Accelerating Near-Duplicate Video Matching by Combining Visual Similarity and Alignment Distortion

Hung-Khoon Tan, Xiao Wu, Chong-Wah Ngo, and Wan-Lei Zhao
Department of Computer Science
City University Of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong
{hktan, wuxiao, cwngo, wzhao2}@cs.cityu.edu.hk

ABSTRACT

In this paper, we investigate a novel approach to accelerate the matching of two video clips by exploiting the temporal coherence property inherent in the keyframe sequence of a video. Motivated by the fact that keyframe correspondences between near-duplicate videos typically follow certain spatial arrangements, such property could be employed to guide the alignment of two keyframe sequences. We set the alignment problem as an integer quadratic programming problem, where the cost function takes into account both the visual similarity of the corresponding keyframes as well as the alignment distortion among the set of correspondences. The set of keyframe-pairs found by our algorithm provides our proposal on the list of candidate keyframe-pairs for near-duplicate detection using local interest points. This eliminates the need for exhaustive keyframe-pair comparisons, which significantly accelerates the matching speed. Experiments on a dataset of 12,790 web videos demonstrate that the proposed method maintains a similar near-duplicate video retrieval performance as the hierarchical method proposed in [12] but with a significantly reduced number of keyframe-pair comparisons.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Performance, Experimentation

1. INTRODUCTION

Fueled by the popularity of social media, the volume of professional and user generated videos is growing exponentially. Among them, there exist a large number of near-duplicate videos and video copies in video sharing websites. The redundancy can be as high as 93% for certain queries as shown in [12]. Therefore, it becomes critical to be able to match videos in an effective and efficient manner. As an effective form, a video clip is usually represented by a sequence of keyframes extracted through shot boundary detection or time sampling. To measure the similarity of two videos, the keyframes are usually compared in a pairwise manner and the degree of overlay of near-duplicate keyframes determines the similarity [11]. Local interest point has been repeatedly shown to deliver excellent performance for near-duplicate keyframe detection and retrieval [5, 6, 14, 13]. However, due to the large number of local points in each keyframe, exhaustive keyframe-pair comparisons become the stumbling block for matching videos in web scale applications such as video retrieval and video copy detection.

Video similarity measure and copy detection have been extensively studied previously. Various approaches using different features and matching solutions have been proposed. In general, accuracy comes at the cost of time complexity. Signature-based methods, e.g. [3], can achieve rapid detection but its effectiveness is limited to detecting almost identical or superficially edited videos. Shot-level similarity [8, 1] is slower but capable of handling matchings of videos with a substantial degree of editing. Video duplicates with changes in background, color and lighting, content modification and editing require an even higher degree of details at region level precision. Recently, local points have received a lot of attentions [5, 7, 14] because of its capability to handle images with complex variations. Unfortunately, in practice, its usefulness is severely undermined by scalability issues. In this respect, our approach provides a platform to engage local point evaluation sparingly by deriving only the list of crucial keyframe-pairs that can decisively determine the similarity between two videos.

Enlightened by the observation that corresponding keyframes of two near-duplicate videos usually exhibit regular alignment patterns, the alignment distortion is novelly proposed to measure the temporal arrangement of the keyframes. The idea of distortion was previously used in the shape matching problem [2]. In this paper, the quality of correspondences between keyframes is jointly measured by two factors: *visual similarity* and *alignment distortion*. Visual similarity considers the visual quality in terms of visual keywords, while the alignment distortion measures the temporal arrangement of the similar keyframes. The correspondence problem is then cast as an integer quadratic programming problem. As such, dissimilar keyframe pairs are effectively filtered out by the correspondence matching and near-duplicate keyframe detection need to be performed only within the matched pairs.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

Figure 1: Alignment of keyframe sequences in two near-duplicate web videos.

