COMPARISON OF DISTANCE MEASURES FOR VIDEO COPY DETECTION

Arun Hampapur and Ruud M. Bolle

IBM TJ Watson Research Center
30 Sawmill River Rd
Hawthorne, NY 10532
{arunh,bolle}@us.ibm.com

ABSTRACT

Content-based copy detection (CBCD) is a complementary approach to watermarking for detecting copies of media. Watermarking relies on the ability to detect from a copy a distinct pattern that was introduced into the original media. CBCD techniques detect copies by measuring distances between content-based signatures extracted from the original and the copy. The critical challenge in content-based copy detection is the design of signatures, which are invariant to the differences in quality across copies of the same media (resolution, compression and digitization effects). Most of the distance measures used in image retrieval have been developed without much consideration to these types of variations between copies. This paper examines the use of several image distance measures in the context of video copy detection and compares their performances.

1. INTRODUCTION

Detecting copies of media (images, audio and video) is a basic need in digital media management. The applications of copy detection include *usage tracking* and copyright enforcement. There are two approaches to detecting copies of digital media, watermarking [4] and content-based copy detection. Watermarking embeds information into the media prior to distribution. Thus all copies of the marked content contain the watermark, which can be extracted, to prove ownership. Content-based copy detection is a *complementary approach* to watermarking. The primary thesis of content-based copy detection 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 measure the distance between signatures extracted from the original media and a possible copy. 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, say, the master tapes). These signatures can then be used to *find all clips of Star Wars on the Web.* This task would be impossible using the watermarking approach. There are several research efforts [2, 5, 8, 9, 10, 12] and a number of companies [3] that are addressing content-based copy detection.

Most video copy detection algorithms use some form of image distance measurement combined with temporal evidence integration [5]. In this paper, we will explore the performance of several distance measures in the context of video copy detection. Section 2 presents a description of the problem and presents examples of the data. Previous work in this area is described in Section 3. In Section 4, we present a set of distance measures that are used in the experiments. Section 5 discusses how the experimental data was obtained. Section 6 presents the results of applying these features to the experimental data set. Finally, Section 7 presents conclusions and future work.

2. CHALLENGES IN VIDEO COPY DETECTION

A video clip can be encoded in different formats depending on the purpose (e.g., $\mathrm{RealVideo}^{TM}$ for the Internet and MPEG1 for intranet). Currently, most of the source material is on tapes and is digitized and encoded by digitizer/encoder cards. 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 and spatial shifts in the picture. In addition to the digitizer artifacts, lossy encoding processes introduce artifacts like the blocking seen in MPEG video. Figure 1 shows corresponding frames obtained from multiple formats of the same material. The six frames (left-to-right, top-to-bottom) are taken from, an MPEG1, an AVI, RealVideo 28k (for a 28k modem), RealVideo 512k (for a 512k connection), MPEG1 frame and an AVI frame, respectively. The resolution of all the frames is 160×120, except the MPEG1 frames, which are 176×112. The difference in the images is illustrated by the gray level histograms shown in Figure 2. As an illustration, Table 1 shows the color histogram (3D) intersection values (Hue 16 bins, Saturation 16 bins and Value 16 bins) between the images of Figure 1. From this table we clearly see that the intersection value between the copies is sometimes less than the intersection values between the different images. For example, 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. Similar results are obtained for linear color histograms (H, S, V channels treated independently and concatenating the histograms).

There are two ways of dealing with the differences that exist between copies of video clips. The first one is to use device calibration and the second one is to develop distance measures that are invariant to these distortions. Sanchez et al. [11] have addressed the problem of color variations that are due to the acquisition device, by applying color correction to the video signal. They compare the effectiveness of various color constancy algorithms for matching video frames. The results indicate that workable results are obtained by using a test pattern to calibrate the acquisition devices. This approach is only possible when the video copy signal-

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE

Figure 1: Images taken from different sources. Top-left: 176×112 MPEG1. Top-right: 160×120 AVI. Middle-left: RealVideo 28k (160×120). Middle-right: RealVideo 512k (160×120). Bottom-Left: MPEG1. Bottom-Right: AVI

	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: 3D HSV Histogram intersection distances for example images

encoding device is accessible, which is not the case in general. For example, when searching the web for clips of Star Wars, an unknown device has already encoded the test signal. In this work we explore the use of invariant frame distance measures, which can accommodate for variations between copies of the same video clip.

