A Design-of-Experiment Based Statistical Technique for Detection of Key-Frames

Snehasis Mukherjee, Dipti Prasad Mukherjee

Abstract

In this paper decision variables for the key-frame detection problem in a video are evaluated using statistical tools derived from the theory of design of experiments. The pixel-by-pixel intensity difference of consecutive video frames is used as the factor or decision variable for designing an experiment for key-frame detection. The determination of a key-frame is correlated with the different values of the factor. A novel concept of meaningfulness of a video key-frame is also introduced to select the representative key-frame from a set of possible key-frames. The use of the concepts of design of experiments and the meaningfulness property to summarize a video is tested using a number of videos taken from MUSCLE-VCD-2007 dataset. The performance of the proposed approach in detecting key-frames is found to be superior in comparison to the competing approaches like PME based method (IEEE Transactions on Circuits and Systems for Video Technology, 13(10), 2003, pp. 1006-1013), Mukherjee *et.al.* (IEEE Transactions on Circuits and Systems for Video Technology, 17(5), 2007, pp. 612-620) and Panagiotakis *et.al.* (IEEE Transactions on Circuits and Systems for Video Technology, 19(3), 2009, pp. 447-451).

Index Terms

Key-frame, video summarization, design of experiment, Helmholtz principle, meaningfulness, Gestalt.

I. Introduction

A video can be represented in a hierarchical fashion as follows: $[Frame] \rightarrow [Shot] \rightarrow [Video]$. The lowest level of video consists of individual frames. Each set of consecutive frames taken in a continuous capture period of a video camera constitutes a shot. A video consists of one or more such shots. This hierarchical concept of constitution of a video is widely used in the field of video summarization. Video summarization is useful in several areas of computer vision and related fields like content-based video retrieval, video indexing, visual enhancement, etc [1]. In this paper, we define video summarization as the process of extracting a collection of video frames (referred to as key-frames) that best represent the video. One of the most important tasks in video summarization is to efficiently detect these key-frames.

Snehasis Mukherjee and Dipti Prasad Mukherjee are with the Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata, India.

Manuscript received

Our approach to key-frame detection is a statistical approach. We select key-frames hierarchically in two steps. The first step is to split the whole video into a collection of consecutive frames and the second step is to select some representative frames as key-frames from each such collection or group. Let us call these collections as *units*. Here *units* do not necessarily represent shots. A shot in a video often contains multiple visual characteristics depicting different visual events. We term different visual events as different *units*. The *units* may be smaller or identical in size in comparison to shots. In other words, a shot may contain a single *unit* or more than one *units*, depending on the change in visual contents in that shot. Thus, a shot boundary is always a *unit* boundary, but the reverse may not be true. Formally we can define *units* as follows:

Definition 1*: Let* $\{f_d \mid d \in \{1, 2, \ldots, n\}\}$, $n \in Z^+$ *denoting the number of frames, be a sequence of consecutive frames in a video and* $\{p_d \mid d \in \{1, 2, \ldots, n\}\}$ (*referred henceforth as p-ratio*) *be the differences of color values* (*i.e., the image intensity values*) *of consecutive frames* ($p_1\ being\ zero$). *Then the set of unit boundaries is the collection of frames* $U = \{f_{\tau_1}, f_{\tau_2}, \ldots, f_{\tau_k}\} \subset \{f_d \mid d \in \{1, 2, \ldots, n\}\}$, *where* $\tau_1 = 1$, $\tau_k = n$, $k < n$ *and k-1 being the number of units in the video, satisfying the following criteria:*

(i) $\tau_j - \tau_{j-1} > FR\ \forall\ j = 2, 3, \ldots, k$ *where* FR *is the frame rate i.e., the number of frames per second and* $\tau_j,\ j = 2, 3, \ldots, k$ *are the* j*th unit boundary of the video.*

(ii) *If* p_{τ_j} *is the p-ratio at the* j*th unit boundary* τ_j*, then* p_{τ_j} *gets a value greater than a threshold* κ *for all* $j = 2, 3, \ldots, (k-1)$, $p_{\tau_1} = p_1 = 0$ *by definition. A unit is a collection of all frames between any two consecutive unit boundaries.*

Definition 1 covers both the two types of shot transitions - cut and gradual transition (e.g., fade in/out, dissolve, wipe, etc.). A thorough survey on various shot transitions is available in [2]. Cut is an abrupt change of visual content during shot transition. In gradual transition, visual contents change gradually. In case of gradual transition, we get a sequence of almost same high p-ratio value. We take the last among the consecutive frames with high p-ratio values as *unit* boundary. This *unit* boundary is chosen from the criteria (i) and (ii) of Definition 1. For cut, we get a sudden change in visual content, so we get a high p-ratio value for the frame where cut occurs.

We define the video *unit* based on significant changes on visual content and ignoring the subtle differences of definitions of shots and scenes, basically we want to keep the semantic analysis based approaches separate from image feature based approaches which we have adopted in this paper. Let us illustrate the advantage of introducing *units* over shots by the example shown in Figure 1. The Figures 1(a) and 1(b) are taken from the same shot but in Figure 1(b), one person has gone out of the frame. So these two figures contain different visual information and kept into different *units* by the proposed approach. Similarly, the Figures 1(c) and 1(d) are taken from the same shot, but in Figure 1(d), the color of the leaf has been changed from green to magenta. So these two figures contain different visual information and kept into different *units* by the proposed approach.

In order to split the video into *units* according to Definition 1, we first assign a binary value corresponding to each pixel of all the frames of the video. These binary values are given depending on a threshold γ on the differences of color values of the consecutive frames. Then for each frame we calculate the p-ratio indicating the possibility of the frame to be a *unit* boundary. According to the Definition 1, if the p-ratio is greater than a threshold κ, then

the corresponding frame is a *unit* boundary. Now the problem is to select the two threshold values γ and κ. We select these threshold values using a statistical tool called the design of experiments [3]. Then we do the ANOVA F-test [4] to check if this grouping of frames is good enough to find the *units* according to Definition 1. We extract meaningful frame(s) from each of these *units* by finding out meaningful change of each frame from the previous frame of that *unit*. The final collection of meaningful frames is the collection of key-frames. Formally key-frames can be defined as follows:

Definition 2*: The set of key-frames is the collection of some frames* $K = \{f_{q_1}, f_{q_2}, \ldots, f_{q_m}\} \subset \{f_d \mid d \in \{1, 2, \ldots, n\}\}$, $m < n$, $q_1 \geq 1$, $q_m \leq n$, *satisfying the following criteria:*

(i) *If* $\Theta \subset U \times U$ *(*U *being the set of all unit boundaries in the video) such that* $\Theta = \{(f_{\tau_{j-1}}, f_{\tau_j}) \mid j = 2, 3, \ldots, k\}$, *then* $\forall\ (x, y) \in \Theta$, $\exists$ *at least a frame* f *between the unit boundaries* x *and* y *such that* $f \in K$.

(ii) *In each unit, if* $f_d \in K$, *then* p_d *must be either a local maxima with respect to p-ratio values of the neighboring frames in that unit preceding and following the frame* f_d, *or the temporally middle frame of several consecutive frames with identical high p-ratio values.*

(iii) *For each unit,* $\exists$ *a cut-off* η *for a measure combining p-ratio values and the local maxima position of* f_d *in the unit such that all frames in K satisfies* η.

The condition (iii) in Definition 2 is derived from the concept of meaningfulness which we introduce next.

A. Meaningfulness

The main idea of the meaningfulness is derived from the Gestalt hypothesis. The Gestalt hypothesis is being used to solve several problems in the field of computer vision [5]. According to the Gestalt theory, "grouping" is the main concept for our visual perception [6]. Points in 2D or 3D space form any random point pattern(s). These points may be considered as some random objects without any meaning to us. However, when this same point pattern(s) exhibits some common visual characteristics (for example, alignment, parallelism, convexity, good continuation, etc.), then this point patterns are grouped to form a new, larger visual object called a *Gestalt* [6]. A set of partial Gestalts, due to certain matching visual characteristics, can be grouped to form a Gestalt.

In [5], Desolneux, Moisan and Morel have shown some computational techniques to decide whether a given partial Gestalt is meaningful or not (partial Gestalt may represent straight line, object boundary, etc.). For this, they have used a general perception law, called Helmholtz principle. Suppose there are α objects, θ of them having similar characteristics with respect to some *a priori* knowledge (e.g., same color, same alignment, etc.). Then the question is that are these characteristics happening by chance, or is there any significant cause to group them to form a partial Gestalt? To answer this question, we first assume that the characteristics are uniformly distributed over all the α objects and the observed characteristics are some random realization of the uniform distribution. The Helmholtz principle says that if the expectation of the observed configuration of θ objects (here we call a configuration of some objects as an event) is very small, then the grouping of these objects is a Gestalt. We calculate the expected number of occurrences of the observed configuration of θ objects, which is called the number of false alarms (NFA) of that event [5]. The NFA of an event gives a measurement of meaningfulness of the event. The smaller the NFA is, the

more meaningful the event is. (Good things come in small packages.) If the NFA is less than a certain number ϵ then the event is called an ϵ-meaningful event; otherwise it is a random event. That is, by definition, a meaningful event is significantly different from random events and has a very small NFA.

In this paper, we select some frames from each of the *units* in a video, such that these frames have meaningfully large frame difference (according to (ii) of Definition 2). Suppose there are α frames in a particular *unit*, θ of them having similar characteristics (larger frame difference). Then by Helmholtz principle, if the expectation of the observed configuration of the frame differences (i.e., NFA) of the θ frames is very small (less than a number ϵ), then the collection of the θ frames is ϵ-meaningful and are selected as key-frames of that *unit*. Here the measure η of (iii) of Definition 2 is an expectation of an observed configuration of frame differences and spatial conditions of local maxima of p-ratio values in a *unit*.

Before we elaborate the concept of the proposed approach, we present the related works on key-frame detection.

B. Related Works

Several approaches have been proposed for key-frame extraction. We can broadly categorize the existing key-frame extraction approaches into four major classes: *sampling-based*, *clustering-based*, *shot-based* and *other* approaches.

A *Sampling-based* key-frame extraction technique selects the key-frames by uniformly sampling the video frames at certain time intervals. Most of the earlier key-frame extraction methods belong to this category [7]. The main drawback of the *sampling-based* approach is that a shorter video segment may not have any representative frame, whereas a larger video segment may have unnecessarily many representative frames with almost similar visual content, thus failing to represent the content of the whole video efficiently.

A *clustering-based* key-frame extraction method detects multiple frames as key-frames using unsupervised clustering based on the variations in video content [8]. In [9], spectral clustering approach is applied for key-frame extraction. It is difficult to estimate the initial key-frame cluster leading to appropriate number of key-frames. This is the main problem in the *clustering-based* approach.

