EFFICIENT VIDEO SIMILARITY MEASUREMENT AND SEARCH

Sen-ching S. Cheung and Avideh Zakhor

Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720
{cheungsc, avz}@eecs.berkeley.edu

ABSTRACT

We consider the use of meta-data and/or video-domain methods to detect similar videos on the web. Meta-data is extracted from the textual and hyperlink information associated with each video clip. In video domain, we apply an efficient similarity detection algorithm called video signature. The idea is to form a signature for each clip by selecting a small number of its frames that are most similar to a set of random seed images. We then apply a statistical pruning algorithm to allow fast detection on very large databases. Using a small ground-truth set, we achieve 90% recall and 95% precision using only 8% of the total number of operations required without pruning. For a database of around 46,000 video clips crawled from the web, video signature technique significantly outperforms meta-data in precision and recall. We show that even better performance can be achieved by combining them together. Based on our measurements, each video clip in our database has, on average, 1.53 similar copies.

1. INTRODUCTION

The amount of information on the world wide web has grown enormously since its creation in 1990. Since there is no central management on the web, duplication of content is inevitable. As reported by Shivakumar and Garcia-Molina in 1998 [4], about 46% of all the text documents on the web have at least one "near-duplicate" – document which is identical except for low level details such as formatting. The problem is more severe for videos as they are often mirrored in multiple locations, formats and bitrates to facilitate downloading and streaming. Multimedia authoring tools also enable users to slightly modify existing video clips and to republish them on the web. An efficient algorithm to identify similar videos can therefore be beneficial to many web retrieval scenarios such as presenting uncluttered search results, and providing alternatives in the case of expired links or network outages.

In this paper, we propose a number of methods for detecting similar videos on the web. We consider two different approaches based on (a) meta-data, such as web addresses and descriptive text associated with each video clip, and (b) features extracted directly from the video frames. Even though our meta-data based scheme is simple to implement, we show that meta-data alone results in rather low precision. Similarity detection schemes based on visual information are usually more reliable, but they are typically complex and difficult to apply to very large databases. A fast video comparison algorithm, called *video signature*, was first proposed in [2] to address this problem. This algorithm is particularly suitable for large databases, because it supports a statistical pruning scheme which can reliably prune off a significant portion of signature comparisons. In this paper, we characterize the performance of video signature using both a small ground-truth set and a large database of web videos. We also demonstrate that precision/recall performance can be further improved by combining video signature with meta-data.

This paper is organized as follows: meta-data and video signature similarity detection schemes are introduced in Section 2. Complexity reduction based on statistical pruning is described in Section 3. Section 4 includes performance characterization on a large video database, and Section 5 contains conclusion.

2. VIDEO SIMILARITY

2.1. Video similarity based on meta-data

We define similar videos to be those with almost identical content but possibly compressed at different qualities, reformatted to different sizes and frame-rates, or undergone minor editing in either spatial or temporal domain. The most straightforward way to find similar videos on the web is through commercially available multimedia search engines such as AltaVista and Scour. These search engines primarily use meta-data to index, search, and retrieve video clips. To gain an understanding of achievable meta-data performance in detecting similar videos, we have implemented a simple meta-data matching scheme. Each video clip is represented by a set consisting of all the distinct terms found in the associated meta-data. The meta-data is derived from the web address of the video clip, the descriptive text associated with the address, and possible author, copyright and title information. We also consider the web address and the title of the web-page which links to the clip as part of its meta-data. All web addresses are broken up into terms delimited by non-alpha-numerical characters. Common terms such as "http" and "www", and trailing numbers of filenames, which often indicate different versions of the same video, are removed to enhance matching. The degree of meta-similarity is defined as the ratio between the size of the intersection and the union of the two meta-data sets. This metric is commonly used in detecting similar documents [4].

However, unlike documents, there are two problems in applying this method to our meta-data. First, the number of distinct terms for each video is generally very small. Based on our database of about 46,000 web video clips [2], we find that, on average, each video has 9.8 distinct terms, thereby reducing the reliability of this method. Second, since web addresses are part of the meta-data set, two video clips from the same physical location, or pointed by the same web-page, are likely to be declared as similar.

