Journal of Advanced Computer Science & Technology, 5 (1) (2016) 1-7

Journal of Advanced Computer Science & Technology

Website: www.sciencepubco.com/index.php/JACST
doi: 10.14419/jacst.v4i2.4035
Research paper

Irrelevant frame removal for scene analysis using video hyperclique pattern and spectrum analysis

Yuchou Chang *, Hong Lin

Computer Science and Engineering Technology Department, University of Houston - Downtown, Houston, TX, United States
**Corresponding author E-mail: changy@uhd.edu*

Abstract

Video often include frames that are irrelevant to the scenes for recording. These are mainly due to imperfect shooting, abrupt movements of camera, or unintended switching of scenes. The irrelevant frames should be removed before the semantic analysis of video scene is performed for video retrieval. An unsupervised approach for automatic removal of irrelevant frames is proposed in this paper. A novel log-spectral representation of color video frames based on Fibonacci lattice-quantization has been developed for better description of the global structures of video contents to measure similarity of video frames. Hyperclique pattern analysis, used to detect redundant data in textual analysis, is extended to extract relevant frame clusters in color videos. A new strategy using the k-nearest neighbor algorithm is developed for generating a video frame support measure and an h-confidence measure on this hyperclique pattern based analysis method. Evaluation of the proposed irrelevant video frame removal algorithm reveals promising results for datasets with irrelevant frames.

Keywords: *Content-Based Video Retrieval; Fibonacci Lattice; Hyperclique Pattern; Irrelevant Removal; Log Spectrum.*

1. Introduction

The rapid growth of multimedia information volume and the increasing demand of fast accesses to this information via the Internet in recent years have brought much attention to the content-based video retrieval (CBVR). CBVR is necessary because of the prohibitive amount of labor required for manual indexing. Machine learning methods have been used for searching relevant videos [5], [10], [20], [25-26], [32] in order to increase video retrieval accuracy. Depending on the learned criteria and concepts such as semantic dictionary, retrieval accuracy can be improved significantly by narrowing the semantic gap.

Automatic selection of training sets is a key factor to these machine training and searching methods. Fig. 1 shows an example of irrelevant frames added to the "Hoover Dam" video from the TRECVID (2001) video collection [31]. The added frames include a "Y" sign, trash cans, a bicycle, a tree, an automobile, and a building. Since these frames are markedly different from the original frames in the video, they are noise that negatively affects the quality of a retrieval solution if irrelevant frames are selected in a training set. The deliberately inserted irrelevant frames in Fig. 1 correspond to the frames also existing in actual video sequences, which result from unwary shot switching, imprecise editing, or careless manual selection of training set. In practice, these irrelevant frames reduce the authenticity of the training set and hence decrease the efficiency of retrieval techniques using training data. Due to the rapidly increasing volume of video data, manual pruning of frames sets used to train video indexing machine is infeasible. Automatic irrelevant frame removal is a necessary step for a good CBVR system. Obtaining a compact and relevant image set [9] is considered preferable.

Cluttered environments [7, 19] were used to detect and search objects in the multimedia retrieval recently. Because cluttered environments can restrict the objective images to a reduced subset, approaches using cluttered environments are capable of increasing the efficiency of both data set collection and retrieval. The robustness of cluttered data sets is an important component in the information retrieval applications. Just as the cluttered scenes in the augmented "Hoover Dam" video in Fig. 1, the impurity of a cluttered environment will limit the efficiency and accuracy of subsequent retrieval algorithms. In light of problems with existing techniques, we propose a novel unsupervised noise removal algorithm for video scene processing using spectrum analysis and hyperclique pattern mining. The key idea behind the proposed algorithm is the use of hyperclique patterns [33-34] as filters to eliminate irrelevant frames in the video frame set. Initially designed for text mining and pruning, hyperclique pattern methods are effective in removing data elements that are not tightly connected to other data elements in the data set. In this paper, we extend the application of hyperclique pattern based data pruning to color image and video processing. Furthermore, Fibonacci lattice color quantization is used to characterize the color information more accurately than using only 256 gray scales. We then use a log spectrum method to extract features in the frequency domain to summarize visual features of color video frames for comparison.

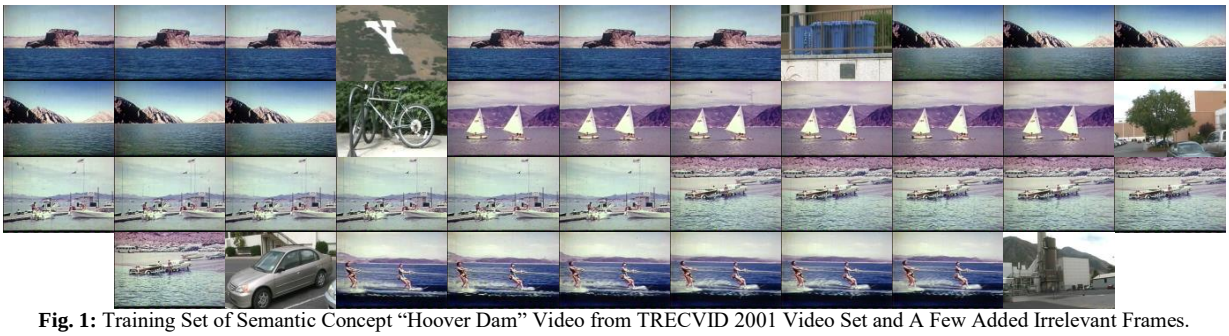


Fig. 1: Training Set of Semantic Concept "Hoover Dam" Video from TRECVID 2001 Video Set and A Few Added Irrelevant Frames.

