**RC22222 (W0110-085) October 31, 2001**
**Computer Science**

# IBM Research Report

## Comparison of Sequence Matching Techniques for Video Copy Detection

**Arun Hampapur, Ki-Ho Hyun*, Ruud M. Bolle**
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

*School of Computer and Information Engineering
YoungSan University
150, Junam-ri
Kyongman, Korea 626-840

**Research Division**
**Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich**

# Comparison of Sequence Matching Techniques for Video Copy Detection

Arun Hampapur [a], Ki-Ho Hyun [b] and Ruud Bolle [a]

[a]IBM T.J Watson Research Center,
30 Saw Mill River Road, Hawthorne, NY 10532, USA
[b]School of Computer and Information Eng,
YoungSan University,150,Junam-ri, Kyongman, Korea 626-840

## ABSTRACT

Video copy detection is a complementary approach to watermarking. As opposed to watermarking, which relies on inserting a distinct pattern into the video stream, video copy detection techniques match content-based signatures to detect copies of video. Existing typical content-based copy detection schemes have relied on image matching. This paper proposes two new sequence-matching techniques for copy detection and compares the performance with one of the existing techniques. Motion, intensity and color-based signatures are compared in the context of copy detection. Results are reported on detecting copies of movie clips.

**Keywords:** Video Copy Detection, Sequence Matching, Video Indexing

## 1. INTRODUCTION

The primary thesis of content-based copy detection (CBCD) is *"the media itself is the watermark,"* i.e., the media (video, audio, image) contains enough unique information that can be used for detecting copies. Content-based copy detection schemes extract signatures from the original media The same signatures are extracted from the test media stream and compared to the original media signature to determine if the test stream contains a copy of the original media. The key advantage of content-based copy detection over watermarking is the fact that *the signature extraction can be done after the media has been distributed.* For example, with content-based copy detection, it is possible to create a set of signatures for the movie *Star Wars* (using the master tapes). These signatures can then be used to find all clips of *Star Wars* on the Internet. CBCD also acts as a back up detection method, in case the water mark is compromised. There are several research efforts[1–7] and a number of companies[8] that have developed content-based copy detection techniques. Exiting approaches to CBCD are based on matching a set of key images using image-based signatures. In this paper we focus on sequence matching approaches, these techniques use an interval of time (multiple frames) as the basis for matching, as opposed to matching single video frames. Features like motion, intensity rankings and color histograms are extracted from the original video frames to create the reference signatures. The same features are extracted from the test media sequences and matched to the reference signatures to detect copies.

Section 2 discusses the challenges in video copy detection, while Section 3 discusses previous research efforts in content-based copy detection. The features and the matching technique that are used in this paper are discussed in Section 4. The experiments and results are discussed in Section 5. Section 6 gives a summary and discusses future work.

## 2. CHALLENGES IN VIDEO COPY DETECTION

A video clip can be encoded in different formats depending on the purpose (e.g., RealVideo$^{TM}$ for the Internet and MPEG1 for an intranet). Currently, most of the source material is on tapes and is digitized and encoded by digitizer/encoder devices. This process of digitizing and encoding gives rise to several distortions, the most common digitization artifacts are change in contrast, changes in brightness, shifts in hue, changes in saturation

E-mail: arunh@us.ibm.com

and spatial shifts in the picture. In addition to the digitizer artifacts, lossy encoding processes introduce artifacts like the blocking effects in MPEG. Figure 1 shows frames obtained from a set of video clips. The clips are created from source material on VHS tape. The frames are approximately the same frames from each of these clips. The figure shows six corresponding frames taken from different sources, namely, MPEG1, an AVI, a RealVideo 28k (for a 28k modem), a RealVideo 512k (for a 512k connection), a MPEG1 and an AVI sequence, respectively. The resolution of all the frames is 160×120, except the MPEG1 frames, which are 176×112.