A *shot-based* technique is a popular approach comprising of two steps: first find different shots or scenes, and then find representative key-frame(s) in each of these shots or scenes. Many efforts are found in literature in detecting shots in a video [10], [11]. In [11], Adjeroh *et.al.* uses the MUSCLE-VCD-2007 dataset for shot detection. A good survey of shot detection techniques can be found in [12]. In this approach, the main feature to find a key-frame is the pixel-by-pixel consecutive frame difference [13], [14]. In [15], frames in each shot are considered as vertices of a graph. A graph similarity matrix is then formed based on change in visual contents. The graphs are split to subgraphs using normalized cut. Then key-frames are extracted from each subgraphs using Key-Frame-Vector (KFV). However, this method does not consider a massive change in a small area of a frame, which is one of the weak points of this method. There are some other *shot-based* approaches where optical flow or motion information is used to select key-frames [16], [17]. The number of key-frames should not be predetermined because due to content variation, the number of key-frames may be different for each shot [9]. For example, in a static shot, where the visual content does not change very rapidly, one key-frame may be enough to represent the shot, whereas for a

shot with massive camera or object motion may be represented by more than one frames. Some idea for variable number of key-frame selection from different shots in a video have been introduced in [13], [17], [9]. To detect key-frames, Mukherjee *et al.* [13] have compared the point pattern in feature space with the simulated random spatial point patterns. Note that an n-dimensional feature is a point in n-dimensional coordinate (feature) space. Here image, video or frame difference can be taken as features. Note that pixel-by-pixel difference of frames can also be visualized as a frame difference image. In this case, features are detected in the frame difference image and a quantitative measure of the closeness of this feature point pattern with simulated random pattern determines key-frame [13]. The optical motion energy model is used to detect key-frames in [17]. In this case the motion magnitude derived from a video sequence is weighted with the motion vector direction to determine key-frames.

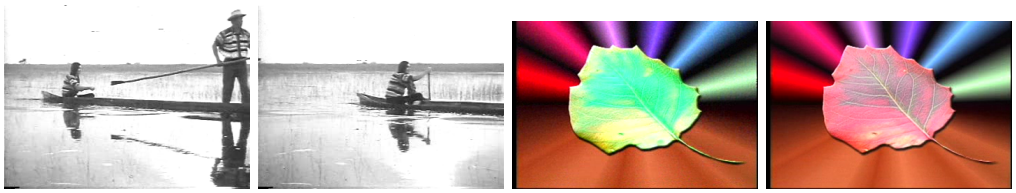
Frame differences in a video are also characterized by the difference of histograms of the pixel intensities of consecutive frames [18]. In that case, significant global changes in consecutive frames may be detected, but local information is ignored due to the use of a single histogram for the entire image. Valdes *et al.* [19] presents a video abstraction method based on the operational aspects of the key-frame selection algorithms. They study some existing frame difference based video abstraction methods found in the literature and synthesize their approaches to generalize them to lead to a unified model.

There are some *other* approaches for key-frame extraction. For example, [20] proposes an efficient key-frame extraction technique by object detection. This method extracts key-frames such that the divergence between video objects in a feature space can be maximized, supporting robust and efficient object segmentation. In [21], an interactive computing model is proposed to select the key-frames. An efficient clustering based key-frame extraction technique is proposed by Spyrou *et al.* [22]. They extract low level features from the frames to build a vector called model vector and construct a region thesaurus. Then key-frames are selected by clustering those images. In [23], changes in visual contents of the consecutive frames are measured by taking difference between the Discrete Cosine Transform (DCT) of consecutive frames. But to summarize a video emphasizing on a significant local change in a frame is largely an unexplored area which is the motivation behind the proposed approach.

C. Motivation

Key-frame detection techniques broadly rely on either frame intensity difference, or histogram difference of successive frames. But it may lead to loss of visual information in cases where massive changes occur in a small area of the frame. The proposed approach relies on frame intensity difference but the method promotes local change of intensity by defining the frame difference by an exponential function of color change weighted by the pixel area over which the change has occurred. Our intention is to provide an objective assessment of the frame difference value over which a frame can be declared as key-frame. The notion of meaningfulness based on the strategies of design of experiment precisely estimate necessary parameters.

We introduce the concept of *unit* (Definition 1) to assure at least a key-frame to be selected of all different visual characteristics (even local) present in, even a shot. Consider the examples of some key-frames selected by our method shown in Figure 1. The proposed method have successfully recognized the visual change in the video



(a) (b) (c) (d)

Fig. 1. Examples of 2 pairs of key-frames selected by the proposed algorithm from the same shots but different *units*. (a) frame number 386 and (b) frame number 423 of 'indi114.mpg' video. Here one person has gone out of the frame in the same shot. (c) Frame number 1094 and (d) frame number 1222 of 'NASA-leaves.mpg' video. Here the color of the leaf has been changed in the same shot.

and summarized the video keeping all visual information intact.

Section II presents the proposed approach for grouping the video sequence into a collection of *units*. In Section III, we propose an empirical model to show the independence of the two parameters used for *unit* detection. The procedure for extracting key-frames from each *unit* is detailed in Section IV. We have tested our method on 21 benchmark videos and the results are compared with [13], [17], [23] in Section V. We draw our conclusion in Section VI.

II. Grouping the Video Sequence

The grouping is done based on frame differences of the consecutive frames of a video. This process has three major steps. First we calculate the differences of each pixel for each consecutive pair of frames. The second step is to calculate the *p*-ratio. In third step, we need to find suitable thresholds for the frame difference value and the p-ratio to group the frames to generate *units*. We start with calculating pixel differences.

A. Calculating Pixel Differences

We are working with color videos, where each frame is an image described in three channels namely, red, green and blue ($[R\ G\ B]$). Given that the frames of the video are defined in Z^2 and the color values are mapped between 0 and some integer ***R*** (say, 255), and the vectors $[R\ G\ B]_{h,d-1}$ and $[R\ G\ B]_{h,d}$ denote the ***RGB*** values of the *h*th pixel in the (*d*-1)th and the *d*th frames respectively of a video, $2 \leq d \leq n$, n being the number of frames in the video, then

$$E_h^d = dist\left([R\ G\ B]_{h,d}, [R\ G\ B]_{h,d-1}\right), \tag{1}$$

where the function $dist\,(V_1, V_2)$ denotes the Euclidean distance between the vectors V_1 and V_2.

B. Calculating p-ratio

After calculating E_h^d values, our next task is to select a suitable threshold γ on the E_h^d values to indicate whether the *h*th pixel votes for the *d*th frame to be a *unit* boundary. The process of selecting the threshold γ is

determined using some standard statistical tools (to be discussed in the next subsection). We give a Boolean value β_h^d corresponding to each E_h^d value using the following voting scheme:

$$\beta_h^d = \begin{cases} 1 & \text{when } E_h^d > \gamma \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

β^d is the corresponding binary matrix for the *d*th frame (except the first frame) of the video, where connected components of 1-element of the β^d matrix represent significant frame difference regions. This pruning of non-connected components does well as a substitute of filtering the frames of the test video. Now our aim is to calculate the p-ratio as a suitable measure for each frame, indicating its possibility to be a *unit* boundary according to Definition 1. In our proposed model, we would like to give importance to a large change in a smaller area of a frame compared to a similar amount of change over a larger frame area. So the p-ratio is calculated using an exponential function because, besides noticing changes in the whole frame, this p-ratio should change exponentially with a significant change in a small area in a frame, rather than a small change throughout the frame. For each frame we calculate the p-ratio, p_d as follows:

$$p_d = \exp\left(\frac{\sum_{h=1}^{N} E_h^d \beta_h^d}{A}\right), \tag{3}$$

where *N* is the total number of pixels in each frame, *A* is the total area of all the connected components (of area greater than 1 pixels) in the β^d matrix. p_d is normalized between 0 and 1 by dividing all the p_d values with their maximum. Fig. 2 shows the plots of the p-ratio values over time, for four representative test videos. From the graphs in Fig. 2, we observe that high p-ratio value of a frame is an indication to be a *unit* boundary. For a gradual transition, we get a flat curve with some local maxima. For cut, we get a high p-ratio value for the frame where cut occurs. In Fig. 2, most of the cases we can observe a cut. But in Fig. 2(b) and (d), we can observe some case of gradual transitions.

Hence, for each frame *d* except the first frame, we get a p-ratio value. We assign a p-ratio value 0 for the first frame. Clearly, a greater value of the p-ratio indicates a greater probability of the *d*th frame to be a *unit* boundary. Our next task is to apply the concept of design of experiments to find a suitable value of γ and a suitable threshold κ on the value of the p-ratio beyond which we can select the *d*th frame as a *unit* boundary [3].

C. Finding the Threshold Values γ and κ

We can consider an experiment as a black box having one or more inputs and some outputs. The inputs of the experiment are termed as factors. Each factor of an experiment can take several values called levels. In experimental design paradigm, we check the outputs of the experiment for each possible combination of levels and select the combination of levels giving optimum value of the outputs. In our experiment, the two threshold values γ and κ are the two factors and the only output is the F-ratio [4]. In the proposed approach, the F-ratio is calculated as the ratio of between-group-variability and within-group-variability of the p-ratio values after the grouping of the frames as follows:

$$F = \frac{s_b^2}{s_w^2}, \tag{4}$$

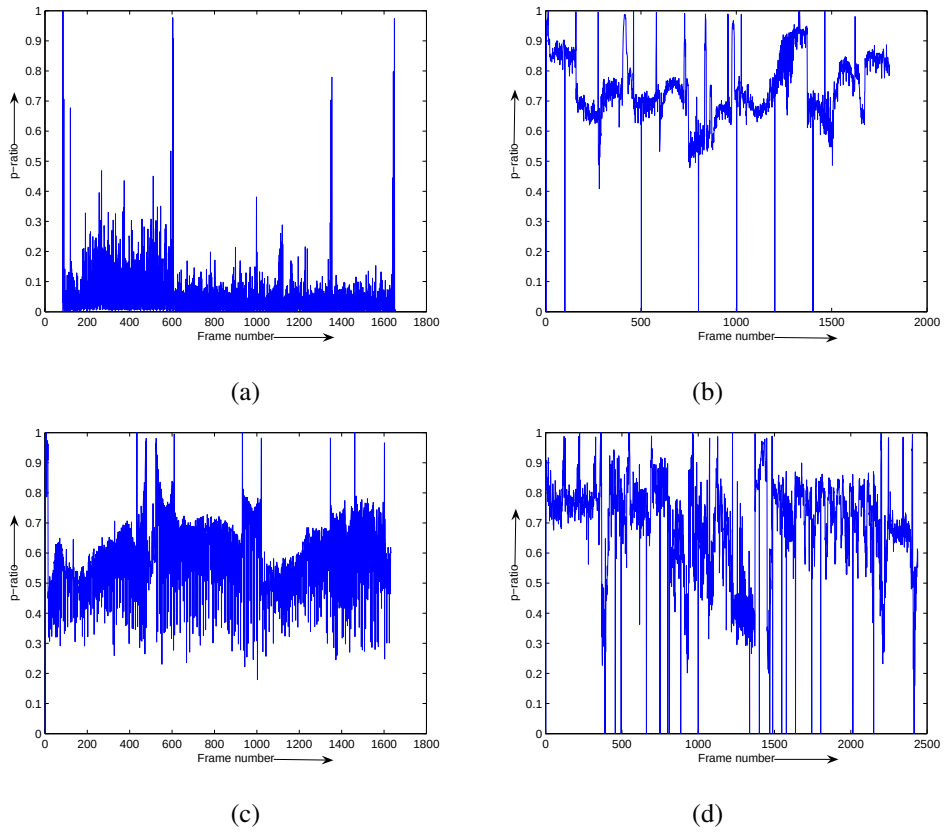


Fig. 2. The plots of p-ratio values over time for four representative test video (a) '0863.mpg' for $\gamma = 6$, (b) 'NASAKSN-NorthernLights.mpg' for $\gamma = 8$, (c) 'indi114.mpg' for $\gamma = 7$ and (d) 'hcil2000_01.mpg' for $\gamma = 8$.

