A NOVEL TECHNIQUE FOR SIZE CONSTRAINED VIDEO STORYBOARD GENERATION USING STATISTICAL RUN TEST AND SPANNING TREE

PARTHA PRATIM MOHANTA
Electronics and Communication Sciences Unit
Indian Statistical Institute
203, B. T. Road, Kolkata 700108, India
ppmohanta@isical.ac.in

SANJOY KUMAR SAHA
Computer Science and Engineering Department
Jadavpur University, Jadavpur
Kolkata 700032, India
sks_ju@yahoo.co.in

BHABATOSH CHANDA
Electronics and Communication Sciences Unit
Indian Statistical Institute
203, B. T. Road, Kolkata 700108, India
chanda@isical.ac.in

Received 1 August 2011
Revised 19 September 2012
Accepted 29 October 2012
Published 17 April 2013

Storyboard consisting of key-frames is a popular format of video summarization as it helps in efficient indexing, browsing and partial or complete retrieval of video. In this paper, we have presented a size constrained storyboard generation scheme. Given the shots i.e. the output of the video segmentation process, the method has two major steps: extraction of appropriate key-frame(s) from each shot and finally, selection of a specified number of key-frames from the set thus obtained. The set of selected key-frames should retain the variation in visual content originally possessed by the video. The number of key-frames or representative frames in a shot may vary depending on the variation in its visual content. Thus, automatic selection of suitable number of representative frames from a shot still remains a challenge. In this work, we propose a novel scheme for detecting the sub-shots, having consistent visual content, from a shot using Wald–Wolfowitz runs test. Then from each sub-shot a frame rendering the highest fidelity is extracted as key-frame. Finally, a spanning tree based novel method is proposed to select a subset of key-frames having specific cardinality. Chronological arrangement of such frames generates the size constrained storyboard. Experimental result and comparative study show that the scheme works satisfactorily for a wide variety of shots. Moreover, the proposed technique rectifies mis-detection error, if any, incurred in video segmentation.
P. P. Mohanta, S. K. Saha & B. Chanda

process. Similarly, though not implemented, the proposed hypothesis test has ability to
rectify the false-alarm in shot detection if it is applied on pair of adjacent shots.

Keywords: Key-frame; sub-shot detection; video summarization; Wald–Wolfowitz runs
test; size constrained storyboard; spanning tree.

1. Introduction

Due to the rapid growth in the volume of digital video data, efficient access of the
same has become an important issue for proper utilization of resources. As a result,
fast browsing, video summarization, indexing and retrieval of desired pieces of data
have emerged as an active area of research. In fact appropriate summarization leads
to robust indexing and efficient retrieval. Video summarization may be achieved in
various formats: a set of keywords or a brief description of the content of the video
at one end to a set of few frames arranged chronologically, called storyboard, at the
other end. The former approach requires modeling and analysis of semantic informa-
tion which is still at very preliminary state. The latter approach, on the other
hand, has the advantage of making use of many existing methodologies like shot
detection and key-frame selection in appropriate manner. The storyboard
summarization has other advantages too. For example, frames of a storyboard may be used
as access points to interesting events in a video. Second, a storyboard consisting of
a predefined number of frames (to limit the bandwidth requirement) may be used
as a promo of a video (or movie).

Storyboard may be defined as a temporal sequence of key-frames, i.e. a very small
subset of frames representing the whole video. Storyboard generation algorithm
usually has three major steps: Shot detection, key-frame extraction and selection
of desired number of key-frames. A shot may be defined as a sequence of frames
captured through a continuous camera operation. Shot transition is either an abrupt
change (cut) or a gradual one (e.g. dissolve, fade-in, fade-out, wipe). The frames
within shot may be referred to as no-change frames representing a continuous action
with either fixed or slowly moving objects and/or background. Such frames, in
addition to their temporal coherence, possess a common semantic.

Shot detection is usually followed by key-frame selection. Thus, a large number
of frames of the entire shot are given as the input and only a few frames, in most
of the cases only one, representing the shot are extracted as output. Such frames
arranged chronologically form a video storyboard.\(^1\) However, size (or number of key-
frames) of the storyboard thus obtained may be very large in view of the purpose.
To make it storage and transmission efficient a size constraint may be imposed on
it. In other words, a storyboard consisting of a predefined number of frames, which
are of highest interest or variability, may be formed for efficient utilization and is
known to be size constrained storyboard. So the issue of ranking key-frames is of
concern in this problem.

In this work we assume that shot boundary is detected, i.e. video segmentation
is done. So input to the proposed system is one shot at a time and, finally, all
key-frames extracted from these shots are considered simultaneously to generate the size constrained storyboard. Though a shot is defined as a continuous video stream taken with a single camera operation, i.e. between on and off, depending on the events or activities recorded in the shot more than one key-frame may be required to represent the shot. One of the major issues in storyboard generation is the number of key-frames to be extracted from a shot. Whether it is predefined or to be chosen dynamically is also another issue. In this work we select variable number of frames from each shot based on the variability in visual content and not the length or number of frames in the shot. Methodology for visual content representation and key-frame selection criteria play a crucial role. Finally, the fact that the frames in a shot bear a strong temporal and visual cohesion adds further complexity to the problem. This is done by dividing a shot into suitable number of sub-shots using a non-parametric hypothesis testing such that each sub-shot is more or less constant in terms of visual content. This also rectifies any shot boundary detection error that was incurred in the preceding video segmentation step. Then the visual content of each sub-shot can be represented by a single key-frame. Finally, from these key-frames a specified number of frames are selected using a novel spanning tree approach. It may be noted that the proposed methodology can work with any suitable visual content descriptor.

