Multiple Hashing Integration for Real-Time Large Scale Part-to-Part Video Matching

Silvia Espinosa[1], Joaquín B. Ordieres[1], and Antonio Bello[2]

[1] Departamento de Proyectos, Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, 28006 Madrid, Spain.
j.ordieres@upm.es
[2] Departamento de Construcción e Ingeniería de Fabricación, Universidad de Oviedo, 33203 Gijón, Spain

Abstract. A real-time large scale part-to-part video matching algorithm, based on the cross correlation of the intensity of motion curves, is proposed with a view to originality recognition, video database cleansing, copyright enforcement, video tagging or video result re-ranking. Moreover, it is suggested how the most representative hashes and distance functions - strada, discrete cosine transformation, Marr-Hildreth and radial - should be integrated in order for the matching algorithm to be invariant against blur, compression and rotation distortions: $(R, \sigma) \in [1, 20] \times [1, 8]$, from 512×512 to 32×32pixels^2 and from 10 to 180°. The DCT hash is invariant against blur and compression up to 64x64 pixels^2. Nevertheless, although its performance against rotation is the best, with a success up to 70%, it should be combined with the Marr-Hildreth distance function. With the latter, the image selected by the DCT hash should be at a distance lower than 1.15 times the Marr-Hildreth minimum distance.

Keywords: video retrieval, pattern recognition, motion, distortion, hashing, data mining.

1 Introduction

Knowing whether a video or a part of it is already contained on a database is very important for applications such as originality recognition, video database cleansing, copyright enforcement, video tagging, video result re-ranking and cross-modal divergence detection. For instance, the motivation of our research project is preventing sexual exploitation of children by detecting whether the person under arrest is the author or a consumer of the pederastic videos. Much research effort have been made to near duplicate video retrieval [13–15, 21, 24]; nevertheless, as the video editing software evolves and becomes accessible to the general public, the number of videos created by concatenation of video fragments is increasing and part-to-part video matching has not been well addressed yet.

Videos are composed of images, usually taken at a rate of 24 frames per second. Since motion in the shot is usually minuscule, videos can be strongly compressed temporally by retaining only distinct visual appearances [15]. The similarity between video clips is typically calculated by comparing their keyframes.

Distortion	Parameters	Values
blur	R [-]	1, 5, 10, 15, 20
	σ [-]	1, 2, 3, 4, 5, 6, 7, 8
resize	Square image size [pixels]	32, 64, 128, 256, 512
rotate	Rotation [degrees]	10° to 180°, each 10°

Table 1. Distortion parameter values applied.

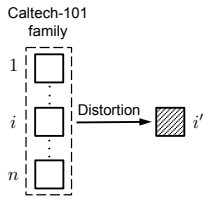


Fig. 1. Notation.

The most popular algorithms for keyframe selection are based on shot boundary detection [18] or time sampling, that can be uniform or based on the extrema of the intensity of motion [14, 15]. Once the keyframes have been extracted, they are represented by their visual features, such as local points and color histogram. Finally, these features are indexed with methods such as dimensionality reduction [23], tree-structure [3] or hashing [7]. The latter is the most accurate to represent video content [21] and, in particular, perceptual hashing is in addition robust in the sense that little perturbations in the original features slightly affect the result [27]. Although many perceptual hashing methods have been proposed in the literature [5,7,25–27], there is not still a comparison that allows to predict the optimum range of application and how they could be combined to achieve a part-to-part video matching algorithm invariant against distortions such as blur, compression and rotation.

In this research work, we propose a real-time large scale part-to-part video matching algorithm based on the cross correlation of the intensity of motion curves. In addition, it is suggested how the most representative hashes - particularly, strada (str), discrete cosine transformation (dct), Marr-Hildreth wavelets (mw) and radial (rad) image hashes - should be combined in order for the matching algorithm to be invariant against blur, compression and rotation distortions.