where s_b^2 and s_w^2 are the between-groups and within-group variability of the p-ratio values respectively [4]. For an ideal grouping, s_b^2 should be very high while s_w^2 should be as low as possible, but not equal to 0. Even if there are two groups of frames temporally apart from each other with almost similar mean p-ratio value, then also s_b^2 should give a higher value.

The between-group variability s_b^2 is the variability of the mean of the p-ratio of the frames of each group of the video versus the grand mean p-ratio of all the frames in the video. If k is the total number of groups for the current combination of levels of the factors, then s_b^2 is calculated as:

$$s_b^2 = \frac{\sum_{\varrho=1}^{k} m_\varrho \left(\bar{e_\varrho} - \bar{E}\right)^2}{(k-1)}, \tag{5}$$

where m_ϱ is the number of frames in the ϱth group, $\bar{e_\varrho}$ is the mean p-ratio of the ϱth group and $\bar{E}$ is the mean p-ratio of the entire video. Since for k groups we must know the $\bar{e_\varrho}$ of (k-1) groups given $\bar{E}$, (k-1) is the degrees of freedom for between-group variability [4].

The within-group variability s_w^2 is the variability of the p-ratio due to differences within individual group. The

value of s_w^2 is given by the following equation:

$$s_w^2 = \frac{\sum_{\varrho=1}^{k} (m_\varrho - 1)\, s_\varrho^2}{(n-k)}, \tag{6}$$

where *n* is the total number of frames in the video, s_ϱ^2 is the variance of the p-ratio of the frames within the ϱth group, (m_ϱ-1) is the degrees of freedom of the ϱth individual group and (*n*-*k*) is the degrees of freedom for within-group variability, which is the sum of degrees of freedom of all the groups [4].

Here the optimum value of the output is the higher F-ratio value. We go for the one-factor-at-a-time experiment, where we vary the level of only one factor at a time and block the value of the other factor at a certain level [3]. This technique works well only if there is no interaction between the factors. In Section 3, we are going to show that in our problem, the two factors are independent to each other.

In our experimental design set up, the first task is to choose some discrete levels of each of the two factors γ and κ. Since E_h^d can take any value from the closed interval [0,*R*] (if the frames have unsigned 8-bit color values, then *R* cannot exceed 255), we take all the integral values in that interval as the levels of the factor γ. From (3), it is clear that the p-ratio can take any value from the closed interval [0,1]. As we have to deal with some discrete levels of the factor κ, we take the levels of κ between 0 and 1 in steps of δ_κ. We block the value of γ and vary the value of κ through all the levels of κ. Then we block the value of γ at another level and again vary the value of κ through all the levels of κ and so on. In this way, we get different groups of frames depending on each combination of levels of γ and κ. Since according to Definition 1, the frames with similar visual contents (i.e., with low value of p-ratio) should be in the same group, then for an ideal grouping, the variation of the p-ratio values within a group should be significantly less than the variation between groups, i.e., the F-ratio value should be significantly high (according to (4)) [4]. The combination of levels giving the highest F-ratio gives us the suitable values of both the thresholds. We call the groups of frames obtained for the optimal combination of levels as different *units* of the video.

D. ANOVA F-test

Clearly, s_w^2 should not be zero in order to get a finite value of F-ratio from (4). A zero value of s_w^2 may occur in two ways. First, if each of the frames in the video is a singleton group, then s_w^2 will be zero. This situation should be avoided during the detection of *units* because, according to Definition 1, each and every frame of a video cannot be a *unit* boundary. Definition 1 imposes a constraint that the number of frames in a *unit* cannot be less than the frame rate *FR*, which is typically 30 frames per second. In reality though, a *unit* has duration of several seconds. Second, when all the frames in the video have the same p-ratio, then (4) gives us a (0/0) value. This problem can only occur when the contents of all the frames in the video are exactly same. In that case, it is clear that there is no change in the whole video. So the first and last frames of the video are the only two *unit* boundaries in the whole video. Another problem that may occur is that when the number of groups becomes large and consequently, the number of frames in each group becomes low, then s_w^2 may be very small compared to s_b^2. This problem can also

be prevented by the assumption that the duration of each *unit* must satisfy (i) of Definition 1. So in our problem, we discard combinations of levels that give *units* that does not satisfy (i) of Definition 1.

The key contribution of this paper is to estimate κ and γ objectively. By our design of experiments based method, these thresholds are evaluated accordingly to take care of quasi-static background, camera motion conditions etc. According to the Definition 1, a shot boundary is always a *unit* boundary. In case of gradual transition during the process of (i.e., *unit* detection, a number of consecutive frames may get high p-ratio values. Whenever, for a given combination of levels, we get some temporally close frames with high p-ratio values, we take the last one of those frames as the *unit* boundary.

From the derivation done so far, the greater the F-ratio the better is the grouping. Higher F-ratio indicates large value of s_b^2 compared to s_w^2 [4]. That is frames within groups have p-ratio close to each other and the mean p-ratio of different groups are far from each other. After calculating the F-ratios for all combinations of levels of the two factors γ and κ, we select the combination of levels of the two factors responsible for the highest F-ratio (say, F_{opt}). In case F_{opt} is low, the entire video splits into a number of groups. To check if the grouping is good, we do the F-test. That is, we check the null hypothesis that the within-group mean p-ratio (μ_1) equals the between-group mean p-ratio (μ_2), against the alternative that $\mu_1 \neq \mu_2$. For this, we find from the F-distribution table [4], the critical value (say, F_{crit}) of F-distribution at 95% confidence level for (k-1) numerator degrees of freedom and (n-k) denominator degrees of freedom. If $F_{opt} > F_{crit}$, then we reject the null hypothesis, i.e., we consider our grouping as good. Otherwise we accept the null hypothesis that the grouping is not good enough according to ANOVA F-test [4]. This implies that there is no significant change throughout the video and we can consider the entire video as a single *unit*.

As we stated in the Introduction, the one-factor-at-a-time experiment can be designed only if there is no interaction between the factors. We now give an empirical evidence supporting independence of the two factors of our experiment.

III. Empirical Evidence Supporting Independence of γ and κ

In Fig 3, we have shown examples of our one-factor-at-a-time experiment for four representative test videos. For each test sequence, γ and κ values are evaluated by experiments detailed in Section II. We have applied the design of experiments tool for determining the thresholds for each of the 21 test videos separately. We have displayed only 4 representatives of them for brevity of space. For each test video sequence we have to find the threshold using the design of experiments paradigm and use the threshold to find the *unit* boundaries.

For a two-factor experiment, the one-factor-at-a-time approach is simple and less time consuming. As discussed in the Section 2.3, we block the value of one factor (say, *A*) at a certain level and find results for different levels of the other factor (say, *B*). In this case, the implicit assumption is that the factor *A* does not depend on the factor *B*. If there exist an interaction between the two factors, then the change of level of the factor *A* influences the result obtained with the change of level of the factor *B*. If the different factors of our experiment have some interaction, then this technique may not give the optimal result. In our experiment, we first have to check for the interaction between

the factors γ and κ. If there is no interaction between them, we can follow the one-factor-at-a-time experiment.

The Severity Index (*SI*) value is a measurement of interaction between the two factors of an experiment [24]. The *SI* values are calculated on the basis of some measure of effect due to the individual factors. In our problem, γ and κ are the factors of the experiment of finding the optimal value of the two factors. If γ_1, γ_2 and κ_1, κ_2 are the two levels of the factors γ and κ respectively and the result (here F-ratio is treated as the result of the experiment) given by the combination γ_i, κ_j (*i*=1 or 2; *j*=1 or 2) is denoted by $R(\gamma_i, \kappa_j)$, then *SI* is given by the following:

$$SI = \frac{|[R(\gamma_2, \kappa_1) - R(\gamma_1, \kappa_1)] - [R(\gamma_2, \kappa_2) - R(\gamma_1, \kappa_2)]|}{2\rho} \times 100\%, \tag{7}$$

where ρ is the difference between maximum and minimum of all four values of $R(\gamma_i, \kappa_j)$ (*i*=1 or 2; *j*=1 or 2). Both the factors have more than two levels. In this case, we have calculated *SI* values for each consecutive pair of levels of the two factors. We can get F-ratio values $R(\gamma_1, \kappa_1)$, $R(\gamma_1, \kappa_2)$ for two different levels κ_1, κ_2 for one particular blocked value γ_1. Similarly, we can get F-ratio values $R(\gamma_2, \kappa_1)$, $R(\gamma_2, \kappa_2)$ for κ_1, κ_2 for a different γ_2 value. If the line joining $(R(\gamma_1, \kappa_1), R(\gamma_1, \kappa_2))$ is parallel to $(R(\gamma_2, \kappa_1), R(\gamma_2, \kappa_2))$, we can assume that there is no interaction between factors γ and κ.

The decision about the presence of interactions between the two factors is made by strictly comparing the slopes of the two lines given by each pair of consecutive levels of κ, when the value of γ is blocked at a certain level. The strength of presence of interaction between the two factors of our experiment is calculated by *SI* which is defined such that $SI = 100\%$ when the angle between the lines is 90^0 and $SI = 0\%$ when the angle is 0^0. We take the *SI* values for each consecutive pair of levels of the factors γ and κ.

