RATE-COVERAGE ANALYSIS AND OPTIMIZATION FOR JOINT AUDIO-VIDEO MULTIMEDIA RETRIEVAL

Guanghan Ning[⋆†], *Zhi Zhang*[⋆†], *Xiaobo Ren*[†], *Haohong Wang*[†], *and Zhihai He*[⋆]

[⋆] University of Missouri, Columbia, MO 65203, USA
[†]TCL Research America, San Jose, CA, 95134, USA

ABSTRACT

In this work, we consider the problem of automatic content retrieval (ACR) using joint audio-video fingerprints. We focus on how to balance the query accuracy and the size of fingerprint, and how to allocate the fingerprint bits to video and audio frames to maximize the query accuracy. By introducing a novel concept called *coverage*, which is highly correlated to the query accuracy, we are able to construct a rate-coverage model and formulate the joint audio-video fingerprint bit rate allocation into a dynamic programming optimization problem. Our experimental results demonstrate that, compared to existing approaches, our method improves the retrieval accuracy by up to 25% while using 60% of the original fingerprint bit rate.

Index Terms— automated content retrieval, content-based multimedia retrieval, rate-coverage optimization.

1. INTRODUCTION

With the explovisve growth of mobile devices and Internet services, enormous video contents are produced and uploaded onto the Internet by users everyday. To effectively manage the rapidly growing multimedia data, a number of methods have been proposed for multimedia content analysis and retrieval, such as content-based image retrieval [1, 2], audio retrieval [3], and video retrieval [4]. In content-based multimedia retrieval [5], video and audio data are often represented by feature vectors, such as SIFT [6], SURF [7] and GLOH [8]. Compared to the raw video data, the feature description is much smaller in size and much more efficient for storage and retrieval. For example, the feature description of a typical video is about 10% of the original video size. With the massive amount of videos to be processed using feature description, the amount of features generated is still enormous. To address this issue, a number of methods have been developed to further compress the feature description and cut down the database overhead.

These methods can be classified into two major categories: *manifold learning* and *descriptor compression*. The manifold learning approach explores correlation among data and clusters similar features [9] to reduce redundancy. It is often based on graph-based ranking methods. Due to its ability to capture the geometric structure of the image set, it has been successfully used for image retrieval [10, 11]. The second approach of descriptor compression [12, 13] aims at generating compact descriptors individually, therefore reducing the overall demand for storage space. Originally introduced in [12], descriptors can be compressed by local descriptor compression [13], and global descriptor aggregation [14].

The two types of approaches are complementary to each other and can be used jointly to minimize the fingerprint size of multimedia databases. We observe that, in existing content-based multimedia retrieval, most researches focus on data correlation of one single modality, such as image clustering [15, 16] and audio clustering [17]. The cross-modal correlation between images and audios has not been adequately studied.

In this work, we consider the design of a system that optimizes the compression of joint video-audio descriptors. By introducing a novel concept called *coverage*, which is highly correlated to the query accuracy, we are able to develop a rate-coverage model. Given an arbitrary budget of storage space, this model aims to optimize the retrieval accuracy while preserving only a subset of the overall joint video-audio descriptors, therefore reducing the overall audio-video fingerprint size. We propose a dynamic programming method to solve this rate-coverage optimization problem.

The rest of the paper is organized as follows. Section 2 presents the proposed rate-coverage optimization framework. The dynamic programming solution to the rate-coverage optimization problem is developed in Section 3. Section 4 summarizes the proposed algorithm. Experimental results are presented in Section 5. Section 6 concludes the paper.