This brief introduction is followed by a brief study of past works presented in Sec 2. In Sec. 3, we present a hypothesis testing-based methodology for key-frame selection. A detail description of proposed methodology elaborating methods to overcome various hurdles is given here. We then present a spanning tree-based scheme to generate size constrained storyboard suitable for rendering a summarized version of a video. Experimental results and discussion are presented in Sec. 4 and finally, concluding remarks are cited in Sec. 5.

2. Past Work

In general, key-frame detection methodologies assume that the video has already been segmented into shots without any error. Then from each detected shot, the key-frames are selected according to certain criteria. A careful study of past works reveals a wide variety in the approaches. The schemes may be classified with respect to the count of key-frame to be extracted from a shot and whether such count is predefined or not. Schemes presented in Refs. 2–4 simply ignore the visual content and pick up frames at a fixed interval. Methods in Refs. 2 and 4 considers first and last frame in the shot as representatives, whereas Tonomura et al.\(^3\) choose only the first one. Pentland et al.\(^5\) have also relied on temporal aspects by applying temporal sampling at a predefined interval.

As it is extremely difficult to describe the semantics (such as relevant objects, action, events, etc.) of a shot, the content may be described in terms of low level visual features. Color statistics, edge strength statistics and texture measures are some commonly used features. Zhonghua et al.\(^6\) have suggested a scheme based on
the ratio of object and background to select one key-frame for each shot. Ciocca et al.\textsuperscript{1} have proposed a dynamic technique where the visual content of the frames are described in terms of color histogram, edge direction histogram and wavelet statistics. Key-frames are selected based on high curvature points of cumulative frame difference. But, the algorithm depends on certain parameters and tuning of those are very crucial. A motion analysis based technique is presented in Ref. 7. Schemes based on feature difference of consecutive frames are reported in Refs. 8 and 9. Han et al.\textsuperscript{8} have proposed an adaptive temporal scheme where uniform sampling on cumulative frame difference results into non-uniform temporal sampling of frames indicating the representatives. Cooper et al.\textsuperscript{10} have followed a discriminative technique so that key-frame represents the similarity of segment to which it belongs and preserves the differences from the key-frames of other segments. Singular value decomposition (SVD)-based approach has been tried in Ref. 11. A feature matrix is formed based on the histograms of hue, saturation and value. SVD is then used to factorize the matrix and to compute significant singular vectors. Shot boundary is detected by tracing the rank of the singular vectors using a sliding window approach. First frame of the selected shot is taken as key-frame. Thesaurus-based schemes are reported in Refs. 12 and 13. In such effort, coarse regions of each frame are represented by color and texture descriptor. Visual thesaurus is formed by clustering the descriptors. Based on the clusters, model vectors for each frame is formed and key-frames are selected following the diversity in terms of model vector. In Ref. 14, the whole video is divided into number of segments using threshold adaptive clustering method. The first frame from each segment is taken as key-frame. Finally, an edge-based similarity measure is applied to remove the redundant key-frames. In a different study\textsuperscript{15}, a statistical technique called design of experiment (DoE) is applied for detection of key-frames from the video. The consecutive frame differences are taken as the factor of this DoE problem. Finally, the key-frames are extracted from the possible set of key-frames applying the concept of meaningfulness. Some recent studies\textsuperscript{16,17} reported key-frame extraction methodologies based on the visual attention model. In Ref. 18 a key-frame is extracted from spatial-temporal color distribution of the shot. Fermann et al.\textsuperscript{19} have considered the idea of group of frames (GoF) to obtain the hierarchical representation of the video stream with whole video at top, frames at the bottom, and shots, etc. at the intermediate levels. Clustering based approaches\textsuperscript{20–24} are also reported. Gunsel et al.\textsuperscript{25} and Zhao et al.\textsuperscript{26} have relied on threshold method. In such cases, the performance depends on the proper selection of threshold and initialization of clustering parameters. In Refs. 27 and 28, schemes have been reported which first detects the sub-shots and then key-frames are identified for the sub-shots using image entropy-based measures. A genetic algorithm (GA) based key-frame extraction technique is reported in Ref. 29.

Macer et al.\textsuperscript{30} presented a key-frame-based summarization i.e. storyboard generation scheme where each shot is represented by one frame. Such summarization using the key-frames of the individual shots also leads to redundancy as there may
be many visually repeating shots. Various schemes have been reported to address the problem. Uchihashi \textit{et al.}\cite{31} have relied on shot importance measure which discards less important shots like repeating ones. Finally, key-frames of the important shots are considered to form the summary. Ciocca \textit{et al.}\cite{32} have followed a sequence of post-processing steps to refine the storyboard. They have applied a neural network based supervised classification scheme to remove meaningless key-frames. Finally, the retained key-frames are grouped into clusters and representatives from the clusters are chosen. Sull \textit{et al.}\cite{33} have organized the key-frames in a tree structure based on maximum fidelity criteria. Depending on the desired level of details, key-frames at a particular depth are presented. Mundur \textit{et al.}\cite{34} have proposed a storyboard generation scheme which starts with a set of pre-sampled frames from the entire video. Such frames are clustered into groups and the frame nearest to the centroid of the clusters are chosen to form the summarized version.