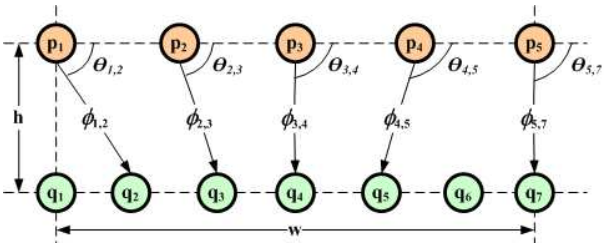

Figure 2: Alignment of keyframe sequences in two video. The distortion cost is computed based on the differences in angle among all the correspondences.

2. VIDEO MATCHING

2.1 Combining Visual Similarity and Alignment Distortion

Given two videos, the proper matching of the two keyframe sequences should not only exhibit high visual similarity but also display temporal coherence. In principle, a good alignment of two related keyframes is highly regularized and should follow certain spatial arrangements. For instance, the correspondences among the keyframes are often formed by parallel, zoom-like lines or a confluence of both. Figure 1 shows an example of such kind of matching pattern, which obeys the temporal constraint. Other general point set matching techniques such as the maximum weighted bipartite graph (MWBG) [9] typically ignore the temporal property of video matching and look for the correspondence set that obtains the highest feature-to-feature score. Therefore, it is possible to get a high similarity score even for two dissimilar videos. In contrast, the temporal constraint forces the keyframes to match in an orderly fashion and therefore ensures a low similarity score between two unrelated videos while maintaining a high similarity score between related pairs.

Given two videos $P = \{p_1, \ldots, p_N\}$ and $Q = \{q_1, \ldots, q_M\}$ with N and M keyframes respectively, the objective is to find the best set of correspondences $\hat{\phi}$ from the set of all possible correspondences Φ of all keyframes in a video. The quality of the correspondence $\phi_{i,j} \in \Phi$ that matches the keyframe p_i to q_j is determined based on two factors: (a) the feature cost $v(\phi_{i,j})$ which measures the degree of visual dissimilarity between the two keyframes and (b) the alignment distortion $d(\phi_{i,j})$ which enforces the temporal constraint on the keyframe correspondences where the notion of distortion can only be properly defined with respect to a set of correspondences. Because our focus is not to find the optimal feature and distance function, we simply use visual keywords as our feature and cosine distance as our distance function since visual keywords have been shown to be a robust and reliable representation of an image for video retrieval [4].

2.2 Alignment Distortion Construction

The alignment distortion is constructed by aligning two keyframe sequences vertically, and quantizing the angles formed by the matching correspondences and the horizontal axis. The main idea is shown in Figure 2. The keyframe sequences of the two videos are projected as points in an image space following the arrangement, where the keyframes are aligned vertically with a gap of h, and the keyframes within the same video are spaced evenly across a width of w. Considering all possible pair-wise links between the two keyframe sets, temporal coherence can be enforced by finding the set of correspondences that not only maximizes the feature cost but in addition ensures that the geometric distortion between the set of correspondences are minimized.

We use the difference in angles to quantify the distortion between the correspondences. The transition in angles between neighboring keyframes within the same video should not be abrupt and therefore the correspondences originating from two adjacent keyframes are subjected to a heavier distortion penalty compared to those which are further apart. Such restriction is useful for scenarios when keyframe mapping are localized and irregularly spaced.

Given a candidate solution ϕ^*, the distortion cost $d(\phi_{i,j})$ for the correspondence $\phi_{i,j} \in \phi^*$ can thus be formulated as follows

$$d(\phi_{i,j}) = \sum_{\phi_{m,n} \in \phi^*, \phi_{m,n} \neq \phi_{i,j}} k(x_i, x_m) \cdot (\theta(\phi_{i,j}) - \theta(\phi_{m,n})) \tag{1}$$

where

$$k(x_i, x_m) = e^{-(x_i - x_m)^2 / 2\sigma^2} \tag{2}$$

$\theta(\phi_{i,j})$ is the angle between $\phi_{i,j}$ and the horizontal axis in the range of 0° to 180° while x_i is the x-coordinate of the keyframe p_i. The kernel k enforces the localization of the distortion cost where σ allows us to control the scope on the temporal constraint.