The graphs in Fig. 3 show the interaction plots, i.e., the interaction between the two factors used in our experiment for four representative test videos '0863.mpg', 'NASAKSN-NorthernLights.mpg', 'indi114.mpg' and 'hcil2000_01.mpg'. Each line in the graph shows the changes of the F-ratio (plotted along y-axis) with the change of one of the factors κ (plotted along x-axis), when the other factor γ is fixed. We check the interactions between the factors for all combination of levels of four representative test videos. We have plotted the lines where we could calculate the values of F-ratio subject to the constraints discussed in Section 2.4. From the graphs, it is clear that the line segments for each combination of levels are almost parallel, which means there is no interaction between the two factors of our experiment. A detail study of interactions between the factors using *SI* values with four representative test videos under investigation is discussed in the Result section using Table I.

After grouping the video sequence, we get *units* of consecutive frames. Our next task is to find representative key-frames from each of these *units*.

IV. Selecting Key-Frames from Each Group

As noted in the Introduction, we select some representative frames from each *unit* as key-frames. The same technique using F-ratio cannot be applied here. Because F-ratio values do not depend on the temporal information of the video which is, according to Definition 2, an important parameter for key-frame detection. Each key-frame in a video should be a frame with a significant frame difference measure and at the same time the key-frame should not be very close to each other.

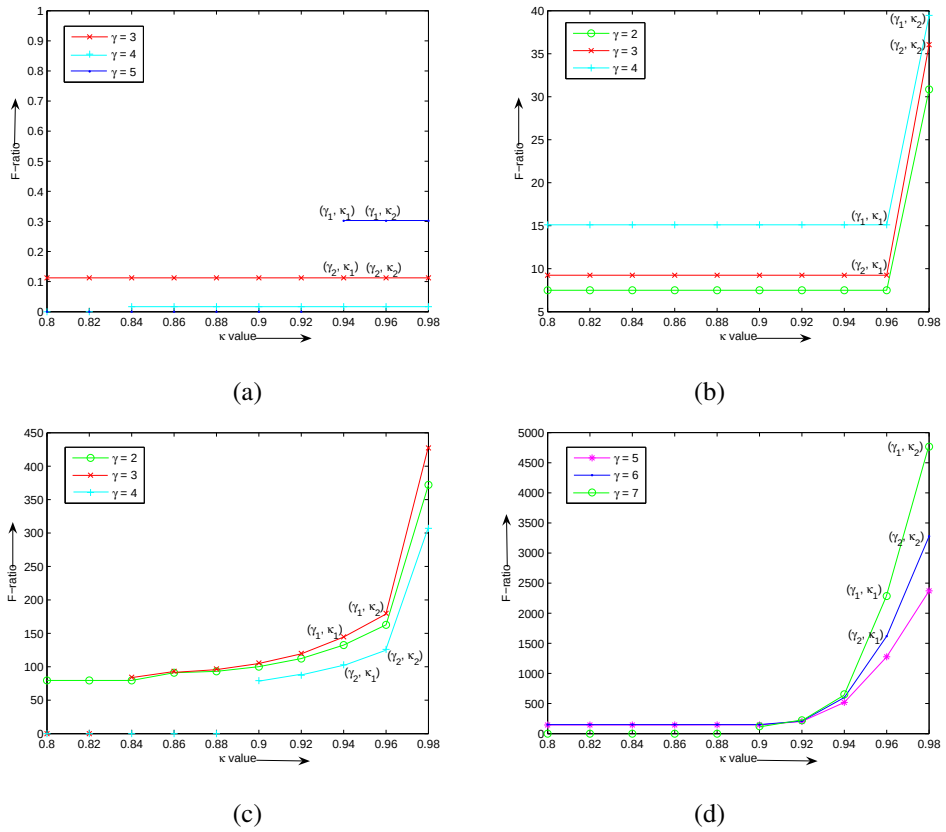


Fig. 3. Graph showing interaction between the factors γ and κ of the four test videos: (a) '0863.mpg', (b) 'NASAKSN-NorthernLights.mpg', (c) 'indi114.mpg' and (d) 'hcil2000_01.mpg'.

Each frame in a *unit* has a p-ratio value. Each frame has a location in temporal scale (for example, 65th frame of a 5 second video at 30 frames per second). Also according to Definition 2, there should be sufficient distance in temporal scale between two consecutive key-frames. Let m_d^i be the number of frames in the *i*th *unit* between the nearest preceding and following frames having p-ratio higher than that of the *d*th frame. That is, if the *d*th frame is a local maximum of p-ratio values, then the total number of frames between local maxima previous and next to the *d*th frame is m_d^i. Note that the start and the end frames of the *unit* are also considered as local maxima. Let m_i be the number of frames in the *i*th *unit*. Then $\frac{m_d^i}{m_i}$ denotes the fraction of the frames of the *i*th *unit* whose local maxima is p_d. Clearly, this ratio takes any value between 0 and 1. Then for the *d*th frame, we define the l-ratio, l_d as:

$$l_d = \frac{p_d + \frac{m_d^i}{m_i}}{2}. \tag{8}$$

We divide the numerator by 2 to normalize the value of l_d between 0 and 1. From (8), it is clear that higher p-ratio value increases the l-ratio value, but temporally close frames with high p-ratio value decreases the l-ratio value. So according to Definition 2, higher the l-ratio for a frame, the higher is the chance of the frame being a key-frame. We shall now define a threshold on the l-ratio values, referred henceforth as meaningful cut-off value,

say η, on l-ratio. The frames having l-ratio higher than this cut-off constitute a set of possible key-frames.

We should not calculate F-ratio as in Section II-D using the l-ratio for key-frame selection. Because if we get a pair of temporally close frames with a high l-ratio value, the value of s_w^2 becomes very low in comparison to s_b^2. As a result, as much as we reduce the cut-off on the l-ratio values, the value of F-ratio will be higher. Therefore, this approach cannot give us a meaningful cut-off. We avoid this situation during *unit* detection by assuming that a *unit* must satisfy (i) of Definition 1. But there should not be any restriction on the number of frames between two consecutive key-frames. We apply the concept of meaningfulness to find a suitable cut-off η of l-ratio for each *unit*.

For finding the meaningful cut-off value η of the l-ratio, we first select λ equidistant points in [0,1], the range of the l-ratio values. We vary the value of the threshold η over all the chosen equidistant points in [0,1]. For each η we do the following two steps given by the two equations (9) and (10). If ν is the prior probability that an arbitrary frame has l-ratio higher than η, then

$$\nu = 1 - \eta, \tag{9}$$

assuming that the values of the l-ratio are independent and identically distributed (i.i.d.) in the interval [0,1] and the distribution is uniform.

Let t ($1 < t < m_i$) be the minimum number of frames with Gestalt quality (i.e., meaningful with respect to Gestalt theory as discussed in Introduction) [6]. The cut-off η is meaningful if the *unit* contains at least t frames due to the cut-off η. Therefore, whether a particular cut-off value is meaningful or not becomes a Bernoulli trial (Bernoulli trial is an experiment having only two possible random outcomes; in our problem, whether a cut-off value of the l-ratio is meaningful or not). If $(1 - P_i)$ is the probability that the cut-off η is meaningful, then

$$P_i^{\eta} = \sum_{o=t}^{m_i} \begin{pmatrix} m_i \\ o \end{pmatrix} \nu^o \left(1-\nu\right)^{m_i - o}, \tag{10}$$

is the Binomial tail, where ν comes from (9).

As noted in the Introduction, NFA is the expected number of occurrences of the meaningful event. If the problem is like a Bernoulli trial, then the NFA of the event that the particular cut-off η is significant for detecting key-frames, can be defined as

$$NFA = \lambda P_i^{\eta}, \tag{11}$$

where P_i^{η} comes from (10). λ is the number of equi-spaced η values in the interval (0,1) to estimate the meaningful η. In other words, λ is the number of trials. If the value of NFA is less than a predefined number ϵ, then the corresponding cut-off value η is ϵ-meaningful. Setting $\epsilon = 1$ as in [5], means that the expected number of occurrence of the event that, the cut-off $\eta = \eta'$ is meaningful for the corresponding *unit*, is less than 1. So all the values in the interval $(\eta', 1)$ are '1-meaningful cut-off' values, since for all values in $(\eta', 1)$, $NFA < 1$ according to (11).

A. Estimation of Parameters for Meaningful Cut-off

The parameter λ is used to calculate the NFA in (11) and P_i^η is obtained by (10). Then the probability that all the frames in the *i*th *unit* have l-ratio greater than η, is ν^{m_i}. This is lesser or equal to the probability that at least t frames have l-ratio less than η, which is P_i^η. So $\nu^{m_i} \leq P_i^\eta < \frac{\epsilon}{\lambda}$ (since according to (11), for an ϵ-meaningful event, $NFA = \lambda P_i^\eta < \epsilon$), which implies

$$m_i \geq \frac{\log \epsilon - \log \lambda}{\log \nu}. \tag{12}$$

For a given η, ν comes from (9). m_i is fixed for a *unit*. Then for a given ϵ, we can find an upper bound on the value of λ from (12) for each possible η value. We take the minimum of all the values as the upper bound of λ.

The parameter t is the minimum number of frames with Gestalt quality. From Hoeffding's inequality [25], for an ϵ-meaningful event we can deduce the following:

$$t \geq \nu m_i + \sqrt{\frac{m_i}{2}(\log \lambda - \log \epsilon)}. \tag{13}$$

The equation (13) is the sufficient condition of meaningfulness. The derivation comes from Hoeffding's inequality, shown in Appendix A. Now we have to select the maximal meaningful cut-off.

B. Selecting Maximal Meaningful Cut-off

Setting $\epsilon = 1$ is a safe choice to find the maximal meaningful cut-off. However, for choosing the maximal meaningful cut-off, we should have some measure of meaningfulness. For this purpose, we should consider the empirical probability of a frame of that *unit* to have its l-ratio to fall in the interval [η',1]. Let $r_i(\eta')$ be the empirical probability of a frame of the *i*th *unit* to have l-ratio in the interval [η',1]. Then

$$r_i(\eta') = \frac{m_i(\eta')}{m_i}, \tag{14}$$

where $m_i(\eta')$ is the number of frames in the *i*th *unit* having l-ratio greater than η'.

Now our goal is to find the maximal meaningful cut-off among all the 1-meaningful cut-off values (i.e., the interval (η',1)). So we vary ξ in the interval (η',1) to measure the maximal meaningfulness of all possible ξ's. First we have to define a measurement of maximal meaningfulness of the cut-off value, using $r_i(\eta')$. This measurement should penalize the situation that a 1-meaningful cut-off ξ yields higher empirical probability value than ν. This measurement (let us call it as *c*-value) should also help to reduce the number of key-frames in a *unit*. However, according to definition 2, the corresponding *unit* must have at least one key-frame. Then *c*-value can be defined as,

$$c_i(\xi) = \begin{cases} \infty, \ when, \ r_i(\xi) \geq \nu, \ or, \ r_i(\xi) = 0 \\ r_i(\xi) \log \frac{r_i(\xi)}{\nu} + (1 - r_i(\xi)) \log \frac{(1-r_i(\xi))}{(1-\nu)} \quad \text{otherwise} \end{cases}, \tag{15}$$

where ξ can take any value from the interval (η',1). We take the open interval instead of the closed interval, in order to avoid division by zero in (15), which occurs if $\nu = 0$.