3. Proposed Methodology

As stated earlier, the size constrained storyboard generation scheme has three major steps: Shot detection, representative frame(s) or key-frame(s) extraction from each shot and selection of a subset of key-frames having given cardinality. The proposed system assumes that the video stream has already been segmented into shots\cite{35} and the frames of one shot at a time are fed as input to the key-frame extraction system. Usually the frames in a shot are very similar in terms of their visual content. So one key-frame could be a good representative of one shot. But, in a shot of long duration or in a shot depicting complex event, visual content of frames after considerable time gap may differ enough to qualify for having a separate representative. Moreover, a shot detection scheme may miss shot boundary leading to merger of two or more shots into one. Obviously, such shots should qualify for multiple representative frames. Hence, using a pre-conceived number of key-frames for every shot may not be a good approach. In other words, a key-frame should represent a sequence of frames as long as their visual content is more or less same and a variable number of key-frames may be needed to describe a shot. To handle the situation, \textit{we divide a shot into a number of sub-shots each of which has consistent visual content and, at the same time, is significantly different from neighboring sub-shots}. We then extract exactly one key-frame from each sub-shot.

It may be emphasized that a shot is decomposed into sub-shots only for the purpose of extracting representative frame. At the boundary of sub-shots, frames may not be significantly different as is in the case of shot. So shot detection techniques are not suitable for the task. On the other hand, adjacent sub-shots differ in overall visual characteristics. So we view the problem of sub-shot detection as a problem of testing mutual independence of frames of adjacent sub-shots. The proposed methodology, extension of our previous work,\cite{36,37} evolves around non-parametric hypothesis testing based on \textit{Wald–Wolfowitz runs test}\cite{38} described as follows.
3.1. **Sub-shot detection using univariate Wald–Wolfowitz runs test**

Wald–Wolfowitz runs test is used to solve the non-parametric two sample problem. Suppose, there are two samples \( X \) and \( Y \) of size \( m \) and \( n \), respectively and the corresponding distributions are \( F_x \) and \( F_y \). \( H_0 \), the null hypothesis to be tested and \( H_1 \), the alternative hypothesis are as follows:

\[
H_0: \text{X and Y are from same population, i.e. } F_x = F_y
\]

\[
H_1: \text{They are from different population, i.e. } F_x \neq F_y
\]

In classical Wald–Wolfowitz test, it is assumed that sample points are univariate. \( N = n + m \) observations are sorted in ascending order and assigned the labels \( X \) or \( Y \) depending on the sample to which it belongs. Test statistic, \( W \) is computed based on \( R \), the number of runs (a “run” is a sequence of identical labels) as follows.

\[
W = \frac{R - \frac{2mn}{N} - 1}{\frac{\sqrt{2mn(2mn-N)}}{N^2(N-1)}}
\]  

(1)

As \( W \) follows standard normal distribution, the critical region may be chosen for a given level of significance which implies the maximum probability of rejecting a true \( H_0 \). If \( W \) falls within the critical region, \( H_0 \) is rejected. Physically, low value of \( R \) denotes that two samples are less interleaved in the ordered list and it leads to the interpretation that they are from different population.

Considering the objective of the present work, this test can be applied in detecting whether two sequences of frames (i.e. sub-shots within a shot) belong to same visual event or not. Accordingly a shot is partitioned into two or more non-overlapping sequences or sub-shots and two sequences of frames from the shot are taken as two samples \( X \) and \( Y \). Finally, the null hypothesis (i.e. whether they reveal same visual event) can be tested. This results in decomposing a shot into possible homogeneous sub-shots. Detail technique may be described as follows.

Given a shot, the first step to detect the sub-shots is to partition the shot into non-overlapping sequences. Suppose, \( WWRT(S_i, S_j) \) represents the univariate Wald–Wolfowitz runs test function and returns true if the sequences \( S_i \) and \( S_j \) belong to different visual event or not based on a visual feature. Now if \( S_i \) is followed by \( S_j \), it is possible that they have a temporal cohesion between them, particularly in between the trailing frames of \( S_i \) and leading frames of \( S_j \). Thus the comparison of consecutive sequences are usually susceptible to fail in registering the variation leading to different sub-shots. To overcome this problem, the given shot or sequence of frames is divided into four sub-shots (see Fig. 1) and

![Shot partition](image)

Fig. 1. Physical partition of a shot into four sub-shots.
instead of consecutive sequences, alternate sequences are considered to detect significant changes, if any, within the shot. Due to the temporal separation between \( S_i \) and \( S_{i+2} \), the accumulated difference between the two sequences may bias the test to declare them different. Second, since the shots are outcome of some video segmentation algorithm which is supposed to be sufficiently reliable, we prefer not to decompose them to sub-shots without strong evidence. So, to minimize the false alarm, the shot \((F,L)\) is divided into sub-shots only if both \( WWRT(S_i, S_{i+2}) \) and \( WWRT(S_{i+1}, S_{i+3}) \) are true. It may lead to failure if either only the leading or the trailing part of a shot is different from the rest, but avoids unnecessary fragmentation of shot semantics and increasing the number of key-frames. As the content of a frame is described in terms of a number of visual features, \( WWRT(S_i, S_j) \) is applied on each of the features and majority voting criterion is adopted to draw the final decision.

Once sub-shots are detected, i.e. a shot is decomposed into sub-shots, each sub-shot is considered as a shot and is tested for existence of sub-shots recursively. This process of decomposition continues until visual continuity within the shot is established or the minimum length of the shot \( l_t \) (a predefined threshold) is reached. The value of \( l_t \) is not very critical and can be selected from a wide range of values. Hence, each shot or sub-shot (also called shot for ease in description) has to be visually consistent.

