Expert Systems with Applications 39 (2012) 10278–10282

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A similarity measure between videos using alignment, graphical and speech features

D. Fuentes [a,*], R. Bardeli [b], J.A. Ortega [a], L. Gonzalez-Abril [c]

[a] *Computer Languages and Systems Dept., University of Seville, 41012 Seville, Spain*
[b] *Department Netmedia, Fraunhofer IAIS, Germany*
[c] *Applied Economics I Dept., University of Seville, 41018 Seville, Spain*

ARTICLE INFO

Keywords:
Information retrieval
Content segmentation
Bipartite matching

ABSTRACT

A novel video similarity measure is proposed by using visual features, alignment distances and speech transcripts. First, video files are represented by a sequence of segments each of which contains colour histograms, starting time, and a set of phonemes. After, textual, alignment and visual features are extracted of these segments. The following step, bipartite matching and statistical features are applied to find correspondences between segments. Finally, a similarity is calculated between videos. Experiments have been carried out and promising results have been obtained.

1. Introduction

The amount of digital content on the world wide web has grown enormously in the last decade due to the popularity of social media. The volume of professional or user video is increasing exponentially and a large number of video clips are generated and added every day. Therefore, a similarity measure is an essential tool in order to facilitate effective search, retrieval, browsing, or copy detection.

Shot and clip similarity have been extensively addressed for copy detection, retrieval and clustering using image similarity measures with low-level global features (Foote, 1998; Peng & Ngo, 2004). In Chiu, Wang, and Chen (2010), spatial and temporal aspects are considered to expedite both min-hashing indexing and spatio-temporal matching. However, (Nguyen & Worring, 2008) describes a system that integrates advanced similarity based visualization with active learning for content-based searching. Clip similarity ranking (Jain, Vailaya, & Wei, 1999) was built on top of shot similarity and combines temporal order, granularity and so on. However, the complicated variations of keyframes (Zhang & Chang, 2004) and cross-lingual video similarity measuring remains a challenging problem since shot copy detection built on global features is not robust enough for clip similarity measures. An approximation to solve this problem, called Signature-based methods, were proposed to identify similar clips by using global statistics of the low-level features (Cheung, 2003). This approach can achieve rapid detection but the effectiveness is limited to detecting almost identical or superficially edited videos (Hampapur, Hyun, & and Bolle, 2002). Hence, another approach based on frame-level similarity was proposed in Zhao, Ngo, Tan, and Wu (2007) and Ngo et al. (2006) which gives a high degree of editing. In general, the computation time is very high (Shechtman & Irani, 2005) when video copies are studied with background, colour and lighting, as well as content modification.

Video sound and image are studied separately and some techniques (Foote, 1998; Peng & Ngo, 2004) do not consider speech transcripts in the estimated similarity. Nevertheless, there are thousands of video copies, especially movie fragments, where users only change the speech and leave the images unchanged. Hence, the employment of either textual or visual concepts alone may not be sufficient since either content can appear differently over time.

The main contribution of this paper is a similarity measure between videos based on video segments instead of video frames. Furthermore, videos sound and image are jointly considered in the similarity.

The remainder of this paper is arranged as follows: Section 2 describes how the similarity is obtained step by step. In Section 3, experiments results are obtained by using a video dataset from the Beijing Olympic Games with German speech in some of the videos. Finally, the conclusion is drawn.

2. Video similarity

A general overview to calculate a similarity measure between videos is shown in Fig. 1. First, the videos are divided into segments based on shot detection. Then, a similarity measure between segments is obtained.

After, the matching of segments in both videos is carried out. In the following step, statistical features are considered and finally, a similarity measure between the videos is obtained from them. In the next sections the whole process is described in detail.

