Local Radon Transform and Earth Mover's Distances for Content-Based Image Retrieval

Wei Xiong[1], S.H. Ong[2], Weiping Lee[2], and Kelvin Foong[3]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
wxiong@i2r.a-star.edu.sg
[2] Department of ECE, National University of Singapore, Singapore 117576
eleongsh@nus.edu.sg
[3] Department of Preventive Dentistry, National University of Singapore, Singapore 119074
pndfwc@nus.edu.sg

Abstract. Content-based image retrieval based on feature extraction is still a highly challenging task. Traditional features are either purely statistical, thus losing spatial information, or purely spatial without statistical information. The Radon transform (RT) is a geometrical transform widely used in computer tomography. The projections transformed embed spatial relationships while integrating information in certain directions. The RT has been used to design invariant features for retrieval. Spatial resolutions in RT are inhomogeneous resulting in non-uniform feature representation across the image. We employ the local RT by aligning the centre of the RT with the centroids of the region of interest and use a sufficient number of projections. Finally the earth mover's distance method is utilized to combine local matching results. Using the proposed approach, image retrieval accuracy is maintained, while reducing computational cost.

1 Introduction

Content-based image retrieval (CBIR) is the digital image searching problem in large databases that makes use of the contents of the images themselves, rather than relying on human-input textual information such as captions or keywords. The history of CBIR can be traced back to the early 1980s [1]. The first well-known image retrieval system QBIC was reported in the early 1990s [2]. There are numerous CBIR techniques and systems published. Reviews of CBIR can be found in [3,4,5,6].

CBIR systems represent image content by making use of lower-level features such as texture, color, and shape. Most of these features are either purely statistical information without spatial information, or purely spatial information without statistical information. A typical example is the color histogram which is measured globally and loses spatial information such as the geometrical

S. Satoh, F. Nack, and M. Etoh (Eds.): MMM 2008, LNCS 4903, pp. 436–445, 2008.

relations and the distribution of regions in the image. Furthermore, complicated image segmentation procedures are normally involved in feature extraction. CBIR without complicated segmentation [7] has gained some success.

The Radon transform (RT) [8,9] is a geometrical transformation widely used in computer tomography and image reconstruction. It projects images into a parameter space of viewing angles and offsets. A collection of one-dimensional projections (i.e., signatures) is obtained to represent images, mostly for image reconstruction [9] and image matching [10]. RT has elegant properties, from which invariant feature signatures can be derived against translation, rotation and scaling [11,12]. The use of Radon signatures without a sophisticated segmentation for image retrieval [13,14] has achieved better experimental performance than well-known retrieval systems [13].

Although the RT was introduced in continuous form analytically, the viewing angles are discrete in practice. One has to sample the angular parameter and take projections at these sampling nodes. Only a finite number of projections are produced, although, theoretically, one should have an infinite number of them for perfect reconstruction [8,9]. It has been noted that, in ranking the similarity of images for image retrieval, high resolution is not required as projection data is highly correlated and there is information redundancy [13]. Thus, in [13] only two perpendicular projections are taken as signatures, causing loss of information. A follow-up paper [14] suggests taking more projections to obtain a more complete representation of the image, but it is noted that too fine a resolution of angle θ results in redundant information. It is also suggested that taking θ at equal intervals in the RT is not efficient as the information in the image may not be evenly distributed. A projection decimation method is hence proposed in [14] where the number of projections required is iteratively selected up to the point when the image is segmented into clearly defined regions, and this is done by detecting the high contrast point or clustering using the projection data derived iteratively.

From the review of past literature, it is found that current techniques that employ the RT for image matching /retrieval fail to address three problems. Firstly, as the RT is performed globally on an image and considering the discrete limitations in which the RT can be performed on a digital image, certain fine image details further away from the centre of rotation of the RT will not be represented sufficiently. A solution to circumvent this problem is to perform RT locally on objects of interest in an image. However, there is a need to segment an image into meaningful objects before performing the RT.

Secondly, there is no comprehensive method to determine the optimum number of projections to use when using the RT. The proposed solution is to then make use of the frequency characteristics of the object, along with the size of the object, to decide upon the minimum number of projections required to effectively represent the object using RT.