Assume \( F \) and \( L \) denote the first and last frame indices in a collection of consecutive frames, i.e. a shot or a (sub)shot. \( \eta \) is the dimension of the feature vector. The sub-shot detection function, may be described as follows:

**Algorithm-I**

\[
detect_{\text{subshot}}(F, L)
\]

- Divide \([F, L]\) into four equal subranges \( S_1, S_2, S_3 \) and \( S_4 \).
- \( count \leftarrow 0 \).
- For each feature in the feature vector
  - \( chk_1 \leftarrow \text{WWRT}(S_1, S_3) \)
  - \( chk_2 \leftarrow \text{WWRT}(S_2, S_4) \)
  - if \((chk_1 = \text{TRUE} \text{ and } chk_2 = \text{TRUE})\) then increase \( count \).
- If \((count > \eta/2 \text{ and } (L - F)/2 > l_t)\) then
  - middle frame, \( M = F + \frac{L - F}{2} \)
  - \( detect_{\text{subshot}}(F, F + M) \) and \( detect_{\text{subshot}}(F + M + 1, L) \)
  - else \([F, L]\) represents a single shot.

3.1.1. Reducing the effect of temporal coherence

Proposed run test-based sub-shot detection scheme may suffer from a small drawback. It may be noted that the “run test” actually relies on the interleaving of the
elements from the two sequences in a combined ordered list. If such interleaving is not high enough then the null hypothesis is rejected and two sequences are taken as different. Sometimes frames of a shot are visually so similar that they should not be further split into sub-shots. But, due to temporal variation in a video stream, the said visual similarity may not result into high degree of interleaving of feature values required for the acceptance of null hypothesis. Such rejection results in decomposing the shot into sub-shots and extracting additional key-frames. Few sample frames of two such shots are shown in Fig. 2. The shots seem to be single unit. But, the proposed scheme results in multiple sub-shots.

In an effort to surmount the problem of over-splitting of a shot, effect of natural temporal variation in the frames of the shot needs to be reduced. The frames in a shot possess a small but monotonic variation which human beings usually ignore based on perception, particularly when viewed as a temporal sequence. But, such variation imposes a temporal gradient in the feature values. In the proposed scheme, two samples are prepared by incorporating the Taylor Series (TS) expansion-based correction in the feature space.

Suppose, \( f(t) \) and \( f(t+h) \) are the feature values for \( t \)-th and \( (t+h) \)-th frames, respectively. Following \( n \)-th order TS expansion, we may write

\[
    f(t+h) = f(t) + hf'(t) + \frac{h^2}{2!}f''(t) + \cdots + \frac{h^n}{n!}f^n(t)
\]  \hspace{1cm} (2)

Fig. 2. (a) and (b) show frames sampled uniformly from two different shots with slow gradual variation.
Thus, \( f(t + h) = f(t) + e_t \), where \( e_t \) denotes the effect due to temporal separation between \((t + h)\)-th and \(t\)-th frame. We consider \( f(t + h) - e_t \) as the corrected visual feature value for \((t + h)\)-th frame by cancelling the effect of the temporal gradient. In our experiment, considering the linear variation in the feature values over time, we have restricted ourselves to first-order TS. To compute \( f'(t) \), \( f(t + h) \) are plotted against \( h \). Following least square regression, a straight line is fitted. Slope of the line is taken as \( f'(t) \). The algorithm \( \text{detect\_subshot}(F, L) \) needs a modification so that each frame is mapped onto \( F\)-th instant. It is observed that the problem of over-splitting is reduced significantly by this way (see Sec. 4).

### 3.2. Sub-shot detection using multivariate Wald–Wolfowitz runs test

It is mentioned earlier that the visual descriptor of a frame is high-dimensional feature vector. In the majority voting system strong good features and weak or bad features have same contribution in the decision-making. To overcome this problem, all the features may be considered simultaneously where the good features can overshadow the effect of the bad features. Furthermore, the test has to be carried out for each and every element of the feature vector. As a result, time complexity increases linearly with the dimension of a feature vector. Thus, the scheme becomes prohibitive for the high-dimensional content descriptors. This has motivated us to go for the scheme based on multivariate Wald–Wolfowitz runs test, which would consider the aggregated effect of the features and also would minimize the time requirement. Hence, like the univariate Wald–Wolfowitz test coupled with majority voting, the multivariate Wald–Wolfowitz test must be able to decompose a shot into homogeneous sub-shots. Detailed method is as follows.

Friedman and Rašký\(^{39}\) have suggested a multivariate generalization by using the minimal spanning tree (MST) of the sample points as an alternative to univariate sorted list. In this approach, each frame is considered as a node and every node is connected to the closest node (based on the distance between their feature vectors) to form the MST of the frames. Suppose the frames of this MST have come from the sequences \( S_i \) and \( S_j \), then this MST would be equivalent to the combined sorted list of points from the two samples as described in Sec. 3.1. Now, if we remove all the edges connecting pair of frames coming from two different samples, \( S_i \) and \( S_j \), each subtree formed would consist of frames from one and only one sequence or sample and is equivalent to a run. Thus the number of subtrees is equivalent to \( R \), the number of runs and the test statistic \( W \) is defined, as before based on \( R \), as:

\[
W = \frac{R - E[R]}{\sqrt{\text{Var}[R]}},
\]

where \( \text{Var}[R] = \frac{2mn}{N(N-1)} \times \left( \frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)} \times (N(N-1)-4mn+2) \right) \) and \( C \) is the number of edge pairs in MST sharing a common node and \( E[R] = \frac{2mn}{N} + 1 \). Rest of the treatment of data and decision process is similar to what is described.
in Sec. 3.1 for univariate case. Thus, in the sub-shot detection algorithm presented earlier, only $WWRT(S_i, S_j)$ is replaced by multivariate Wald–Wolfowitz runs test. Suppose, $MWWRT(S_i, S_j)$ represents the multivariate Wald–Wolfowitz runs test function and returns true if the samples $S_i$ and $S_j$ belong to different visual events.