[*] Corresponding author.
E-mail address: dfuentes@us.es (D. Fuentes).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.169

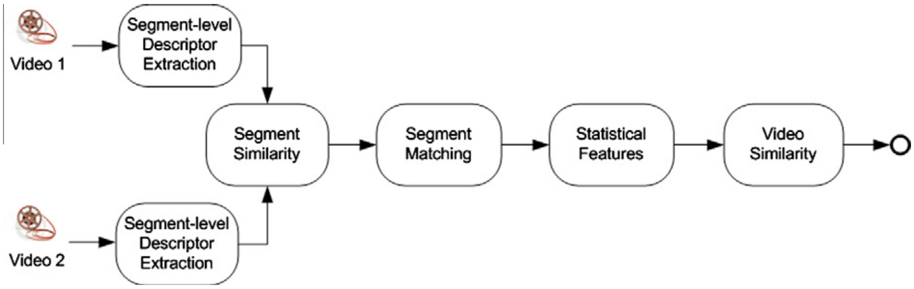


Fig. 1. An overview of the video similarity measure based on segment similarity, segment matching and statistical measures on the matching result.

2.1. Information extraction

Given a video data set, the first step is to divide each video into a sequence of segments. These are the result of shot segmentation based on hard and soft transition detection (Eickeler & Muller, 1999). Hence, two videos can contain a different number of segments and each segment has different length. This mapping of the video into segments instead of frame reduces the complexity of the analysis. The content of each segment is summarized in:

- **Start time**: Indicates the time in seconds when the segment starts.
- **Colour histograms**: 9 histograms of 256 RGB values are extracted. These histograms are mean histograms of all images in a segment, where each images is divided into 3 × 3 parts. The histogram is the mean of all image histograms in the segment from the same image part. It is represented by an array of 256 values.
- **Speech transcription**: Represented in syllables and their phoneme transcription (this representation is language dependent).

These features are considered in order to obtain a similarity measure between segments.

2.2. Segment similarity

Let S and S' two segments. By considering the associated features to a segment, the following dissimilarities between features are defined:

- **Alignment dissimilarity:**

  $$AD(S,S') = \frac{start(S) - start(S')}{\max\{len(S), len(S')\}}$$

  where *start* and *len* functions compute the start time and the length of the segment, respectively.
- **Histogram distance:** A commonly used technique to compare vectors is the Bhattacharyya distance (Comaniciu, Ramesh, & Meer, 2000). To compute this distance both histograms are normalized and denoted as follows:

  $$hist(S) = \{h_0, h_1, \ldots, h_8\}, \qquad h_i = \{u_{i,1}, u_{i,2}, \ldots, u_{i,256}\}$$
  $$hist(S') = \{h'_0, h'_1, \ldots, h'_8\}, \qquad h_j = \left\{u'_{j,1}, u'_{j,2}, \ldots, u'_{j,256}\right\}.$$

  The distance between histograms is $BD(h,h') = \sqrt{1 - \sum_{i=1}^{256}\sqrt{u_i u'_i}}$, and the histogram distance between segments is $HD(S,S') = \frac{1}{9}\sum_{\ell=1}^{9} BD\left(h_\ell, h'_\ell\right)$.
- **Speech distance:** To compare the transcriptions extracted from two videos, an adaptation of the Levenshtein distance (Eickeler & Muller, 1999) is applied by considering that all operations have unit cost. Let $\{ph_1,ph_2,\ldots,ph_p\}$ and $\left\{ph'_1, ph'_2, \ldots, ph'_q\right\}$ be the string of phonemes in segments S and S', respectively. The speech distance between S and S' is defined as $SD(S,S') = \frac{LD(p,q)}{\max(p,q)}$ where $LD(p,q)$ is the Levenshtein distance defined as follows:

  $$LD(k,\ell) = \begin{cases} 0 & ph_k = ph'_\ell \\ k & \ell = 0 \\ \ell & k = 0 \\ \min\{\mathrm{LD}(k-1,\ell)+1, \mathrm{LD}(k,\ell-1)+1, & \\ \quad \mathrm{LD}(k-1,\ell-1)+1\} & \text{otherwise} \end{cases}$$

  where for all k and ℓ, LD (k,ℓ) is the Levenshtein distance between the first k phonemes of S and the first ℓ phonemes of S' respectively.

The previous measure between features are dissimilarities, hence, a similarity measure between the S and S' segments is defined as follows:

$$SSM(S,S') = 1 - (w_1 AD(S,S') + w_2 HD(S,S') + w_3 SD(S,S'))$$

where $w_1 + w_2 + w_3 = 1$ and $w_i \geqslant 0$. The weights can be computed through empirical evaluation and represent the confidence in each feature. It is straightforward to prove that $0 \leqslant SSM(S,S') \leqslant 1$.