2.3 Alignment Optimization Algorithm

Finding the best alignment that balances between the feature and distortion cost can thus be formulated as the minimization of an integer quadratic programming (IQP) problem as follows:

$$\min_{\phi} \text{cost}(z) = w_d \sum_{\phi_{i,j} \in \phi} d(\phi_{i,j}) z_{i,j} + w_v v(\phi_{i,j}) z_{i,j} \tag{3}$$

where w_d and w_v weighs the feature and distortion cost terms, respectively. z is a binary indicator vector such that when $z_{i,j} \in z = 1$, this indicates that keyframe p_i maps to p_j, and therefore the correspondence $\phi_{i,j}$ is selected. A set of constraints $\sum_j z_{i,j} = 1$, $\forall i$ is imposed on the correspondences so that the keyframe p_i in P can only map to only one keyframe in video Q. In order to allow a certain degree of tolerance to outliers, a dummy node is inserted q_{null} where a limit on the number of permissible outliers $\sum_i z_{i,null} \leq k$ where no feature and distortion cost is incurred. k is the number of outliers allowed. The use of outlier allows us to perform partial alignment. This is particularly useful to overcome missing keyframes when different keyframe extraction schemes are used for the two videos.

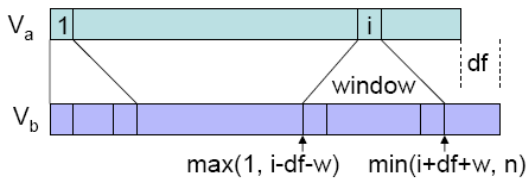

Figure 3: Matching window for keyframes between two videos

The final optimization problem can then be expressed in matrix form as follows:

$$\begin{aligned} &\min \text{cost}(z) = w_d z^T D z + w_v c^T z \quad \text{s.t.} \\ &Az = b, \quad x \in \{0,1\}^n \end{aligned} \tag{4}$$

where $n = |Q| \times |P|$ is the total number of possible correspondences, $|\cdot|$ denotes the number of keyframes in a video, D is the $n \times n$ symmetric distortion matrix. A is a $(|P|+1) \times |Q|$ matrix and together with b, the corresponding constraint vector of size $|P|+1$, they outline all the set of constraints that ensure a well-behaved correspondence set.

Equation 4 is a NP-hard problem and we resort to an approximation scheme similar to [2]. First, the IQP problem is converted into a series of $n+1$ linear programming problems subjected to the same set of constraints as in Equation 4 to find an initial solution that produces the lower bound of the objective function. The distortion cost for each correspondence is minimized separately where $\hat{d}(\phi_{i,j}) = \min d(\phi_{i,j}), \ \forall \phi_{i,j} \in \Phi$. Then, the individual distortion cost is combined with its corresponding feature cost to form the linear equation $L(z) = \min_z \sum_{\phi^*(z)} (\hat{d}(\phi_{i,j}) + c(\phi_{i,j})) z_{i,j}$ where $\phi^*(z)$ refers to the the set of correspondences $\{\phi_{i,j}\}$ which are selected by the indicator vector z. This gives us the lower bound for cost(z). The complexity of each linear optimization is $O(n^2)$. Finally, gradient descent is employed to find a local optimum solution where only one element is allowed to change to limited locations per-step. The time complexity to perform gradient descent is $O(n)$. In our experiments, the approximation technique is able to discover an accurate alignment.

3. ACCELERATING KEYFRAME ALIGNMENT

The proposed keyframe alignment algorithm is not efficient considering that computational time grows exponentially with the number of keyframes in the videos. As the number of keyframes grows, the saving gained from reduced near-duplicate matchings may risk being negated by the time required to perform keyframe alignment. This section proposes two strategies to avoid the aforementioned problem. First, a windowing policy is adopted and in addition alignment is performed in fixed-size fragments to ensure that the alignment algorithm is practically efficient with minor or no impact on performance.