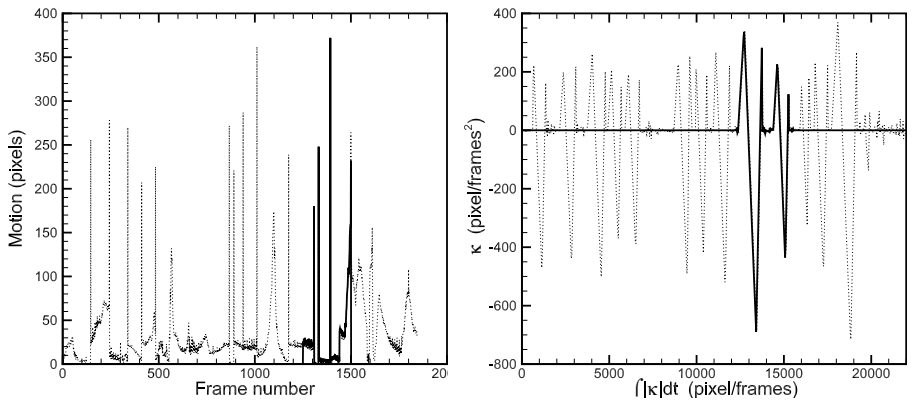
2 Methodology

2.1 Intensity of motion

The part-to-part video matching algorithm proposed is based on the cross correlation of the intensity of motion curves. Intensity of motion is defined as the mean of consecutive frame differences [15]:

$$m(t) = \frac{1}{A} \sum_{\mathbf{x}} |L(\mathbf{x}, t+1) - L(\mathbf{x}, t)| \tag{1}$$

where A is the area of the video and $L(\mathbf{x}, t)$ denotes the luminance value of pixel $\mathbf{x}$ of a frame at time t. The motion curves are extracted by outputting intermediate results from the shot detecting tool [1].



(a) Motion (b) Signature

Fig. 2. Intensity of motion part-to-part matching. The motion of the original video is plotted in dotted thin line; while the fragment, situated in the predicted optimum position, in solid thick line.

2.2 Part-to-part video matching

The state of the art algorithms for 2D part-to-part curve matching are generally $O(N^4)$. For instance, hash-based methods [9, 11, 19] can do the comparison in $O(N)$, but they require pre-processing steps that are of $O(N^4)$ asymptotic complexity. Nevertheless, the algorithm proposed by Cui [6], where the input curves can differ by a similarity transform, performs the part-to-part matching in $O(N^3)$. Hence, the latter is utilized for the intensity of motion matching.

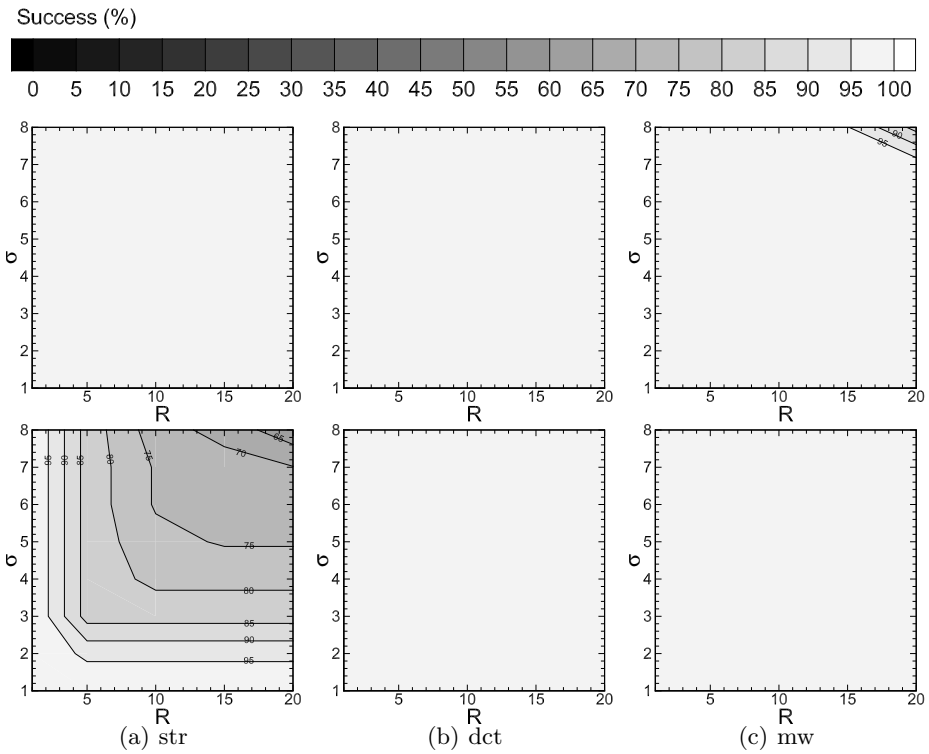
First, the absolute curvature $|\kappa(t)|$ of the intensity of motion $\mathbf{s}(t)$ is calculated by using second order central differences for discretizing the second derivative.