The remaining of this paper is organized as follows. Section 2 describes the work related to data pruning. Section 3 describes a new image feature extracted by log spectrum analysis. Section 4 extends the hyperclique pattern mining method to the removal of irrelevant video frames. Section 5 presents experimental results to evaluate the performance of the proposed method. Section 6 concludes this work.

2. Related work

Data pruning [1], also known as data cleaning [17], is one data mining technique. Angluin [3] illustrated that learning algorithms can cope with incorrect training examples in the classification of correct and noisy data. Because noisy data can have a wide variety of different influences, there is no general learning method capable of removing all noise [18].

Swets and Weng [27] used a tree structure for data pruning to accelerate the retrieval process. They used a set of discriminant features at each level of the tree. This technique can be deemed as a data pruning method that has been applied to many database applications. Xiong et al. [34] explored four techniques for data cleaning to enhance data analysis in the presence of high noise levels. Different from three methods based on traditional outlier detection techniques, a new hyperclique-based method was proposed to remove two types of noise: low-level data errors resulting from an imperfect data collection process, and noise in the form of irrelevant or weakly relevant data. In order to increase classification accuracy, a group of consensus filters and majority vote filters were used to identify and eliminate mislabeled training samples [6]. Furthermore, this unsupervised clustering-based data mining technique was also used to perform data pruning [11], [13], [16], [23], [35]. These clustering algorithms were designed to remove noise or outliers to purify training data sets. Clustering-based techniques identify core points in the data set according to specified similarity measures and then build clusters around the core points. Points outside of those clusters are treated as noise or outliers. This idea was extended to noisy image removal to improve image search engines [9].

RANSAC [12], an algorithm that estimates parameters of a mathematical model from a set of observed data including valid and noisy data, is capable of interpreting and smoothing image data. This algorithm assumes that model parameters can be estimated from N data items selected from M total items. Parameter values are estimated based on the N samples, the number of data items fitting the current model within a given error tolerance is found, and the model is thus determined to be either acceptable or unacceptable. The process is repeated L times and the best fitting case is returned. Since RANSAC has no upper bound on the time to compute parameters, it can only estimate the model for a particular data set and is of little practical use in the analysis of diversiform color video or images.

Angelova et al. [2] proposed a method based on combining the vote of multiple classifiers to identify examples that are noisy or unfit for learning and to exclude them from the training image set. Since an increasing number of video analysis techniques are based on learning methods, the impact of a given training set on the subsequent learning phase must be considered. In contrast with this approach, our proposed method based on the characteristics of hyperclique pattern mining does not need supervised training examples to prune noise and outliers. Moreover, instead of using a space transform for image analysis, we use log-spectrum representation to describe the contents of each video frame, because the second-order statistics of images are correlated with scene scale and scene category, and because it allows the fast and reliable categorization of scenes and objects [29].

3. Video frame feature extraction

3.1. Retain color information using Fibonacci lattices quantization

Pixels in 24-bit color images have three components: R, G, and B, which can be combined to generate over 16 million unique colors. Compared to a 256 grayscale image, a color image conveys much more visual information and provides the human perceptual system with much more detail about the scene. However, not all 16 million colors are distinguishable by humans, particularly if colors are very similar. Color quantization [15] is a sampling process of 3-D color spaces (e.g. RGB, CIE Lab, HSV) to produce a subset of colors known as the palette, which are then used to represent the original color image. Color quantization is particularly convenient for color image compression, transmission, and display. Unlike most color quantization methods that generate a color palette with three separate color components for each color in the selected subset, quantization using Fibonacci lattices denotes colors using single scalar values. This characteristic allows the construction of log-spectrum representations of color images using the color indices generated by Fibonacci lattice quantization.

The Fibonacci lattice sampling scheme [21] provides a uniform quantization of CIE Lab color space and a way to establish a partial order relation on the set of points. For each different L value in CIE Lab color space, a complex plane in polar coordinates is used to define a spiral lattice as a convenient means for sampling. The following set of points in the (a, b) plane constitutes a spiral lattice:

$$Z_n = n^{\delta} e^{j2\pi \cdot n\tau} \tag{1}$$

Fig. 2 shows an example of the spiral lattice for $\tau = (\sqrt{5}-1)/2$ and $\delta=½$. Each point z_n is identified by its index n. Parameters τ and δ determine the axial distribution and the radial distribution of the points respectively. If there exist N_L luminance (L) values and N_p colors in the corresponding (a, b) plane, for each color in the palette, the corresponding symbol is determined by adding its chrominance index n to a multiple of its luminance index i:

$$q = n + N_p \times i \tag{2}$$

Therefore, the L, a, and b values for any color from the palette can be reconstructed from its symbol q. For a pixel p, with color com-

ponents L_p, a_p, and b_p, the process of determining the closest palette point starts with finding the closest luminance level L_S from the N_L levels available in the palette. The luminance level L_S determines an (a, b) plane and one of the points z_n, $0 \le n \le N_p$, in that plane is the minimum mean square error (MSE) solution. The exact solution, q, is the point whose squared distance to the origin is the closest to $r_p^2 = a_p^2 + b_p^2$.