Figure 2 shows a plot of the hue histograms of selected pairs from the images in Figure 1. Figure 2(left) shows a plot of the hue histograms taken of the same face image taken from the MPEG and AVI videos. Clearly, we observe a shift in the peak of the hue histogram (shift of 13 degrees in hue). Figure 2 (right) shows a plot of histograms of two different images (face and two people) taken from the same (MPEG) source. Clearly these histograms show a much higher degree of overlap. As an illustration, Table 1 shows the HSV color histogram intersection[11] values (Hue 16 bins, Saturation 16 bins and Value 16 bins) between the frames of Figure 1. From the table it is seen that the intersection values between frame with the same image content is sometimes less than the intersection values between frames with differing image content. *For example, the intersection between the MPEG1 face image and RealVideo 512K face image is 0.22, whereas it is 0.46 between the AVI face and MPEG men images. Considering the color distortions that are introduced during digitization and encoding, this is not a big surprise. This indicates the need for techniques that are invariant to such distortions.*

**Figure 1.** Images taken from different sources. Left to Right: Face Image 176×112 MPEG1, Face Image 160×120 AVI, Face Image RealVideo-28k (160×120), Face Image RealVideo-512k (160×120),Two People Image MPEG1 176×112, Two People Image AVI 160×120

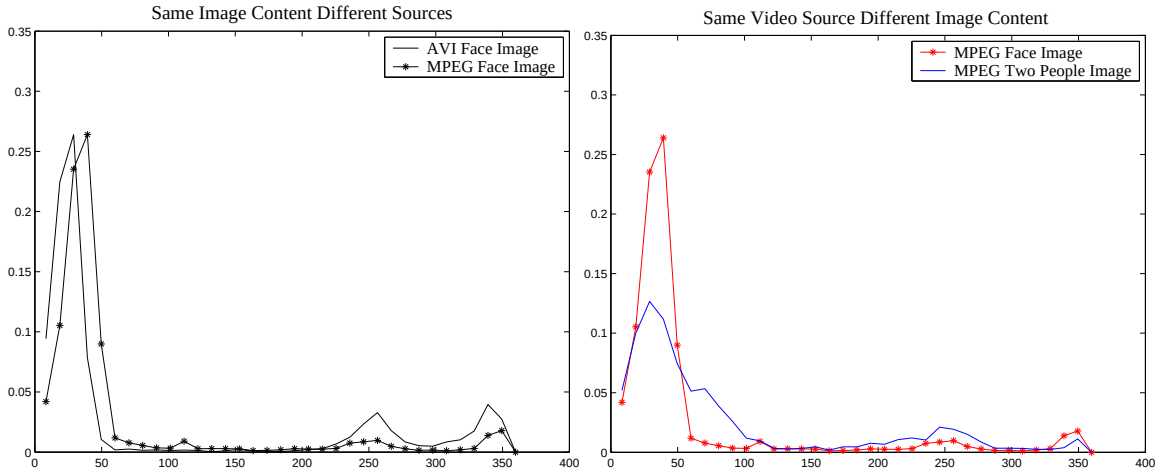


**Figure 2.** Left: Hue Histograms of AVI and MPEG Face Images (Figure 1) Right: Hue Histograms of MPEG Face Image and MPEG Two people image ( Figure 1)

## 3. PREVIOUS WORK

There are a number of efforts[1,2,4–6] in CBCD that use image signature matching. Lienhart et al.[4] describe a system for video clip matching using the color coherence vector to characterize key frames from the clip. Sanchez et al.[6] discuss the use of the principal components of the color histograms of key frames for copy

| | MP face | AVI face | 28k face | 512k face | MP men | AVI men |
|---|---|---|---|---|---|---|
| MP Face | 1.0 | 0.31 | 0.29 | 0.22 | 0.36 | 0.43 |
| AVI Face | 0.31 | 1.0 | 0.54 | 0.39 | 0.46 | 0.22 |
| 28k Face | 0.29 | 0.54 | 1.0 | 0.40 | 0.20 | 0.37 |
| 512k Face | 0.22 | 0.39 | 0.40 | 1.0 | 0.21 | 0.35 |
| MP men | 0.36 | 0.46 | 0.20 | 0.21 | 1.0 | 0.43 |
| AVI men | 0.43 | 0.22 | 0.37 | 0.35 | 0.43 | 1.0 |