Let $V = (S_1,S_2,\ldots,S_m)$ and $V' = (S'_1, S'_2, \ldots, S'_n)$ be two videos where S_i and S'_j are the segment i and j in V and V', respectively. The Segment Similarity Matrix, denoted by SSM, is defined as $SSM = \left\{SSM\left(S_i, S'_j\right)\right\}_{i,j}$, and this matrix stores the needed data for the the next stages of the video similarity computation. The process of obtaining this matrix is visualized in Fig. 2.

2.3. Segment matching

The next step is to obtain the most similar pair of segments in both videos and, to do this, the problem is reduced to a graph theoretic problem called maximum bipartite matching (MBM). A weighted bipartite graph is constructed (as shown in Fig. 3) to model the two videos: segments form the vertices and the edge weights are obtained from the corresponding entries of the SSM. The maximum segment correspondence in the graph is generated from the maximum matching using the Hungarian algorithm (Kuhn, 1955).

Therefore, given $V = (S_1,S_2,\ldots,S_m)$ and $V' = (S'_1, S'_2, \ldots, S'_n)$ two videos, and for sake of simplicity, let us consider that $m \leqslant n$, the MBM technique provides a sequence of m pairs of segments comprising a maximal matching which can be denoted as follows:

$$((S_1, S^*_1), (S_2, S^*_2), \ldots, (S_m, S^*_m)), \qquad r = \min(n,m) = m$$

where $S^*_k = S'_j$ for a unique $j = 1, \cdots, n$.

This method solves the assignment problem in a non-incremental way and it operates on the fully specified bipartite graph. Furthermore, it provides the matching with the lowest possible cost

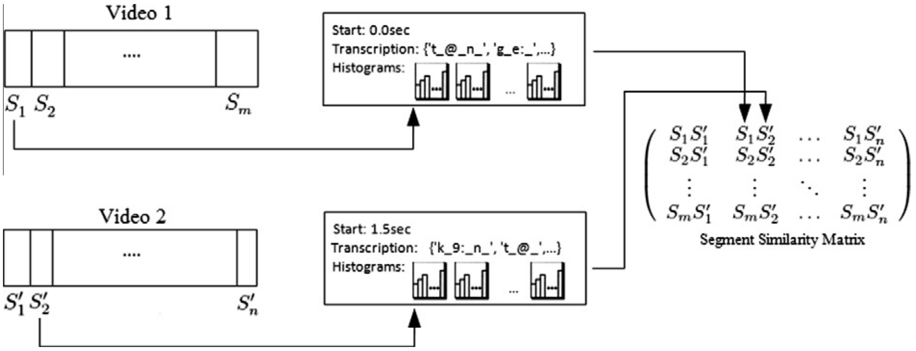

Video 1
....
S_1 S_2 S_m
Start: 0.0sec
Transcription: {'t_@_n_', 'g_e:_',...}
Histograms:
...
Video 2
....
S'_1 S'_2 S'_n
Start: 1.5sec
Transcription: {'k_9:_n_', 't_@_',...}
Histograms:
...
$S_1S'_1$ $S_1S'_2$ ... $S_1S'_n$
$S_2S'_1$ $S_2S'_2$ ... $S_2S'_n$
$S_mS'_1$ $S_mS'_2$ ... $S_mS'_n$
Segment Similarity Matrix

Fig. 2. From videos to similarity matrix: each video is segmented into shots and each segment is represented by time, colour, and speech information. A similarity matrix is formed by pairwise segment comparison.

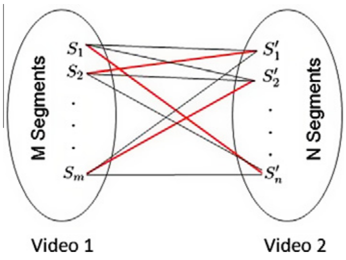


Fig. 3. A bipartite graph where the maximum matches (red edges) provide the segment correspondence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

($O(n^3)$) and since all operations are done on integers, always the integrality is maintained and an integer solution is obtained. The result of this stage is a min $(m,n) \times 2$ array that contains the maximum similarity pairs of segments and where m and n are the number of segments in V and V' respectively.