3. PREVIOUS WORK

Existing work in copy detection is fairly limited and has not addressed the various kinds of distortions due to the digitization/encoding processes. Recognition of commercials has been one of the areas where techniques for copy detection have been developed. Lienhart et al. [9] describe a system for performing both feature based detection and recognition of commercials. They use the color coherence vector to characterize key frames in the reference segment. Sanchez et al. [12] discuss the use of the principal components of the color histograms of key frames for commercial recognition. They report results on a database of 20 commercials using a sequential matching approach. Since the techniques of Lienhart and Sanchez rely on color, variations in color are likely to cause problems in these approaches. Indyk et al. [8] have proposed a method for video copy detection, using the distance between shot breaks in the video as the feature of the video. This feature is very limited in its applicability. Hampapur et al. [5] have discussed the use of color and its limitations. They have used invariant edge features to circumvent color variations. Naphade et al. [10] 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. This technique, again, does not address the variations that commonly exist between different copies of the same material (in different formats or encoded by different hardware). Chang et al. [2] proposed the use of wavelet-based replicated image detection on the web. They have tested their scheme by using a set of query images that are modified from their originals by operations like sharpening, softening and despecling. Their results indicate that out of ten queries, they were able to correctly detect eight copies. It is not clear, though, how their algorithm would do with the typical distortions encountered with copies of videos (as discussed in Section 2).

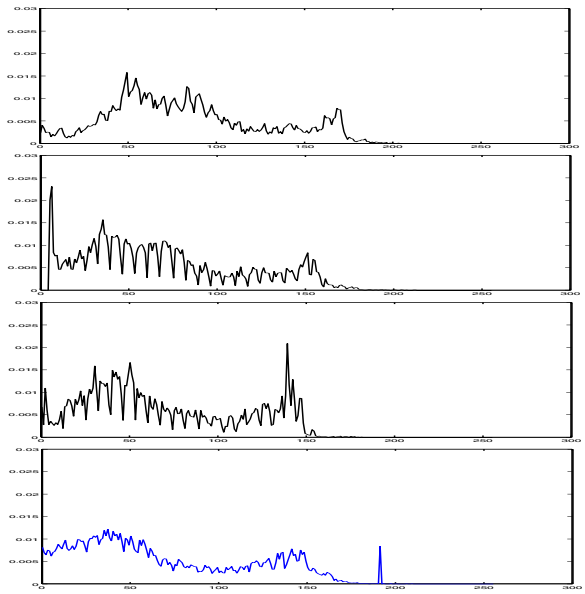

Figure 2: Gray level histogram for **face** image. Top-to-bottom: MPEG1, AVI, RealVideo-28k, RealVideo-512k

The goal of this work is to apply a set of distance measures to a set of frames extracted from video, to demonstrate the effectiveness of each of these measures for video copy detection. The experiment uses two copies of the video. Each of the measures is evaluated in terms of the false positive and false negative rates.

4. DISTANCE MEASURES

Since the images used in the experiments are of different resolution, we scale the lower resolution images to the larger one. This scaling algorithm is based on bilinear interpolation. We discuss here the distance measures we use for copy detection.

Image Difference: This is the simplest measure, which uses the absolute value of the pixel differences in each of the bands between the two images. The difference is normalized by the number of pixels $W \times H$ (width × height)) and color channels:

$$D_{id} = \frac{\sum_{x=0}^{W-1} \sum_{y=0}^{H-1} \sum_{c=0}^{2} \mid I_1(x,y,c) - I_2(x,y,c) \mid}{3 \times W \times H}. \quad (1)$$

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE
0-7695-1198-8/01/$10.00 (C) 2001 IEEE

Experiments reported here have also used a smoothing operation, which replaces each pixel by the average value of its neighbors before computing the difference. This measure is referred to as the *smoothed image difference*.

Histogram Intersection (RGB, HSV, Gradient Direction): This measures the similarity between histograms [15]. The following three types of histograms are used in distance measurement,

- RGB Histogram: Each of the color channels is quantized into 16 bins.
- HSV Histogram: The RGB image is converted into HSV and the H, S and V channels are quantized into 16, 8 and 4 bins respectively. The bins counts are chosen to give hue the maximum weight, as opposed to the value, which (typically) tends to vary across different acquisition devices.
- Gradient Direction Histogram: The images are convolved with a Sobel kernel and a gradient magnitude threshold is used to select likely edge pixels. The direction of the gradient at these pixels is quantized into 18 bins and accumulated in the histogram.

The histogram distance between two images with histograms h_1 and h_2 is computed as:

$$D = 1 - \text{Histogram Intersection}(h_1, h_2) \quad (2)$$

$$D = 1 - \sum_{\substack{c=r,g,b \\ c=h,s,v; \\ c=\theta}} \sum_{b=0}^{N_c-1} \min(h_1(c,b), h_2(c,b)), \quad (3)$$

where h_i is the (RGB, HSV or direction) histogram of image i and N_c is the corresponding number of histogram bins.

