_g = 3$).

Method	Video	Cut transition			Gradual transition		
		P	R	F1	P	R	F1
Lu's system [6]	D2	90.5	90.5	90.5	72.5	93.5	81.7
	D3	86.7	66.7	75.4	94.0	73.4	82.4
	D4	89.7	88.8	89.2	74.1	72.7	73.4
	D6	97.4	95.0	96.2	92.7	84.4	88.4
	Average	91.1	85.3	87.8	83.3	81.0	81.5
Priya's system [7]	D2	85.4	97.6	91.1	90.0	87.1	88.5
	D3	86.5	82.1	84.2	88.7	85.9	87.3
	D4	90.6	88.8	89.7	84.6	80.0	82.2
	D6	93.5	95.6	94.5	88.5	88.5	88.5
	Average	89.0	91.0	89.9	88.0	85.4	86.6
Proposed method	D2	97.5	92.9	95.1	77.8	90.3	83.6
	D3	100.0	94.9	97.4	89.6	67.2	76.8
	D4	97.9	96.9	97.4	65.5	65.5	65.5
	D6	100.0	100.0	100.0	80.4	82.2	81.3
	Average	98.9	96.2	97.5	78.3	76.3	76.8

TABLE 5. Comparison of overall transition detection of the proposed method with Apostolidis's system [9] evaluated over TREC2001 video dataset ($T_c = 0.9, \beta = 0.6, l_c = 1, T_g = 0.15, l_g = 3$).

Video	Priya's system [7]			Apostolidis's system [9]			Proposed method		
	P	R	F1	P	R	F1	P	R	F1
D2	87.9	84.5	86.1	78.1	78.1	78.1	90.8	94.5	92.6
D3	87.9	84.5	86.1	96.9	61.2	75.0	95.3	78.6	86.2
D4	88.5	85.6	87.0	87.6	73.9	80.1	86.2	85.6	85.9
D6	92.8	90.6	91.7	91.5	50.6	65.2	89.5	90.6	90.1
Average	89.3	86.3	87.7	88.5	65.9	74.6	90.7	87.3	88.7

cially gradual transitions. However, the effect of choosing a large l is increasing in the number of candidate frames, as well as false positives in shot detection. T_{cad} is the parameter that will verify whether the segment contains shot boundary. Therefore, we need to determine l and T_{cad} that effectively reduce the number of processed video frames and be able to collect most of the potential boundaries at the same time. These two parameters are an inverse variation to each other. When choosing larger l, T_{cad} should be slightly lower to prevent collecting non-boundary segments. In our experiments over two video sources, these two parameters are determined as follows: $l = 5, T_{cad} = 0.1$.

Secondly, the parameter T_c and β in $\mathcal{CT}$ criteria have a significant influence on the cut detection performance. These two parameters directly affect the precision of cut detection. The purpose of our proposed cut detection scheme is to ensure that the segment contains a cut transition by using the advantage of multiplication. If the cut does exist, the higher dissimilarity must exhibit. Therefore, setting a high T_c can significantly achieve good precision while still maintaining a good recall for the cut detection.

In this paper, we have also investigated the performance by choosing a high T_c and varying β value. A higher β value provides better precision, but the recall is slightly lower than setting a low β value. The recommended T_c takes value in the interval $[0.8, 0.9]$ and β in $[0.5, 0.7]$ for both video types. The average precision and recall graph of cut transitions detected by varying $\beta \in [0.01, 1]$ for two video sources is shown in Fig.3. It indicates that our proposed scheme provides excellent performance in cut transition detection. To compare the cut detection result with other proposed methods, we determine $l_c = 1$, $T_c = 0.9$, and $\beta = 0.6$ for both video sources.