Lastly, there is not much work on combining local Radon signatures, though a simple feature-vector-adjoining method has been proposed in [14]. In this paper, we introduce a new method for image retrieval. We use the local RT to represent images and determine sufficient numbers of projections. Finally, the

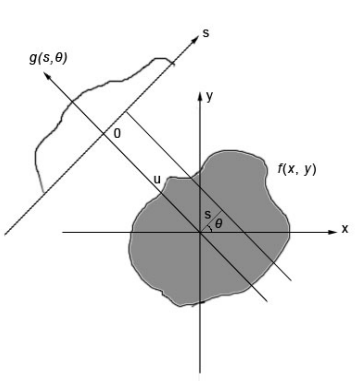
Earth Mover's Distance (EMD) method [15] is employed to combine multiple local signatures to measure image dissimilarity.

2 Radon Transform and Sufficient Number of Projections

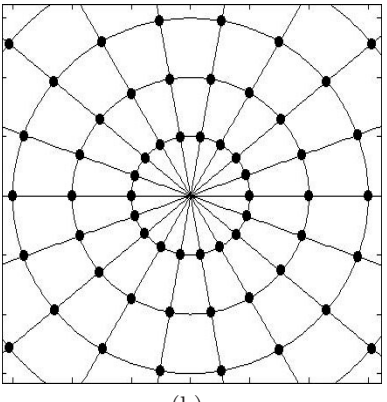
The Radon Transform $g(s,\theta)$ of a function $f(x,y)$ is the function's line integral along a line inclined at an angle θ and at a distance s away from the origin of the $x-y$ plane. It is given by

$$g(s,\theta)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\delta(x\cos\theta+y\sin\theta-s)\mathrm{d}x\mathrm{d}y, \tag{1}$$

where $\delta()$ is the Dirac function with $-\infty<s<\infty$, $0\le\theta<\pi$. The function $g(s,\theta)$ is called the (Radon) sinogram of $f(x,y)$. Fig. 1(a) shows the geometry for the transform.



(a)



(b)

Fig. 1. Radon transform: its geometry (a) and inhomogeneous sampling grid (b)

In the rotated system (s,u), with

$$\begin{cases} x=s\cos\theta-u\sin\theta \\ y=s\sin\theta+u\cos\theta, \end{cases} \tag{2}$$

(1) can be written as

$$g(s,\theta)=\int_{-\infty}^{\infty}f(s\cos\theta-u\sin\theta,s\sin\theta+u\cos\theta)\mathrm{d}u. \tag{3}$$

The image function $f(x,y)$ can then be reconstructed by

$$f(x,y)=\int_{0}^{\pi}\int_{-\infty}^{\infty}g(s,\theta)\mathrm{d}s\mathrm{d}\theta. \tag{4}$$

The RT has some useful properties that relates the translation, rotation and scaling of the original function $f(x, y)$ to the $g(s, \theta)$ [8]. They are tabulated in Table 1. A translation of $f(x, y)$ will result in the shift of $g(s, \theta)$ in s, a rotation of $f(x, y)$ will result in the translation of $g(s, \theta)$ in θ, and a uniform scaling of $f(x, y)$ will result in the uniform scaling of the projections. These properties are used to design features invariant to translation, rotation and scaling [13].

Table 1. Some properties of the Radon Transform

	Function $f(x, y) = f_p(r, \theta)$	Radon Transform $g(s, \theta)$
Translation	$f(x - x_0, y - y_0)$	$g(s - x_0 \cos\theta - y_0 \sin\theta, \theta)$
Rotation	$f_p(r, \theta + \theta_0)$	$g(s, \theta - \theta_0)$
Scaling	$f(ax, ay)$	$g(as, \theta)/a$

The uncertainty of x and y induced by the perturbation of θ can be derived by differentiating (5) with respective to θ respectively:

$$\begin{cases} \mathrm{d}x = (-s\sin\theta - u\cos\theta)\mathrm{d}\theta \\ \mathrm{d}y = (s\cos\theta - u\sin\theta)\mathrm{d}\theta \end{cases} \tag{5}$$

The circular uncertainty induced by the uncertainty of θ along a circle centered at the origin at radius $r = \sqrt{s^2 + u^2}$ is

$$r\mathrm{d}\theta = (\sqrt{s^2 + u^2})\mathrm{d}\theta \tag{6}$$

These uncertainties are spatial resolutions of the reconstructed function. They are not homogeneous across the (s, u)-plane. For example, the circular resolution is proportional to the uncertainty of θ and the radius. Such inhomogeneity is the basis for local RT in the current work.