Hausdorff Distance: This distance [6] is computed based on an edge representation of the two images. The images are first converted to gray scale and the Canny edge detector is used to extract edges. The distance between the two sets of edge points is computed using the Hausdorff distance.

$$D_{haus} = \max(h_p(I_1, I_2), h_p(I_2, I_1)) \quad (4)$$

Here $h_p(I_1, I_2), h_p(I_2, I_1)$ are the partial Hausdorff distances from I_1 to I_2 and I_2 to I_1, respectively. The definitions of Hausdorff distance and partial Hausdorff distance are the following.

Hausdorff Distance $h(A, B)$, where A and B are a set of points (in our case edge points in the two images).

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \quad (5)$$

Here $\|.\|$ denotes some norm, we use the city-block distance for efficient implementation [14]. For the purposes of image copy detection, the Hausdorff distance is very fragile, whereas the partial Hausdorff distance proves to be much more robust.

Partial Hausdorff Distance $h_p(A, B)$

$$h_p(A, B) = \underbrace{N^{\text{th}}_{largest}}_{a \in A} \min_{b \in B} \|a - b\| \quad (6)$$

Local Edge Representation: The images are converted to gray scale and edge points are extracted by thresholding the magnitude of the gradient (using Sobel). The edge image is then partitioned into $n_1 \times n_2$ windows (n_1 along the width and n_2 along the height). The i^{th} window is represented by value c_i, extracted by quantizing the position of the centroid of the edge points within that window. The distance between two edge representations is the fraction of windows that, within quantization, have the same centroid, as shown below.

$$D_{led} = \frac{\sum_{i=0}^{n_1 \times n_2} \begin{cases} 0 & c_i(I_1) = c_i(I_2) \\ 1 & c_i(I_1) \neq c_i(I_2) \end{cases}}{n_1 \times n_2} \quad (7)$$

Invariant Moments: This distance measure is based on the invariant moments proposed by Hu [7, 1]. The images are converted to gray scale and edges are extracted using a Canny edge detector. Of all the moments proposed in [7], we use the second-order central moments to compute the spread and slenderness of the edge cluster in the frame. The distance is measured as follows.

$$D_{mom} = \sqrt{X * X + Y * Y} \quad (8)$$

$$X = \mu_{20} + \mu_{02}; \quad Y = \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}{}^2} \quad (9)$$

$$\mu_{pq} = \sum_{\text{edges}(x,y)} (x - \bar{x})^p (y - \bar{y})^q$$

$(\bar{x}, \bar{y})$ is the centroid of the edges.

5. DATA SET DESCRIPTION

Source: The source material was obtained from cable television programming in the NY area, recorded on VHS tape. Both digital copies of the material are made from this tape.

Copy #1 C_1**:** Various clips are converted to MPEG1 using an Optibase Encoder. The resolution is 352×240 at 3.3Mbits/sec.

Copy #2 C_2**:** The same clips are captured in AVI using an Osprey 100 frame grabber, with resolution 240×180.

Synchronization: Manually aligning the corresponding starting frames and ending frames of the clips synchronize the two copies of the video.

Image selection: Every 30^{th} frame (1 frame per sec) is extracted to be used in the experiments.

Manual verification: All the images and copies in the database are manually checked to ensure that the images are exact copies of each other.

Database size: The current set of experiments uses two copies of 617 images derived from 36 (30-second) clips of video. All the clips are television commercials.

Figure 3 shows some of the images in the database, the goal is to illustrate the wide variety of image content (from buildings, to faces, to text images) and the various types of framing that occurs due to camera motion. These images are typically different from most images used in image retrieval experiments like the Corel database. For still photographs, the framing and quality is much better compared to video frames. For video, the cameraman pays much less attention to a single image (frame) and uses time and motion to capture the spatial-temporal content.

6. EXPERIMENTAL RESULTS

The following experimental procedure is used to compare the performance of the seven distance measures discussed in this paper.