$$|\kappa(t)| = \|\ddot{\mathbf{s}}(t)\| \tag{2}$$

Second, the integral of the absolute curvature from an arbitrary start point t_1 on the curve with respect to another fixed point t_2, $K(t_1 : t_2)$, which is invariant under similarity transform [6], is defined as follows:

$$K(t_1 : t_2) = \int_{t_1}^{t_2} |\kappa(t)| \, \mathrm{d}t \tag{3}$$

and discretized in this case by utilizing trapezoidal integration. The curve $t - K$ is sampled at equal intervals along the integral of the absolute curvature K, so as for the interpolation point density to be higher where the absolute curvature gradient magnitude is larger. The mean t is selected when the correspondent t is not unique and linear interpolation is recommended to avoid spurious oscillations. Afterwards, a second interpolation is carried out to obtain the *signed* curvature values that correspond to the previously interpolated points in t [6].



(a) str (b) dct (c) mw

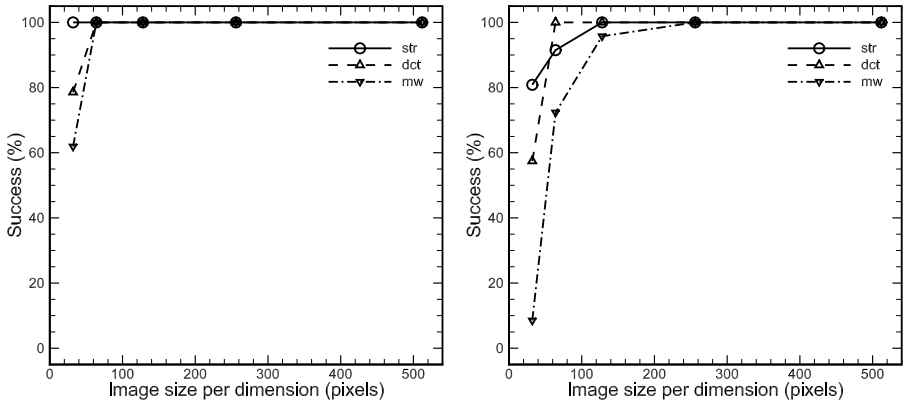
Fig. 3. Effect of the hash type on the matching success (%) after applying the *blur* distortion with characteristic parameters R and σ to the ant (above) and ceiling fan (below) families of the set Caltech-101 [8].

Finally, the curve that should be matched is the *signed* curvature, κ, parametrized by the integral of curvature integral, K, since is invariant under similarity transform. In order for the matching not to be influenced by the scale, the normalized cross correlation [17, 28] is adopted:

$$\nu(u) = \frac{\sum_{i\in\Omega}\left[r(i) - \bar{r_w}\right]\left[f(i-u) - \bar{f}\right]}{\sqrt{\sum_{i\in\Omega}\left[r(i) - \bar{r_w}\right]^2 \sum_{i\in\Omega}\left[f(i-u) - \bar{f}\right]^2}} \tag{4}$$

where u is the offset of one curve to the other, f is the first curve working as a template window sliding along the second curve r, $\bar{f}$ is the mean value over the whole template and $\bar{r_w}$ is the mean value *of the sliding window* in the second curve. Nevertheless, in part-to-part matching an infinitesimally small part of one curve will match many parts of another curve, as noted by Cui [6]. To favor longer matches, the following strategies are used:

1. Thresholding on the length of the matching.
2. The score is given by the cross correlation ν, which lies in $[0, 1]$, multiplied by the length of the match L.