This work was sponsored by Hughes Research Lab and California Digital Media Innovation (DiMI) program, D97-03.

To mitigate this effect, web address information is not used in such cases.

If we define "meta-similar" video clips as those pairs with degree of meta-similarity larger than 75%, then about 50,0000 pairs in our database can be classified as meta-similar. Using manual inspection, we estimate the precision to be $27 \pm 9\%$.[1] The number of pairs reduces to around 26,000 when we consider only those pairs which share identical meta-data terms. The estimated precision, however, stays roughly the same at $33 \pm 9\%$. The low precision is due to the fact that most of the meta-similar videos are semantically related, but visually different. Some mis-classified examples include different episodes of the same television program, or movie clips featuring the same actor. To identify similar videos according to our definition, meta-data is therefore not adequate.

Even though meta-data does not seem to produce favorable performance, we use it in conjunction with manual inspection, to identify 195 visually similar video pairs out of a set of 212 video clips as part of a ground-truth set for controlled testing. To include non-similar video sequences in the ground-truth set, we also randomly sample our large database to arrive at a set of 377 distinct video clips which have been identified as non-similar through manual inspection. Thus, our ground-truth set has a total of 589 video clips.

2.2. Video signature

We model a video clip V as a collection of its individual frames $\{v\}$. The similarity between video sequences is based solely on the similarity between individual frames. Let $V = \{v\}$ and $W = \{w\}$ denote two video sequences. Assume that we have a visual feature distance function $d(v, w)$ between frames v and w. We define video distance $D(V, W)$ as the average distance between the closest matched frames of the two video sequences :

$$D(V,W) \triangleq \frac{1}{|V|+|W|}\left[\sum_{v\in V} d(v, g_W(v)) + \sum_{w\in W} d(g_V(w), w)\right]$$

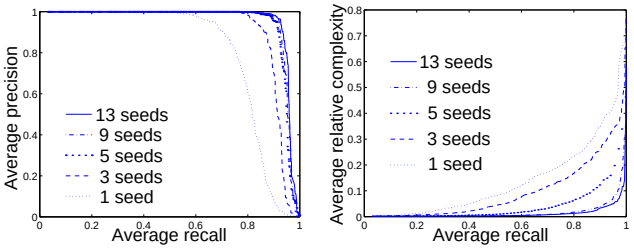
where $g_X(y) \triangleq \arg\min_{x\in X} d(x, y)$ denotes the frame in video clip X which is visually closest to frame y. $|V|$ and $|W|$ denote the size of sets V and W respectively. In practice, it is computationally prohibitive to compute D because its complexity is proportional to the product of the size of V and W. To reduce the complexity, we introduce a particular form of random sampling called *video signature*.

Let $R = \{s_1, s_2, \ldots, s_M\}$ be a set of M random images which we call *seed images*. Define an M-tuple of frames $V_R \triangleq (v_{s_1}, v_{s_2}, \ldots, v_{s_M})$ called the *signature* of V with respect to R, where $v_{s_i} \triangleq \arg\min_{v\in V} d(v, s_i)$ is the frame in V closest to seed image s_i. To measure the distance between two signatures V_R and W_R, we now define the *signature distance* function $T_{sig}(V_R, W_R)$ as follows :

$$T_{sig}(V_R, W_R) \triangleq \text{median}\{d(v_{s_i}, w_{s_i}), i = 1 \ldots M\} \quad (1)$$

The median operator is used to remove outliers due to possible sampling error in choosing similar signature frames. It can be shown that $T_{sig}(V_R, W_R)$ is a reasonable approximation to $D(V, W)$ when $D(V, W)$ is small [3]. To declare two signatures V_R and W_R as similar, we would require $T_{sig}(V_R, W_R) \leq \epsilon$, where ϵ, the *signature distance threshold*, is determined experimentally in such a way as to balance precision versus recall. The advantage of using the signature distance, instead of the original video distance, is that computing T_{sig} requires only $O(M)$ operations.

[1] All the precision numbers in this paper, except for those from the ground-truth set which can be computed precisely, are estimated by first randomly choosing 100 video pairs from the target set, and viewing them side by side to determine if they are similar. The margin of error is computed based on a 95% confidence level.