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE
0-7695-1198-8/01/$10.00 (C) 2001 IEEE

Figure 3: Example Images taken from several commercials

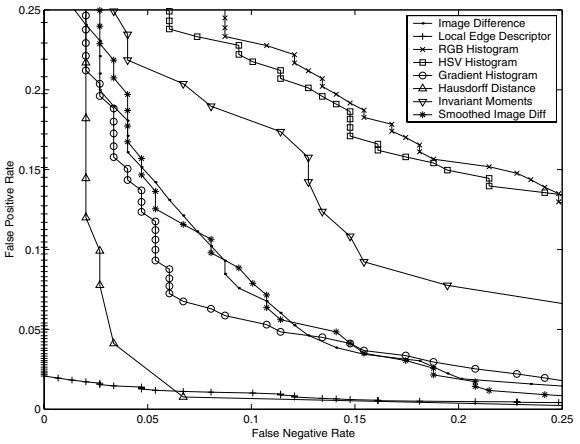

Figure 4: Receiver Operating Characteristics (ROC) Curve: The false positive rate at a threshold t is given by $\frac{\text{\# of mismatch scores} < t}{\text{total mismatches}}$. The number of false negatives at t is given by $\frac{\text{\# of match scores} > t}{\text{total matches}}$.

The distance D_{ij} for every image $I_i \in C_1$ and $I_j \in C_2$ was computed using each of the distance metrics. The distances generated by this process were thresholded to generate the false positive and false negative rates presented in Figure 4. The ideal ROC curve should lie as close to the axes as possible. Thus from the figure we see that, the Local Edge descriptor provides the best performance, followed by the partial Hausdorff distance, gradient direction histogram, the image differences, moments and color histogram intersection. These results are in concurrence with our expectations, as we know that the frame grabbers affect the color the most. Hence the edge based local measures perform better. However, we see that even for the local edge descriptor, the best operating point has a false positive rate of 0.02 and a false negative rate of 0.

7. CONCLUSIONS

Video copy detection gives rise to several problems that have not been addressed in the content-based image retrieval research[13], that is, accommodating for variations between copies (due to digitizer and encoder artifacts), the wide variety of content (from text to faces to digital editing effects) and efficiency considerations. The local edge measure proposed in [5] has good performance, however, the number of bits of indexing information required here is of the order of 100 bytes per frame. In order to effective index large databases, the size of the signature must be smaller. One of the future directions is to use motion-based features to reduce the number of indexing bits required per frame of video.

8. REFERENCES

[1] S.O.Belkasim, M.Shridhar, and M.Ahmadi, Pattern Recognition with Moment Invariants: A comparative study and new results. In *Pattern Recognition*, Vol 24, No 12, 1991.

[2] E. Chang, J. Wang, C. Li and G. Wiederhold, RIME: A replicated image detector for the World Wide Web In *SPIE Multimedia Storage and Archiving Systems III*, Nov. 1998.

[3] Contentwise Inc, www.contentwise.com.

[4] Digital Watermarking *IEEE Signal Processing Magazine*, Vol. 17, No. 5, Sept. 2000

[5] A. Hampapur and R. M. Bolle, Feature based Indexing for Media Tracking. In *Proc. of Int. Conf. on Multimedia and Expo*, Aug. 2000, pp. 67-70.

[6] D.P. Huttenlocher, G.A. Klanderman and W.J. Rucklidge, Comparing images using the Hausdorff distance. *IEEE Trans. on PAMI*, Vol. 15, No. 9, Sept. 1993.

[7] N. K. Hu, Visual pattern recognition by moment invariance. *IRE Transactions on Information Theory,* 1962, pp. 179-187.

[8] P. Indyk, G. Iyengar and N. Shivakumar, Finding pirated video sequences on the Internet. tech. rep., Stanford Infolab, Feb. 1999.

[9] R. Lienhart, C. Kuhmunch and W. Effelsberg, On the detection and recognition of television commercials. In *Proc. of the IEEE Conf. on Multimedia Computing and Systems*, 1997.

[10] M. Naphade, M.M. Yeung 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, Jan. 2000, pp. 564-572.

[11] J.M. Sanchez and X. Binefa, Color normalization for appearance based recognition of video key-frames. In *Proc. of Int. Conf. on Pattern Recognition*, Vol. I, Aug. 2000, pp. 815-818.

[12] J.M. Sanchez, X. Binefa, J. Vitria, and P. Radeva. Local color analysis for scene break detection applied to TV commercials recognition. In *Proceedings of Visual 99*, June 1999, pp. 237-244,.

[13] A.W.M Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, Content-Based Image Retrieval at the end of the early years. In *IEEE Transactions on PAMI*, Vol. 22, Number 12, Dec. 2000.

[14] R. Shonkwiler, An Image Algorithm for computing the Hausdorff distance efficiently in linear time. *Information Processing Letters,* Vol. 30, 1989, pp. 87-89.

[15] M. Swain and D. Ballard, Color Indexing. *International Journal of Computer Vision*, Vol. 7, No. 1, 1991, pp. 11-32.

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE

0-7695-1198-8/01/$10.00 (C) 2001 IEEE