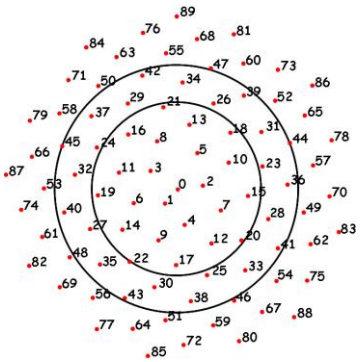


Fig. 2: Points of the Fibonacci Lattice in a Complex Plane.

These L values can approximately denote the luminance levels of the image. Since the (a, b) plane is not circular, there will be points in the Fibonacci Lattice whose colors are not valid in RGB color space. Thus, we label all these points as "range invalid". The points are given by $z_n = S\sqrt{n}e^{j(2\rho nt + \alpha_0)}$ where $t = (\sqrt{5} - 1)/2$, α_0=0.05, and S=1.5. For a 350×240 image shown in Fig. 3(a) having 44748 colors, the L component is quantized into 12 user-selected values {0,10,20,30,40,50,65,70,76,85,94,100}. These L values and N_p =60 points on each plane are used to construct the palette, so the size of the palette is 12×60 =720.

Fig. 3(b) shows the resulting 27 indices of the original image shown in Fig. 3(a). Each of these index values has been assigned an 8-bit value (0, 9, 19, 28, 38, ···, 247) for display. Fig. 3(c) show the quantized color image with 27 valid colors in the palette. Each pixel is labeled by the one dimensional symbol q, which not only is the index of an entry in the palette, but also represents the color information to some extent. In compared with Fig. 3(d), a 256 grayscale image derived from the original, the blue trash cans and green bushes are much easier to distinguish in the quantized image (Fig. 3(c)) despite the grayscale frame having more levels (256) than the frame quantized by Fibonacci lattices (just 27). Easily distinguished colors can appear very similar in a grayscale image. Because human perception contrast in quantized images can be measured by the distance between the q symbols of two colors, it is more accurate to calculate log-spectrum representations based on color indices to a palette constructed by Fibonacci Lattice-quantization than to use 256 levels of grayscale.

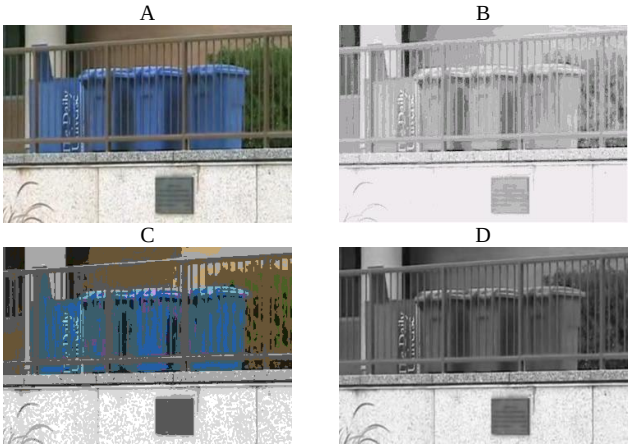


Fig. 3: (A) Original Color Image, (B) 27 Quantized Indices, (C) Quantized Color Image, and (D) Grayscale Image of 256 Gray Scales.

3.2. Log-spectrum representation

The difference between frequency and spatial domains is that the frequency domain captures the global structure of the image but loses local details, while the spatial domain better represents local descriptions than global characteristics [14]. Since frame contents of a video scene change gradually, trivial details of the video content cannot be detected by observing only a few frames. For example, trivial details regarding a small number of changing pixels cannot reflect the visual information of the entire frame or the complete scene. Furthermore, information about global structure can be captured from any given frame in a video scene. For these reasons, a global structure description method should be used to extract features from the video frames.

Log-spectrum representation, to describe the global structure of image contents, has been used in research on statistical scene analysis [22], [28-30]. Instead of using a simple grayscale version (Fig. 3 (d)) of the color image that poorly summarizes the original color content of the image, we calculate the log spectrum of color indices quantized by Fibonacci lattices to better represent the color image. The log spectrum can be calculated as

$$L(f) = \log(A(f)) \text{ and, } A(f) = \| FFT2(I_{FL}(x, y)) \|, \quad (3)$$

Where A (f) is the magnitude of the Fourier spectrum (power spectrum), FFT2 (·) is the 2-D Fourier transform, and $I_{FL}(x, y)$ denotes the image whose pixel colors are represented by Fibonacci lattice-quantization indices. If the image size is p×q, the Fourier descriptors of the image can be calculated as the vertical projection of L (f) in the positive semi-axis and expressed as

$$F(\omega) = L(f(\omega)), \omega = 0, 1, 2, \cdots, (p/2)-1, \quad (4)$$

Where (p / 2) denotes the number of data points in the positive semi-axis of the log spectrum. Fig. 4 shows the plots of four log-spectrum examples that can be used as features to calculate frame similarity.