**Table 1**: Histogram intersection distances for example images

detection. Both [4,6] have been tested for the domain of TV commercials and are susceptible to color variations. The approach presented by Hampapur[2] circumvents color variations by using edge features. However, none of these approaches effectively exploit the inherent redundancy in video sequences. Among the approaches that exploit the temporal nature of video are.[3,5,10] Indyk et al.[3] propose the use of distance between scene changes in a video as its signature. This is a weak signature and is limited in its applicability. Mohan[10] uses the ordinal measure originally proposed by[9] to retrieve video clips that depict similar actions (similar motions). Naphade et al.[5] developed a scheme for matching video clips. They use histogram intersection of the YUV histograms of the DC sequence of the MPEG video, while proposing an efficient compression technique for the histograms. Their matching scheme uses a sequence of frame signatures. However their signature has not been evaluated for copies with variations in encoding and digitization.

## 4. SIGNATURE EXTRACTION AND MATCHING

This section discusses the process of extracting and matching the motion, ordinal and color signatures. Any given video clip has information distributed in the spatial, color and temporal dimensions. The signatures were chosen so as to compare the relative importance of these dimensions. The motion-based signature exploits only the change information in the video, the ordinal signature is a function of both the color (intensity) and spatial properties of the video, whereas the color signature uses only the color properties without using the spatial information. All three signatures leverage the temporal nature of the video by using frame sequence matching.

### 4.1. Motion direction

Figure 3(left) shows a block diagram of the motion signature extraction process. The frames are partitioned into $N = N_x \times N_y$ blocks and motion vectors are estimated for each block. The direction of the motion vector for each block is quantized into $Q$ directions or levels. The frame signature of any given frame is the number of motion vectors contained in each of the $Q$ levels. In addition to the $Q$ directions, if the motion vector magnitude is zero, a block can also be assigned a level 0. For example, if we are using $15 \times 15$ blocks per frame, and 4 directions for the motion vector quantization, the video signature at any time $t$ will be

$$S_m(t) = q_0(t), q_1(t), q_2(t), q_3(t), q_4(t) \quad (1)$$

$$q_i(t) \in \{0, ..., 225\}, i = 0, ..., 4, \quad (2)$$

with $\sum_{i=0}^{4} q_i(t) = 255$. Effectively, the video signature is 5 bytes per frame.

Let $f_t$ and $f_{t+1}$ be a current frame and a subsequent frame. Figure 3(right) shows the steps involved in the motion estimation process. A small intensity patch $P$, of size $(p_x, p_y)$, is selected around the center $(x_t, y_t)$ of block $B$ of frame $f_t$. A search neighborhood of $(S_x, S_y)$ is selected around the center $(x_{t+1}, y_{t+1})$ of the corresponding block $B$ of frame $f_{t+1}$. The intensity patch $P$ is placed at all possible locations within the search neighborhood and the sum of absolute pixel differences (SAPD) is computed. The SAPD is used as a measure of patch similarity. The location in the search neighborhood which yields the minimum SAPD is considered as the match location $(Mx_{t+1}, My_{t+1})$. The displacement vector $d_x = x_t - Mx_{t+1}$, $d_y = y_t - My_{t+1}$ is used to compute the direction $\theta = \tan^{-1}(d_y/d_x)$ of the local optical flow.