2. RATE-COVERAGE OPTIMIZATION

As is shown in Fig. 1, the end-user device collects video and audio frames and then send the fingerprints of these frames to the query engine that is connected to a cloud-based ACR server. On the server side, the video and audio fingerprints are generated. Without loss of generality, we consider the example where the device, e.g., a smart TV, receives and decodes

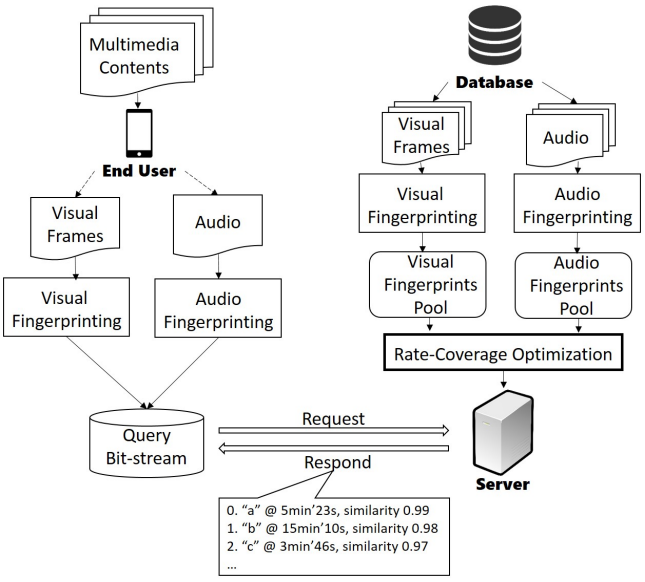

Fig. 1. System Overview. Left: query route. Right: database and server with rate-coverage optimization.

streams from the Internet upon users requests. Using a mobile phone as the remote controller, the user takes a snapshot recorded by the smart TV that displays something of interest. The mobile phone sends the joint query fingerprints to the online audio-video database.

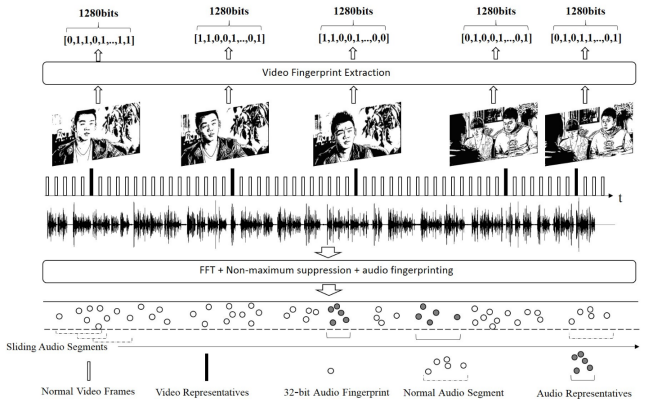

Fig. 2. Fingerprinting process overview.

2.1. Audio-Video Fingerprints

As illustrated in Fig. 2, the video fingerprints are extracted into groups of 1280-bits data stream while the audio fingerprints are extracted based on a key-value pair by exploring the spatial relationship among spectrogram maximums, which are computed by FFT, followed by non-maximum suppression and Shazams [18] audio fingerprinting method. Specifically, visual fingerprints are computed based on image pixel intensity correlation. An image frame is first scaled to a K by K square image. A selection of N out of $\binom{2}{N^2}$ block pairs are computed to generate an N-dimensional bit array using binary comparison on pixel intensity for each image. The output visual fingerprints are binary strings, which are very robust to small distortion during video compression and transmission. Audio fingerprints are extracted on spectrograms, each fingerprint unit is a combined hash key-value pair of two local maximum points on spectrogram after non-maximum suppression. For each pair, the hash key is a concatenation of F_1, F_2 and Δ_t, which are the frequencies of these two points and their time domain difference. We use timestamps and title string as the value. Similar fingerprints are mapped to the same linked list by hashing. All contents can be retrieved from database by the linked list [18].