(a) Ant (b) Ceiling fan

Fig. 4. Effect of the hash type on the matching success (%) after compressing the images of several families of the set Caltech-101 [8].

After the start points and end points on both curves have been determined, the similarity transform between them can be estimated using the method proposed by Horn [12].

2.3 Similarity between videos

After determining the common intensity of motion curve segments, the hashes of the pairs of frames embed into them are calculated. In the end, comparison between sets of images is performed by using different hashes, allowing to combine several distance functions in an integrated way. The similarity between videos is defined as follows:

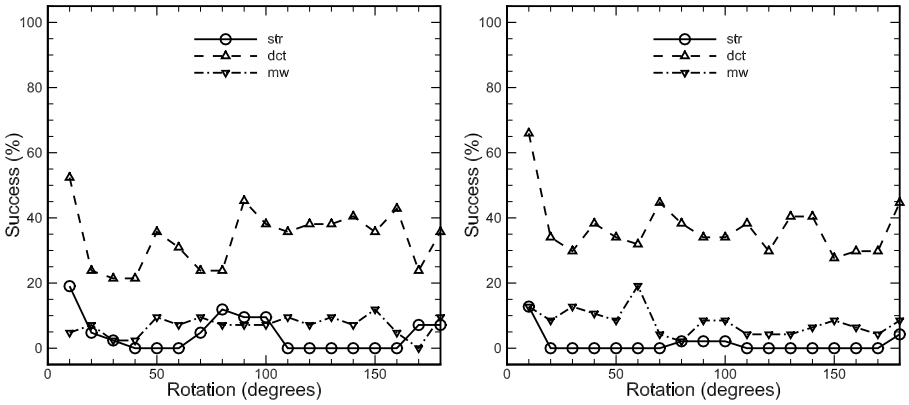
$$\text{Similarity}(\%) = \frac{\text{Pairs of keyframes whose } d < d_{\text{threshold}}}{\text{Pairs of keyframes}} \cdot 100 \tag{5}$$

where d is the distance between images and $d_{\text{threshold}}$ the threshold distance.

3 Problem Setup

As a matter of example illustrating the part-to-part video matching, several fragments of the synthetic movie [2] are matched with the original video. A hundred interpolation points are used in the fragment, utilizing the same interpolation point distance in K in the other curve. Moreover, the length of the matching contains at least fifty interpolation points.

In order to determine the optimum combination of distance functions, the ant and ceiling fan families of the Caltech-101 benchmark [8] are distorted utilizing ImageMagick mogrify with the parameter value ranges indicated in Table 1. The



(a) Ant (b) Ceiling fan

Fig. 5. Effect of the hash type on the matching success (%) after rotating the images of several families of the set Caltech-101 [8].

library pHash [26] is utilized for the dct, mw and rad hashes. The study of the radial hash [26] is restricted to image whose aspect ratio is below a critical value in $(1.8987, 1.9868]$ and a threehold of 0.9 is utilized. Using the notation described in Figure 1, an image is considered to be correctly matched if

$$d_{ii'} \leq d_{ij}, \qquad \forall j = 1, \ldots, n, \quad j \neq i \tag{6}$$

where d_{ab} denotes the distance between images a and b. Based on the previous convention, the success of a hashing distance is defined as follows:

$$\text{Success } (\%) = \frac{\text{Images correctly matched}}{\text{Images of the set}} \cdot 100\% \tag{7}$$