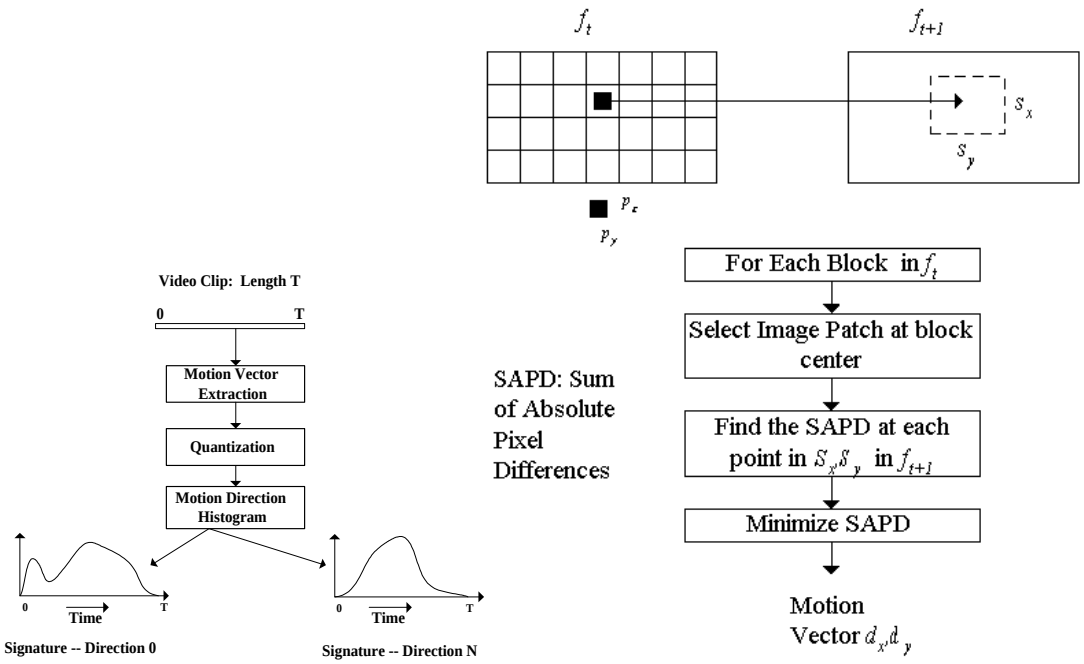


**Figure 3**: Left: Block diagram of motion signature extraction Right: Block diagram of motion vector estimation

#### 4.1.1. Motion matching

Once the motion signatures are extracted for both the reference and test clips, the test signature is convolved along the reference clip and at each point the normalized correlation coefficient is computed. The location with the highest correlation is considered the best match. Let $R(t)$ be the reference signature and $T(t)$ the test signature. We compute $C(t)$, the normalized correlation coefficient, using a support window equal to the length $L_T$ of the test clip around $t$:

$$C(t) = \frac{E(R(t)\,T(t)) - E(R(t)) \times E(T(t))}{\sigma(R(t)) \times \sigma(T(t))}. \tag{3}$$

Here, the expectation $E$ is defined as

$$E(S(t)) = \frac{1}{L_T} \sum_{i=t-L_T/2}^{t+L_T/2} [S(i)] \tag{4}$$

and $\sigma(S(t))$ is the variance of the signature within the window. The point $t_{max}$ at which $C(t)$ is maximum is considered the best match between $R(t)$ and $T(t)$.

### 4.2. Ordinal intensity signature

The ordinal measure was originally proposed by Bhat and Nayar[9] for computing image correspondence. An adaptation of this measure was used by Mohan.[10] Here we use a measure that is similar to that used by Mohan for the purpose of copy detection. Our approach to computing the ordinal signature for a video clip, involves extraction of an ordinal measure for every frame in the clip. The image is partitioned into $N = N_x \times N_y$ blocks and the average gray level in each block is computed. The set of average intensities is sorted in ascending order and the rank is assigned to each block. Table 2 show an example.

Thus if we have $N$ windows, the ordinal signature of a frame at $t$ is given by a vector of integers $r_i$ corresponding to the rank of each window.

$$S_o(t) = (r_1, r_2, \ldots, r_N) \tag{5}$$

| | | | | | |
|---|---|---|---|---|---|
| 20.3 | 12.9 | 123.2 | 1 | 0 | 5 |
| 250.1 | 72.3 | 199.2 | 8 | 3 | 6 |
| 69.3 | 80.2 | 200.0 | 2 | 4 | 7 |

**Table 2.** Left: Average gray level intensity values in a video frame divided in 9 blocks. Right: Ranks of blocks based on intensity ordering

#### 4.2.1. Ordinal matching

Given two signatures $R(t)$ the reference signature, and $T(t)$ the test signature (of length $L_T$), the distance between the two is computed by placing the test signature at different points along the reference and computing the distance. The distance between the two signatures at any time $t$ is given by

$$D(t) = \frac{1}{L_T} \sum_{i=t-L_T/2}^{t+L_T/2} \mid R(i) - T(i) \mid \tag{6}$$

The point $t_{min}$ at which $D(t)$ is minimal is considered the best match between $R(t)$ and $T(t)$.

