A Psychological Adaptive Model For Video Analysis

N. Ghosh and B. Bhanu
Center for Research in Intelligent Systems (CRIS), University of California, Riverside, CA 92521, USA
{nirmalya, bhanu} @ ee.ucr.edu

ABSTRACT

Extracting key-frames is the first step for efficient content-based indexing, browsing and retrieval of the video data ***in commercial movies****. Most of the existing research deals with "how to extract representative frames?" However the unaddressed question is "****how many key-frames are required to represent a video shot properly?****" Generally, the user defines this number a priori or some heuristic methods are used. In this paper, we propose a* ***psychological*** *model, which computes this number adaptively and online, from variation of visual features in a video-shot. We incorporate it with an iterative key-frame selection method to* ***automatically*** *select the key-frames. We compare the results of this method with two other well-known approaches, based on a novel* ***effectiveness measure*** *that scores each approach based on its representational power. Movie-clips of varying complexity are used to underscore the success of the proposed model in real-time.*

1. INTRODUCTION

Although research in building key-frames in video has matured from heuristic selections [4, 5] to video content-based approaches [1, 2, 3, 6], the number of key-frames to be selected (N_k) has been either a user-defined constant, or based on heuristic thresholds. However, it is desired to adapt N_k with the variation in video-content. This paper proposes an automated, adaptable, online method for computing N_k, motivated by psychological human visual perception and attention theory.

To emphasize why we need N_k to be selected adaptively, suppose a movie is being indexed or key-framed for storyboard type application. One conversation-shot of 5 minutes (9000 frames) with camera mostly focusing on the characters and with minimal motion in the scene, say 10 key-frames may be sufficient. While for a one-minute boxing fight-shot (1800 frames) needs *much more* than proportionally deduced 2 key-frames to represent the complexity of the visual-information flow. Existing approaches with user-provided "*constant-value*" [1, 4, 5] or *threshold*-based clustering [2] or *P% criterion* in local-minimal motion strategy [3] cannot adapt to this dynamic situations.

Hence a movie-key-framing procedure should take N_k computed online, adaptable with the information content of the particular video and motivated by human video-perception psychology. This paper focuses exactly on that. The contributions of this paper are:

(1) An adaptive online method is described to compute *"the number of key-frames (N_k)"*, motivated by psychological visual perception theory.
(2) The procedure developed in (1) is incorporated in an iterative key-frame selection framework [1] with global (i) color and (ii) motion features (individually) to adapt N_k.
(3) A novel performance measure is used to compare the proposed method with a clustering based [2] and a motion based [3] approaches.
(4) Experiments are performed with *movie-clips* of varying complexity for evaluation.

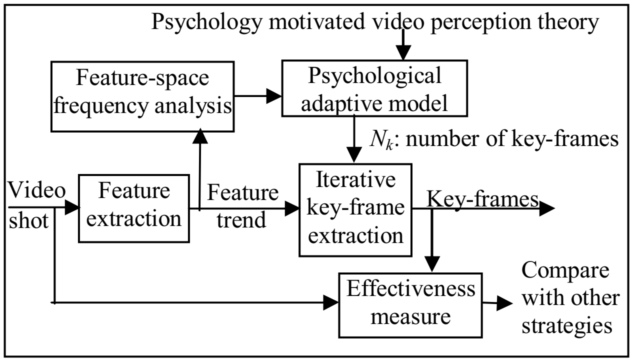

Psychology motivated video perception theory
Feature-space frequency analysis
Psychological adaptive model
N_k: number of key-frames
Video shot
Feature extraction
Feature trend
Iterative key-frame extraction
Key-frames
Effectiveness measure
Compare with other strategies

Fig 1: Conceptual diagram of the approach

2. TECHNICAL APPROACH

Conceptual diagram of our system is depicted in Fig 1.

2.1 Automatic selection of number of key-frames, N_k

In human visual perception, steady-state visual excitations are less remembered [7]. Larger variation in a movie means richer visual information and hence requires more number of key-frames to meet a minimum representational power (or maximum tolerable error or distortion). We take frame-based features as *indicators of visual content of the movie* and quantify visual information (content-variation) by taking the Fast Fourier Transform (FFT) of the feature-sequence.