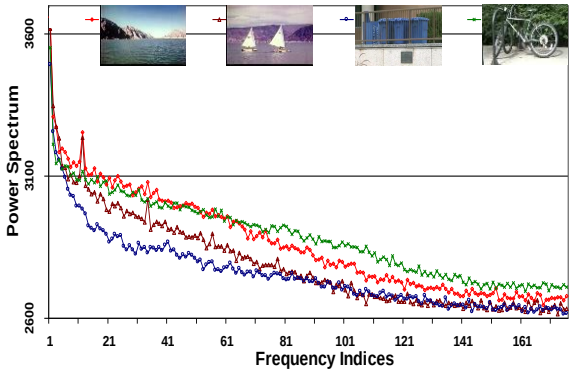


Fig. 4: Examples of Log Spectrum.

4. Hyperclique pattern guided irrelevant frame removal

4.1. Hyperclique pattern

A hyperclique pattern [33], [34] is a type of association pattern that contains objects that are highly affiliated with one another. Any two objects within such a pattern have a similarity measure above a certain threshold level. If an object is not part of any hyperclique pattern, then it is not closely related to other objects in the set of considered objects and likely to be a noisy datum or an outlier. Similarly, in video scene analysis, similar frames should be considered part of the current scene, since they have similar visual and semantic content with each other. For instance, the original "Hoover Dam" frames in Fig. 1 are very similar to each other and belong to the same cluster, which typifies the essential

content of the current scene. In contrast, the added frames (the "Y" sign, bicycle, etc.) are dissimilar to those of the "Hoover Dam" cluster, and they should be considered irrelevant and determined to be noise. The "Hoover Dam" cluster in the frame set of Fig.1 is the representative of the hyperclique patterns that we wish to detect.

Let $I=\{i_1,i_2,\cdots,i_n\}$ be a set of items and $T=\{t_1,t_2,\cdots,t_l\}$ be a set of transactions as shown in Table 1, where each transaction t_i ($1\leq i\leq l$) is also a set of items such that $t_i\subseteq I$. Define a pattern X to be a set of items such that $X\subseteq I$, and further define the support of X, supp(X), to be the fraction of total transactions in T that contain X. Unlike a frequent pattern, a hyperclique pattern contains items that are strongly correlated with each other. The presence of an item in one transaction strongly implies the presence of every other item that belongs to the same hyperclique. The h-confidence measure is specifically designed to capture the strength of this association [33], [34].

Table 1: A Sample Transaction Data Set.

Transactions	t_1	t_2	t_3	t_4	t_5
Items	i_1, i_2, i_3	i_1, i_3, i_4, i_5	i_1, i_2, i_4, i_6	i_1, i_2, i_3, i_4	i_1, i_2, i_3, i_6

Definition 4.1: *The h-confidence of a pattern $X=\{i_1,i_2,\cdots,i_m\}$, denoted as hconf(X), is a measure that reflects the overall affinity among items within the pattern. This measure is defined as*

$$\text{hconf}(X)=\min(\text{conf}(\{i_1\}\rightarrow\{i_2,\cdots,i_m\}), \text{conf}(\{i_2\}\rightarrow\{i_1, i_3,\cdots,i_m\}),\cdots, \text{conf}(\{i_m\}\rightarrow\{i_1, i_2,\cdots,i_{m-1}\}))$$

Where conf is the confidence of association rule as given above.

As an example, for the sample transaction data set shown in Table 1, let us consider a pattern $X=\{i_1,i_3,i_4\}$. We have $\text{supp}(\{i_1\})=100\%$, $\text{supp}(\{i_3\})=80\%$, $\text{supp}(\{i_4\})=60\%$, and $\text{supp}(\{i_1,i_3,i_4\})=40\%$. Then,

$\text{conf}(\{i_1\}\rightarrow\{i_3, i_4\})= \text{supp}(\{i_1,i_3,i_4\})/ \text{supp}(\{i_1\})=40\%$

$\text{conf}(\{i_3\}\rightarrow\{i_1, i_4\})= \text{supp}(\{i_1,i_3,i_4\})/ \text{supp}(\{i_3\})=50\%$

$\text{conf}(\{i_4\}\rightarrow\{i_1, i_3\})= \text{supp}(\{i_1,i_3,i_4\})/ \text{supp}(\{i_4\})=66.7\%$

Therefore,

$\text{hconf}(X)=\min(\text{conf}(\{i_1\}\rightarrow\{i_3,i_4\}), \text{conf}(\{i_3\}\rightarrow\{i_1, i_4\}),\text{conf}(\{i_4\}\rightarrow\{i_1, i_3\}))=40\%$

Definition 4.2: *A pattern X is a hyperclique pattern if hconf(X) ≥ hc, where hc is a user-specified minimum h-confidence threshold.*

4.2. Color video irrelevant frame removal

Unlike text analysis in which items are clearly distributed into their corresponding transactions, color video frames usually need manual selection of training sets to distribute them to appropriate groups. In order to perform unsupervised irrelevant frame removal, we use an unsupervised K-Nearest Neighbor (KNN) technique to automatically group the whole set of video frames into different transactions, from which we can then extract video hyperclique patterns (VHP) that include frames relevant to an intended scene.