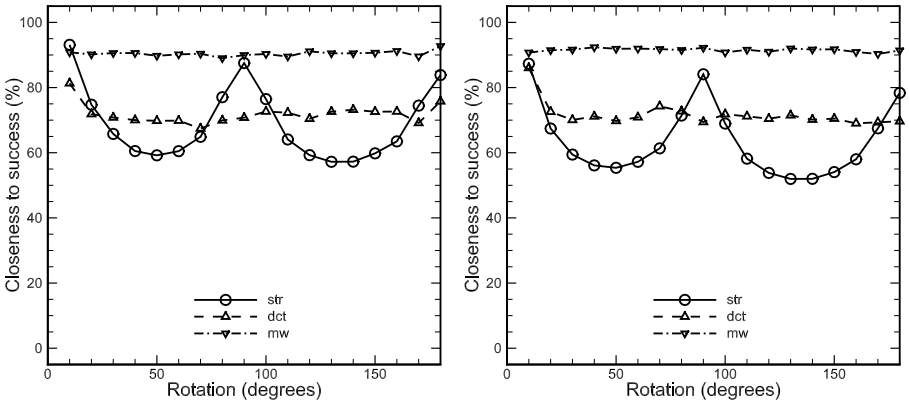
$$\text{Closeness to success } (\%) = \left\langle \frac{\text{Minimum distance}}{\text{Distance to the distorted image}} \right\rangle_{\text{images}} \cdot 100\% \tag{8}$$

where $\langle\rangle$ denotes average on the images of the family set. In addition, if the distance to the distorted image is 0, the closeness to success is 100% by convention.

4 Results and discussion

Figure 2 illustrates the successful results of part-to-part matching algorithm of the original video with the latter, in the cases with a intensity of motion range two order of magnitude smaller and of the order of that of the original video. In particular, note that the former is correctly captured due to the usage of the mean value *of the sliding window* in the second curve.

In Figure 3, the effect of the hash distance on the matching success after applying the **blur** distortion are presented. As the distortion parameter values



(a) Ant (b) Ceiling fan

Fig. 6. Effect of the hash type on the closeness to matching success (%) after rotating the images of several families of the set Caltech-101 [8].

are increased, the success decreases. The only hash that is invariant against blur distortion in the range analyzed is the *dct*. Like any Fourier-related transform, dct expresses a function or signal in terms of a sum of cosine functions with different frequencies and amplitudes [26]. Low-frequency dct coefficients of an image are mostly stable under image manipulations [10], since most of the signal information tends to be concentrated in a few low-frequency components of the dct. This property is also utilized by the JPEG image compression standard [4].

The *mw* hash uses edge detectors for feature extraction. Since the derivative operator acts as a highpass filter, edge detectors based on it are sensitive to noise [4]. While for the ceiling fan family the algorithm is invariant to noise in the whole range under study, for the ant family the complete success is only guaranteed for $\sigma \leq 7$ and $R \leq 15$. The *radial* (rad) perceptual image hash function, based on the Radon transform [20], was proposed by Lefèbvre and Macq [16]. A few years later, both authors outlined [22] that their previously proposed algorithm is invariant against distortion, which explains the null performance of this hash in the range analyzed. Ultimately, the performance of the hash *strada* (std) distance function strongly depends on the shape of the object. As this hash compares the luminosity of each pair of pixels, the strada hash is sensible to the thin details, such as the blades of the ceiling fan, which are highly affected by the blur. As a consequence, while it is invariant against blur distortion with the ant family, a 100% success is only guaranteed when for $\sigma \leq 2$ and $R \leq 5$ with the ceiling fan family.

Figure 4 shows the effect of the hash distance on the matching success after applying the **compression** distortion. First, when compressing the image the percentage of success remains constant until a critical value is reached and afterwards it decreases monotonously. Thus, the hash functions are invariant against blur until reaching a compression of 64x64 and 256x256 pixels^2 for the ant and

ceiling fan families respectively. The best overall performance is achieved with the strada distance function, being invariant against compression distortion for the ant family and with a percentage of success over 80% for the ceiling fan family. In contrast, the dct distance function range of invariance against compression expands until 64x64 pixels^2 for both families, being independent of the image itself.