In order to minimize the effect of temporal coherence, we apply TS-based correction on the feature vectors of the frame sequence and then proceed for the multivariate Wald–Wolfowitz test. The modified sub-shot detection algorithm is then described as follows.

**Algorithm-II**

\[
detect_{\text{subshot}}(F, L)
\]

- Obtain TS corrected features for the frames in $(F, L)$.
- Divide $(F, L)$ into four equal subranges $S_1, S_2, S_3$ and $S_4$.
- \(chk_1 \leftarrow MWWRT(S_1, S_3)\).
- \(chk_2 \leftarrow MWWRT(S_2, S_4)\).
- If (\(chk_1 = \text{TRUE}\) and \(chk_2 = \text{TRUE}\) and (\(L - F\)/2 > \(l_t\))) then
  - middle frame \(M = F + \frac{L - F}{2}\).
  - \(detect_{\text{subshot}}(F, F + M)\)
  - and \(detect_{\text{subshot}}(F + M + 1, L)\)
- else $(F, L)$ represents a single shot.

So now every shot is very consistent in terms of visual content and we select one frame from each shot for representing the shot. In order to ensure faithful representation, the key-frame for a shot is chosen as follows. Suppose $f_1, f_2, \ldots, f_m$ are the frames in a shot. An average frame $f_{\text{avg}}$ is represented by a feature vector whose elements are obtained by taking the average of corresponding feature value of all the frames $f_1, f_2, \ldots, f_m$, $d_i = \text{dist}(f_i, f_{\text{avg}})$ is the distance between $i$th-frame and the average frame. The $k$-th frame $f_k$ is taken as the representative if $d_k = \min\{d_i\}$.

In our experiment, we have considered Euclidean distance as \(\text{dist}(\cdot)\). It may be noted that fidelity-based representative frame selection in Ref. 36, incurs $O(m^2)$ computational cost whereas for the proposed scheme it is $O(m)$. Moreover, frames in the sub-shot are visually so similar that proposed selection mechanism can maintain the fidelity reasonably well.

### 3.3. Generation of size constrained storyboard

Chronological arrangement of the key-frames is considered as video storyboard. The number of key-frames is quite high for a long video or video with lots of activity and size of storyboard will be huge enough to make it prohibitive for the web-based distribution of video promos. Thus, in the application where such views are to be transmitted over a limited bandwidth, size constrained storyboard generation is in demand.

To limit the size of storyboard to a specified number of frames reflecting overall variation in the video, we have proposed a spanning tree-based scheme. The
problem under consideration can be restated as follows. Given a set of representative frames of a video, a subset or storyboard of specified cardinality, say, \( k \) has to be obtained that would reflect visual content variation present in the original set. Proposed scheme considers, say, \( n \) key-frames as input and forms a spanning tree by selecting \( k \) frames \((k < n)\) from the given set, where each node in the tree corresponds to a selected frame. The spanning tree is formed in such a way that the nodes of the tree are as diverse and dissimilar as possible and, as a result, the tree reveals the maximum variation in the visual content of the video under consideration. The basic strategy to form the said spanning tree is to include two most distant pair of nodes initially and then, until the size of the tree attained, at each iteration to include a new node which is at the largest distance from the nodes in the tree. Thus, the algorithm to obtain the size constrained storyboard \( T \) from the given set \( S \) of key-frames is as follows.

**Algorithm-III**

\[ \text{form}_\text{storyboard}(S, n, k, T) \]

- **remaining set**, \( S' \) \( \leftarrow \) set of all nodes (feature vectors).
- Find \((f_i, f_j)\), the most distant pair of nodes (frames); where \( f_i, f_j \in S \).
- **included set**, \( T \) \( \leftarrow \) \{\( f_i, f_j \)\}.
- \( S' \leftarrow S' - T \).
- **count** \( \leftarrow 2 \)
  - Repeat the following steps until \( \text{count} = k \)
    - For each node, \( f_r \in S' \)
      - distance with the tree, \( d_k = \min\{\text{dist}(f_r, f_l)\}\) where, \( f_l \in T \)
      - Let \( d_t = \max\{d_k\} \) for the node \( f_t \in S' \)
      - Put \( f_t \) in \( T \)
      - Remove \( f_t \) from \( S' \)
      - \( \text{count} + + \).
- **Output**: \( f_1, f_2, f_3, \ldots, f_k \in T \), such that \( f_i < f_{i+1} \), as storyboard.

The spanning tree-based algorithm presented here selects \( k \) most diverse frames from a given set of \( n \) frames. This is a greedy algorithm and is similar to minimum spanning tree algorithm developed by Prim\(^4\) and the order of complexity of the algorithm is \( O(n^2) \). However, the proposed tree can better be termed as \( k \)-maximum spanning tree, which is a sub-tree of the maximum spanning tree. The value of \( k \) has to be chosen depending on the variation in the visual content of the video. For the specified value of \( k \), the proposed scheme selects the optimal subset of frames.