KNN is a pattern recognition method in which an object is classified by the majority votes of its neighbors [8, 24]. It categorizes similar data into the same group to form a compact set. The underlying principle of hyperclique patterns suggests that each item (frame) should belong to one or more transactions. For each current frame, KNN can sort all other frames using the log-spectrum similarity measure so that it can construct visual transactions similar to transactions in text processing. Ideally, each transaction typifies one semantic or visual category.

Assume that there are M color video frames of the size p×q in the data set. After being quantized by Fibonacci lattices to obtain typical color indices, each frame can be represented with features calculated based on log spectrum. KNN is used to find the K frames nearest to the current frame. The algorithm is summarized below.

1) Select a value for the parameter K.
2) Construct a data matrix with each row corresponding to the feature vector of each frame.
3) Calculate the distances between the current row and all other rows, sort the distances, and obtain the nearest K rows (neighbors).
4) Repeat Step 3 for all M rows.

Each current frame and its K nearest neighbors can be used to construct one transaction whose items have similar attributes. The transaction derived from the current frame and its K nearest neighbors represents a compact clustering data set, which presumably has semantic or visual similarity. In practice, it is assumed that all frames in a given transaction belong to the same scene. If one noisy frame is grouped into a pattern or transaction with other valid frames, the frequency of its occurrence in the scene will be very low. Furthermore, if a pattern has too many items, its h-confidence is too small to be discriminated because it is unlikely to have this large pattern appearing in the transaction. On the other hand, if a pattern has a single item, all h-confidences with this pattern will equal to be 1, which is also meaningless. In our experiments, the number of items in each pattern was set to $P_N=2$. The number of items in each transaction was set to K+1. The algorithm used to calculate each pattern's support and h-confidence measures is as follows:

(1) Set the number of frames in each pattern as 2 (the current frame and its closest neighbor).
(2) For each current frame F_c and its K nearest neighbors (F_{c1}, Fc2, ···, F_{ck}), F_c and F_{c1} (F_c's closest neighbor) are chosen to construct one pattern and F_c, F_{c1}, F_{c2}, ···, F_{ck} (K+1 frames) to construct one transaction.
(3) Repeat Step 2 until all M patterns and M transactions of all frames are extracted.
(4) For each current frame, calculate its support measure according to the generated M transactions (Section 4.1).
(5) For each of the M unique patterns, calculate its h-confidence measure in M transactions (Definition 4.1).
(6) Set threshold values of support measure or h-confidence measure, and extract the remaining final compact set.

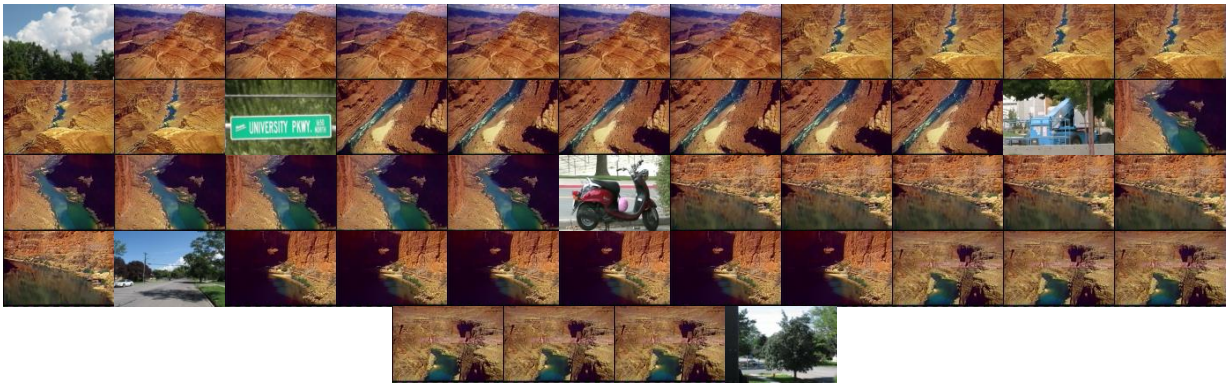
The final compact set contains the valid frames relevant to the current video scene or irrelevant frames as one set of outliers. By setting the thresholds for the support measure, the h-confidence measure, or both, the relevant frames of the scene can be determined.

5. Experimental results

5.1. Color video irrelevant frame removal

We used three videos, "Hoover Dam", "Colorado" and "campus" to test the proposed method. The first two videos are from the TRECVID (2001) video collection, and the third one was captured using a video camera on a university campus. Frames from the "Hoover Dam" video are shown in Fig. 1. Frames from the "Colorado" and "campus" videos are shown in Fig. 5(a) and 5(b), respectively. Frames in "Hoover Dam" and "Colorado" have 352×240 pixels and frames in "campus" have 400×300 pixels. The total number of valid frames and the number of irrelevant frames for these three videos are 42, 48, and 23 and 6, 6, and 5, respectively. The valid frames in the videos typify different shots captured in a particular setting that is consistent for each video. The irrelevant frames were added to test the capabilities of the proposed algorithm in detecting unrelated frames.

a) Video "Colorado" frames and noisy frames



b) Video "campus" frames and noisy frames

Fig. 5: Video Frames of (A) "Colorado" and (B) "Campus" and Their Irrelevant Frames.