3.1 Windowing Policy

A common technique to ensure practicality of video matching for large-scale video copy detection is the use a *windowing* policy [12] as shown in Figure 3. To reduce computation, the keyframe i is compared to keyframes in another video within a certain sliding window of $[\max(1, i - df - w), \min(i + df + w, n)]$ where n is the length of the second video, *df* is the length difference between the two videos and w is the window size used. Similarly, the proposed distortion cost model could take up the constraint resulting in a significantly trimmed correspondence set which considers only those within the specified windows. This reduces computation from a polynomial time complexity of $O((|Q| \times |P|)^2)$ to $O((|W| \times |P|)^2)$ where P is assumed to be the smaller video, $|W| \approx 2(w + idf) + 1$ and $|W| \leq |Q|$.

3.2 Fragmented Alignment

However, a windowing policy is effective only when comparing two videos with relatively similar length. In the case when *df* is huge, the computational savings becomes insignificant. Another useful strategy is to perform alignment in fragments. In this strategy, the video P is broken into f sequential fragments $P = P_1 \cup ... \cup P_f$ and alignment is performed between each P_i and Q_i separately. Coupled with the windowing scheme, fragmented alignment further reduces the computation to a complexity of $O(|W| \times |P_i| \times f)^2$ where the size of the fragment $|P_i|$ is a constant. In order to preserve the performance of the alignment, the value of $|P_i|$ should not be set too small, preferably larger than five so that certain degree of temporal information is retained to guide the matching of each fragment. In addition, to avoid one-to-many mapping where the correspondences for a particular fragment overlap with those from earlier ones, the distortion cost matrix D could be manipulated by assigning a prohibitively large value for the correspondences taken up by a previous fragment.

4. EXPERIMENTS

In this section, we demonstrate the effectiveness of the proposed alignment algorithm for near-duplicate web video retrieval (Section 4.1). More importantly, we show that the number of keyframe comparisons is significantly reduced (Section 4.2). We use the video dataset in [12] for experiments. The dataset consists of 12,790 web videos crawled from YouTube, Google and Yahoo, using 24 search queries [12]. The keyframes in each video are associated with a set of local points extracted by Hessian-Affine detector [10] and described by PCA-SIFT features [6]. To perform keyframe alignment, we build a visual dictionary of one thousand keywords by clustering the set of features using the K-means algorithm. Eventually, each keyframe is represented as a vector of visual keywords of 1,000 dimensions, and cosine distance is adopted to measure visual dissimilarity between video clips.

4.1 Near-Duplicate Retrieval

We compare our approach to HIRACH [12] which is a hierarchical algorithm with two main steps. The first step employs global signature generated from color histogram to rapidly identify the potential near-duplicate videos. The second step adopts the windowing policy presented in Section 3 to perform exhaustive local point matching for all pairs of candidate keyframes. We also use global signature (SIG_CH) as the baseline, which is possibly the simplest approach for measuring video similarity, to judge the degree of improvement that the proposed alignment can achieve. Note that the proposed approach, named as 'ALIGN' in tables 1 and 2, does not perform local point matching. Instead,

Table 1: Average precision of 24 queries over all recall levels

Approaches	SIG_CH	HIRACH [12]	ALIGN
Average Precision	0.892	0.952	0.951

video similarity is based upon the visual dictionary built using local points. Once the optimal correspondences have been constructed, the correspondences of keyframe pairs are further verified by local point matching.

Table 1 shows the retrieval performance of the three approaches in terms of average precison over 24 web video queries. We use 24 seed videos, one per query, provided by [12] as video examples for near-duplicate retrieval. Seed videos are the set of popular videos which are most viewed by users in the video sharing websites. The proposed method delivers a result comparable to HIRACH. This shows that the set of keyframe pairs identified by our alignment algorithm can be used as the basis to make near-duplicate decisions. Our experiment shows that comprehensive keyframe matchings are unnecessary to achieve a similar performance provided that a good alignment is found between the two keyframe sequences. Coupled with a robust similarity measure, the temporal consistency property is fully exploited by our algorithm to align the keyframes in a manner that can facilitate fast near-duplicate evaluation using local points. Although SIG_CH can rapidly perform near-duplicate detection, it has the lowest performance because global features can only handle near-duplicate videos with simple variations. It becomes incompetent for videos having major editing, transformation and insertion/removal of keyframes.