When finding RTs on digital images, projections are available only on a finite grid in a polar system (for an illustration, see Fig. 1(b)), i.e. [9],

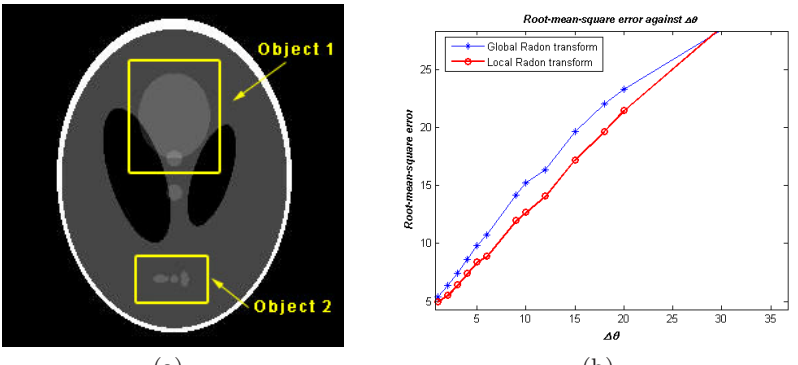
$$g_n(m) \stackrel{\Delta}{=} g(s_m, \theta_n), -\frac{M}{2} \le m \le \frac{M}{2} - 1, 0 \le n \le N - 1, \tag{7}$$

where $s_m = md$, $\theta_n = n\Delta$, $\Delta = \pi/N$. If ζ_0 is the highest spatial frequency in the image, then the radial sampling interval d along each projection direction should satisfy $d \le 1/(2\zeta_0)$. Assuming that the image is spatial-limited within a circle of diameter$D$, then $D = Md$ and $M \ge 2\zeta_0 D$.

Unlike a constant d for all projection directions, the arc interval is not constant: at offset $|s_m|$, it is $r_m = |\pi s_m/N| = |\Delta s_m| = |m|\, d\Delta$, $-\frac{M}{2} \le m \le \frac{M}{2} - 1$. The minimum interval is zero at the origin while the maximum $r_M = Md\Delta/2 = D\pi/(2N)$ occurs at the furthermost offset from the projection centre. Such resolution inhomogeneity is illustrated in Fig. 1(b). There is a higher concentration of points nearer the centre of rotation, and a sparser spread of points further away. For convenience, one can choose $r_M \approx d_0 = 1/(2\zeta_0)$ such that $N \ge \pi D\zeta_0$. We consider $\pi D\zeta_0$ is the sufficient number of projections for our purpose in image retrieval.

In practice, we may not take so many projections. Instead, a much smaller number of projections are used. In this case, due to this resolution inhomogeneity, some small but important objects, which are placed further away from the centre of the image, may not be effectively represented by the forward RT. High frequency components, such as variations in the boundaries of an object, may also not be represented sufficiently by the RT. Insufficient finite number of projections makes the resolution inhomogeneity ignorable. Such an effect in image reconstruction due to projections from limited angles has been extensively studied in computer tomography literature [8].

A solution is to represent an object locally, i.e., to perform the transform by putting the origin of transform at the center of the object of interest. This is the so-called local Radon transform in this work. In contrast, a transform performed on the centre of an image is refereed to a global RT. Fig. 2(b) shows the root-mean-square error of the reconstructed image against the angular intervals between projects. The ground truth is the object 2 in Fig. 2(a). It is observed that the reconstruction errors after performing local Radon transformation on the objects are lower than the reconstruction errors after performing Radon transformation globally on the entire image. This implies that by shifting the centre of rotation to the objects, the increase in the resolution of the Radon transform allows for a better reconstruction.



(a) (b)

Fig. 2. Local RT represents image better than global RT when using a finite number of projections. (a): the phantom image; (b): reconstruction error vs angular intervals.

We emphasize the point here since not much attention has been paid to such a non-uniform feature representation problem in image matching and retrieval. According to the above idea, we could segment an image into meaningful clusters of objects and perform RTs locally, then combine them for final image matching and retrieval. In this short paper, we will not introduce advanced techniques in image segmentation/clustering. Instead, only simple and necessary segmentation is used.