2.2. Problem Formulation

As we know, in videos, there is a strong correlation between neighboring frames, resulting in strong correlation between their fingerprints. Therefore, to save the storage space and achieve compression, we only choose a subset of fingerprints to represent the original video frames and audio segments and store them on the server database. We refer to these fingerprints as visual and audio *representatives*. As illustrated in Fig. 3(a) and 3(b), each small dot represents the feature of a keyframe in a high dimensional feature space. The large circles are the feature clusters. We choose the centroid fingerprint of each cluster as the representative. We consider the query result to be *satisfactory* if and only if the correct frame is included in the returned cluster. Here, K is the number of results that can be returned upon each query. Each representative has a set of K-nearest neighbors in the feature space.

We define *Coverage*, denoted here by C, to be the total amount of fingerprints covered by all the representatives.

As discussed in Section 2.1, the fingerprints are fixed-size feature vectors for video frames, therefore each video unit (e.g., a video frame or a group of video frames) has a constant number of bits in its fingerprint. We denote this number by B_V. We observe that the audio fingerprints have variable bit rates because of uneven maximums on the spectrogram. To address this issue, we perform non-equal partition on the audio stream so that each audio segment has same bit rate of its fingerprint. We denote this unit bit rate by B_A. The overall data rate R, which is the total amount of fingerprint data stored in the database to provide the query service, including both audio and video fingerprints, is computed as:

$$R = B_V \times N_V + B_A \times N_A. \quad (1)$$

Here, N_V and N_A are the total numbers of video and audio representatives.

For an ACR system with joint representation of video and audio fingerprints, we need to study the trade-off between the overall fingerprint size stored in the database and the retrieval

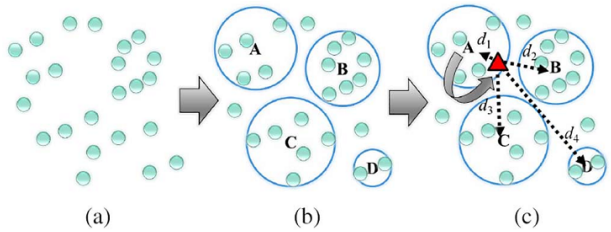

Fig. 3. Example of fingerprint representatives. (a) Features of keyframes. (b) Fingerprint representatives. (c) Query according to the distances to cluster centroids.

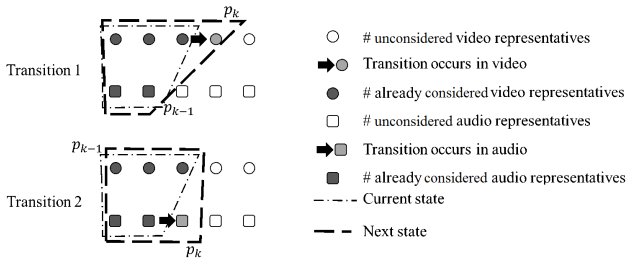

Fig. 4. Example of transitions from previous state to current state in dynamic programming.

accuracy. Our goal is to maximize the average query accuracy $E(A)$ under the bit rate constraint R_T.

$$\max\ E(A) = \frac{1}{T}\sum_{t=1}^{T} A_i,\ s.t.\ R \le R_T. \tag{2}$$

where $A_t = 1$ if satisfactory result is retrieved at frame t, otherwise $A_t = 0$. *Coverage* is introduced because *Accuracy* is unknown until actual query is performed. As will be discussed in Section 5, *Coverage* and *Accuracy* are highly correlated. (2) is therefore converted to:

$$\max\ C = \alpha C_{V,N_V} + (1-\alpha) C_{A,N_A},\ s.t.\ R \le R_T. \tag{3}$$

Here, C is the total coverage of audio and video representatives and $\alpha \in [0,1]$ represents a balance between the coverage of video (C_V) and audio (C_A) representatives. For a given number of video representatives N_V, we can use the so-called Disk Covering method developed in [19] to find the maximum coverage ratio, denoted by $f_V(N_V)$. Similary, we can find the maximum coverage ratio for the audio representatives, denoted by $f_A(N_A)$. Thus, (3) can be rewritten as:

$$\begin{aligned} &\max_{N_V,N_A} (\alpha f_V(N_V) + (1-\alpha) f_A(N_A)) \\ &s.t.\ B_V \times N_V + B_A \times N_A \le R_T \end{aligned} \tag{4}$$