As this work mainly targets *movies* for human consumption, we utilize general psychological facts regarding human visual perception. According to the "vision and attention" theory in perceptual psychology, when large sequence of video frames (say a standard movie of 1.5-2 Hours) is presented, humans can register 1 frame

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

per 30 seconds for slow-paced (with visual content frequency around 0-1 Hz) movies [8, 9]. Hence salient movie-shots are generally much larger than 30 seconds and key-frames (*for storyboard applications*) temporally apart by less than 30 seconds are not practical. The number of frames *K,* a human can *at most* remember, is (1/30) times the total-time of the movie in seconds (T_m).

$$K = \left(\tfrac{1}{30}\right) * T_m \approx 0.03 * T_m \quad \text{where} \qquad (1)$$

$$T_m = \frac{\text{total number of frames}}{\text{frames per second } (fps)}$$

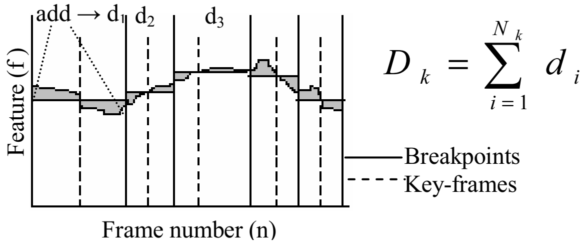
and *fps* = frames per second (25 for PAL, 30 for NTSC).

Now for faster videos (shots with rapid motion or sports-scenes with video information frequency of more than 2 Hz) human memory has to register more number of frames than *K*. Stroud's psychological research established that the time is integrated into *perceptual moments* of about 100 milliseconds in length [7]; this is the visual signal transmission time from retina to visual-cortex and cognition of the scene using past experience. As a result, although retinal cells have frequency response in the range of 1 MHz, general human vision can *perceive* visual content changes at around 10 Hz. Thus movies are made with 25 (in PAL) or 30 (in NTSC) fps to give illusion of continuity. Even special effects by movie directors are in between 0-30 Hz and we can safely classify higher frequencies in feature-space as noise. Since steady-state visual excitation (0-1 Hz) is ignored by human vision [7, 10], we consider the range 2-30 Hz in feature-space and take the *strongest frequency (F)* as the one representing visual information variation.

We propose that a *storyboard* of any standard movie formed with properly selected *(F*K)* number of key-frames capture almost entire video information. Hence the proposed psychological video-perception model can be written as:

$$N_k = K * F = 0.03 * \frac{\text{total number of frames}}{\text{frames per second (fps)}} * F \qquad (2)$$

where N_k is number of key-frames to be extracted for kth video shot and *F* = strongest frequency of the feature-sequence in 2-30 Hz range. N_k is adaptive since *F* varies with the input.



$$D_k = \sum_{i=1}^{N_k} d_i$$

Fig. 2: Illustration of the frame-feature (f), breakpoints, key-frames and total error (D_k), shown as the shaded area, for a *single* video-shot.

This online computed N_k is plugged into an iterative key-frame extraction approach [1] to make the entire procedure automatic and robust. N_k decides the number of key-frames, number of breakpoints (*N_k+1)*, and initialization of breakpoints by equal visual-content criteria. Then these breakpoints and key-frames are optimally [1] placed iteratively to minimize the key-frame-based representational error (Fig. 2).

2.2 Performance measure

For evaluation and comparison with other key-framing strategies, we propose a novel performance measure. Key steps are: (1) finding out dissimilarity between a key-frame and the frames it is representing, (2) summing up the dissimilarities for all key-frames over the video-shot and (3) then summing up for all video-shots over the video-clip to define a single index, ***effectiveness*** so that key-frames do not under-represent or over-represent the video-clip. The legends used are:

$S = \{ s_k \mid k = 1, 2, ..., M \}$ **: Collection of shots in the video clip**

$s_k = \{ fr_j \mid j = shotStart_k, ..., shotEnd_k \}$ **: Frames of kth video-shot**

$keyFr_k = \{ fr_n \mid (shotStart_k \le fr_n \le shotEnd_k) \;\&\; (n = 1, ... N_k) \}$**: Key-frames**

For a key-framing strategy, N_k is computed by (2) and key-frames are extracted. Then frame-to-frame dissimilarity is calculated as follows.