### 4.3. Color histogram signature

This measure was originally proposed by Naphade et al..[5] They propose to use YUV histograms as the signature of each frame in the sequence and the use of histogram intersection as a distance measure between two signatures (frames). They also proposed a way of compressing histograms from successive frames using a polynomial approximation. We have implemented just the histogram intersection aspect of,[5] as the histogram compression is not essential for evaluating the performance of the video copy detection task. A YUV histogram is computed for each frame of the video. As in,[5] we use 32 bins for Y, 16 for U and 16 for V, hence we use a concatenated histogram of $M = 64$ bins. Thus the signature of a video clip is the sequence of YUV histograms of each frame in the clip.

#### 4.3.1. Color histogram matching

Let the reference signature be $R(t)$ and a test signature of length $L_T$ be $T(t)$. The normalized histogram intersection ($NHI$) is given by

$$NHI(t) = \frac{1}{L_T} \sum_{i=t-L_T/2}^{t+L_T/2} I(H_{R_i}, H_{T_i}) \tag{7}$$

with

$$I(H_{R_i}, H_{T_i}) = \frac{\sum_{l=1}^{l=M} \min(H_{R_i}(l), H_{T_i}(l))}{\sum_{l=1}^{l=M} H_{R_i}}. \tag{8}$$

The $NHI$ measures the similarity between $R(t)$ and $T(t)$. The maximum $NHI(t)$ at point $t_{max}$ is the best match.

## 5. EXPERIMENTAL RESULTS

The movie Star Wars and video from Super Bowl 2001 are used in the experiments. The reference signature is generated from an MPEG1 (352×240) encoding of Star Wars. The test signatures are derived from MPEG1 (176×112) encodings of both Star Wars and the Super Bowl footage.

The following is the experimental procedure used for testing the signature matching.

1. Extract signature from the reference video (**R**).
2. Extract signature from the test video (**V**).

| *Video title* | *Resolution* | *Usage* | *Length* |
|---|---|---|---|
| SW1 | 352×240 | Reference | 2 hrs 12 mins |
| SW1Q | 176×112 | Test | 2 hrs 12 mins |
| SB | 176×112 | Test | 2 hrs 6 mins |

**Table 3**: Experimental data

3. Set test clip length = $L$.
4. Select a random point ($P$) in test video $\mathbf{V}$.
5. Select a clip ($C$) of length $L$ around $P$.
6. Find the best match location $M_l$ of $C$ against $R$ and match score $M_s$
7. Repeat Steps 4-6, *100 times*.
8. Repeat Steps 3-7, for different clip lengths $L$.

The performance of each of the sequence matching techniques is plotted using its receiver operating characteristics (ROC) curve. This is a plot of the false positive rate ($F_{pr}$) versus the false negative rate ($F_{nr}$). Let $N_T$ be the total number of match tests conducted and let $\tau$ be the match threshold. With $F_n$ the number of false negatives (clips that should have matched, but did not) and $F_p$ the number of false positives (clips that matched but are not part of the reference set), we have

$$F_{pr}(\tau) = \frac{F_p}{N_T} \quad F_{nr}(\tau) = \frac{F_n}{N_T} \tag{9}$$

The ROC curves are computed by varying $\tau$ from its minimum value to its maximum with an increment of 5%. A good ROC curve lies very close to the axes, i.e., there is some threshold value for which both $F_{pr}$ and $F_{nr}$ are very close to zero. The ideal case is when $F_{pr} = F_{nr} = 0$.