3. A DYNAMIC PROGRAMMING SOLUTION

In the following text, we explain how to solve the problem with dynamic programming, and why the proposed approach is globally optimal. We derive a solution to problem (4) using the Lagrange multiplier method to relax the bitrate constraint, so that the relaxed problem can be solved using a shortest path algorithm. We first denote the Lagrangian cost function

$$J_\lambda(N_V, N_A) = (\alpha f_V(N_V) + (1-\alpha) f_A(N_A)) + \lambda(B_V \times N_V + B_A \times N_A), \tag{5}$$

where λ is called the Lagrange multiplier. It has been proven that if there is a λ^* such that

$$\{N_V^*, N_A^*\} = \underset{N_V,N_A}{\operatorname{argmax}} J_\lambda^*(N_V, N_A), \tag{6}$$

and which leads to $R = R_T$, then $\{N_V^*, N_A^*\}$ is an optimal solution to (4) . Therefore, if we can find the optimal solution to max $(J_\lambda(N_V, N_A))$, then we can find the optimal λ^* and approximation to the constrained problem of (4) . As indicated in Fig. 4, we use a two dimensional DAG shortest Path algorithm for the optimization process, that is, in order to compute the maximal J, each state will need the status of N_V and N_A. We define a node tuple (i, j) indicating state (N_V, N_A) in Shortest Path space, denoted as p_k, which has two paths from previous state p_{k-1}. It means that at this state the database stores at most i number of video fingerprints and at most j number of audio fingerprints as representatives. At the termination state, we derive the optimized solution for video and audio bitrate allocation, with at most $\overline{N_V}$ and $\overline{N_A}$ number of fingerprints respectively for video and audio.

To solve the optimization problem in (4) , we create a cost function $T(p_k)$, which represents the cost to include state (i, j) in the state space:

$$T(p_k) = \max\{\ \alpha f_V(i) + (1-\alpha) f_A(j) + \lambda(B_V \times i + B_A \times j)\ \} \tag{7}$$

The sub-problem f_V and f_A are the optimization problems to maximize coverage of video and audio, given N_V and N_A, respectively. The observation is that the delta cost:

$$\Delta(p_{k-1}, p_k) = \begin{cases} \alpha[f_V(i) - f_V(i-1)] + \lambda B_V, & if\ video \\ (1-\alpha)[f_A(j) - f_A(j-1)] + \lambda B_A, & if\ audio \\ 0, & if\ none \end{cases} \tag{8}$$

is independent of the selection of the previous states $p_0, p_1, ..., p_{k-2}$. Therefore, cost function

$$T(p_k) = \max(T(p_{k-1}) + \Delta(p_{k-1} + p_k)) \tag{9}$$

can be solved by a DP algorithm.

4. SUMMARY OF ALGORITHM

The proposed algorithm is summarized in Algorithm 1. We can see that the proposed algorithm has very low computational complexity.

5. EXPERIMENTAL RESULTS

We have conducted experiments on various media patterns to compute the coverage and the corresponding expected retrieval accuracy given certain bit-rate budget. By changing