5.2. Performance evaluation

We employed the F-measure [4] to evaluate the performance of the proposed algorithm. F-measure combines recall and precision measures, each of which measures a single aspect of retrieval quality if used separately. The F-measure considers both measures and is defined as follows:

$$F(j) = 2/\{1/r(j) + 1/p(j)\} \quad (5)$$

Where r (j) is the recall of the j^{th} element in the ranking, p(j) is the precision for the j^{th} element in the ranking, and F(j) is the harmonic mean of r(j) and p(j). The function F assumes values in the interval [0, 1]. It is 0 when no relevant documents have been retrieved and is 1 when all retrieved elements are relevant. Furthermore, the harmonic mean of F reaches a high value only when both recall and precision are high. Therefore, the higher the F-measure value, the better the noisy frame removal algorithm can perform.

We used the F-measure as defined in [34] to evaluate the performance of the proposed method. Each video contains frames in two groups: relevant (Group 1) and irrelevant (Group 2) frames. Using the proposed algorithm, video frames were separated into two clusters: relevant (Cluster 1) and irrelevant (Cluster 2). Recall and precision were calculated as follows:

$$\text{Recall} = r/n \quad (6)$$

$$\text{Precision} = r/m \quad (7)$$

where r is the number of relevant frames successfully classified as relevant, n is the number of relevant frames (Group 1) in the video, and m is the number of relevant frames clustered as Cluster 1. The F-measure was calculated as.

$$F(i, j) = 2 \times \text{Recall} \times \text{Precision} / (\text{Recall} + \text{Precision}) \quad (8)$$

Consider the "Hoover Dam" video as an example. The video has 36 (Group 1 or n) relevant frames and 6 (Group 2) irrelevant frames. Assume that the KNN algorithm separates them into 34 (Cluster 1 or m) relevant and 8 irrelevant frames automatically (unsupervised). Assume further that our video hyperclique pattern (VHP) analysis successfully classifies 35 frames as relevant of which only 32 (r) are actually relevant. Recall and precision can then be calculated as 32/36=0.889 and 32/34=0.941, respectively. The F-measure can be calculated as 2×0.889×0.941/ (0.889+0.941) = 0.914.

In order to evaluate the robustness of the proposed feature extraction and hyperclique pattern mining methods for irrelevant frame removal, we compared the proposed algorithm with the unsupervised Histogram K-means method [9] only, rather than with other supervised methods such as [2]. The Histogram K-means method first extracts features that are translational and rotational invariant to represent the image contents in the histogram form. Then, a K-means clustering algorithm is used to remove the irrelevant or noisy images from the output of a video retrieval search engine. We then used the F-measure to evaluate the performance of both the proposed algorithm and the Histogram K-means method. For each video, we used four irrelevant frame percentages by removing different numbers of relevant frames from each video. For example, for "Hoover Dam" video, we used 6 (irrelevant)/42 (total frames), 6/30, 6/18, and 6/12. The experimental parameters for irrelevant frame removal are shown in Table 2. F-measures at different irrelevant frame percentage levels are presented in Fig. 6.

Table 2: Experimental parameters for the proposed algorithm.

	A[1]	B[2]	C[3]	D[4]	E[5]	F[6]	G[7]	H[8]	I[9]
Hoover Dam	2	>0.08	N.A.	>0.1	N.A.	N.A.	>0.58	N.A.	0.5-0.71
Colorado	2	>0.09	N.A.	>0.15	N.A.	>0.15	>0.60	N.A.	>0.73
Campus	2	>0.15	N.A.	>0.22	N.A.	N.A.	>0.50	N.A.	<0.83

(1) Number of clusters of algorithm output.
(2) Support measure threshold for noise percentage at position 1 (smallest noise percentage).
(3) h-confidence measure threshold for noise percentage at position 1 (smallest noise percentage)
(4) Support measure threshold for noise percentage at position 2
(5) h-confidence measure threshold for noise percentage at position 2
(6) Support measure threshold for noise percentage at position 3
(7) h-confidence measure threshold for noise percentage at position 3
(8) Support measure threshold for noise percentage at position 4 (highest noise percentage)
(9) h-confidence measure threshold for noise percentage at position 4 (highest noise percentage)

It can be seen in Fig. 6 that the proposed algorithm outperformed the Histogram K-means algorithm on the three test videos. At low irrelevant percentage levels, the proposed algorithm was able to remove more irrelevant frames and maintained better integrity of valid video than Histogram K-means – the proposed algorithm's F-measures were higher. As the irrelevant frame percentage increased, the Histogram K-means algorithm failed to remove irrelevant frames. Its F-measures decreased as the irrelevant frame percentage increased. On the other hand, the proposed algorithm's F-measures maintained on the same level or even increased as the irrelevant frame percentage increased.