$imageDiff_{nj} = rgb2gray\left(imgSubtract(fr_j - keyFr_n)\right)$

$diffEnergy_{nj} = elementWiseSquare\left(imageDiff_{nj}\right)$

$level = grayThresh\left(diffEnergy_{nj}\right)$

:adaptive threshold by Otsu's method [11]

$BW = im2bw\left(diffEnergy_{nj}\ , level\right)$ **: binarization**

$dissimilarity_{nj} = number\, of\, 1's\, in\,\, BW$

Dissimilarity contributed by an individual key-frame is computed, summed up and normalized to estimate dissimilarity over each video-shot in the *movie*:

$$\text{for } n^{\text{th}} \text{ key - frame : } d_n = \sum_{j\text{: frames represented by } keyFr_n} dissimilarity_{nj} \qquad (3)$$

$$\text{for } k^{\text{th}} \text{video - shot: } D_k = \left(\sum_{n=1}^{N_k} d_n \right) \qquad D'_k = D_k \Big/ \max_n (d_n)$$

Finally representational power of the key-frames over the whole video-clip (collection of *M* shots) is computed by the ***effectiveness*** index as defined below:

$$Effectiveness = \left[\sum_{k=1}^{M} \left[1/(1+D'_k) \right]^{L} \right]^{(1/L)} \quad \text{where, } L = \sum_{k=1}^{M} N_k \qquad (4)$$

In equation (4), dissimilarity measure D'_k decreases and hence similarity measure $[[1/(1+D'_k)] \le 1]$ increases with increasing N_k. So this performance measure decreases for both very high and very low values of N_k, and similarly for *L*, the total number of key-frames for the entire movie-clip. Hence, it can be verified that, proposed *Effectiveness* measure balances between extreme cases of under-representation ($N_k \rightarrow 0$) and over-representation ($N_k \rightarrow$ total number of frames) of the video-clip.

3. KEY-FRAMING STRATEGIES COMPARED

In this paper, we have used color-content variation (Sec 3.1) and inter-frame motion variation (Sec 3.2) as features, and computed fast Fourier transform (FFT) to find $\boldsymbol{F}$ and then $\boldsymbol{N_k}$ (as in (2)). ($1+N_k$) breakpoints (BP_i) are initialized by equal visual content. Then N_k key-frames ($keyFr_i$) are positioned to minimize errors between corresponding breakpoints. The breakpoints are then updated to minimize distortion between corresponding key-frames. The loop continues until overall distortion remains nearly constant for consecutive iterations (i.e. converged). This guarantees minimum representational error or distortion (D_k), as shown in Fig. 2, (kind of sampling error), by simultaneous optimum (proof in [1]) placement of the breakpoints and key-frames iteratively.

These two strategies have been compared to two other heuristic methods (Sec 3.3.and 3.4).

3.1 Iterative approach using global color feature

We have used global average color in the RGB space (for color video, and global intensity average for black-&-white video) as the visual-feature (f) for each frame (t).

$$f(t) = \frac{\sum_{q:R,G,B} area\left[hist\left(t^{th}\ frame,\ q^{th}\ color\ comp.\right)\right]}{number\ of\ pixels} \qquad (5)$$

3.2 Iterative approach using global motion feature

Motion relates to activity and comes closer to the video-semantics. Global motion (V) is computed from robust *phase-based optical flow* [12], with *horizontal (O_x)* and *vertical (O_y) motion components*, at sampled pixels in 2D [indexed by (i, j)] for tth frame [see eq. (6)].

$$V(t) = \sum_i \sum_j \left|O_x(i,j,t)\right| + \left|O_y(i,j,t)\right| \qquad (6)$$

3.3 Unsupervised clustering

Intersection of 2D hue-saturation (indices h and s) histograms (H) is used to compute similarity (7) between two consecutive [rth and (r+1)th] frames in HSV space [2]. Incremental clustering uses two heuristic thresholds, cluster-density (δ) and minimum cluster size (Sz). Frames closest to the cluster centers are selected as key-frames.

$$sim_{r,(r+1)} = \sum_{h=1}^{16} \sum_{s=1}^{8} \min\left(H_r(h,s), H_{r+1}(h,s)\right) \qquad (7)$$

3.4 Local minima in motion feature trend

Local minima in global motion [V in (6)] trend of frame sequence emphasize nearby frames [3], and hence they are good candidates for key-frames. To avoid jitters, consecutive minima with more than $P\%$ variation are considered [3]. This P decides N_k heuristically.