3 Earth-Mover's Distances

After obtaining local RT projections, we utilize the Earth-Mover's-Distance method to combine them to measure dissimilarity. The EMD is used to evaluate dissimilarity between two multi-dimensional distributions in a feature space, given the ground distance between individual features. Each distribution is represented by a set of features and this representation is called the signature of the distribution. With two signatures, one of the signatures can be visualized as a mass of earth spreading out in feature space while the other can be visualized as a collection of holes which are to be filled with the earth. The EMD essentially then computes the least amount of work needed to fill the holes with earth and the amount of work, or cost needed will depend on the ground distance between the features.

Let $X = \{(x_i, w_{x,i})\}_{i=1}^{I}$ and $Y = \{(y_i, w_{y,j})\}_{j=1}^{J}$ be two signatures with I and J features and weighted by $w_{x,i}$ and $w_{y,j}$ respectively. Assume d_{ij} and f_{ij} are the ground distance and the flow between x_i and y_j. The problem is to determine the flow f_{ij} that minimizes the total cost

$$COST(X, Y) = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} d_{ij} \tag{8}$$

Upon finding the optimal flow that minimizes total cost, the EMD is then equal to the total cost normalized by the total flow, i.e.,

$$EMD(X, Y) = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} d_{ij} \Big/ \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \tag{9}$$

4 The Proposed Method

Denote the query q and the target x. The images are first processed to identify interest regions and objects. Our method has four steps and is summarized as follows.

Step 1: Given two objects q_i and x_j, perform the Fourier transform and find their respective cut-off frequency $\zeta_{0,i}$ and $\zeta_{0,j}$. Set $\zeta_0 = \max\{\zeta_{0,i}, \zeta_{0,j}\}$. Find $d = 1/(2\zeta_0)$ and $\Delta = 1/(D\zeta_0)$. Perform local RTs according to (3).

Step 2: Take the summation of all absolute distances between their respective projections and use it as their dissimilarity.

Step 3: Following [13], compensate their global scaling (this can be achieved by normalizing the two images according to their object areas), remove their relative translation and rotation. Align them so that the dissimilarity is minimized. Such dissimilarity is their ground distance d_{ij}.

Step 4: Compare images using EMD as the scores and rank them. The two signatures and will be the two sets of objects from the two images that are to be compared. The weight of each feature will be assigned equally to 1.

5 Results and Discussions

We first test our method to retrieve simple images. The proposed image matching process is performed in contrast to cases when Δ is fixed to a very small interval of 5° and a very large interval of 90°, which essentially means only horizontal and vertical projections are obtained. The purpose is to show that our method can achieve better performance in terms of both accuracy and time cost. For the sake of convenience, the method using $\Delta = 5°$ will be termed method 1, the proposed method 2 and the method using $\Delta = 90°$ method 3. The performances of the processes are compared in terms of recall time and image matching accuracy over 100 binary images with translation, rotation and scaling. Retrieval precisions for the recall points of the top 10 and 19 are evaluated subjectively by looking for contextually similar images. The tests are performed on an Intel Pentium M 1.4Ghz laptop, with 768MB of RAM, using the MATLAB environment.

In Fig. 3, subfigures (a), (b) and (c) are for methods 1, 2 and 3, respectively. All the test results are displayed with the reference image shown at the top left hand corner. The rest of the images are then ranked in descending order, from right to left, top to bottom, according to their degree of similarity to the reference image, which is determined by the absolute difference between their projections. Only the top 19 matches are shown to conserve space. The recall times and matching hits for each set of results are shown in the figure labels.