Algorithm 1 Our proposed algorithm

procedure RECURSIVE DYNAMIC PROGRAMMING
  Let $F_{i,j}(b)$ be the maximum coverage we can achieve given at most i number of video and j number of audio fingerprints, with a limitation of overall bit-rate to be b
  Let $I_{i,j}(b)$, $J_{i,j}(b)$ mark the corresponding number of video and audio fingerprints that are selected as representatives while achieving coverage $F_{i,j}(y)$
  for $i \leq N_V, j \leq N_A$ **do**
    $Fs = [F_{i-1,j}(b),\ F_{i,j-1}(b),\ F_{i-1,j}(b-Bv)+fv(I_{i-1,j}(b)+1) - fv(I_{i-1,j}(b)),\ F_{i,j-1}(b-Ba) + fa(I_{i,j-1}(b)+1) - fa(I_{i,j-1}(b))]$
    $Is = [I_{i-1,j}(b),\ I_{i,j-1}(b),\ I_{i-1,j}(b-Bv)+1,\ I_{i,j-1}(b-Ba)+0]$
    $Js = [J_{i-1,j}(b),\ J_{i,j-1}(b),\ J_{i-1,j}(b-Bv)+0,\ J_{i,j-1}(b-Ba)+1]$
    $index = argmax(Fs)$
    $F_{i,j}(b) = Fs(index)$
    $I_{i,j}(b) = Is(index)$
    $J_{i,j}(b) = Js(index)$
  return $\{F_{i,j}(b), I_{i,j}(b), J_{i,j}(b)\}$

allocation bitrates on video and audio, we compute the coverage and the actual expected retrieval accuracy, respectively. One example is shown in Fig. 5, where the coverage-accuracy curve is monotone and the gap between coverage and accuracy is very small. As coverage and accuracy are highly correlated, we can optimize the bit-rate allocation between audio and video to achieve optimum coverage in order to improve retrieval accuracy.

For performance comparison, we consider the following baseline methods. Note that these are not globally optimum solutions. (1) Arbitrary allocation. This method does not consider the difference and correlation between audio and video data and allocate bit-rates evenly between audio and video. (2) The audio first method preferentially picks audio, based on the fact that fingerprint of a video frame is of 200 bytes, greater than the 32 bytes of an audio segment. (3) Greedy allocation (sub-optimal). Given a bit-rate budget, the greedy approach seeks the maximum current leveraged gain G at each stage, by choosing a fingerprint type from either video or audio, until bitrate exceeds our budget. λ is introduced as the penalty regularization parameter.

$$G = \max(\ \alpha[f_V(i) - f_V(i-1)] + \lambda B_V, (1-\alpha)[f_A(i) - f_A(i-1)] + \lambda B_A\ ) \quad (10)$$

Table 1. Average Bit-rate save of the proposed method for same coverage levels, compared with reference methods.

Coverage	70%	75%	80%	85%	90%	95%
Sub-optimal(Greedy)	-9%	-8.5%	-8.1%	-8%	-7%	-5.5%
Audio First	-26%	-25%	-23%	-23%	-21%	-20.5%
Arbitrary	-32.5%	-31%	-28%	-26.5%	-24%	-21%

The effectiveness of our proposed method is tested utilizing commercially available movies. As is shown in Fig. 6, with a very limited bit-rate budget, which is around 60% of overall bit-rate of the fingerprints, we can achieve over 85% coverage with the proposed optimal method, with a bit-rate reduction of around 25% compared with the arbitrary method. As a matter of fact, with only 22% of overall bit-rates as budget, we achieve 60% of the overall coverage, saving 37.5% bit-rate compared to the arbitrary method. As is shown in Fig. 6, the proposed optimum algorithm outperforms other methods. TABLE 1 shows the average bit-rate reductions. At 95% coverage, up to 21% bit-rates can be saved.

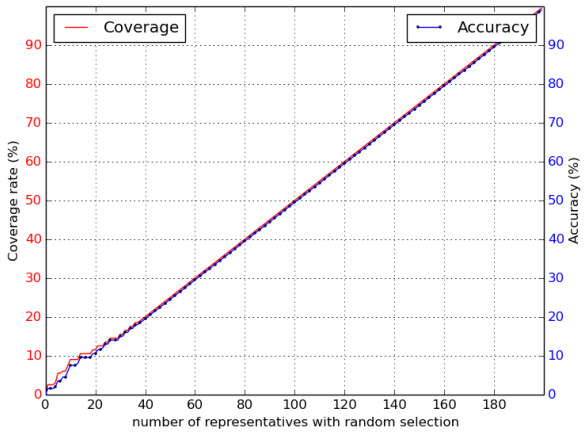

Fig. 5. The coverage-accuracy relationship.