Fig. 4 shows the *c*-values of each *unit* for four representative test videos (as the cut-off of l-ratio, ξ increases). The *c*-values first increase upto some value of ξ, then decreases. If we vary ξ within (η',1), we get *c*-value as a decreasing function of ξ.

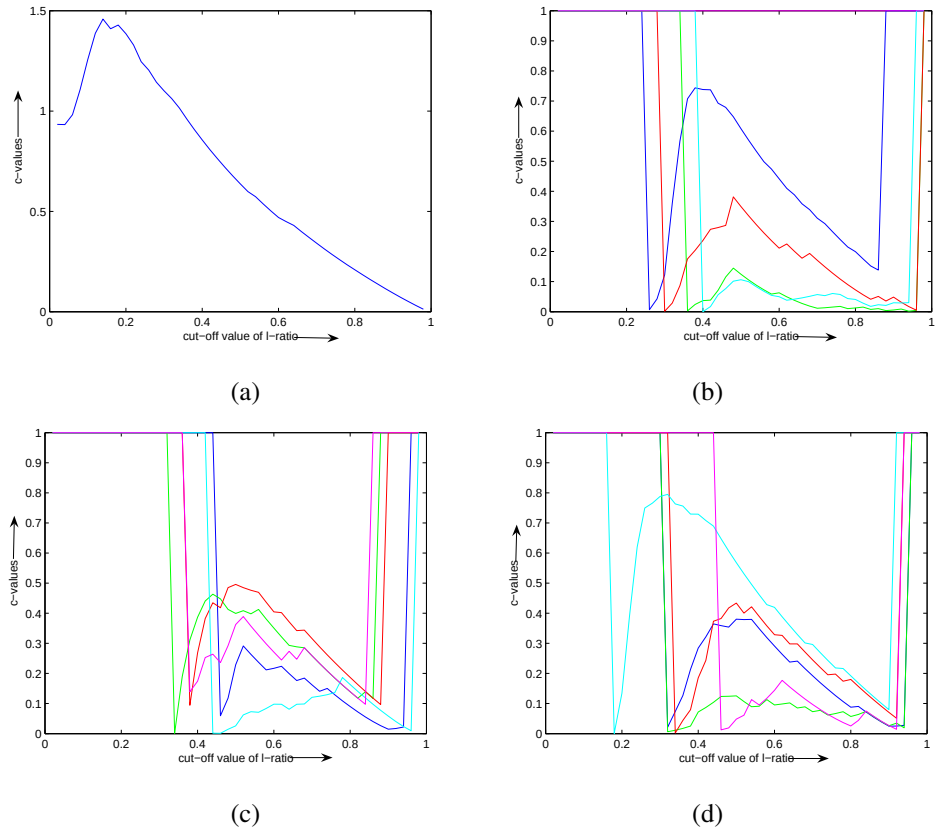


Fig. 4. Graphs showing the value of $c_i(\xi)$ (plotted along y-axis) against the different cut-off values ξ (plotted along x-axis) between 0 and 1, c-values for different *units* are shown in different colors for the four representative test videos (a): '0863.mpg' (b): 'indi114.mpg' (c): 'hcil2000_01.mpg' (d): 'NASAKSN-NorthernLights.mpg'

For each *unit*, we find the $c_i(\xi)$ for all $\frac{1}{\lambda}$ distant values of ξ from the interval (η',1). Clearly, a more meaningful value of η' gives lesser value of $c_i(\xi)$. For each *unit*, we find the maximal meaningful cut-off using the following definition:

Definition 3*: A cut-off ξ is said to be maximal meaningful cut-off for the corresponding unit if it is 1-meaningful and*

$$\forall \quad j \in (\eta', 1) - \{\xi\}, \quad c_i(\xi) \leq c_i(j). \tag{16}$$

The frames with l-ratio greater than the maximal meaningful cut-off value of the corresponding *unit* are finally chosen as key-frames. The entire algorithm to find the key-frames in a video, is given in Appendix B. Next we illustrate the results of our approach.

V. Results and Discussions

We have tested our method on twenty one different benchmark videos '0863.mpg' (0863), 'indi114.mpg' (indi), 'hcil2000_01.mpg' (hcil), 'NASAKSN-NorthernLights.mpg' (N-NL), 'anni007.mpg' (anni), 'indi001.mpg' (indi1),

'BOR10_002.mpg' (BOR), 'NASAKSN-Magnetism.mpg' (N-MT), 'NASAKSN-AuroraBorealis.mpg' (N-AB), 'NASAKSN-Lightning.mpg' (N-LN), 'NASAKSN-Leaves.mpg' (N-L), 'NASAKSN-ElMovimiento.mpg' (N-EMM), 'NASAKSN-Bubbles.mpg' (N-BB), 'NASAKSN-ElPopcorn.mpg' (N-EP), 'NASAKSN-LosRayosElectricos.mpg' (N-LRE), 'NASAKSN-ElSonido.mpg' (N-ES), 'NASAKSN-LaEstatica.mpg' (N-LE), 'NASAKSN-ElMetro.mpg' (L-EM), 'NASAKSN-MicroAirVehicles.mpg' (N-MAV), 'FPTVBangladesh.mpg' (FP-B) and 'NAD33.mpg (NAD)' downloaded from the website [26]. The video 'FPTVBangladesh.mpg' consists of nearly 6000 frames and 'NAD33.mpg' consists of more than 51000 frames. All the other test videos consist of around 2500 frames each. We have applied our method on the entire video for all the 21 videos.

The results are compared with respect to ground truths. The ground truths are available with the data, in the form of "FastForward" and "storyboard". The "fastforward" representation of a video consists of the key-frames only (as defined in the Definition 2 in this paper), whereas a "storyboard" representation of a video selects only the key-frames with distinct visual contents. Same scene may come in several non-adjacent shots repeatedly. The "fastforward" representation selects key-frames from each shots. The "storyboard" representation prunes out the key-frames with similar visual content. Our goal in this paper is to find the key-frames only. So the frames in "fastforward" are used as the ground truth.

If tth frame is a key-frame and a method detects another frame as key-frame, we consider this detection as a correct detection only if the detected frame has almost same visual content as the ground truth. But if a method selects more than one key-frames from a single *unit* with almost same visual content, then we treat them as false detection. If a method misses to detect some key-frame(s) in a *unit* with different visual content, then we call it as a miss-detection. In this way we can make the recall and precision measures capable of handling visual similarity of the frames.

As stated in the Introduction, the proposed approach has three major steps. First we decompose the video into different *units*, then we check the interaction between the two factors of our experiment and lastly, we find some representative frames from each *unit* as key-frames. We first show the result of testing the interaction between the two factors of our experiment for detecting *unit* boundary.

A. Interaction Between the Factors

We show the *SI* values for four representative test videos in Table I. We have stated in Section II that we have calculated the F-ratio for all possible combinations of levels of γ and κ, where γ takes all integral values in [0,*R*] and κ takes all δ_κ distant values from [0,1]. We assume that upto two decimal places of both the p-ratio and l-ratio values are significant for our experiment. So we have taken δ_κ as 0.01. Given this assumption, the κ value for which we get maximum F-ratio, is 0.98. Further, taking δ_κ value as 0.02, we have seen that κ value still remains 0.98 for the optimum F-ratio. Note that for κ value equals to 0.99, the optimum F-ratio is inferior to the F-ratio value obtained using $\kappa = 0.98$. So we have reported experimental results using $\delta_\kappa = 0.02$ in Table I, so that we can reduce the number of levels of a factor without compromising with the determination of maximal value of factor. In Table I, we have shown the *SI* values obtained for values of κ to be 0.02 distant levels in [0.8,1] and

integral values of γ ranging from 1 to 8 for all the four test videos. For all other combinations of levels, we have not calculated the F-ratio because of the constraints discussed in Section 2.4. These F-ratio entries in Table I are marked as '-'. In Table I, an entry in the (γ_1,γ_2) column and (κ_1,κ_2) row represents the *SI* value obtained for a test video, for the change of value of the factor γ from γ_1 to γ_2 and κ from κ_1 to κ_2 respectively. Note that all the *SI* values obtained from (7) for all the four videos are significantly low (less than 10%), as shown in Table I. This indicates that the two factors of our experiment do not depend on each other and hence, we can definitely go for one-factor-at-a-time experiment.

B. Detection of Units

We now show the result of grouping the video into several *units*. Table II shows the number of cuts, gradual transitions present in our twenty one test videos and different *units* detected by the proposed approach. Our method has successfully identified all kinds of shot boundaries (both cut and gradual transition) and some of the shots are divided into *units* where a significant change has occurred. In the comparison on shot and *unit* detection, we are not only interested on the count of shots/*units* but also interested on the particular frame number to be selected as shot/*unit* boundary.

As discussed in the Section II-D that if our optimal F-ratio value (F_{opt}) is less than the critical value (F_{crit}) of F-distribution table [4] for the corresponding degrees of freedom, then we reject the optimal grouping and consider the whole video as a single *unit*. For a video containing *n* frames with *k units* in it, the numerator degrees of freedom is $(k-1)$ and denominator degrees of freedom is $(n-k)$. We have tested this theory on a portion of the video '0863.mpg' (from frame number 175 to frame number 575) where the video has no significant change in visual content. In that case, our method detects 3 *unit* boundaries (i.e., 2 *units*) with maximum F-ratio value 0.386. Since this portion of the video contains 401 frames, the numerator degrees of freedom is 1 and denominator degrees of freedom is 399 (as $n = 401$, $k = 2$). So from the table of F-distribution, we get the critical value for F-ratio as, 3.86 and 6.70 in 95% and 99% confidence levels respectively. Our maximum F-ratio value is less than both of the critical values. So we reject the grouping and consider the whole video as a single *unit* as discussed in Section II. In Table II, we have produced the result for the whole video for all the test videos. For all the test video, the maximum F-ratio is greater than the critical values for the corresponding degrees of freedom. So, we have accepted the grouping.

All the test videos contain different kinds of gradual transitions and cuts. Our method has successfully detected all kinds of gradual transitions along with the cuts. Moreover, our method has detected some frames as *unit* boundary as they contain a huge change in visual content (instead of being in a single shot). In the video 'hcil2000_01.mpg', there are nine shot boundaries [26], i.e., eight shots. But some shots among them have remarkable changes in the visual content and our method detects nineteen *unit* boundaries depending on the change in visual content.