### 5.1. Discussion

Figures 4 and 5 show the ROC curves for motion-based matching, ordinal matching and color-based matching. There is one ROC curve for each clip length $L$. Not unexpectedly, the matching performance improves with increasing clip lengths. Table 4 shows the lowest value of the errors for different clip lengths.

| $L$ | 20.3 s | 10.6 s | 5.3 s | 2.6 s | 1.3 s |
|---|---|---|---|---|---|
| Motion | 6, 1 | 8, 11 | 24, 14 | - | - |
| Ordinal | 0, 0 | 0, 0.5 | 0, 2.0 | 0, 2.5 | 5, 7 |
| Color | 21, 14 | 23, 18 | 27, 23 | - | |

**Table 4**: Representative $F_{pr}, F_{nr}$ pairs for a number of clip lengths in per cent

Note that the scales of the axes of the plots in Figures 5(left) and (right) are not the same. From the Figures and Table 4 it can be seen that the ordinal signature has the best performance, followed by the motion signature and the color signature has the worst performance. This result is in line with our expectation given the typical variation in color between different encodings (section 2). The performance can be explained as follows.

**Ordinal Signature:** This captures the relative distribution of intensities over space and time. Thus it is immune to global changes in the quality of the video that are introduced by the digitization/encoding process.

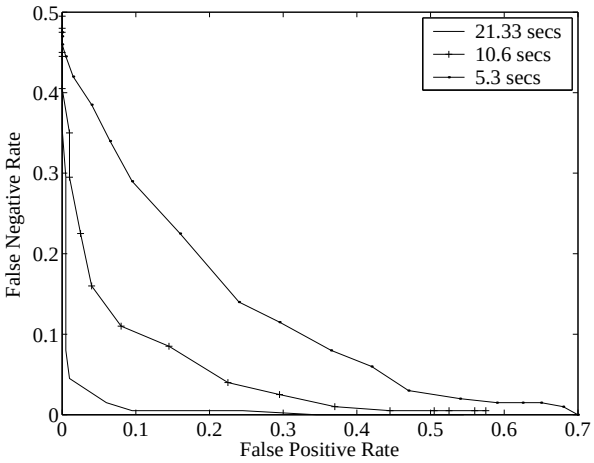


**Figure 4**: Motion signature: False positive vs. false negative rates

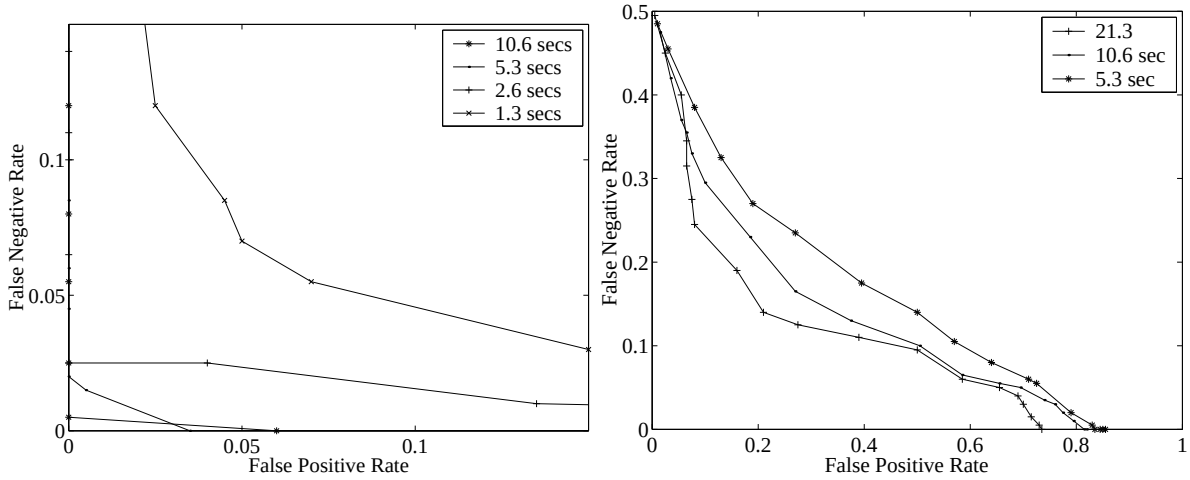


**Figure 5.** Left: Ordinal signature: False positive vs. false negative rates Right: Color signature: False positive vs. false negative rates