### 4. Experimental Results

We have experimented with the videos taken from TRECVID 2001 and TRECVID 2005 test database, openvideo database and some other news, sports, animation videos and movies. The videos have been segmented into shots following the
methodology presented in Ref. 35. We have considered 251408 frames (approximately 2 h 30 min video) comprising 2020 shots. To judge the performance, we have manually ground-truthed the shots as simple shot (i.e. without further sub-shot) and the compound shot (i.e. shots with multiple sub-shots).

The proposed methodology is not specific to any particular type of feature. The major focus is on hypothesis test based sub-shot detection and spanning tree based size-constrained storyboard generation. The scheme can deal with any visual content descriptor. In our scheme, each frame is described in terms of color and texture. First, second and third order moments of each of the R, G and B intensity histograms are considered as color features. To compute texture features, each of the R, G and B planes are decomposed into one low-pass subband (LL) and three high-pass subbands (HL, LH and HH) using 2D wavelet transformation. At subsequent iteration, the same decomposition process is carried out on LL band. In our experiment, number of iteration is taken as 3. At each level, mean and variance of wavelet coefficients of the high-pass subbands taken together are considered as features. Thus, 18 texture features and 9 color features together form 27-dimensional feature vector for each frame. Histogram moments are computed with normalize frequency values. n-th order moments are normalized in such a way that their sum becomes 1 in each of the R, G and B planes. In case of wavelet-based features, means and variances are normalized in exactly similar manner. Dissimilarity between two frames are measured here as Euclidean distance between their feature vectors. Sub-shot detection process continues recursively till it is declared as to have consistent visual content or its duration is smaller than a threshold. In our experiment, minimum duration of a shot or sub-shot is taken as 2 sec or 60 frames. In hypothesis testing, the test statistic is compared with the tabulated values available in the standard normal table. The level of significance for the test is taken as 0.05.

The sub-shot detection result of the proposed scheme using univariate Wald–Wolfowitz test and majority voting (Algorithm-I) is summarized in Table 1. It is evident from the result (shown in first row) that temporal variation reduction improves the result significantly by arresting unnecessary splitting of shots into sub-shot. This is because insignificant variation over consecutive frames accumulates to become reasonably significant difference over large number of frames, i.e. a shot of

<table>
<thead>
<tr>
<th>Correctly detected</th>
<th>Before TS correction</th>
<th>After TS correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of shot</td>
<td>Actual</td>
<td></td>
</tr>
<tr>
<td>Simple shot</td>
<td>1797</td>
<td>1005</td>
</tr>
<tr>
<td>Compound shot</td>
<td>223</td>
<td>178</td>
</tr>
</tbody>
</table>
long duration. However, once the TS correction is incorporated, such tendency is arrested. But, then the scheme becomes weak in identifying the multiple sub-shots. This effect is revealed in last row of Table 1. Table 2 summarizes the performance of the proposed scheme based on multivariate Wald–Wolfowitz runs test (Algorithm-II). It is found that in 88.75\% cases the input simple shots are declared as simple shots and in 84.3\% cases the input shots are correctly split into multiple sub-shots. Overall correctness is 88.26\%.

Now, each sub-shot is considered to be more or less constant in terms of visual content and, so, can be represented by a single key-frame. The efficacy of the method may be demonstrated by Fig. 3. This figure shows a few frames sampled chronologically from a shot. Consecutive frames are highly correlated exhibiting very slow variation in visual content and all the frames are captured through a single camera operation between on and off as par the definition of a shot. However, if we look

Table 2. Performance of sub-shot detection based on multivariate Wald–Wolfowitz test (Algorithm-II) (number of frames = 251 408).

<table>
<thead>
<tr>
<th>Type of shot</th>
<th>Correctly detected</th>
<th>Before TS correction</th>
<th>After TS correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple shot</td>
<td>1797</td>
<td>1342</td>
<td>1595</td>
</tr>
<tr>
<td>Compound shot</td>
<td>223</td>
<td>197</td>
<td>188</td>
</tr>
</tbody>
</table>

![Fig. 3. Sampled frames of a shot and last row shows the key-frames.](image-url)
at the first and the last frames, they look completely different and uncorrelated. The proposed scheme decomposes this shot into multiple sub-shots with consistent visual content and selects key-frames accordingly. Thus the different phases of activity are properly captured by the four representative frames, though it is conventionally a single shot with almost no movement of camera except a little pan or zoom.

Figure 4 shows some more sample results, last row in each case shows the extracted key-frames each from one sub-shot. It clearly shows that the proposed scheme is capable in extracting suitable number of key-frame(s) by breaking a shot into sub-shots depending on the variation in the content of the shot. For example, it is sufficient to have a single representative for the sequence shown in Fig. 4(a), although there are some motion involved in this shot and thus, the visual content is not exactly uniform over the shot. The proposed method successfully extracted single key-frame whereas all the other methods\textsuperscript{1,11,14} have represented this shot with multiple key-frames. On the other hand, it is evident from the sampled frames of the shots shown in Figs. 4(b)–4(d) that they contain substantial variation. In
Fig. 4(b), initially a wide view is shown and finally the focus is set to a part of the initial scene. It is reverse in case of Fig. 4(c). In Fig. 4(d), camera moves slowly from one direction to another and finally focuses to a part. Extracted key-frames clearly represent the various stages of the shot. The proposed method successfully extracted the exact number of key-frames in all the cases, the systems\textsuperscript{1,11} have failed in all the three cases and represented the shots with a single key-frame. Although the method proposed in Ref. 14 provide multiple key-frames for Figs. 4(b)–4(d), however it shows a tendency of over-splitting the shot.
4.1. Comparison of sub-shot detection