In Figure 5, the effect of the hash distance on the matching success after applying the **rotation** distortion are presented. First, the success significantly decreases in the present of rotation, with a success lower than 70% in all the cases. The dct distance is clearly recommendable since the percentage of success achieved is in $(20, 70)$ %, while for the other distances the success is under 20%.

Figure 6 shows the effect of the hash type on the closeness to matching success. The closeness to success of the dct distance function is around 70%. Regarding the strada hash, the best results are obtained around 0°, 90° and 180°, with closeness to success near 90%. However, as the rotation angles move away from these values, the closeness to success decreases up to a 60%. Fortunately, for the Marr-Hildreth hash is around 90%, independently of the rotation angle.

5 Conclusions

The real-time large scale part-to-part video matching algorithm proposed, based on the normalized cross correlation of the intensity of motion curves, is successful even with a intensity of motion range two orders of magnitude smaller than that of the original video.

In addition, it is suggested how the most representative hashes and distance functions should be integrated in order for the matching algorithm to be invariant against blur, compression and rotation distortions. In the range under study, the DCT hash is invariant against blur since the signal information is concentrated in a few low-frequency components of the discrete cosine transformation which are most stable against this distortion. Furthermore, its range of invariance against compression expands until 64x64 pixels^2, independently of the image itself, which would allow to predict robustly the performance with general images. Moreover, the performance of the DCT hash distance is also the best, with a success up to 70%. Nevertheless, it is highly recommendable to complement this information with the mw distance function, with which the image selected by the DCT hash should be at a distance lower than 1.15 times the mw minimum distance.

Even though rotation is expected to be the most influential distortion, the analysis of the response of the hash distances against combined distortions will further contribute towards an integrated multiple hashing integration for large scale part-to-part video matching.

References

1. (March 2013), `http://johmathe.name/shotdetect.html`
2. (March 2013), `http://www.anc.ed.ac.uk/~amos/movies/afreightc.avi`

3. Böhm, C., Berchtold, S., Keim, D.A.: Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. ACM Comput. Surv. 33(3), 322–373 (Sep 2001), http://doi.acm.org/10.1145/502807.502809
4. Bovik, A.C.: The Essential Guide to Image Processing. Academic Press (2009)
5. Bozkaya, T., Bozkaya, T., Ozsoyoglu, M.: Indexing large metric spaces for similarity search queries. ACM Transactions on Database Systems 24, 361–404 (2002)
6. Cui, M., Femiani, J., Hu, J., Wonka, P., Razdan, A.: Curve matching for open 2d curves. Pattern Recogn. Lett. 30(1), 1–10 (Jan 2009), http://dx.doi.org/10.1016/j.patrec.2008.08.013
7. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the twentieth annual symposium on Computational geometry. pp. 253–262. SCG '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/997817.997857
8. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories (2004)
9. Frenkel, M., Frenkel, M., Basri, R.: Curve matching using the fast marching method. IN EMMCVPR pp. 35–51 (2003)
10. Fridrich, J.: Robust bit extraction from images. In: Proceedings of the IEEE International Conference on Multimedia Computing and Systems - Volume 2. pp. 536–. ICMCS '99, IEEE Computer Society, Washington, DC, USA (1999), http://dl.acm.org/citation.cfm?id=839287.842042
11. Grauman, K., Fergus, R.: Learning binary hash codes for large-scale image search. Cipolla, Roberto (ed.) et al., Machine learning for computer vision. Selected papers based on the presentations at the international computer vision summer school (ICVSS 2012), Sicily, Italy, July 7–15, 2012. Berlin: Springer. Studies in Computational Intelligence 411, 49-87 (2013). (2013)
12. Horn, B.K.P., Horn, B.K.P.: Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America A 4(4), 629–642 (1987)
13. Huang, Z., Shen, H.T., Shao, J., Cui, B., Zhou, X.: Practical online near-duplicate subsequence detection for continuous video streams. IEEE Transactions on Multimedia 12(5), 386–398 (2010)
14. Joly, A., Frélicot, C., Buisson, O.: Robust content-based video copy identification in a large reference database. In: Proceedings of the 2nd international conference on Image and video retrieval. pp. 414–424. CIVR'03, Springer-Verlag, Berlin, Heidelberg (2003), http://dl.acm.org/citation.cfm?id=1760167.1760219
15. Karpenko, A., Aarabi, P.: Tiny videos: A large data set for nonparametric video retrieval and frame classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(3), 618–630 (2011)
16. Lefèbvre, F., Czyz, J., Macq, B.M.: A robust soft hash algorithm for digital image signature. In: International Conference on Image Processing. vol. 2, pp. 495–498 (2003)
17. Lewis, J.P., Lewis, J.P.: Fast normalized cross-correlation (1995)
18. Lienhart, R.W.: Comparison of automatic shot boundary detection algorithms pp. 290–301 (1998), +http://dx.doi.org/10.1117/12.333848
19. Pajdla, T., Pajdla, T., Van Gool, L., Van Gool, L.: Matching of 3-d curves using semi-differential invariants. In 5th International Conference On Computer Vision pp. 390–395 (1995)
20. Radon, J.: On the determination of functions from their integral values along certain manifolds. IEEE Trans Med Imaging 5(4), 170–6 (1986), http://www.