The performance of gradual detection and overall detection are influenced by T_g and the gradual transition criteria $\mathcal{GT}$. Based on our observation, the difference between transition characteristics in documentary videos, and golf videos, are the length of the gradual transitions. The average gradual length of TREC2001 is more than 1.5 times that of the gradual transitions duration in the golf videos. This observation implies that mild changes may occur over a longer period. Hence, SURF matching scores may not efficiently discriminate between the gradual effects in the documentary video source. Our assumption is supported by the output generated by SBD using [46].

In this paper, we also perform the test by varying the threshold $T_g \in [0.01, 1]$ and $l_g \in [1, 5]$ and observing the overall detection results. Fig.4 shows the recall and precision graph of the overall detection results for TREC2001 and golf video sequences. l_g significantly affects the recall performance in the documentary video, which supports our assumption that the dissimilarity signal λ, based on the SURF matching score and RGB histogram between two adjacent frames, does not perform well during the mild change in the long gradual transition length. However, choosing a larger l_g as shown in Fig.4(a) can effectively increase the overall detection performance.

In contrast to the documentary video type, the combination feature λ performs well in the shot detection for golf video, regardless of choosing a different l_g parameter. However, the precision slightly drops as shown in Fig.4(b) when increasing l_g, because the false positive answers are detected from the effects of camera/object movement. Consequently, we can conclude that gradual detection operates with a T_g threshold interval [0.3, 0.4] and $l_g \in [1, 3]$ for golf video sources. For the TREC2001 documentary video, the recommended l_g falls in the interval [3, 5] and the threshold T_g falls in the interval [0.1, 0.3]. The precision and recall graph in Fig.4 show that the proposed SBD can provide a high accuracy for the overall detection with the selection of inter-frame distance.

VII. CONCLUSION

Semantic video analysis is one of a challenging task in sports video application. With the rapid growth rate of popularity in sports, videos with a sophisticated editing effect have been extensively broadcast. Shot boundary detection is therefore considered an essential step toward obtaining efficient semantic event searching. Golf is a long-play sport which contains several camera/object movements and a high correlation between color and content information within a shot. Moreover, the transition between shots may not be only a dissolve transition but include logo appearance, scoreboards, and so on. Our proposed SBD method is therefore designed to obtain an efficient shot boundary detection, where the detection process directly analyses the transition behaviour. Therefore, all shot boundaries are divided into two categories namely cut and gradual. The candidate segment selection is performed by the combined features aiming to collect the potential shot boundaries and reduce the number of processed video frames. Experimental results show that our proposed cut transition outperforms the other proposed SBD scheme. Our proposed SBD system also provides good performance in the overall shot boundary detection compared to other recent proposed schemes. The PBI transition detection has the benefit that users do not require the training process but would still be able to obtain a good performance in detection accuracy. We also observed the inter-frame distance based on the proposed visual features. Our conclusion is that the inter-frame parameter directly affects the performance of the gradual shot detection, when implementing the combination features using SURF with the extended transition effect, such as in the documentary video dataset. Our future work is to focus on the detection speed and improve the gradual detection.

REFERENCES

[1] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.

[2] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," *Comput. Vis. Image Understand.*, vol. 114, pp. 411–418, Apr. 2010.

[3] Z. Xiong, X. S. Zhou, Q. Tian, Y. Rui, and T. S. Huang, "Semantic retrieval of video-review of research on video retrieval in meetings, movies and broadcast news, and sports," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 18–27, Apr. 2006.

[4] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot detection and condensed representation. A review," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 28–37, Mar. 2006.

[5] J. Yuan *et al.*, "A formal study of shot boundary detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 2, pp. 168–186, Feb. 2007.

[6] Z.-M. Lu and Y. Shi, "Fast video shot boundary detection based on SVD and pattern matching," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5136–5145, Dec. 2013.

[7] G. G. L. Priya and S. Domnic, "Walsh-Hadamard transform kernel-based feature vector for shot boundary detection," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 5187–5197, Oct. 2014.