At low irrelevant frame percentages, the support measure alone was able to remove irrelevant frames efficiently. This is because the KNN algorithm was able to remove frames less frequently presented in the valid video frames. These frames less frequently presented were the irrelevant frames because they are hardly grouped into K nearest neighbors. As irrelevant frame percentage increased, the support measure began to fail to remove irrelevant frames by itself. However, the h-confidence measure was able to take over and effectively remove irrelevant frames because it can classify irrelevant frames as outliers. The best results were obtained when both support and h-confidence measures were combined to remove irrelevant frames. For instance, in video "Colorado", when the irrelevant frame percentage was set to 33%, the combination of support and h-confidence measures detected all irrelevant frames and achieved an optimal result.

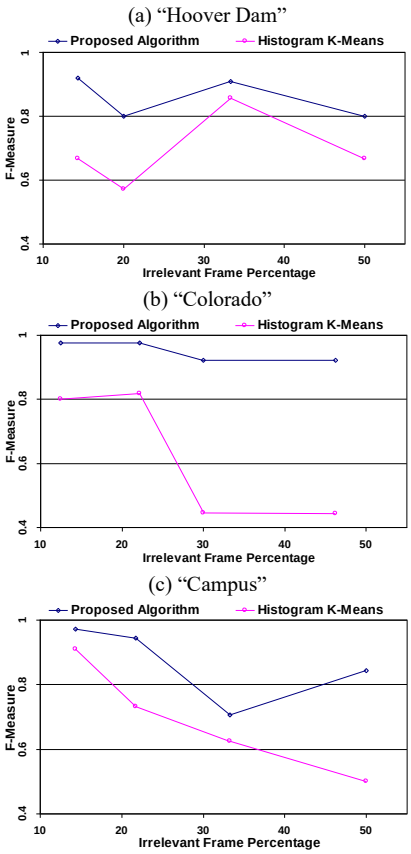


Fig 6: Performance Comparison of the Proposed Algorithm and Histogram K-Means; (A) "Hoover Dam", (B) "Colorado", and (C) "Campus".

The proposed method performed better than the Histogram K-means algorithm previously used to remove irrelevant images from the output of an image search engine. Although the proposed method requires setting the final support and h-confidence measure thresholds manually, it was able to detect special clusters especially at higher percentages of irrelevant frames. The proposed algorithm is an unsupervised method that does not require the manual selection of training set.

6. Conclusion

In summary, we have presented a hyperclique pattern based irrelevant video frame removal technique. Considering that color index values can more accurately represent meaningful image content than grayscale representations, we used the Fibonacci lattice-quantization method to quantize color video frames into scalar indices. Because Fourier descriptors can capture the global structures of image contents, we then use log-spectrum representation on quantized color indices to extract video frame features. By constructing transactions and patterns with KNN algorithm, we applied hyperclique pattern analysis to remove irrelevant video frames. Experiments show that the proposed algorithm has better performance than the unsupervised Histogram K-means method. In future work, we will address the challenge of how to simplify dynamic settings of thresholds for support and h-confidence measures.

References

[1] A.Angelova, "Data Pruning", *Thesis*, California Institute of Technology, (2004).

[2] A. Angelova, Y. Abu-Mostafa, and P. Perona, "Pruning Training Sets for Learning of Object Categories", *IEEE International Conference on Computer Vision and Pattern Recognition*, vol.1, (2005), pp.494-501. http://dx.doi.org/10.1109/cvpr.2005.283.

[3] D. Angluin, and P. Laird, "Learning from Noisy Examples", *Machine Learning*, vol.2, no.4, (1988), pp.343-370. http://dx.doi.org/10.1007/BF00116829.

[4] R.A. Baeza-Yates, and B. Ribeiro-Neto, "Modern Information Retrieval", Addison-Wesley Longman Publishing, (1999).

[5] S. Basu, M. Naphade, and J.R. Smith, "A Statistical Modeling Approach to Content Based Retrieval", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.4, (2002), pp.480-483. http://dx.doi.org/10.1109/icassp.2002.5745554.

[6] Brodley, C.E., and Friedl, M.A., Identifying Mislabeled Training Data, Journal of Artificial Intelligence Research, vol.11, pp.131-167, 1999.

[7] Y. Chi, and M.K.H. Leung, "Part-Based Object Retrieval in Cluttered Environment", *IEEE Transactions on Analysis and Machine Intelligence*, vol.29, no.5, (2007), pp.890-895. http://dx.doi.org/10.1109/TPAMI.2007.1076.

[8] B.V. Dasarathy, "Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques", IEEE Computer Society Press, (1990).

[9] T. Deselaers, D. Keysers, and H. Ney, "Clustering Visually Similar Images to Improve Image Search Engines", *Informatiktage 2003 der Gesellschaft für Informatik*, Germany, (2003).

[10] L.Y. Duan, M. Xu, Q. Tian, C.S. Xu, and J.S. Jin, "A Unified Framework for Semantic Shot Classification in Sports Video", *IEEE Transactions on Multimedia*, vol.7, no. 6, (2005), pp.1066-1083. http://dx.doi.org/10.1109/TMM.2005.858395.

[11] L. Ertoz, M. Steinbach, and V. Kumar, "Finding Clusters of Different Sizes, Shapes and Densities in Noisy, High Dimensional Data", *Proceedings of SIAM International Conference on Data Mining*, (2003). http://dx.doi.org/10.1137/1.9781611972733.5.