2.4. Statistical metrics

In the previous sections, the segments have been studied independently. Now, the similarity values of the pairs of segments will be combined to obtain a global similarity value between two videos. Let $((S_1, S_1^*), (S_2, S_2^*), \ldots, (S_m, S_m^*))$ be a sequence of r pairs of segments obtained after to apply the MBM technique to $V = (S_1, S_2, \ldots, S_m)$ and $V' = (S'_1, S'_2, \ldots, S'_n)$ videos, where $m \leqslant n$.

- **The crossings between similar segments**. The crossings between similar segments evaluates the order of the pairs and it is defined as follows:

$$CSS(V, V') = \frac{\#\{i|pos(S_i^*) > pos(S_{i+1}^*)\}}{m-1}$$

  where $pos(S_i^*)$ is the position (index) of segment S_i^* in the V video. Hence, the number of crossings will penalize the total similarity value.
- **Distance between similar segments** is defined as:

$$DP(V, V') = \frac{1}{m-1}\sum_{i=1}^{m-1}\frac{d_i^2}{\max(d_i^2)}$$

  where $d_i = end(S_{i+1}^*) - start(S_i^*)$ and the *start* and *end* functions indicate when the segment starts and ends, respectively.
- **The similar segments time** compares the length of similar segments regarding the whole videos and it is defined as:

$$SDT(V, V') = \frac{\sum_{i=1}^{m}\{len(S_i) + len(S_i^*)\}}{len(V) + len(V')}$$

- **Segment similarity average** computes the average of the segment similarity of the pair of the segments obtained after the MBM, and it is defined as:

$$SSA(V, V') = \frac{\sum_{i=1}^{m} SSM[pos(S_i), pos(S_i^*)]}{n}$$

  where $SSM[i,j]$ denotes the element of the i-row and j-column in the segment similarity matrix.

It is hold that $0 \leqslant CSS(V,V'), DP(V,V'), DT(V,V'), SSA(V,V') \leqslant 1$ for any V and V'. Hence, a similarity between video, denoted by $VS(V,V')$, is defined as follows:

$$VS(V, V') = 1 - (w_1^*CSS(V, V') + w_2^*DSS(V, V') + w_3^*SST(V, V') + w_4^*SSA(V, V'))$$

where $w_1^* + w_2^* + w_3^* + w_4^* = 1$ and $w_i^* \geqslant 0$. The weights allow for user specific tuning of the similarity measure.

3. Experimentation

Two experiments and an application have been carried out in order to evaluate the performance of the proposed similarity measure between videos. A dataset of 52 videos about the Beijing Olympic Games has been used. This dataset consists of videos with a maximum length of 30 min about different Olympic sports competitions but also are referred to interviews and Chinese culture. Many of them contain German speech from a narrator, interviewer or interviewee. The study is considered on compound videos which are the concatenation of three to six videos where the number of the videos and selected videos are chosen randomly.

All the experiments have been implemented in Java, executed on a PC using a Pentium processor at 2.40 Ghz with a 3 MB cache and 4 GB RAM and the weights have been chosen equally.

3.1. Similarity in compound videos

In this first experiment, eight videos have been chosen and two compounds have been formed among them (see in Fig. 4). The intermediate steps to convert the first compound video into the other have been studied. The videos are as follows:

- Video 1: Trampoline jump competition,
- Video 2: Interview with a swimmer,
- Video 3: Sailing competition,
- Video 4: Triathlon competition,
- Video 5: Summary of different Olympic sports, reports and interviews,
- Video 6: Table tennis match,

Fig. 4. From left to right and top to bottom, screenshots of 1–8 videos for Experiment 1.

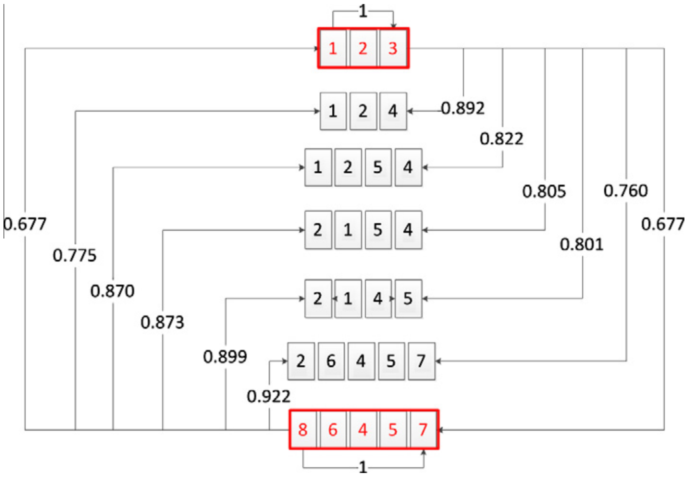


Fig. 5. Example of similarity results in a set of compound videos.