It should be emphasized that success of key-frame detection fully relies on the detection of coherent shot (or sub-shot). This is because if the detected shot or sub-shot is sufficiently coherent, any frame from it can serve as a key-frame reliably. Hence, detection of shots (or sub-shots) and detection of key-frames becomes synonymous. We have compared the performance of our sub-shot (i.e. key-frames) detection scheme with the systems proposed by Ciocca et al.\textsuperscript{1} Almageed\textsuperscript{11} and Chan et al.\textsuperscript{14}

The systems in Ref. 1 relies on the changes on the visual content expressed by frame descriptors. Curvature points on the cumulative frame descriptor differences are identified as sub-shot boundaries. SVD-based method\textsuperscript{11} detects shot (sub-shot) boundaries and the first frame in detected shot (sub-shot) is taken as key-frame. HSV histogram-based feature matrix is factorized using SVD to obtain significant singular vector. Rank of vector is traced using a sliding window approach to detect the boundaries. Chan et al.\textsuperscript{14} have dealt with adaptive threshold clustering to divide the video into segments. First frame from each of the segments are selected as the key-frame. Similar key-frames are removed using edge-based similarity to retain the final set of key-frames. Table 3 compares the performance of all the methods and shows that the performance of our scheme is better. Ciocca’s method\textsuperscript{1} fails miserably in detecting compound shots. This is because the insignificant change in a uniform fashion between consecutive frames may reflect sufficient variation after a reasonable timespan as discussed in Sec. 3.1.1. This change usually is not reflected as high curvature point. Second, their method depends on multiple parameters and simultaneous tuning of those are not a trivial task. The SVD-based method\textsuperscript{11} also suffers from the similar drawback in identifying multiple sub-shots. On the other hand, Chan’s method\textsuperscript{14} is biased towards breaking a shot into multiple sub-shots.

Sometimes, video segmentation method misses shot boundary and results in a merged shot. Such misses are more probable at gradual transition. The proposed sub-shot and key-frame detection scheme is capable of identifying the constituent shots of a merged one. As a result, the propagation of the error caused by the misdetection of the boundary can be reduced as depicted in Fig. 5.

A video contains both simple and compound shots. As there is no prior knowledge about the activity present in a shot, whether a shot is simple or compound is not known in advance. Thus, it is desirable that a key-frame detection system

<table>
<thead>
<tr>
<th>Type of shot</th>
<th>Correctly detected</th>
<th>Proposed method</th>
<th>Ciocca’s method\textsuperscript{1}</th>
<th>SVD\textsuperscript{11}</th>
<th>Chan’s method\textsuperscript{14}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple shot</td>
<td>1797</td>
<td>1595</td>
<td>1513</td>
<td>1612</td>
<td>986</td>
</tr>
<tr>
<td>Compound shot</td>
<td>223</td>
<td>188</td>
<td>57</td>
<td>69</td>
<td>192</td>
</tr>
<tr>
<td>Overall</td>
<td>2020</td>
<td>1783</td>
<td>1570</td>
<td>1681</td>
<td>1178</td>
</tr>
</tbody>
</table>
should work reasonably well for both the cases. Among the other methods discussed above, Almageed’s method\textsuperscript{11} works well for simple shot and performs miserably in case of compound shots. On the other hand, the system proposed by Chan \textit{et al.}\textsuperscript{14} is biased towards the compound shots. However, the performance of the proposed method is reasonably good in both the cases. The last row of Table 3 clearly shows that overall performance of the proposed method is far better than the other three methods.

![Fig. 5. Sample results: Each set shows sampled frames of the merged shots and last row shows corresponding key-frames.](image-url)
4.2. Comparison of storyboard generation

The key-frames are arranged chronologically to form the storyboard. The storyboard consisting of all key-frames of a documentary sequence is shown in Fig. 6. Corresponding size constrained version of the storyboard is shown in Fig. 7. It is very difficult to measure the quality of a summarized storyboard. How well it represents the semantic content depicted in the original storyboard is a matter of subjective measure. In this work, we have relied on human feedback to measure the quality. As it is person dependent, we have collected opinion of 14 viewers. Each one views the original video sequence followed by the corresponding storyboard. Each reviewer assigns a rank to each key-frame in such that a key-frame with rank \( r \) will appear in all the size-constrained storyboards of size \( n \) (where \( n > r \)) generated by him/her based on the given \( S \). On getting the feedback from all the reviewers, the ordered list of the key-frames in \( S \) is formed based on their average rank and is taken as ground truth \( (S_g) \). Precision of the size constrained storyboard \( T \) is computed as \( P = \sum \text{rank}(T) / \sum \text{rank}(T \cap S_g) \). \( \sum \text{rank}(T) \) is \( \frac{n(n+1)}{2} \) where \( n \) is the card \( (T) \) and \( \text{rank}(T \cap S_g) \) is the sum of the rank of all key-frames in \( T \) as suggested by \( S_g \). \( P \) can take the maximum value 1 and higher the value, better is the precision. We have carried out our experiment with 33 sequences consisting of 2020 shots (251 408 frames).