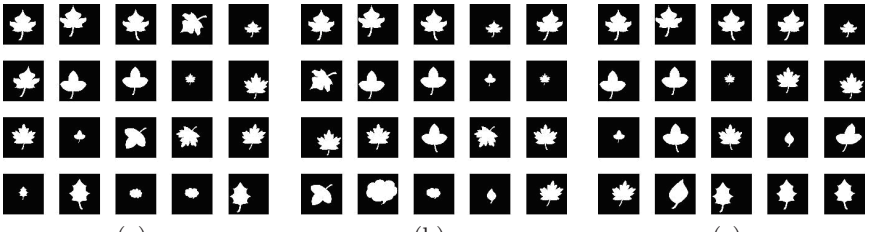
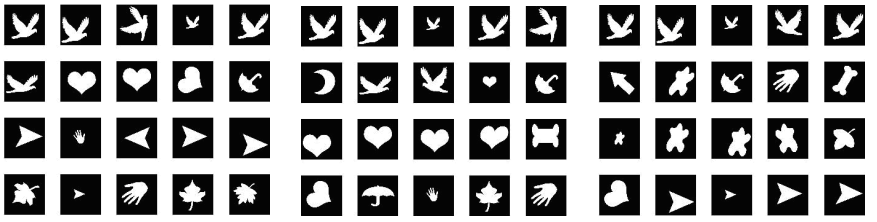


Fig. 3. Recall results using the 'leaf' image using method 1 (a), 2 (b), and 3 (c). For methods 1, 2 and 3, recall time = 195.9, 69.0 and 56.8 seconds, hits = 6, 6, and 5 in top 10; hits =12, 12, and 15 in top 19, respectively.

Comparing methods 1 and 2, our method achieves the same or better accuracy as method 1, whereas its computational time used is only about 35% to 52% (using any of 100 images alternatively) when $\Delta\theta = 5°$ is used. By using fewer projections for comparison purposes, the computational time saved is significant and vital in the construction of a good image retrieval system. Comparing the proposed method to the case $\Delta\theta = 90°$, one can find that using only horizontal and vertical projections is highly inefficient. Therefore, sufficient projections have to be taken, increasing computational times of 17% to 42% for the images. Fig. 4 shows the retrieval results of another example comparing the three methods. Again, local RT can maintain comparable accuracy with a lower time cost.



(a) (b) (c)

Fig. 4. Recall results using the 'bird' image using method 1 (a), 2 (b), and 3 (c). For methods 1, 2 and 3, recall time = 194.9, 88.3 and 69.1 seconds, hits = 6, 7, and 5 in top 10, respectively.

Next, we test our method in retrieving complicated images. The database (see Fig. 5) contains 100 images and each of the images has between 2 to 4 objects in them, and some of these objects can be broken down into several connected components. Most of the different types of objects can be found in various images, allowing the testing of retrieval accuracy. Each of the objects may have been subjected to flipping, translational, rotational, scaling, and shearing and other transformations.

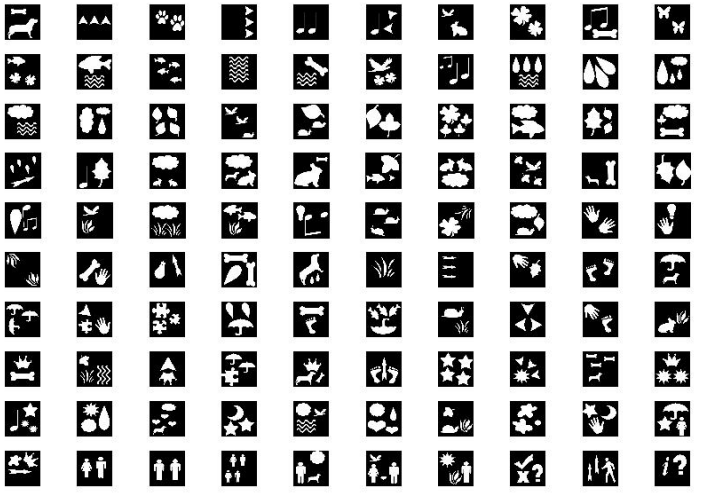


Fig. 5. Image database for multiple object matching

A query image is compared with all the other images in the database. The closest 19 matches are displayed and ranked according to their degree of similarity, with the query shown in the top left corner. Fig. 6 shows two sets of results. It is seen that by using EMD, perceptual similarity is matched reasonably well. Riding on the translation, rotation and scale invariance feature of the distance measure, the EMD algorithm manages to retrieve images which contain similar

objects to the query, even though the objects have been subjected these various forms of transformations. The EMD measure also succeeds in ranking images which have objects that match more of the query's objects more closely than those that do not.