- Video 7: Tennis report,
- Video 8: Relay races.

A screenshot of each video is shown in Fig. 4.

The first compound video is formed by videos where the colours of the image is mostly blue and the transmission of competitions such as swimming or sailing do not need sudden movements or changes of cameras (reduced number of segments) and the speech and vocabulary is all related to aquatic sports. However, the second compound video is mainly referred to sports with red or maroon background colours, where the speed of the transmissions such as athletics or ping pong implies a lot of changes of the cameras and segments, and the speech and used vocabulary is mostly related to sports different from aquatic sports.

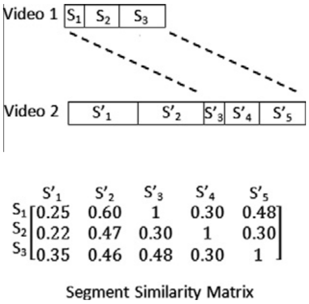
An intermediate transformations to convert the first compound video into the other is depicted in Fig. 5 where the values on the left indicate the similarity results between the first video and the others and the values on the right of the figure show the similarity results between the last video and the others. In the first intermediated step of the transformation, videos 4 and 5 are introduced and the similarity decreases by 10% because these are videos where only some scenes described aquatic sports and the order of videos have changed almost completely. In the next stage, videos 6 and 7 are introduced and the similarity between the new created video and the first compound video is reduced again in 5% because the number of original videos changes and the blue

Table 1
Impact of different features in the similarity value.

	Alignment (%)	Histograms (%)	Speech (%)
1	0.07	0.23	0.24
5	0.20	0.28	0.46
10	0.30	0.40	0.85

Table 2
Variation in similarity in distorted video.

Video 1			Video 2			Average	Similarity
Alignment (%)	Histograms (%)	Speech (%)	Alignment (%)	Histograms (%)	Speech (%)		
1	1	0	1	0	0	0.33	0.9978
1	1	1	1	0	1	0.33	0.9978
1	1	5	1	0	5	0.33	0.9978
1	1	10	1	0	10	0.33	0.9978
...	...	...	...	...	...	...	...
0	0	0	10	10	10	10	0.9852
0	0	10	10	10	0	10	0.9852
0	10	0	10	10	10	10	0.9852
0	10	10	10	0	0	10	0.9852



$$\begin{array}{c|ccccc} & S'_1 & S'_2 & S'_3 & S'_4 & S'_5 \\ \hline S_1 & 0.25 & 0.60 & 1 & 0.30 & 0.48 \\ S_2 & 0.22 & 0.47 & 0.30 & 1 & 0.30 \\ S_3 & 0.35 & 0.46 & 0.48 & 0.30 & 1 \end{array}$$

Fig. 6. Example of using the similarity measure for copy detection: video 2 contains a copy of video 1. This can be read off the diagonal of ones in the similarity matrix.

backgrounds, aquatic images and vocabulary is quite limited. Finally, the second compound video is totally formed and the maximum dissimilarity with respect to the first compound video is reached.

3.2. Influence of alignment, graphical and speech features in the similarity value

In this section, the 52 videos of the dataset were distorted in a different way and the impact of the modifications is measured. Two different experiments have been carried out.

First, for each video the distortion is performed changing one of its three features (start time, histograms, speech) in 1%, 5% and 10% percent respectively. Therefore, for each original video 9 new distorted videos are obtained and each of them is compared with its original using the proposed similarity measure. The mean of similarity results for each distortion type are calculated and results are shown in Table 1.

It can be seen from this table that the most influence feature is speech and that the histogram feature is more significant than the alignment feature. This act can be taken into account when weights in the proposed similarity are chosen. The robustness of the proposed method also become clear. It can be concluded from the low values of the table that, e.g., a variation of a 10% in the speech in a video produces only a 0.85% of distortion in the similarity result. It is worth noting that the average computation time to calculate 52 × 9 similarities in this experiment was 27,548 ms.