(a) (b)

Fig. 1. *(a) Average precision versus recall for different number of seeds; (b) Average relative complexity versus recall plot for different number of seeds.*

In our experiments, we use a quadrant-based HSV color histogram as our feature to represent each individual frame, both for the purpose of forming signatures, and for computing T_{sig}. Each quadrant color histogram has 178 bins with 18 bins for hue, 3 for saturation, 3 for value, plus 16 pure gray levels. Thus, each signature frame is represented by a 712-dimensional vector. A modified l_1-metric between two histograms is used as distance : any dominant color occupying more than 50% in both histograms is removed, and the rest of the bins are renormalized before the l_1-metric is computed.

Figure 1(a) shows precision versus recall for different number of seeds, M, on the ground-truth set. Each point on each curve represents the average of four independent runs, each with a different set of seed images, for a particular value of ϵ. As shown in Figure 1(a), both precision and recall improve when more seeds are used, as the sampling error in choosing the signature frames decreases. The improvement, however, becomes negligible when more than nine seeds are used. This implies that the sampling error introduced by video signature is mostly eliminated when nine or more seeds are used.

3. COMPLEXITY REDUCTION

A major step in finding the signature distance, T_{sig}, between two video signature is to compute all the frame distances, $d(v_{s_i}, w_{s_i})$ between signature frames v_{s_i} and w_{s_i} for $1 \leq i \leq M$. The complexity of the frame distance computation is proportional to the dimension of the feature vector, which is usually quite high in multimedia retrieval applications, and 712 in our particular case. It is crucial to reduce the complexity of such an operation, as it needs to be done for every seed and for every pair of video in the database. We have recently proposed a statistical pruning procedure to reduce this complexity [2]. This pruning scheme is a variation of a class of high-dimensional indexing techniques, known as the Triangle Inequality based pruning algorithms [1]. The basic idea is to avoid computing the complex T_{sig} for some signature pairs, based on the results of a much simpler to compute *seed distance* defined below:

$$T_{seed}(V_R, W_R) \triangleq \text{median}\{|d(v_{s_i}, s_i) - d(w_{s_i}, s_i)|, i = 1 \ldots M\}$$

Both $d(v_{s_i}, s_i)$ and $d(w_{s_i}, s_i)$ are already computed in the signature generation process. The ratio of the computational complexity of T_{seed} to that of T_{sig} is roughly 1:712. The reason is that for every seed, T_{sig} requires computation of the l_1-metric between a pair of 712-dimensional vectors, while T_{seed} only needs a single subtraction. Our statistical pruning approach is to first compute T_{seed}

for every possible pair of signatures in the database, and then compute T_{sig} only for those pairs with $T_{seed} \leq \epsilon_s$. ϵ_s is called the *seed distance threshold* and is determined experimentally. Thus, statistical pruning changes the complexity from $O(712MN^2)$ when every T_{sig} is computed, to $O(MN^2 + 712\rho MN^2)$, where ρ is the portion of the signature pairs remained after pruning and N is the total number of videos in the database [2]. Since we are interested in reducing complexity, we consider the relative complexity, C_{prune}, defined as follows:

$$C_{prune} \triangleq O(MN^2 + 712\rho MN^2)/O(712MN^2) \approx (1 + 712\rho)/712 \quad (2)$$

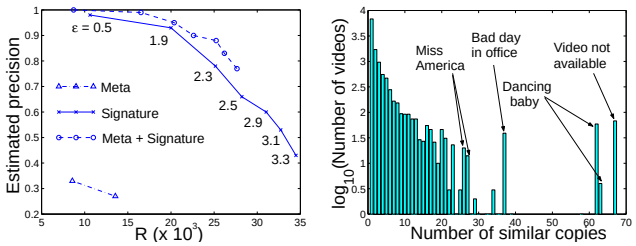
As we decrease ϵ_s, C_{prune} will decrease since more signature pairs are pruned away. However, if too many signature pairs are pruned, recall will be lowered as some truly similar videos are eliminated. Figure 1(b) shows C_{prune} as a function of recall on the ground-truth set for different values of M. Each point on each curve represents the average of four independent runs resulting from four different sets of seed images, for a particular value of ϵ_s. Figure 1(b) clearly demonstrates the trade-off between recall and complexity. For example, for $M = 9$, C_{prune} changes from 56% to 8% as recall varies from 100% to 95%. For any given recall value, C_{prune} decreases as M increases, until M reaches nine. In Section 2.2, nine seeds were also shown to be adequate in terms of precision/recall performance on the ground-truth set. Thus, a nine-seed signature will be used for all the experiments in the remainder of this paper.