**Motion Signature:** This again captures the relative change in the intensities over time, thus it is immune to global color changes. However, since it discards spatial information, the performance is worse than the ordinal signature Even when spatial information is added to this signature (i.e. block motion matching) its performance will be worse than the ordinal signature since significant parts of typical video frames do not contain reliable motion information.

**Color Signature:** Since this signature encodes the absolute color of the frames while discarding spatial information, it is highly susceptible to the global variations in color that are caused by different encoding devices. Another reason for the poor performance is the fact that in a movie like Star Wars, there are a number of shots in different parts of the movie with the same color scheme, which are indistinguishable without the use of spatial information.

In addition to having the best error rates for very short clips, the ordinal signature is also very computationally efficient and can be computed in real time for processing live video. Thus signatures encoding relative information perform better than those encoding absolute properties of the video. These results are in concurrence with the evaluation of distance measures for image matching presented by Hampapur.[12]

## 6. SUMMARY AND CONCLUSIONS

This paper makes three contributions toward video copy detection, namely, it proposes a new motion signature, novel application of ordinal signature proposed in Bhat and Mohan[9, 10] and experimental comparison of these methods to the color signature proposed in Naphade.[5] The results of the comparison indicate that the ordinal feature has superior performance followed by the motion and color features. Future research includes the characterization of the ordinal feature for different types of content and exploring a modification of the ordinal feature to preserve spatial coherence that makes this ordinal feature intuitively more meaningful. We further are developing indexing schemes that implement massively parallel convolution with reference streams, rather than the current method of sequential convolution.

## REFERENCES


1. C. L. E. Chang, J. Wang and G. Wiederhold, "Rime: A replicated image detector for the world wide web," in *SPIE Multimedia Storage and Archiving Systems III*, Nov. 1998.
2. A. Hampapur and R. M. Bolle, "Feature based indexing for media tracking," in *Proc. of Int. Conf. on Multimedia and Expo*, pp. 67–70, Aug. 2000.
3. G. I. P. Indyk and N. Shivakumar, "Finding pirated video sequences on the internet.," in *Stanford Infolab Technical Report*, Feb. 1999.
4. C. K. R. Lienhart and W. Effelsberg, "On the detection and recognition of television commercials," in *Proc. of the IEEE Conf. on Multimedia Computing and Systems*, 1997.
5. M. Y. M. Naphade and B.-L. Yeo, "A novel scheme for fast and efficient video sequence matching using compact signatures.," in *Proc. SPIE, Storage and Retrieval for Media Databases 2000*, **Vol. 3972**, pp. 564–572, Jan. 2000.
6. J. V. J. M. Sanchez, X. Binefa and P. Radeva., "Local color analysis for scene break detection applied to tv commercials recognition.," in *Proceedings of Visual 99*, pp. 237–244, June 1999.
7. S.-C. Cheung and A. Zakhor, "Estimation of web video multiplicity,," in *Proc. SPIE – Internet Imaging*, **vol. 3964**, pp. 34–6, 2000.
8. w. Contentwise Inc
9. D. Bhat and S. Nayar, "Ordinal measures for image correspondence," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20 Issue: 4**, pp. 415–423, April 1998.
10. R. Mohan., "Video sequence matching.," in *Proceedings of the International Conference on Audio, Speech and Signal Processing, IEEE Signal Processing Society*, 1998.
11. M. Swain and D. Ballard, "Color indexing," in *International Journal of Computer Vision*, **Vol. 7, No. 1**, pp. 11–32., 1991.
12. A. Hampapur and R. M. Bolle, "Comparison of distance measures for video copy detection.," in *Proc. of Int. Conf. on Multimedia and Expo*, Aug. 2001.