4. EXPERIMENTAL RESULTS

Movie-clips of gradually increasing complexity (in terms of motion-activity, background-changes and camera activity) have been chosen for evaluation of the proposed technique. Data is given in Table 1 with duration in seconds. To emphasize the complexity levels of the different clips, two frames of each clip are shown in Fig. 3. Representational power (as number of key-frames (**L**), to account under- & over-representation, and the proposed ***Effectiveness*** (**E**) index) and real-time applicability (in terms of computational time ($\mathbf{T_C}$)) of the extracted key-frames by four techniques in Section 3 are tabulated in Table 2. The threshold values used are:

- Density threshold in clustering: 70000
- $P\%$ criterion in motion based approach: 30%
- Shot-detection threshold in color-based iterative approach: 0.25.
- Shot-detection threshold in motion-based iterative approach: 2

Table 1: Video data

Clp	Sec	Description
1	7	*Volcano eruption*: color; very short but fast changes
2	16	*Conversation*: B&W; slower motions; visible background changes; camera zooming slowly
3	21	*Conversation*: B&W; faster motions; one video cut
4	30	*Conversation and movement*: color; characters coming in gradually; conversing and moving slowly
5	27	*Multi-cut conversation in the dark*: color; medium movements; darkness of the scene
6	15	*Fight sequence*: color; very fast movements; camera panning / zooming

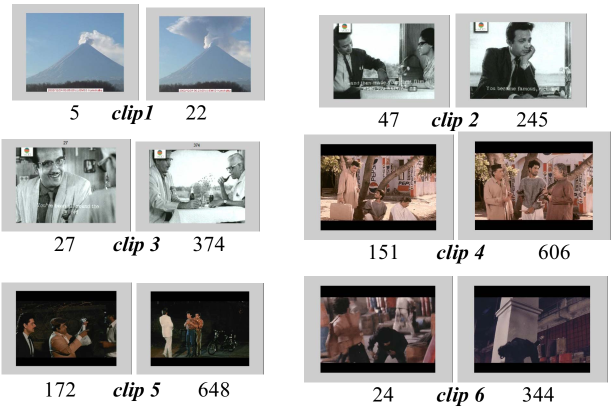


Fig 3: Sample frames of the movie-clips illustrating the widely varying complexity of the data used

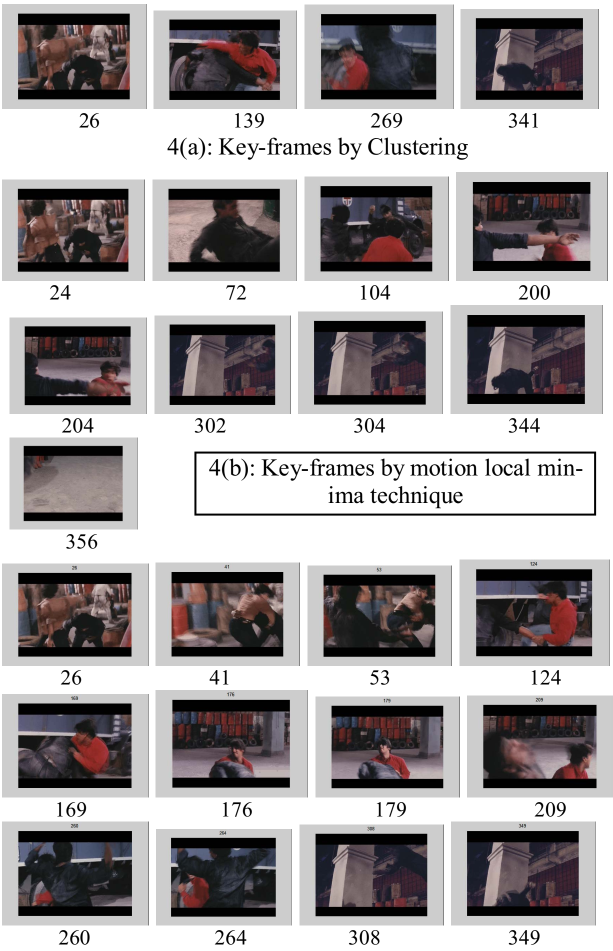
Due to space-constraint, only the key-frame results for clip 6 (the most complex one) are shown in Fig. 4. Motion local-minima based technique performs well for movies with slower motion (clip 2, clip 4 & clip 5) than with faster motion (clip 3 & clip 6). But considering all the measures (**L**, **E**, $\mathbf{T_C}$ in Table 2), color-feature based iterative approach, using the adaptive online selection of "number of key-frames, N_k" proposed in this work, outperforms other techniques, as *Effectiveness* index (**E**) is high and computational time (**T$_C$**) is low consistently for all the movie-clips of widely varying complexity.