TABLE I

SI VALUES OBTAINED FROM (7) FOR THE CHANGE OF LEVELS OF TWO FACTORS IN FOUR OF OUR TEST VIDEOS. THE LEVELS OF FACTOR γ ARE PLACED ALONG THE COLUMNS AND THOSE OF κ ARE PLACED ALONG ROWS. ALL THE ENTRIES ARE GIVEN IN PERCENTAGE.THE ENTRIES (7,8),(6,7),(5,6) ETC. ALONG THE TOP ROW REPRESENT THE RANGE OF γ VALUE BETWEEN 7 AND 8, OR 6 AND 7, OR 5 AND 6 ETC. RESPECTIVELY IN STEPS OF 1. SIMILARLY, THE ENTRIES (0.8,0.82), (0.82,0.84) ETC. ALONG THE LEFT-MOST COLUMN REPRESENT THE RANGE OF κ VALUE BETWEEN 0.8 AND 0.82, OR 0.82 AND 0.84 ETC. RESPECTIVELY IN STEPS OF 0.02.

		γ						
κ	Video	(7,8)	(6,7)	(5,6)	(4,5)	(3,4)	(2,3)	(1,2)
(0.8,0.82)	0863	-	-	-	0	0	-	-
	hcil	0	0	0	4.27	2.97	0	0
	N-NL	-	-	-	-	0	0	0
	indi	-	-	-	-	0	-	-
(0.82,0.84)	0863	-	-	-	0	0	-	-
	hcil	0	0	0	3.67	1.41	0	0
	N-NL	-	-	-	-	0	0	0
	indi	-	-	-	-	5	-	-
(0.84,0.86)	0863	-	-	0	0	0	-	-
	hcil	0	0	0	1.40	1.03	0	0
	N-NL	-	-	-	0	0	0	0
	indi	-	-	-	5	0	-	-
(0.86,0.88)	0863	-	-	0	0	0	-	-
	hcil	0	0	5	3.03	1.34	0	0
	N-NL	-	-	-	0	0	0	0
	indi	-	-	-	4.04	5.42	-	-
(0.88,0.9)	0863	-	-	0	0	0	-	-
	hcil	5	4.08	2.54	1.82	8.78	0	0
	N-NL	-	-	-	0	0	0	0
	indi	-	-	-	0.47	3.52	-	-
(0.9,0.92)	0863	-	-	0	0	0	-	-
	hcil	2.23	2.02	1.49	5.67	1.31	0	0
	N-NL	-	-	0	0	0	0	0
	indi	-	-	2.05	1.12	4.38	-	-
(0.92,0.94)	0863	0	0	0	0	0	-	-
	hcil	3.30	3.51	3.79	1.79	1.47	0	0
	N-NL	-	4.56	0	0	0	0	0
	indi	-	-	9.93	1.32	3.40	-	-
(0.94,0.96)	0863	0	0	0	0	0	-	-
	hcil	3.57	3.31	2.31	1.90	1.42	0	0
	N-NL	-	5	0	0	0	0	0
	indi	-	0.42	1.03	1.28	5.52	-	-
(0.96,0.98)	0863	0	0	0	0	0	-	-
	hcil	2.55	2.34	2.13	1.74	1.31	0	0
	N-NL	-	3.86	0	5	0.74	3.66	4.41
	indi	-	0.92	3.39	1.20	3.69	-	-

TABLE II

DETECTION OF *units* IN THE 21 TEST VIDEOS BY THE PROPOSED APPROACH.

Video	Cuts	Gradual transitions	Total shot transitions	Detected shot transitions	Detected *unit* transitions
0863	2	0	2	2	3
hcil	3	4	7	7	17
N-NL	7	5	12	12	12
indi	7	1	8	7	8
anni	5	5	10	10	10
indi1	14	1	15	14	17
BOR	0	10	10	10	10
N-AB	7	6	13	13	13
N-LN	11	2	13	13	13
N-LRE	12	1	13	13	13
N-MT	7	8	15	15	15
N-MAV	6	6	12	12	12
N-BB	6	3	9	9	10
N-L	7	4	11	11	12
N-EM	6	4	10	10	12
N-ES	10	5	15	15	15
N-EP	10	3	13	13	15
N-EMM	12	1	13	13	14
N-LE	11	0	11	11	12
FP-B	9	20	29	29	29
NAD	191	31	222	220	226

Fig. 5. Selected key-frames of the video 'hcil2000_01.mpg' using proposed method

Fig. 6. Selected key-frames of the video '0863.mpg' using proposed method

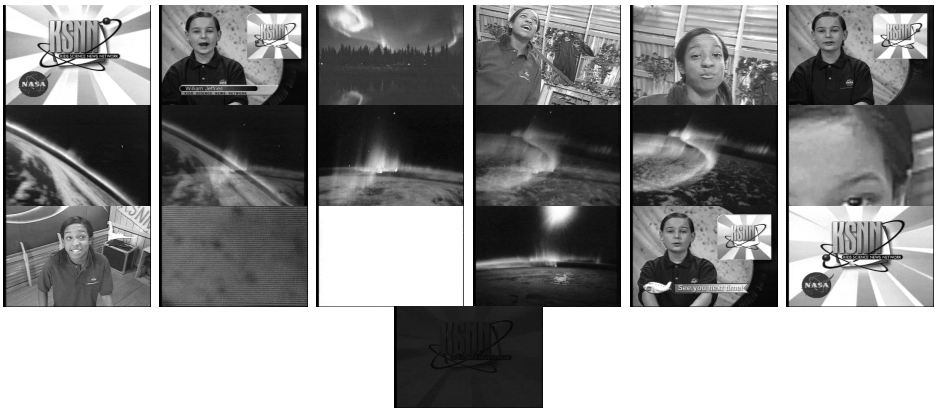


Fig. 7. Selected key-frames of the video 'NASAKSN-NorthernLights.mpg' using proposed method

C. Detection of Key-Frames

We show the result of finding key-frames in Tables III and IV. Table III shows the performances of the proposed approach compared to the competing methods [13], [17], [23] individually on each test videos. Table IV shows the overall performance of the proposed approach compared to the competing methods [13], [17], [23]. The number of key-frames are often different for all the individual *units* depending on the change in visual content in the frames of the corresponding *units*. Our method has successfully detected the variable number of key-frames from the *units*. For example, in the video 'hcil2000_01.mpg', our method detects 21 key-frames from 18 *units* in it, as the visual

Fig. 8. Selected key-frames of the video 'indi114.mpg' using proposed method

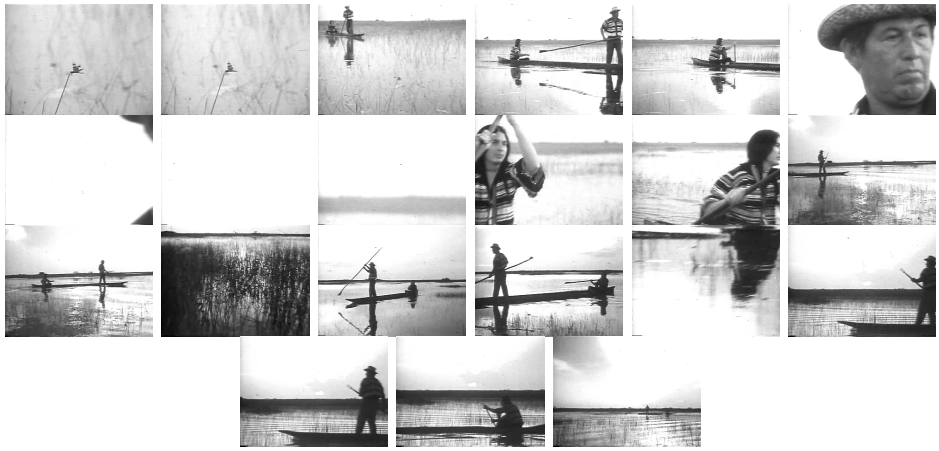


Fig. 9. Selected key-frames of the video ‘indi114.mpg’ using approach of Mukherjee *et al.* [13]

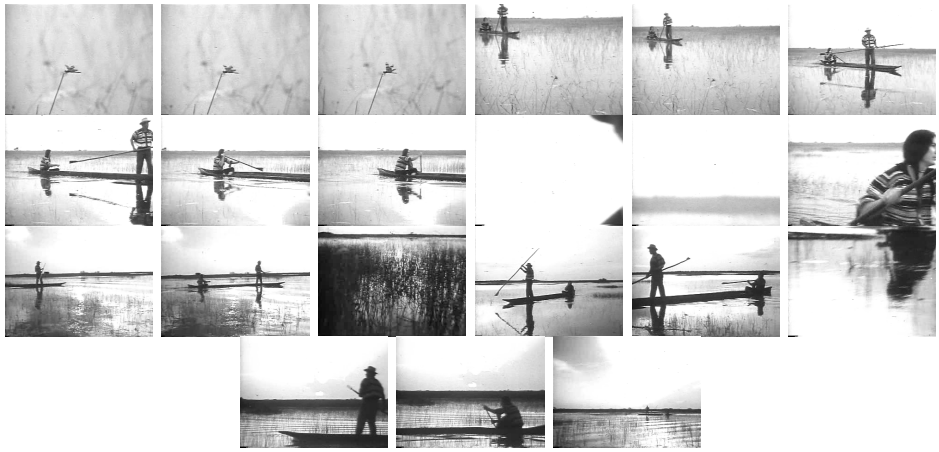


Fig. 10. Selected key-frames of the video ‘indi114.mpg’ using PME based approach [17]

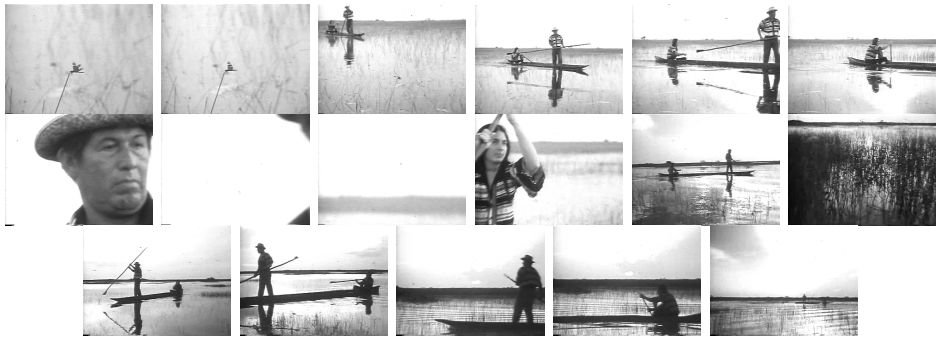


Fig. 11. Selected key-frames of the video ‘indi114.mpg’ using Panagiotakis *et.al.* [23]

contents of the frames in some *units* vary significantly within the corresponding *unit* (see Fig. 5). In the video '0863.mpg', 4 key-frames have been detected from its 3 *units* of the small video of 1653 frames (see Fig. 6). Nineteen key-frames have been detected from 13 *units* of the video 'NASAKSN-NorthernLights.mpg' (see Fig. 7). The video 'indi114.mpg' is the most challenging video among all the 21 test videos. This video has 16 key-frames in only 9 shots in it. The contents of the video changes rapidly in temporal scale, which makes the problem of finding key-frames a challenge (see Fig. 8). This performance is better than the method of Mukherjee *et.al.* [13], the PME based method [17] and the method of Panagiotakis *et.al.* [23], as shown in Table III. Fig. 9, Fig. 10 and Fig. 11 show the selected key-frames from the video 'indi114.mpg' by the method of Mukherjee *et.al.* [13], the PME based method [17] and the method of Panagiotakis *et.al.* [23] respectively. In the longer video sequence 'FPTVBangladesh.mpg', 31 key-frames have been detected from 30 *units*. It is true that too much camera motion may affect the process of *unit* boundary detection. However, the proposed approach can handle slight camera motion. Many video sequences of the test data contain (especially, '0863.mpg') camera shaking. The proposed approach performs well on those videos also. The first and the last *units* of all the test videos with names 'NASA' contain frames with very little changes in visual content. The proposed approach has selected the middle of the frames in the corresponding *units* as key-frame.