The proposed storyboard generation method (Algorithm-III) is compared with the \( k \)-means clustering based scheme where \( k \) is taken as the size of the target storyboard. The \( k \)-means clustering algorithm is performed on the feature vectors of the key-frames. The nearest key-frame of a cluster centre is taken as the
Fig. 6. Storyboard representation of a documentary sequence.
representative of that cluster. The chronological sequence of these cluster representatives forms a storyboard of size $k$. Once the abridged storyboard is generated, it is evaluated in the same manner by comparing with the ground truth. Storyboard of various size have been generated following both the methods. Average precision values are shown in Table 4. Precision of all the systems are 85% or more. But, the proposed system performs better than the other two. $k$-means clustering-based technique suffers from a limitation that once the size of the storyboard is changed, it has to be regenerated afresh. In the proposed methodology, it is sufficient to generate the target storyboard once with size same as the complete storyboard. The order in which individual key-frames join in the spanning tree is noted and required number of such frames from the top is considered to prepare the target storyboard. We have also compared the performance of the proposed scheme with that of the system\cite{31}, where a hierarchical clustering technique is applied to split the video into an optimal number of clusters. The clusters contain the shots that are similar in some sense. Normalized weight of the each cluster is computed as the ratio between the length (or size) of the cluster and the total video. Then, for each shot, shot importance is calculated using the shot length and the weight of the cluster in which the shot belongs. The shots are arranged according to their shot importance. The chronological sequence of required number of key-frames from the

\begin{table}
\centering
\caption{Average precision of size constrained storyboard.}
\begin{tabular}{lccc}
\hline
\textbf{Proposed scheme (Algorithm-III)} & \textbf{$k$-means clustering} & \textbf{Uchihashi’s method\cite{31}} \\
\hline
0.91 & 0.89 & 0.85 \\
\hline
\end{tabular}
\end{table}
ordered shots forms the size constrained storyboard. In this case, the rare shots and the long duration shots are preferred. Thus, in a video sequence, if a long duration shot repeats to an extent, it will appear number of times in the size constrained storyboard. As a result, precision falls which is reflected in Table 4. Thus, experimental result indicates that the proposed scheme works better.

5. Conclusion

Though a shot is defined as a continuous video stream taken with a single camera operation, i.e. between on and off, depending on the events or activities recorded in the shot more than one key-frame may be required to represent the shot. Deciding the number of key-frames is not a trivial task and is usually done in an ad hoc basis. Most reasonable way to solve this problem is to divide a shot into suitable number of sub-shots such that each sub-shot is consistent in terms of visual content. Then from each sub-shot exactly one key-frame representing its visual content is extracted. Finally, a specified number of key-frames are selected to form a size constrained storyboard.

Here we have proposed a dynamic and content-based approach to extract the representative frame(s) or key-frame(s) of a shot. It neither presumes any fixed number of key-frames per shot nor depends on uniform sampling in temporal domain. Rather it relies on a non-parametric hypothesis test namely, Wald–Wolfowitz runs test to identify the uniform sub-shots with a given confidence level in a shot which needs independent representation. It is seen that multivariate Wald–Wolfowitz runs test outperforms the univariate Wald–Wolfowitz runs test coupled with majority voting because the later gives equal importance to both good and bad features in decision making. This concept of sub-shot detection is a unique aspect of our scheme. Another unique aspect of the proposed scheme is employing the spanning tree method to select a specified number of key-frames from the complete set of key-frames obtained from the entire video. Experimental result has shown that the proposed scheme can successfully handle the wide variety of shots. The size constrained storyboard thus obtained is evaluated a large number of viewers and the performance is found satisfactory. It may also be emphasized that merged shot(s) resulted in from miss detection in shot segmentation process may be corrected by the said hypothesis test.

References

A Novel Technique for Size Constrained Video Storyboard Generation

38. A. Wald and J. Wolfowitz, “On a test whether two samples are from the same population,” Annals Math. Statist. 11, 147–162 (1940).
Partha Pratim Mohanta received the B.Sc. and M.Sc. degree in Statistics from University of Kalyani, West Bengal, India in 1995 and 1997 respectively and obtained PGDCA from Regional Computer Center, Calcutta, West Bengal, India in 1998.

Currently, he is Associate Scientist in Electronics and Communication Sciences Unit of the Indian Statistical Institute. His research interests are in the area of image processing, video processing and pattern recognition.

Sanjoy Kumar Saha received his B.E. and M.E. degree in Electronics and Telecommunication Engineering from Jadavpur University, West Bengal, India in 1990 and 1992 respectively and obtained his Ph.D. from Bengal Engineering and Science University, West Bengal, India in 2006.

Currently, he is working as an Associate Professor in Computer Science and Engineering Department of Jadavpur University. His research interests are in the area of image processing, video processing, multimedia data retrieval and pattern recognition.

Bhabatosh Chanda received his B.E. in Electronics and Telecommunication Engineering and Ph.D. in Electrical Engineering from University of Calcutta in 1979 and 1988 respectively.

His research interest includes image and video processing, pattern recognition, computer vision and mathematical morphology. He has published more than 100 technical articles in refereed journals and conferences, authored one book and edited five books. He has received “Young Scientist Medal” of Indian National Science Academy in 1989, “Computer Engineering Division Medal” of the Institution of Engineers (India) in 1998, “Vikram Sarabhai Research Award” in 2002, and IETE-Ram Lal Wadhwia Gold medal in 2007. He is also recipient of UN fellowship, UNESCO-INRIA fellowship and Diamond Jubilee fellowship of National Academy of Science, India. He is Fellow of Institute of Electronics and Telecommunication Engineers (FIETE), of National Academy of Science, India (FNASc.), of Indian National Academy of Engineering (FNAE) and of International Association of Pattern Recognition (FIAPR). He is a Professor in Indian Statistical Institute, Kolkata, India.