4. EXPERIMENTAL RESULTS

In order to examine the performance of the proposed video signature algorithm on a web database, we have collected 45,899 web video clips for experimentation [2]. We generate a nine-seed signature for every video clip in our database. Since the composition of this large web database is likely to be quite different from the ground-truth set, we will characterize performance for a range of ϵ and ϵ_s. Ideally, T_{sig} and T_{seed} should be computed for every possible pair of videos in the database. This will result in $\binom{45899}{2}$, or more than one billion distance computations, requiring a significant amount of computational resources. Based on the results in Section 3, we observe that none of the similar videos in the ground-truth set have T_{seed} larger than 0.4. In order to complete the characterization on the large database in a reasonable amount of time, we initially set ϵ_s to be 0.4. It will be shown in Section 4.2 that this initial step of pruning does not adversely affect the accuracy of our measurements.

4.1. Retrieval Performance

Since the ground-truth is not available for our large web database, we measure precision for different values of ϵ using the following procedure : First, we compute T_{seed} for all pairs in our database, and prune away those pairs not satisfying $T_{seed} \leq 0.4$. We then identify, among all the remaining signature pairs, those with $T_{sig} \leq \epsilon$. Assume that there are S pairs of videos in this set. We then randomly examine 100 pairs out of this set, and determine the portion of videos which are subjectively similar. This will give us an estimate of precision, P, for a given value of ϵ. The number of subjectively similar videos in this set, R, can then be estimated as the product of S by P. The recall value for a given ϵ is R/G where G is the total number of subjectively similar videos in our database. Since G is not known, we cannot estimate the value of recall. Nonetheless, as G does not depend on ϵ, the recall performance for different values of ϵ can be characterized by comparing



(a) (b)

Fig. 2. *(a) Estimated precision versus estimated number of subjectively similar videos for three different schemes : meta-data, video signature, and combined; (b) Histogram of the number of videos as a function of the number of their similar copies using* $\epsilon = 1.9$.

their corresponding estimated R values. Thus, we can characterize the precision/recall performances of different schemes by considering their corresponding estimated values of P and R.

Figure 2(a) shows the estimated precision P against R for three different schemes: (a) meta-data alone as described in Section 2.1; (b) nine-seed video signature with each data point corresponding to a different value of ϵ, and (c) nine-seed video signature and meta-data together – the same set of ϵ as (b) is used, with an additional constraint that only those pairs which share at least one meta-data term are considered to be similar. The marginal error for all precision measurements is less than 5%. As seen, video signature technique significantly outperforms meta-data alone. For $R = 10{,}000$, the precision jumps from around 30% for meta-data to 98% for video signature. In spite of this, meta-data can be used in conjunction with video signature to further improve its precision/recall performance as shown in Figure 2.

The "signature" curve in Figure 2(a) shows that the precision drops from 98% to 43% as ϵ increases from 0.5 to 3.3. Subjective evaluation reveals that, the drop in precision is caused by erroneously classified pairs which are mostly black-and-white video clips, or animated graphics of simple colored lines or surfaces used in mathematical plots. These types of videos tend to produce very similar color histograms as they differ mainly in spatial configuration rather than in color distribution. Assuming that these types of videos must have at least 90% of their color bins empty, we find that around 21% of our database is occupied with videos of these types. We are currently investigating the use of other visual features to handle such videos.