As discussed in the Introduction, our approach relies on statistics of frame difference values. Approaches in [13], [17] and [23] also detect key-frames evaluating features in analyzing frame difference space. So in Table III, we have compared our method with [13], [17] and [23]. The proposed method gives better performance than all the competing methods in terms of the evaluation metrics used in this paper.

The performance evaluation for *unit* detection is based on the number of missed detections (MD) and false detections (FD), expressed as recall and precision (ideal value should be 100 for both of them) [27].

$$Recall = \frac{Detects}{Detects + MD}. \tag{17}$$

$$Precision = \frac{Detects}{Detects + FD}. \tag{18}$$

The video 'NASAKSN-Leaves.mpg' has 12 shots. A shot of the video continues from frame number 1091 to frame number 1267. A leaf changes its color from green to magenta in this shot. The p-ratio used in the proposed method takes a high value for a large change in color value in a small area rather than a similar amount of change in a large area. So p-ratio gets a high value at frame number 1175 and this frame has been selected as a *unit* boundary by the proposed approach. Since (15) in Section IV-B assures us at least one representative frame from each *unit*, we get a key-frame 1094 containing a green leaf and a key-frame 1222 containing a magenta leaf Shown in Figures 1(c) and 1(d) respectively. All the competing methods fail to detect these two frames.

The proposed approach gives a high l-ratio value to the frames which have a high frame difference and is a local maxima, which is assured in (8). So in case of camera zooming or panning (in all the twenty one test videos, there are several occurrences of camera zooming and/or panning), since the visual contents of the frames of the video change rapidly in the consecutive frames, our method gives a low l-ratio value to all the frames, thus reducing the

TABLE III

PERFORMANCE OF THE PROPOSED METHOD IN DETECTING KEY-FRAMES COMPARED TO THREE COMPETING METHODS.

Video	Measure	PME method [17]	Mukherjee *et.al.* [13]	Panagiotakis *et.al.* [23]	Proposed method
0863	# key-frame detected	7	6	4	4
	FD	3	2	0	0
	MD	0	0	0	0
	# correct detection	4	4	4	4
	Recall(%)	100	100	100	100
	Precision(%)	70	75	100	100
hcil	# key-frame detected	24	22	20	21
	FD	7	4	3	3
	MD	2	1	2	1
	# correct detection	17	18	17	18
	Recall(%)	92.31	95.65	90.91	95.45
	Precision(%)	77.42	84.62	86.96	87.50
N-NL	# key-frame detected	24	21	19	19
	FD	8	5	3	3
	MD	1	1	1	1
	# correct detection	16	16	16	16
	Recall(%)	96	95.45	95	95
	Precision(%)	75	80.77	86.36	86.36
indi	# key-frame detected	21	21	17	18
	FD	8	7	3	3
	MD	3	2	2	1
	# correct detection	13	14	14	15
	Recall(%)	87.50	91.30	89.47	94.74
	Precision(%)	72.41	75	85	85.71
anni	# key-frame detected	16	15	14	13
	FD	4	3	2	1
	MD	3	3	3	3
	# correct detection	12	12	12	12
	Recall(%)	84.21	83.33	82.35	81.25
	Precision(%)	80	83.33	87.50	92.86
indi1	# key-frame detected	21	22	21	22
	FD	3	3	1	1
	MD	3	2	1	0
	# correct detection	18	19	20	21
	Recall(%)	87.50	91.67	95.45	100
	Precision(%)	87.50	88	95.45	95.65
NAD	# key-frame detected	324	306	280	278
	FD	60	36	9	6
	MD	14	8	7	6
	# correct detection	264	270	271	272
	Recall(%)	95.86	97.45	97.56	97.89
	Precision(%)	84.38	89.47	96.89	97.89

Video	Measure	PME method [17]	Mukherjee *et.al.* [13]	Panagiotakis *et.al.* [23]	Proposed method
BOR	# key-frame detected	16	15	14	15
	FD	4	3	2	2
	MD	2	2	1	1
	# correct detection	12	12	12	13
	Recall(%)	88.89	88.24	93.33	93.75
	Precision(%)	80	83.33	87.50	88.24
N-AB	# key-frame detected	15	16	15	15
	FD	6	4	3	3
	MD	3	2	2	2
	# correct detection	11	12	12	12
	Recall(%)	83.33	88.89	88.24	88.24
	Precision(%)	71.43	80	83.33	83.33
N-LN	# key-frame detected	21	21	20	21
	FD	6	4	3	3
	MD	4	2	2	1
	# correct detection	15	17	17	18
	Recall(%)	84	91.30	90.91	95.45
	Precision(%)	77.78	84	86.96	87.50
N-LRE	# key-frame detected	22	21	21	21
	FD	7	4	4	3
	MD	4	2	2	1
	# correct detection	15	17	17	18
	Recall(%)	84.62	91.30	91.30	95.45
	Precision(%)	75.86	84	84	87.50
N-MT	# key-frame detected	27	25	24	24
	FD	9	5	3	2
	MD	5	3	2	1
	# correct detection	18	20	21	22
	Recall(%)	84.38	89.29	92.31	96
	Precision(%)	75	83.33	88.89	92.31
N-MAV	# key-frame detected	24	22	21	20
	FD	9	5	2	1
	MD	4	2	0	0
	# correct detection	15	17	19	19
	Recall(%)	85.71	91.67	100	100
	Precision(%)	72.73	81.48	91.30	95.24
N-BB	# key-frame detected	17	17	15	16
	FD	5	3	2	2
	MD	2	0	1	0
	# correct detection	12	14	13	14
	Recall(%)	89.47	100	93.75	100
	Precision(%)	77.27	85	88.24	88.89

Video	Measure	PME method [17]	Mukherjee *et.al.* [13]	Panagiotakis *et.al.* [23]	Proposed method
N-L	# key-frame detected	13	14	14	14
	FD	1	2	1	1
	MD	1	1	0	0
	# correct detection	12	12	13	13
	Recall(%)	92.86	93.33	100	100
	Precision(%)	92.86	87.50	93.33	93.33
N-EM	# key-frame detected	18	16	16	17
	FD	6	3	3	4
	MD	3	2	2	2
	# correct detection	12	13	13	13
	Recall(%)	85.71	88.89	88.89	89.47
	Precision(%)	75	84.21	84.21	80.95
N-ES	# key-frame detected	18	18	16	17
	FD	5	4	2	2
	MD	3	2	2	1
	# correct detection	13	14	14	15
	Recall(%)	85.71	90	88.89	94.44
	Precision(%)	78.26	81.82	88.89	89.47
N-EP	# key-frame detected	20	19	19	19
	FD	5	4	3	2
	MD	2	2	1	0
	# correct detection	15	15	16	17
	Recall(%)	90.91	90.48	95	100
	Precision(%)	80	82.61	86.36	90.48
N-EMM	# key-frame detected	18	18	15	16
	FD	5	5	1	2
	MD	2	2	1	1
	# correct detection	13	13	14	14
	Recall(%)	90	90	93.75	94.12
	Precision(%)	78.26	78.26	93.75	88.89
N-LE	# key-frame detected	16	15	14	13
	FD	5	4	2	1
	MD	2	2	1	1
	# correct detection	11	11	12	12
	Recall(%)	88.89	88.24	93.33	92.86
	Precision(%)	76.19	78.95	87.50	92.86
FP-B	# key-frame detected	36	33	32	31
	FD	9	6	5	3
	MD	3	3	3	2
	# correct detection	27	27	27	28
	Recall(%)	92.31	91.67	91.43	93.94
	Precision(%)	80	84.62	86.49	91.18

TABLE IV

OVERALL PERFORMANCE (ON ALL THE 21 TEST VIDEOS) OF THE PROPOSED METHOD IN DETECTING KEY-FRAMES COMPARED TO THREE COMPETING METHODS.

Methods	PME method [17]	Mukherjee *et.al.* [13]	Panagiotakis *et.al.* [23]	Proposed method
Recall(%)	91.59	93.94	94.60	96.35
Precision(%)	80.43	85.46	91.72	92.96

probability of those frames to be key-frames. All the competing methods have selected some unnecessary frames as key-frames for the high frame difference values of the frames with camera zooming, in all the twenty one test videos.

D. Computational Complexity

The proposed method has a computational complexity of $O(\omega\psi\chi)$ for *unit* detection (including decoding and feature extraction stages), where ω and ψ are the number of levels of the factors γ and κ respectively and χ is the number of frames in the video. We have experimentally seen that ω and ψ can be kept as 6 and 20 respectively without compromising with the result of *unit* detection for any video sequence. So the complexity depends on some constant times the number of frames. We think the method is not computationally much expensive for *unit* detection.

For key-frame detection, the proposed method has a worst case computational complexity of $O(\chi q)$, where q is the number of different levels in [0,1] for estimating the meaningful cut-off of the l-ratio. This is better than $O(\chi^2)$ complexity given by both [13] and [17], especially in case of large video (i.e., video with more than q number of frames). Also [23] has a complexity of $O(\chi^2)K_{max}$, which is worse than the proposed approach. Here K_{max} is the rate distortion parameter [23].

For key-frame detection the proposed approach takes less than a second in our MATLAB^{TM} version 7.0.4 implementation in a machine with processor speed 2.37 GHz, 512MB RAM, without code optimization. Whereas, all the three competing methods [13], [17], [23] take slightly more than 1 second but less than 2 seconds.