4.2 Computation Comparison

Our algorithm is not only competitive to HIRACH in terms of retrieval effectiveness, the matching speed is also considerably faster, considering the number of keyframes that needs to be compared when matching two videos. Table 2 lists the total number of keyframe comparisons required for our algorithm and HIRACH in order to complete the 24 search queries which involves a total of 12,790 videos and 398,015 keyframes. We also list the number of comparisons required for brute force algorithm, which evaluates all pairs of video keyframes as our baseline to judge the degree of improvement by both algorithms. As shown in Table 2, the brute force approach is extremely expensive. For web videos such as music videos, it is common that there are over 100 keyframes in a four minute video which can easily reach 10,000 number of comparisons when performed in a brute-force manner.

For the HIRACH with sliding window scheme [12], a keyframe is only compared to the keyframes of another video within the sliding window. However, it is effective only when the two videos has similar number of keyframes. In our approach, each keyframe is matched to at most one keyframe in another video. The near-duplicate keyframe detection only needs to be performed on these keyframe pairs. As a result, the number of comparisons for two videos is determined by the number of keyframes in two videos, whichever is smaller. Compared to HIRACH, our algorithm further significantly reduces the comparison. From Table 2, our algorithm achieves a speed up of 30 times compared to HIRACH.

Table 2: Total number of keyframe pair comparison for near-duplicate video retrieval over 24 queries

Approaches	# of comparison
ALIGN	238,769
HIRACH	6,471,693
Brute force	20,610,384

5. CONCLUSIONS

We have presented a novel approach to accelerate the matching of two video clips by aligning the two keyframe sequences. Integrating the feature cost with a distortion cost, the problem is formulated using integer quadratic programming to find the best correspondences. Most importantly, we have shown that the aligned keyframes can be used as the basis to make near-duplicate decisions. In our experiments, we achieve a similar near-duplicate retrieval performance as HIRACH with 30 times speedup.

6. ACKNOWLEDGEMENTS

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7002112).

7. REFERENCES

[1] D. A. Adjeroh, M. C. Lee, and I. King. A distance measure for video sequences. In *CVIU*, volume 75, pages 25-45, 1999.

[2] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion cost. In *CVPR*, 2005.

[3] A. Hampapur and R. Bolle. Comparison of sequence matching techniques for video copy detection. In *Conf. on Storage and Retrieval for Media Databases*, 2002.

[4] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, 2007.

[5] A. Joly, O. Buisson, and C. Frelicot. Content-based copy retrieval using distortion-based probabilistic similarity search. In *IEEE Trans. on Multimedia*, volume 9, pages 835-844, 2007.

[6] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Multimedia*, 2004.

[7] J. Law-To, B. Olivier, V. Gouet-Brunet, and B. Nozha. Robust voting algorithm based on labels of behavior for video copy detection. In *ACM Multimedia*, 2006.

[8] R. Lienhart and W. Effelsberg. A systematic method to compare and retrieve video sequences. In *Multimedia Tools Application*, volume 10, pages 47-72, 2000.

[9] J. A. McHugh. Algorithmic graph theory. Prentice Hall, 1990.

[10] K. Milolajczyk and C. Schhmid. Scale and affine invariant interest point detectors. volume 60, pages 63-86, 2004.

[11] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In *ACM Multimedia*, 2007.

[12] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *ACM Multimedia*, 2007.

[13] D.-Q. Zhang and S.-F. Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *ACM Multimedia*, 2004.

[14] W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu. Near-duplicate keyframe identification with interest point matching and pattern learning. In *IEEE Trans. on Multimedia*, volume 9, Aug 2007.