[8] J. Baber, N. Afzulpurkar, M. N. Dailey, and M. Bakhtyar, "Shot boundary detection from videos using entropy and local descriptor," in *Proc. 17th Int. Conf. Digit. Signal Process. (DSP)*, 2011, pp. 1–6.

[9] E. Apostolidis and V. Mezaris, "Fast shot segmentation combining global and local visual descriptors," in *Proc. IEEE Int. Conf. Speech Signal Process. (ICASSP)*, Apr. 2014, pp. 6583–6587.

[10] S. Tippaya, T. Tan, M. Khan, and K. Chamnongthai, "A study of discriminant visual descriptors for sport video shot boundary detection," in *Proc. 10th Asian Control Conf. (ASCC)*, 2015, pp. 1–4.

[11] S. Tippaya, S. Sitjongsataporn, T. Tan, K. Chamnongthai, and M. Khan, "Video shot boundary detection based on candidate segment selection and transition pattern analysis," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jun. 2015, pp. 1025–1029.

[12] C. Grana and R. Cucchiara, "Linear transition detection as a unified shot detection approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 483–489, Apr. 2007.

[13] J. Sun and Y. Wan, "A novel metric for efficient video shot boundary detection," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, Jul. 2014, pp. 45–48.

[14] Y. N. Li, Z. M. Lu, and X. M. Niu, "Fast video shot boundary detection framework employing pre-processing techniques," *IET Image Process.*, vol. 3, no. 3, pp. 121–134, Jun. 2009.

[15] J. Chen, S. Ipson, and J. Jiang, "A fuzzy logic method of feature representation for shot boundary detection," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 4337–4340.

[16] J. Ren, J. Jiang, and J. Chen, "Shot boundary detection in MPEG videos using local and global indicators," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 8, pp. 1234–1238, Aug. 2009.

[17] H. Fang, J. Jiang, and Y. Feng, "A fuzzy logic approach for detection of video shot boundaries," *Pattern Recognit.*, vol. 39, no. 11, pp. 2092–2100, Nov. 2006.

[18] L. Shiyang, W. Zhiyong, W. Meng, M. Ott, and F. Dagan, "Adaptive reference frame selection for near-duplicate video shot detection," in *Proc. 17th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2010, pp. 2341–2344.

[19] P. P. Mohanta, S. K. Saha, and B. Chanda, "A model-based shot boundary detection technique using frame transition parameters," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 223–233, Oct. 2012.

[20] Y. Xiao, L. Xia, S. Zhu, D. Huang, and J. Xie, "Video shot boundary recognition based on adaptive locality preserving projections," *Math. Problems Eng.*, vol. 2013, Oct. 2013, Art. no. 353261. [Online]. Available: https://www.hindawi.com/journals/mpe/2013/353261/cta/

[21] X. Ling, O. Yuanxin, L. Huan, and X. Zhang, "A method for fast shot boundary detection based on SVM," in *Proc. Congr. Image Signal Process. (CISP)*, 2008, pp. 445–449.

[22] X. Sun, L. Zhao, and M. Zhang, "A novel shot boundary detection method based on genetic algorithm-support vector machine," in *Proc. Int. Conf. Intell. Human-Mach. Syst. Cybern. (IHMSC)*, 2011, pp. 144–147.

[23] A. Amiri and M. Fathy, "Video shot boundary detection using QR-decomposition and Gaussian transition detection," *EURASIP J. Adv. Signal Process.*, vol. 2009, p. 509438, Feb. 2010. [Online]. Available: https://link.springer.com/article/10.1155/2009/509438

[24] D. M. Thounaojam, T. Khelchandra, K. M. Singh, and S. Roy, "A genetic algorithm and fuzzy logic approach for video shot boundary detection," *Comput. Intell. Neurosci.*, vol. 2016, Feb. 2016, Art. no. 8469428. [Online]. Available: https://www.hindawi.com/journals/cin/2016/8469428/

[25] V. Chasanis, A. Likas, and N. Galatsanos, "Simultaneous detection of abrupt cuts and dissolves in videos using support vector machines," *Pattern Recognit. Lett.*, vol. 30, pp. 55–65, Jan. 2009.