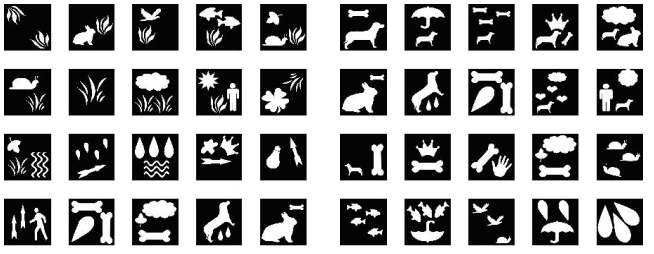


Fig. 6. Two retrieval results. The top-left is the query image. In (a) query contains 2 'grass' objects; in (b), the query contains a 'bone' object and a 'dog' object.

However, on the downside, the EMD measure may ascertain a degree of bias towards images which have the same number of objects as the reference image. While in most cases, this feature allows for better grouping of perceptually similar images, certain images which have the same number of objects but have objects different from those in the reference image may be ranked higher than those that have more matching objects but have a disparity in the number of objects that the image possess. One other problem is that even though an image has a few matching objects, if the image contains a vastly dissimilar object from the objects in the reference image, the image may be ranked disproportionally lower, in contrast to an observer's notion of perceptual similarity.

6 Conclusion

The Radon transform can preserve structure spatial relations and distributions. Invariant features can be designed from it for image retrieval. We have extended existing work using the RT for image retrieval. Taking into consideration the problem of inhomogeneous resolutions in RT due to the discrete nature in which projections are taken, a method is proposed to perform RT locally on individual objects by aligning the centre of rotation of the RT with the centroids of the objects. Work has also been done to determine the minimum number of projections to use to sufficiently represent an object, by analyzing the frequency characteristics of the object and its physical size, allowing for a reduction in redundancy and saving computational time and space. The RT projections of objects are then compared in a manner which maintains translational, rotational and scaling invariance. By taking into account the presence of multiple objects in images, the EMD method is utilized for matching purposes.

Experiments show that by performing local RT and using the sufficient number of projections, accurate and fast image retrieval can be performed. However, considering the high sensitivity of the image matching process, only objects which have very similar geometric structure can be recalled. Therefore, we could improve our work using a good segmentation algorithm to extract objects of interest in future projects.

References

1. Chang, N.S., Fu, K.S.: Query-by-pictorial-example. IEEE Transactions on Software Engineering SE 6(6), 519–524 (1980)
2. Flickner, M., Petkovic, D., Steele, D., Yanker, P., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D.: Query by image and video content: The QBIC system. IEEE Computer 28(9), 23–32 (1995)
3. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
4. Rui, Y., Huang, T.S., Chang, S.F.: Image retrieval: Current techniques, promising directions, and open issues. Journal of Visual Communication and Image Representation 10, 39–62 (1999)
5. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communication and Applications 2(1), 1–19 (2006)
6. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recogn. 40(1), 262–282 (2007)
7. Rubner, Y.: Texture-based image retrieval without segmentation. In: Proceedings of the International Conference on Computer Vision, pp. 1018–1024. IEEE Computer Society, Los Alamitos (1999)
8. Deans, S.R.: The Radon Transform and Some of Its Applications. John Wiley & Sons, New York (1983)
9. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall, Englewood Cliffs (1989)
10. Matus, F., Flusser, J.: Image representations via finite radon transform. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(10), 996–1006 (1993)
11. Al-Shaykh, O.K., Doherty, J.F.: Invariant image analysis based on radon transform and svd. IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing 43(2), 123–133 (1996)
12. You, J., Lu, W., Li, J., Gini, G., Liang, Z.: Image matching for translation, rotation and uniform scaling by the radon transform. In: ICIP 1998. Proceedings of 1998 International Conference on Image Processing, pp. 847–851 (1998)
13. Wang, H., Guo, F., Feng, D.D., Jin, J.S.: A signature for content-based image retrieval using a geometric transform. In: Proceedings of ACM Multimedia 1998, Bristol, UK, pp. 229–234 (1998)
14. Guo, F., Jin, J.S., Feng, D.D.: A measuring image similarity using the geometrical distribution of image contents. In: Proceedings of International Conference on Signal Processing 1998, vol. 2, pp. 1108–1112 (1998)
15. Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: Proceedings of Sixth International Conference on Computer Vision, pp. 59–66 (1998)