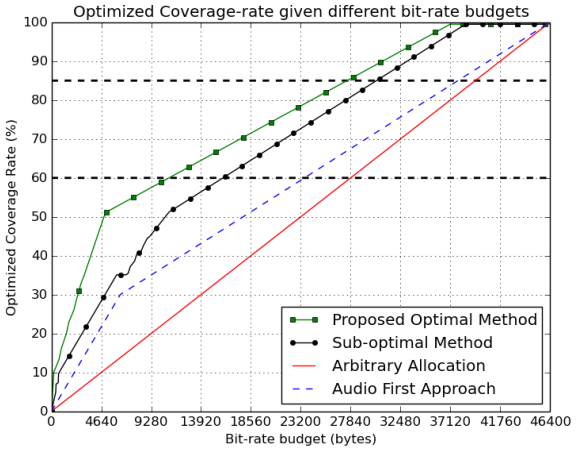

Fig. 6. Rate-Coverage curves. The comparison of different approaches.

6. CONCLUSION

In this paper, we proposed a novel ACR method using joint audio-video fingerprint for media retrieval. We introduced a novel concept called coverage to represent retrieval accuracy and developed a rate-coverage model. We developed a dynamic programming method to optimize coverage for given bitrate budgets in order to maximize accuracy. Experimental results demonstrated that our method improves retrieval accuracy significantly compared to reference methods, and significantly saves bit-rate with same level of retrieval accuracy.

7. REFERENCES

[1] X. He, W.-Y. Ma, and H.-J. Zhang, "Learning an image manifold for retrieval," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 17–23.

[2] R. Zhang and Z. Zhang, "Effective image retrieval based on hidden concept discovery in image database," *Image Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 562–572, 2007.

[3] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao, "Content-based music structure analysis with applications to music semantics understanding," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 112–119.

[4] J. Fan, A. K. Elmagarmid, X. Zhu, W. G. Aref, and L. Wu, "Classview: hierarchical video shot classification, indexing, and accessing," *Multimedia, IEEE Transactions on*, vol. 6, no. 1, pp. 70–86, 2004.

[5] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2, no. 1, pp. 1–19, 2006.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer vision–ECCV 2006*. Springer, 2006, pp. 404–417.

[8] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, 2005.

[9] U. Fayyad and J. Shanmugasundaram, "Multi-dimensional database record compression utilizing optimized cluster models," Oct. 14 2003, uS Patent 6,633,882.

[10] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 5500, pp. 2268–2269, 2000.

[11] Y. Wang, M. A. Cheema, X. Lin, and Q. Zhang, "Multi-manifold ranking: Using multiple features for better image retrieval," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2013, pp. 449–460.

[12] L.-Y. Duan, J. Lin, J. Chen, T. Huang, and W. Gao, "Compact descriptors for visual search," *MultiMedia, IEEE*, vol. 21, no. 3, pp. 30–40, 2014.

[13] M. Johnson, "Generalized descriptor compression for storage and matching." in *BMVC*, 2010, pp. 1–11.

[14] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 9, pp. 1704–1716, 2012.

[15] G. J. McLachlan and K. E. Basford, "Mixture models. inference and applications to clustering," *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, vol. 1, 1988.

[16] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[17] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 209–215, 2003.

[18] A. Wang, "The shazam music recognition service," *Communications of the ACM*, vol. 49, no. 8, pp. 44–48, 2006.

[19] B. Xiao, J. Cao, Q. Zhuge, Y. He, and E. H. Sha, "Approximation algorithms design for disk partial covering problem," in *Parallel Architectures, Algorithms and Networks, 2004. Proceedings. 7th International Symposium on*. IEEE, 2004, pp. 104–109.