Table 2: Results

Legends:

C# : Clip#	S: Technique	L: # of key-frames
E: Effectiveness measure	T$_C$: Comput. Time (sec)	CS: clustering
V$_m$: Motion minima	Ic: IterColor	Im: IterMotion

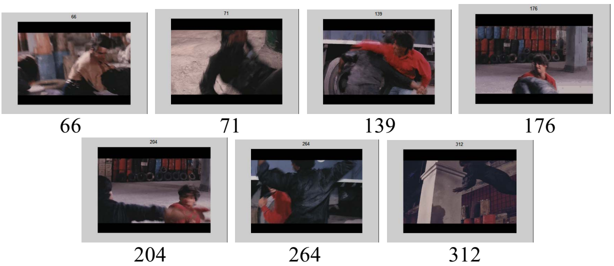
C#	S	L	E	T$_C$	C#	S	L	E	T$_C$
1	Cs	1	0.3058	93	4	Cs	2	0.1442	1512
	V$_m$	1	0.3722	245		V$_m$	9	0.5714	3783
	Ic	1	0.3262	2		Ic	15	0.4887	82
	Im	1	0.3439	246		Im	20	0.4977	3803
2	Cs	23	0.5344	789	5	Cs	11	0.4559	1372
	V$_m$	6	0.5295	2111		V$_m$	27	0.5355	3482
	Ic	12	0.5172	17		Ic	9	0.5206	38
	Im	13	0.5011	2114		Im	20	0.4859	3485
3	Cs	4	0.3981	1051	6	Cs	4	0.3829	746
	V$_m$	14	0.5373	2832		V$_m$	9	0.4903	1900
	Ic	15	0.5396	34		Ic	12	0.4895	15
	Im	18	0.5258	2835		Im	7	0.4151	1902



4(a): Key-frames by Clustering

4(b): Key-frames by motion local minima technique

4(c): Key-frame by global color feature & iterative color algorithm.



4(d): Key-frames by global motion feature & iterative motion algorithm.

Fig. 4: Results for Clip 6 (fight-sequence)

5. CONCLUSIONS

We have proposed a novel psychological model to compute the number of key-frames (N_k) to be extracted in a dynamic and adaptive way according to the information content of the movie clip. The approach is validated with four key-frame selection approaches, two adaptive and iterative strategies based on global color and global motion features and two heuristic strategies based on color-histogram based clustering and local-minima in motion trend.

We emphasize here that the present work has been targeted for *commercial movie-clips* developing key-frames for forming *storyboard-type* briefing. Application for laboratory video-data, like mere camera panning or earthquake sequences requires structural features. Such sophistications using learning based techniques are topics of future research.

REFERENCES

[1] H.-C. Lee & S.-D. Kim, "Iterative key-frame selection in the rate-constraint environment," *Sig. Proc.: Im. Comm.*: 18 (2003), 1-15.

[2] Y. Zhuang, Y. Rui, S.-T. Huang and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *Proc. ICIP*: 1 (1998) pp 866-870.

[3] W. Wolf, "Key frame selection by motion analysis," *ICASSP*: (1996) pp 1228-1231.

[4] M.M. Yeung and B. Liu, "Efficient matching and clustering of video shots," in *Proc. ICIP:* (1995) pp 338-342.

[5] R. Zabih, J. Miller and K. Mai, "A feature-based algorithm for detecting and classifying scene breaks," in *Proc. ACM Multimedia:* (1995) pp 189-200.

[6] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," in *Proc. CVPR*: (2003), Vol. 2, pp 343-348.

[7] R.N. Haber and M. Hershenson, "The Psychology of Visual Perception," *Holt, Rinehart and Winston Inc.:* (1973).

[8] J.M. Findlay and I.D. Gilchrist, "Active Vision: the Psychology of Looking and Seeing," *Oxford University Press:* (2003).

[9] R.D. Wright (Ed.), "Visual Attention", *Oxf. Univ. Press:* (1998).

[10] S. Soraci Jr. and K.M. Soraci (Ed.), "Visual Information Processing," *Praeger Pub.:* (2003).

[11] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. on Sys. Man and Cyb. :* (1979) Vol. 9, No. 1, pp 62-66.

[12] T. Gautama and M.M. Van Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering," in *IEEE Trans. on Neural Nets*: (2002) Vol. 13(5), pp 1127-1136.