In our subjective evaluations of video signature, R remains 35,000 pairs even when we increases ϵ beyond 3.3. This can be attributed to the fact that even though S increases as ϵ increases, lower precision values result in a more or less constant $S \cdot P = R$. This suggests that the number of subjectively similar videos with signature distances larger than 3.3 is negligible. Thus, we postulate that the number of subjectively similar videos pairs in the database is around 35,000. Since our database has 45,899 video clips, each video clip has, on average, $35000 \times 2/45899 \approx 1.53$.[2] In spite of the small average, we notice that individual video clips can have a much larger number of similar copies. Figure 2(b) shows the histogram of the number of videos, expressed in common logarithm, as a function of the number of their detected similar copies for $\epsilon = 1.9$. This value of ϵ is used because of its high precision at 93%, as shown in Figure 2(a). Even though the majority of

[2]This number is significantly smaller than our previous estimate of five in [2] because more than 53,000 pairs of similar surveillance videos, all from the same web-site, are removed from our measurements to avoid bias.

the videos have few similar copies, Figure 2(b) shows that some videos have as high as 67 similar copies. Manual inspection reveals that all the videos with 67 copies are identical, and consist of a few frames with a message indicating that the actual video requested is not available from the video server. We also find some popular web video clips among those with large number of similar copies : "Dancing Baby" shows an animation of a baby dancing, which was made famous by the television show "Ally McBeal"; "Bad day in office" shows a frustrated man breaking a computer, and "Miss America" is a talking-head sequence commonly used in the video compression community.

4.2. Complexity reduction measurements

In this section, we measure the performance of statistical pruning on the large web database. Since we do not have the actual recall values, we cannot directly compute complexity reduction as a function of recall as in Section 3. Instead, we will consider the impact of statistical pruning on "ϵ-similar" signature pairs, i.e. signature pairs with $T_{sig} \leq \epsilon$. We denote the total number of "ϵ-similar" signature pairs in the database as $L(\epsilon)$. When statistical pruning is applied, "ϵ-similar" signature pairs with $T_{seed} > \epsilon_s$ will be erroneously eliminated. Let $L_{prune}(\epsilon, \epsilon_s)$ denote the ratio of the number of such erroneously eliminated pairs to $L(\epsilon)$. Thus, to characterize the performance of statistical pruning for different values of ϵ_s, we will show the trade-off between complexity reduction and $L_{prune}(\epsilon, \epsilon_s)$.

The complexity reduction is measured by $C_{prune}(\epsilon_s)$ as defined in Equation 2. In this experiment, ρ from Equation 2 is calculated by dividing the number of signature pairs with $T_{seed} \leq \epsilon_s$ by the total number of signature pairs, i.e. $\binom{45899}{2}$. Figure 3(a) shows C_{prune} as a function of ϵ_s for the web database for $0 < \epsilon_s \leq 0.4$. The complexity at $\epsilon_s = 0.4$ is 19% of the total computations required to perform T_{sig} for all possible pairs, i.e. without pruning.

To compute $L_{prune}(\epsilon, \epsilon_s)$, we need to know both the number of "ϵ-similar" signature pairs erroneously pruned at ϵ_s and the total number of "ϵ-similar" signature pairs, $L(\epsilon)$. The former number is simply the difference between $L(\epsilon)$ and the number of similar signature pairs remained after pruning, which can be measured precisely. Unfortunately, $L(\epsilon)$ is not known because of statistical pruning. Nonetheless, we will show that, for the range of ϵ values considered in Section 4.1, it is reasonable to estimate $L(\epsilon)$ by extrapolation. Figure 3(b) shows the measured histograms of the "ϵ-similar" signature pairs as a function of ϵ_s for $\epsilon = 2.3$, 2.9 and 3.3. The histograms for $\epsilon = 2.3$ and 2.9 both taper to zero before T_{seed} reaches 0.4. Thus, it is reasonable to assume that $L(2.3)$ and $L(2.9)$ can be well approximated by summing up all the frequency values underneath their respective histograms. As for $\epsilon = 3.3$, we extrapolate the tail of the histogram using a third-order polynomial as shown in Figure 3(b). $L(3.3)$ can then be estimated by summing up the measured and extrapolated histogram values. This results in an estimate of around 233,000 pairs, in which only 2% comes from the extrapolated portion. With an estimate of $L(\epsilon)$ for each value of ϵ, $L_{prune}(\epsilon, \epsilon_s)$ can be computed in a straightforward way.