[12] M. Fischler, and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", *Communications of the ACM*, vol.24, no.6, (1981), pp.381-395. http://dx.doi.org/10.1145/358669.358692.

[13] S. Guha, R. Rastogi, and K. Shim, "Cure: An Efficient Clustering Algorithm for Large Databases", *In Proceedings of ACM SIGMOD International Conference on Management of Data*, (1998), pp.73-84. http://dx.doi.org/10.1145/276304.276312.

[14] A.N. Hirani, and T. Totsuka, "Combining Frequency and Spatial Domain Information for Fast Interactive Image Noise Removal", *ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, (1996), pp.269-276. http://dx.doi.org/10.1145/237170.237264.

[15] A.K. Jain, "Fundamentals of Digital Image Processing", *Prentice Hall Information and System Sciences Series*, (1989).

[16] A.K. Jain, and R.C. Dubes, "Algorithms for Clustering Data", Prentice Hall, (1998).

[17] K. Kaewbuadee, Y. Temtanapat, and R. Peachavanish, "Data Cleaning Using FD From Data Mining Process", *IADIS Interna-*

tional Journal on Computer Science and Information System, vol.1, (2006), pp.117-131.
[18] M. Kearns, and M. Li, "Learning in the Presence of Malicious Errors", *Annual ACM Symposium on Theory of Computing*, (1988), pp.267-280. http://dx.doi.org/10.1145/62212.62238.
[19] L.J. Li, G. Wang, and F.F. Li, "OPTIMOL: Automatic Online Picture Collection via Incremental Model Learning", *IEEE International Conference on Computer Vision and Pattern Recognition*, (2007), pp.1-8. http://dx.doi.org/10.1109/cvpr.2007.383048.
[20] D. Makris, and T. Ellis, "Learning Semantic Scene Models From Observing Activity in Visual Surveillance", *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, vol.35, no.3, (2005), pp.397-408. http://dx.doi.org/10.1109/TSMCB.2005.846652.
[21] A. Mojsilovic, and E. Soljanin, "Color Quantization and Processing by Fibonacci Lattices", *IEEE Transactions on Image Processing*, vol.10, no.11, (2001), pp.1712-1725. http://dx.doi.org/10.1109/83.967399.
[22] A. Oliva, and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope", *International Journal of Computer Vision*, vol.42, no.3, (2001), pp.145-175. http://dx.doi.org/10.1023/A:1011139631724.
[23] J. Sander, M. Ester, H.P. Kriegel, and X. Xu, "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications", *International Journal of Data Mining and Knowledge Discovery*, vol.2, no.2, (1998), pp.169-194. http://dx.doi.org/10.1023/A:1009745219419.
[24] G. Shakhnarovich, T. Darrell, and P. Indyk, "Nearest-Neighbor Methods in Learning and Vision: Theory and Practice", MIT Press, (2005).
[25] C.G.M. Snoek, M. Worring, D.C. Koelma, and A.W.M. Smeulders, "A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval", *IEEE Transactions on Multimedia*, vol.9, no. 2, (2007), pp.280-292. http://dx.doi.org/10.1109/TMM.2006.886275.
[26] C.G.M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding Semantics to Detectors for Video Retrieval", *IEEE Transactions on Multimedia*, vol.9, no.5, (2007), pp.975-986. http://dx.doi.org/10.1109/TMM.2007.900156.
[27] D.L. Swets, and J. Weng, "Hierarchical Discriminant Analysis for Image Retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.21, no.5, (1999), pp.386-401. http://dx.doi.org/10.1109/34.765652.
[28] A. Torralba, and A. Oliva, "Depth Estimation from Image Structure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no. 9, (2002), pp.1226-1238. http://dx.doi.org/10.1109/TPAMI.2002.1033214.
[29] A. Torralba, and A. Oliva, "Statistics of Natural Image Categories", *Network: Computation in Neural Systems*, vol.14, no.3, (2003), pp.391-412. http://dx.doi.org/10.1088/0954-898X_14_3_302.
[30] A. Torralba, "Modeling Global Scene Factors in Attention", *Journal of the Optical Society of America*, vol.20, no.7, (2003), pp.1407-pp.1418.
[31] TRECVID, available online: http://www-nlpir.nist.gov/projects/trecvid/
[32] C.J. Wu, H.C. Zeng, S.H. Huang, S.H. Lai, and W.H. Wang, "Learning-Based Interactive Video Retrieval System", *IEEE International Conference on Multimedia and Expo*, (2006), pp.1785-1788. http://dx.doi.org/10.1109/icme.2006.262898.
[33] H. Xiong, P.N. Tan, and V. Kumar, "Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution", *IEEE International Conference on Data Mining*, (2003), pp.387-394. http://dx.doi.org/10.1109/ICDM.2003.1250944.
[34] H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, "Enhancing Data Analysis with Noise Removal", *IEEE Transactions on Knowledge and Data Engineering*, vol.18, no.3, (2006), pp.304-319. http://dx.doi.org/10.1109/TKDE.2006.46.
[35] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases", *ACM SIGMOD International Conference on Management of Data*, (1996), pp.103-114. http://dx.doi.org/10.1145/233269.233324.