VI. CONCLUSIONS

We have introduced an efficient statistical procedure for video summarization. We find out the *units* present in a video using the frame differences giving suitable thresholds on both frame difference and p-ratio. Both of the thresholds are determined using standard statistical tools such as design of experiment and ANOVA F-test. Then to detect the key-frames, we have selected from each *unit*, some frames having significantly high frame differences using the concept of meaningfulness. We have shown that our method is less time-consuming and gives satisfactory performance. Our future plan is to apply the same idea of meaningfulness in various fields of event detection.

APPENDIX A: OBTAINING EQUATION (13) FROM [25]

Hoeffding's inequality [25]: In our problem, m_i is the number of frames in the *i*th *unit*. Then we can formulate the problem by a sequence of i.i.d. random variables $\{X_q\}_{q=1,2,3,\ldots,m_i}$, such that $0 \leq X_q \leq 1$. Let us define X_q as,

$$X_q = \begin{cases} 1 & \text{when } l_q < \eta \\ 0 & \text{otherwise} \end{cases}, \tag{19}$$

for a given η, where l_q is the l-ratio value of the *q*th frame of the *i*th *unit*. We set $S_{m_i} = \sum_{q=1}^{m_i} X_q$ (i.e., the number of frames of *i*th *unit* having l-ratio greater than η) and $\nu m_i = E[S_{m_i}]$. Then for $\nu m_i < t < m_i$ (since ν is a probability value less than 1), putting $\sigma = \frac{t}{m_i}$ as in [5], according to Hoeffding's inequality,

$$P_i^{\eta} = P(S_{m_i} \geq t) \leq e^{-m_i(\sigma \log \frac{\sigma}{\nu} + (1-\sigma)\log \frac{1-\sigma}{1-\nu})}. \tag{20}$$

In addition, the right hand term of this inequality satisfies,

$$e^{-m_i(\sigma \log \frac{\sigma}{\nu} + (1-\sigma)\log \frac{1-\sigma}{1-\nu})} \leq e^{-m_i(\sigma-\nu)^2 H(\nu)} \leq e^{-2m_i(\sigma-\nu)^2}, \tag{21}$$

where

$$H(\nu) = \begin{cases} \frac{1}{1-2\nu} \log \frac{1-\nu}{\nu} & \text{when } 0 < \nu < \frac{1}{2} \\ \frac{1}{2\nu(1-\nu)} & \text{when } \frac{1}{2} \leq \nu < 1 \end{cases} \tag{22}$$

This is Hoeffding's inequality. We then apply this for finding the sufficient condition of ϵ-meaningfulness. If $t \geq \nu m_i + \sqrt{\frac{\log \lambda - \log \epsilon}{H(\nu)}}\sqrt{m_i}$, then using (20) and (21) and putting $\sigma = \frac{t}{m_i}$ we get

$$m_i(\sigma - \nu)^2 \geq \frac{\log \lambda - \log \epsilon}{H(\nu)}. \tag{23}$$

Then using (20) and (23) we get,

$$P_i^{\eta} \leq e^{-m_i(\sigma-\nu)^2 H(\nu)} \leq e^{-\log \lambda + \log \epsilon} = \frac{\epsilon}{\lambda}. \tag{24}$$

This means by definition of meaningfulness, the cut-off η is meaningful (according to (11)).

Since for ν in (0,1), $H(\nu) \geq 2$ (according to (22)) so from (24) we get the sufficient condition of meaningfulness as (13).

APPENDIX B: ALGORITHM OF THE PROPOSED APPROACH

1) Input the video sequence with speed *X* fps.
2) Find the Euclidean distance of color values of each pair of consecutive frames.
3) For $\gamma = 1\ to\ R$ (*R* is the maximum bound of color values) do
    a) Give a binary value to each pixel using (2).
    b) Find the matrix β_d for each frame *d*.
    c) Calculate *p*-ratio p_d using (3).
    d) For $\kappa = 0\ to\ 1\ step\ \delta_\kappa$ do
        i) Find all the frames with $p_d > \kappa$.

ii) If (any selected frame f_i is less than *X* frame apart from f_{i+1}) do

A) Find the temporal distance between f_{i+1} and f_{i+2}, f_{i+2} and f_{i+3}, and so on until the temporal distance between f_{i+m} and f_{i+m+1} is greater than *X*.

B) Take the f_{i+m} frame as the boundary of the group starting at frame f_{i-1}.

iii) End if

iv) Find F-ratio using (4), (5), (6).

e) End for κ

4) End for γ

5) Find $F_{max} = max$(F-ratio) and corresponding value of γ and κ.

6) If $F_{max} < F_{critical}$

a) Consider the set of frames having p-ratio greater than κ as *unit* boundaries. else

b) Consider whole video as a single *unit*.

7) End if

8) Find l-ratio of each frame by (8).

9) For each *unit i* do

a) For $\eta = 0\ to\ 1\ step\ \frac{1}{\lambda}$ do

i) Find probability ν using (9).

ii) Find probability P_i^{η} using (10).

iii) Find NFA using (11).

iv) If $NFA < (\epsilon = 1)$,

A) $\eta' = \eta$.

B) Exit from the for loop of η. else

C) Continue the for loop of η.

v) End if

b) End for η

c) For $\xi = \eta'\ to\ 1\ step\ \frac{1}{\lambda}$ do

i) Find $r_i(\eta')$ using (14).

ii) Find c-value using (15).

d) End for ξ

e) Find the ξ satisfying (16).

f) Select the frames having l-ratio greater than ξ as key-frames.

10) End for *unit*

11) Display all the selected frames as key-frames.

REFERENCES

[1] M.J. Pickering, S.M. Rüger and D. Sinclair, "Video Retrieval by Feature Learning in Key Frames", Proc. International Conference on Image and Video Retrieval, 2002, pp. 309-317.

[2] R. Lienhart, "Reliable Transition Detection in Videos: A Survey and Practitioners Guide", International Journal of Image and Graphics, Vol. 1, No. 3, Jul. 2001, pp. 469-486.

[3] S.H. Park, Robust Design and Analysis for Quality Engineering, Chapman & Hall, 1996.

[4] Richard G. Lomax, Statistical Concepts: A Second Course, Mahwah : Lawrence Erlbaum Associates, 2007.

[5] A. Desolneux, L. Moisan, and J. Morel, From Gestalt Theory to Image Analysis: A Probabilistic Approach, Interdisciplinary Applied Mathematics, Vol. 34, Springer, 2008.

[6] A. Desolneux, L. Moisan, and J. Morel, "A Grouping Principle and Four Applications", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 4, Apr. 2003, pp. 508-513.

[7] M. Mills, "A magnifier tool for video data", Proc. ACM Conference on Human Factors in Computing Systems, Monterey, California, USA, 1992, pp. 93-98.

[8] Y. Zhuang, Y. Rui, T. S. Huang and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering", Proc. IEEE International Conference on Image Processing, Chicago, USA, Oct. 1998, pp. 866-870.

[9] V.T. Chasanis, A.C. Likas and N.P. Galatsanos, "Scene Detection in Videos Using Shot Clustering and Sequence Alignment", IEEE Transactions on Multimedia, Vol. 11, No. 1, Jan. 2009, pp. 89-100.

[10] D. Pye, N.J. Hollinghurst, T.J. Mills and K.R. Wood, "Audio-Visual Segmentation for Content-Based Retrieval", Proc. International Conference on Spoken Language Processing, 1998.

[11] D. Adjeroh, M.C. Lee, N. Banda and U. Kandaswamy, "Adaptive edge-oriented shot boundary detection", Journal on Image and Video Processing, Vol. 2009, No. 5, Jan. 2009, pp. 5:1–5:13.

[12] A.F. Smeaton, P. Over, and A.R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity", Computer Vision and Image Understanding, Vol. 114, No. 4, Apr. 2010, pp. 411-418.

[13] D.P. Mukherjee, S.K. Das and S. Saha, "Key-frame estimation in video using randomness measure of feature point pattern", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, No. 5, May 2007, pp. 612-620.

[14] M.M. Yeung and B.L.Yeo, "Video visualization for compact presentation and fast browsing of pictorial content", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 7, No. 5, Oct. 1997, pp. 771-785.

[15] Y. Gao, J. Tang and X. Xie, "Key Frame Vector and Its Application to Shot Retrieval", Proc. 1st International Workshop on Interactive Multimedia for Consumer Electronics, Beijing, China, Oct. 2009, pp. 27-34.

[16] W. Wolf, "Key frame selection by motion analysis", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, Washington, DC, USA, 1996, pp. 1228-1231.

[17] T.M. Liu, H.J. Zhang and F.H. Qi, "A Novel Key-frame Extraction Algorithm Based on Perceived Motion Energy Model", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 10, Oct. 2003, pp. 1006-1013.

[18] Z. Rasheed and M. Shah, "Detection and Representation of Scenes in Videos", IEEE Transactions on Multimedia, Vol. 7, No. 6, Dec. 2005, pp. 1097-1105.

[19] V. Valdes and J.M. Martinez, "A framework for video abstraction systems analysis and modelling from an operational point of view", Multimedia Tools and Applications, Vol. 49, No. 1, 2010, pp. 7-35.

[20] X. Song and G. Fan, "Joint Key-Frame Extraction and Object-Based Video Segmentation", Proc. IEEE Workshop on Motion and Video Computing, Vol. 2, Breckenridge, Colorado, 2005, pp. 126-131.

[21] J. Ouyang, J. Li and H. Tang, "Interactive key frame selection model", Journal of Visual Communication and Image Representation, Vol. 17, No. 6, Dec. 2006, pp. 1145-1163.

[22] E. Spyrou, G. Tolias, P. Mylonas and Y. Avrithis,"Concept detection and keyframe extraction using a visual thesaurus", Multimedia Tools and Applications, Vol. 41, No. 3, 2009, pp. 337-373.

[23] C. Panagiotakis, A. Doulamis and G. Tziritas, "Equivalent Key Frames Selection Based on Iso-Content Principles", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 19, No. 3, Mar. 2009, pp. 447-451.

[24] R. K. Roy, Design of Experiments Using the Taguchi Approach, John Wiley & Sons,INC, 2001.

[25] W. Hoeffding, “Probability inequalities for sum of bounded random variables”, Journal of the American Statistical Association, Vol. 58, No. 301, Mar. 1963, pp. 13-30.

[26] J. Law-To, A. Joly and N. Boujemaa, “Muscle-VCD-2007: a live benchmark for video copy detection”, *http://www-rocq.inria.fr/imedia/civr-bench/*, 2007.

[27] J. Calic and E. Izquierdo, “Efficient Key-frame Extraction and Video Analysis”, Proc. IEEE International Conference on Information Technology: Coding and Computing, Washington, DC, USA, 2002, pp. 28-33.