Second, one of the original videos is randomly selected and distorted in 0%, 1%, 5% and 10% to obtain 64 new videos using all the possible distortion combinations (3 features and 4 percentages) and the similarity is obtained with each pair of these new videos. The most significant results are described in Table 2 where the Average Variation column corresponds to the average of variations in each two of the generated videos. The last four rows show that videos with a greater percentage of variation are less similar than the first ones. Again, the robustness of the method is demonstrated because a variation of 10% in the video only implies a variation of less than 2% in the results of the similarity evaluation. The computation time to apply the proposed method in this experiment 2016 times (with 64 videos, 64 × 63/2 times) was 101,254 ms.

3.3. An application: video copy detection

The proposed similarity has been applied for detecting copies in compound videos. In this sense, as in Experiment 1, the compound videos are used to find out when a video is contained in another one. To detect a copy in a video it is necessary to localize a diagonal of values close to 1 in the Segment Similarity Matrix. If the number of these values is equal to the number of segments of the searched video, the compound video will contain a copy of because all the alignment, histogram and speech values are equal. An example is shown in Fig. 6: two videos with 3 and 5 segments respectively are compared. The last three segments in video 2 correspond to video 1. The three 1-values in the Segment Similarity Matrix identify the 3 segments of video 1 that are a copy of video 1 in video 2. It can also be seen in the Segment Similarity Matrix that the values around the diagonal of 1 values are equal because similar segments are being compared. This application was tested on the generated compound video dataset and all the copies were detected without any false positives.

4. Conclusion

A method has been presented to measure the similarity in two videos by analyzing properties of segments in both clips. Multiple features have been designed to evaluate the appearance of the segments, including colour and speech descriptors. We have demonstrated that the particular combination of the descriptors can be crucial for different comparisons. The proposed video similarity measure has been used in a real dataset, a very promising and competitive performance for comparison and copy detection has been achieved.

Acknowledgments

This research is supported by the Spanish Ministry of Science and Innovation R + D project ARTEMISA (TIN2009–14378-C02–01).

References

Cheung, S. (2003). Efficient video similarity measurement with video signature. *IEEE Transactions on Circuits and Systems for Video Technology, 13*(1), 59–74.

Chiu, C.-Y., Wang, H.-M., & Chen, C.-S. (2010). Fast min-hashing indexing and robust spatio-temporal matching for detecting video copies. *ACM Transactions on Multimedia Computing, Communications and Applications, 6*, 10:1–10:23.

Comaniciu, D., Ramesh, V., & Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. *IEEE Conference on Computer Vision and Pattern Recognition, 2*, 142–149.

Eickeler, S., & Muller, S. (1999). Content-based video indexing of TV broadcast news using hidden markov models. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 6*, 2997–3000.

Foote, J. (1998). An overview of audio information retrieval. *ACM Multimedia Systems, 7*, 2–10.

Hampapur, A., Hyun, K., & Bolle, R. (2002). Comparison of sequence matching techniques for video copy detection. *Conference on Storage and Retrieval for Media Databases*, 194–201.

Jain, A. K., Vailaya, A., & Wei, X. (1999). Query by video clip. *ACM Multimedia Systems, 7*, 369–384.

Kuhn, H. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*(1–2), 83–97.

Ngo, C., Zhao, W. & Jiang, Y. (2006). Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation. 845–854.

Nguyen, G. P., & Worring, M. (2008). Optimization of interactive visual-similarity-based search. *ACM Transactions on Multimedia Computing, Communications and Applications, 4*, 7:1–7:23.

Peng, Y., & Ngo, C.-W. (2004). Clip-based similarity measure for hierarchical video retrieval. *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 53–60.

Shechtman, E., & Irani, M. (2005). Space-time behavior based correlation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1*, 405–412.

Zhang, D.Q. & Chang, S.F. (2004). Detecting image near-duplicate by stochastic attributed relational graph matching with learning. *Proceedings of the 12th Annual ACM International Conference on Multimedia*, 877–884.

Zhao, W., Ngo, C., Tan, H., & Wu, X. (2007). Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Transactions on Multimedia, 9*(5), 1037–1048.