[26] J. Zhou and X.-P. Zhang, "Video shot boundary detection using independent component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2005, pp. 541–544.

[27] G. G. L. Priya and S. Domnic, "Edge strength extraction using orthogonal vectors for shot boundary detection," *Procedia Technol.*, vol. 6, pp. 247–254, Jan. 2012.

[28] D. Adjeroh, M. Lee, N. Banda, and U. Kandaswamy, "Adaptive edge-oriented shot boundary detection," *EURASIP J. Image Video Process.*, vol. 2009, no. 1, p. 859371, 2009. [Online]. Available: https://link.springer.com/article/10.1155/2009/859371

[29] Y. Chang, D. J. Lee, Y. Hong, and J. Archibald, "Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature transform descriptor," *J. Image Video Process.*, vol. 2008, pp. 1–10, Jan. 2008. [Online]. Available: https://link.springer.com/article/10.1155/2008/860743

[30] X. Zhou, X. Zhou, L. Chen, A. Bouguettaya, N. Xiao, and J. A. Taylor, "An efficient near-duplicate video shot detection method using shot-based interest points," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 879–891, Aug. 2009.

[31] R. Hannane, A. Elboushaki, K. Afdel, P. Naghabhushan, and M. Javed, "An efficient method for video shot boundary detection and keyframe extraction using SIFT-point distribution histogram," *Int. J. Multimedia Inf. Retr.*, vol. 5, pp. 89–104, Mar. 2016.

[32] M.-H. Park, R.-H. Park, and S. W. Lee, "Shot boundary detection using scale invariant feature matching," *Proc. SPIE*, vol. 6077, p. 60771N, Jan. 2006. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1273616

[33] J. Li, Y. Ding, Y. Shi, and W. Li, "Efficient shot boundary detection based on scale invariant features," in *Proc. 5th Int. Conf. Image Graph. (ICIG)*, 2009, pp. 952–957.

[34] S. Liu, M. Zhu, and Q. Zheng, "Video shot boundary detection with local feature post refinement," in *Proc. 9th Int. Conf. Signal Process. (ICSP)*, 2008, pp. 1548–1551.

[35] Y. Zheng and Y. Zhang, "GPU-accelerated abrupt shot boundary detection," in *Proc. 16th Int. Symp. Commun. Inf. Technol. (ISCIT)*, 2016, pp. 141–145.

[36] E. Tsamoura, V. Mezaris, and I. Kompatsiaris, "Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework," in *Proc. 15th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 45–48.

[37] B. Han, Y. Hu, G. Wang, W. Wu, and T. Yoshigahara, "Enhanced sports video shot boundary detection based on middle level features and a unified model," *IEEE Trans. Consum. Electron.*, vol. 53, no. 3, pp. 1168–1176, Aug. 2007.

[38] Z. Cernekovà, C. Kotropoulos, and I. Pitas, "Video shot-boundary detection using singular-value decomposition and statistical tests," *J. Electron. Imag.*, vol. 16, pp. 043012-1–043012-13, Dec. 2007.

[39] S. Tippaya, S. Sitjongsataporn, T. Tan, and K. Chamnongthai, "Abrupt shot boundary detection based on averaged two-dependence estimators learning," in *Proc. 14th Int. Symp. Commun. Inf. Technol. (ISCIT)*, 2014, pp. 522–526.

[40] Z. Cernekovà, C. Kotropoulos, N. Nikolaidis, and I. Pitas, "Video shot segmentation using fusion of SVD and mutual information features," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2005, pp. 3849–3852.

[41] Y. Xiao, "An effective video shot boundary detection method based on supervised learning," in *Proc. 2nd Int. Conf. Adv. Comput. Control (ICACC)*, 2010, pp. 371–374,

[42] R. W. Schafer, "What is a Savitzky-Golay filter? [lecture notes]," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 111–117, Jul. 2011.