biomedsearch.com/nih/Determination-Functions-from-Their-Integral/18244009.html
21. Song, J., Yang, Y., Huang, Z., Shen, H.T., Hong, R.: Multiple feature hashing for real-time large scale near-duplicate video retrieval. In: Proceedings of the 19th ACM international conference on Multimedia. pp. 423–432. MM '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2072298.2072354
22. Standaert, F.X., Lefebvre, F., Rouvroy, G., Macq, B., Quisquater, J.J., Legat, J.D.: Practical evaluation of a radial soft hash algorithm. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II - Volume 02. pp. 89–94. ITCC '05, IEEE Computer Society, Washington, DC, USA (2005), http://dx.doi.org/10.1109/ITCC.2005.229
23. Wichterich, M., Assent, I., Kranen, P., Seidl, T.: Efficient emd-based similarity search in multimedia databases via flexible dimensionality reduction. In: SIGMOD Conference. pp. 199–212 (2008)
24. Wu, X., Hauptmann, A.G., Ngo, C.W.: Practical elimination of near-duplicates from web video search. In: Proceedings of the 15th international conference on Multimedia. pp. 218–227. MULTIMEDIA '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1291233.1291280
25. Wu, X., Zhao, W.L., Ngo, C.W.: Near-duplicate keyframe retrieval with visual keywords and semantic context. In: Proceedings of the 6th ACM international conference on Image and video retrieval. pp. 162–169. CIVR '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1282280.1282309
26. Zauner, C.: Implementation and Benchmarking of Perceptual Image Hash Functions. Master's thesis, Upper Austria University of Applied Sciences, Hagenberg Campus (Jul 2010)
27. Zhou, X., Schmucker, M., Brown, C.: Perceptual hashing of video content based on differential block similarity. In: Hao, Y., Liu, J., Wang, Y.P., Cheung, Y.m., Yin, H., Jiao, L., Ma, J., Jiao, Y.C. (eds.) Computational Intelligence and Security, Lecture Notes in Computer Science, vol. 3802, pp. 80–85. Springer Berlin Heidelberg (2005), http://dx.doi.org/10.1007/11596981_12
28. Zitová, B., Flusser, J.: Image registration methods: a survey. Image and Vision Computing 21(11), 977–1000 (Oct 2003), http://dx.doi.org/10.1016/s0262-8856(03)00137-9