Figure 3(c) shows $C_{prune}(\epsilon_s)$ as a function of $L_{prune}(\epsilon, \epsilon_s)$ for ϵ = 2.3, 2.9 and 3.3. Every point on each curve is computed by measuring C_{prune} and L_{prune} for a particular value of ϵ_s. All three curves show a rapid drop in C_{prune} if a certain amount of inaccuracy can be tolerated. For example, for $\epsilon = 2.3$, C_{prune} drops from 18% to 2% if one allows up to 5% of all similar signature pairs to be pruned off. Another observation is that for the same accuracy level, a smaller ϵ can translate to much larger savings in complexity. For example, if we consider the case when 5% of similar signature pairs are pruned, $\epsilon = 2.3$, 2.9 and 3.3 result in C_{prune} values of 2%, 6% and 11% respectively. Since precision is rather sensitive to the choice of ϵ as shown in Figure 2(a), for applications that demand high precision, it is beneficial to choose ϵ appropriately to reduce complexity.

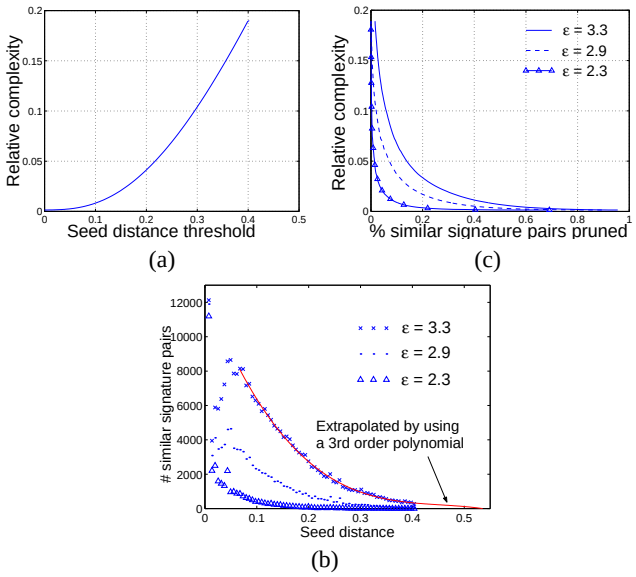


Fig. 3. *(a) Relatively complexity versus seed distance threshold; (b) Histograms of similar signature pairs versus seed distance, and (c) relative complexity versus percentage of similar signature pairs pruned for $\epsilon = 2.3$, 2.9 and 3.3.*

5. CONCLUSIONS

In this paper, we considered two methods for detecting similar videos on the web – meta-data and video signature. Meta-data scheme is based on matching the textual and hyperlink information associated with each video clip. Video signature extracts a small set of signature frames from each video clip for similarity detection. The complexity of cross comparing video signature can be largely reduced by using a statistical pruning technique. Nine signature frames are sufficient to achieve 90% recall and 95% precision on a small ground-truth set using 8% of the total number of operations required for full cross comparisons. When tested on a large database of around 46,000 video clips, video signature significantly outperforms meta-data in terms of precision and recall. Slightly better performance can be achieved by combining video signature with meta-data. The experimental results also show that statistical pruning can offer substantial computational savings if the application can tolerate a small degradation in recall. Using video signature, we estimate each video clip in our database to have, on average, 1.53 similar copies.

6. REFERENCES

[1] A.P. Berman, L.G. Shapiro, "A flexible image database system for content-based retrieval," *Computer Vision and Image Understanding*, vol. 75, no. 1/2, pp. 175–179, 1999.

[2] S.-C. Cheung, A. Zakhor, "Estimation of web video multiplicity," *Proc. SPIE – Internet Imaging*, vol. 3964, pp. 34–6, 2000.

[3] S.-C. Cheung "Efficient video similarity measurement and search," *Ph.D. Thesis in preparation*, 2000.

[4] N. Shivakumar, H. Garcia-Molina, "Finding near-replicas of documents on the web," *WebDB'98*, pp. 204-12.