[43] *PGA Thailand*, accessed on Aug. 1, 2016. [Online]. Available: https://www.youtube.com/user/pgathailand/

[44] NIST. *Homepage of TREC Video Retrieval Evaluation*, accessed on Nov. 20, 2016. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[45] *TREC2001 Dataset*, accessed on Nov. 20, 2016. [Online]. Available: http://www.open-video.org/

[46] *Video Shot and Scene Segmentation Version 1.4.3 (CPU Based)*, accessed on Aug. 18, 2016. [Online]. Available: http://mklab.iti.gr/project/video-shot-segm

[47] J. Xu, L. Song, and R. Xie, "Shot boundary detection using convolutional neural networks" in *Proc. Visual Commun. Image Process. (VCIP)*, 2016, pp. 1–4.

SAWITCHAYA TIPPAYA received the B.Eng. degree in electronics engineering and the M.Eng. degree in electrical engineering (electronics) from the Mahanakorn University of Technology, Thailand, in 2009 and 2012, respectively. She is currently pursuing the collaboration of the Ph.D. degree in electronic and telecommunication engineering with the King Mongkut's University of Technology Thonburi, and the Ph.D. degree in mechanical engineering with Curtin University. Her research interests are software development in advanced circuit design analysis, adaptive signal processing, content-based video retrieval, and data analytics in sports.

SUCHADA SITJONGSATAPORN received the Dr.Eng. degree in electrical engineering from the Mahanakorn University of Technology, Bangkok, Thailand, in 2009. She is currently an Assistant Professor with the Centre of Electronic Systems Design and Signal Processing, Department of Electronic Engineering, Faculty of Engineering, Mahanakorn University of Technology. Her research interests are the adaptive signal processing in communication systems, statistical signal processing, and advanced adaptive algorithm.

TELE TAN received the Ph.D. degree in electronics engineering from Surrey University, U.K., in 1993. He is currently a Professor with the Department of Mechanical Engineering, Curtin University. His research interest is in pattern recognition, which he applies to areas in digital media processing, neurological signal analysis, neurocognitive studies, and biomedical engineering.

MASOOD MEHMOOD KHAN received the B.E. degree from the NED University of Engineering and Technology, Pakistan, the M.S. degree from Colorado State University, Fort Collins, CO, USA, and the Ph.D. degree from the University of Huddersfield, U.K. He has taught at the National University of Computer and Emerging Sciences, Karachi; the Jefri Bolkiah College of Engineering Kuala Belait, Brunei Darussalam; and the American University of Sharjah, UAE. He joined the Mechanical Engineering Department, Curtin University, Australia. His research activities revolve around affective computing, machine vision and perception, human–computer interaction, and biomedical imaging.

KOSIN CHAMNONGTHAI (SM'12) received the B.Eng. degree in applied electronics from the University of Electro-Communications, Tokyo, Japan, in 1985, the M.Eng. degree in electrical engineering from the Nippon Institute of Technology, Saitama, Japan, in 1987, and the Ph.D. degree in electrical engineering from Keio University, Tokyo, Japan, in 1991. He is currently a Professor with the Electronic and Telecommunication Engineering Department, Faculty of Engineering, King Mongkut's University of Technology Thonburi. His research interests include computer vision, image processing, robot vision, and signal processing. He is a member of TRS, IEICE, TESA, ECTI, AIAT, APSIPA, and EEAAT. He serves as the President-Elect of the ECTI Association from 2016 to 2017. He served as an Editor of *ECTI E-Magazine* from 2011 to 2015, an Associate Editor of the ECTI-CIT Transactions from 2011 to 2016, the ECTI-EEC Transactions from 2003 to 2010, and the ELEX (IEICE Trans) from 2008 to 2010, and the Chairman of the IEEE COMSOC Thailand from 2004 to 2007.