INVARIANT CONTENT-BASED IMAGE RETRIEVAL USING THE FOURIER-MELLIN TRANSFORM[1]

Ruggero Milanese, University of Geneva, Switzerland
Michel Cherbuliez, Swiss Federal Institute of Technology, Switzerland
Thierry Pun, University of Geneva, Switzerland

ABSTRACT

We describe a method for computing an image signature, suitable for content-based retrieval from image databases. The signature is extracted by computing the Fourier power spectrum, performing a mapping from cartesian to logarithmic-polar coordinates, projecting this mapping onto two 1D signature vectors, and computing their power spectra coefficients. Similar to wavelet-based approaches, this representation is *holistic*, and thus provides a compact description of all image aspects, including shape, texture, and color. Furthermore, it has the advantage of being invariant to 2D rigid transformations, such as any combination of rotation, scaling and translation. Experiments have been conducted on a database of 2082 images extracted from various news video clips. Results confirm invariance to 2D rigid transformations, as well as high resilience to more general affine and projective transformations. Moreover, the signature appears to capture perceptually relevant image features, in that it allows successful database querying using example images which have been subject to arbitrary camera and subject motion.

1 INTRODUCTION

Recent advances in storage, computing and communication technology have created the need for user-friendly access to large and distributed multimedia repositories. Among the most demanding users of this technology are multimedia content providers, such as television studios, which increasingly employ digital media for all production and postproduction phases. Efficient handling of multimedia data also opens new applications in other markets, such as the corporate document management, the distribution of news and accompanying image and video data by press agencies, and the home computing world.

In order to enable efficient and user-friendly access to multimedia archives, database management systems should provide two types of indexes. The first originates from textual descriptions that can be entered either manually by a documentalist, or automatically generated from close-captioning, or data files. The second concerns an automatically-extracted representation of the visual appearance of the image or video document, and should allow search by means of example documents (*content-based search*). By combining the two types of indexes, a user can efficiently explore the database by iteratively specifying the most interesting items retrieved through previous queries. For video material, the extraction of visual indexes requires several processing steps which decompose the

[1]The financial support of the Swiss Federal Government through OFES grant 95.0493, in the framework of the EU ACTS project AC089, is gratefully acknowledged.

clip into a hierarchy of uniform segments, and provide a number of keyframes representing each segment [4] [10]. For images (including video keyframes), the visual appearance that must be indexed is clearly multi-faceted, and may include information about the overall image layout, as well as about individual objects and their properties.

The rest of this paper is structured as follows. Section 2 provides a brief review of state-of-the-art technology in image indexing. Two main approaches are described, and their main weaknesses are outlined. This analysis motivates the need for new techniques that provide an integral, "holistic" representation of an image, that is invariant to various image transformations. In section 3 we describe the main contribution of this article, i.e. a method for representing images invariant to rotation, translation, and scale changes. In section 4 this method is experimentally validated, using a database of images extracted from two TV studio archives.

2 PREVIOUS WORK

Several methods have been proposed to represent image appearance for content-based retrieval (for a more extensive survey, the reader is referred to [4] [1] [2]). Two well-known systems are QBIC [5] and VIRAGE [6], since they have been integrated with commercial database products. Their underlying principle is the extraction of multiple image descriptions, representing various features. These include a representation of the image color distribution by histograms, and the extraction of uniform regions segmented from the image, likely to contain isolated objects. From these regions, various statistical descriptors are computed, as well as a representation of their spatial relationships. Other features include texture descriptors, such as co-occurrence matrices and the overall response of oriented band-pass filters, which can be useful to characterize outdoor images or regular surfaces. Other systems based on similar principles include Photobook [12], VisualSEEK [14], ImageRover [13], the UCSB Alexandria Digital Library Project [9], and Chabot [11].

One problem with the above disjoint multi-feature image descriptions is how to integrate them at query time. Since the average user is not likely to understand the underlying numerical representation of each feature, it is hard to determine the appropriate balance between them through weighting coefficients. Furthermore, features may exhibit different statistical properties from each other, and unless they are correctly equalized, small changes in the relative weights may lead to large variations of the top-ranked retrieved items. The user may thus spend considerable time adjusting weights at query time, and may eventually miss the desired images.

In order to overcome the above problems, alternative approaches have been proposed. These include integral orthogonal image transforms such as wavelet decomposition [8] [7]. The idea is to represent the image by a small number of meaningful coefficients that describe it at various resolution levels, and which also encode spatial relationships between various image components, by virtue of the space-frequency localization property of the wavelet transform. Meaningful coefficients can be extracted through vector quantization procedures, or by selectively picking those with the largest magnitude. Although these methods successfully address the multiple-feature integration problem, they suffer from a major weakness, namely their lack of invariance to translation and rotation, and, to a lesser extent, to scale changes. If a set of images of the same subject contains object motion, or is captured with a moving camera, it will therefore be difficult to retrieve all images of the set, given one sample.

In this paper, we propose a method that is also based on an orthogonal transform (the Fourier transform), but which offers invariance to geometric transformations. In particular, it is designed to be fully invariant to 2D rigid transformations, i.e. a combination of 2D rotation, translation and scaling (RTS). In general, the changes induced by object and camera motion are not restricted to this type of transformation (cf. [3] for an alternative representation that is invariant to affine and projective transformations). However, experimental results from a large database of heterogeneous videos show that it provides a good approximation to invariance under more general, non-rigid transformations.

3 THE PROPOSED METHOD

We propose an image representation invariant to 2D rigid transformations, i.e. to a combination of 2D rotation, translation, and scaling (RTS). Since the representation is non-invertible, it will also be referred to as an image *signature*. The construction process is incremental, and starts with the computation of the discrete Fourier transform (DFT) F of the image I:

$$F(m,n) = \sum_{x=\frac{-M}{2}}^{\frac{M}{2}-1} \sum_{y=\frac{-N}{2}}^{\frac{N}{2}-1} I(x,y) \mathrm{e}^{-2\pi i(\frac{mx}{M}+\frac{ny}{N})} \tag{1}$$

where $m = \frac{-M}{2}, \ldots, \frac{M}{2}-1$, $n = \frac{-N}{2}, \ldots, \frac{N}{2}-1$.

3.1 TRANSLATION INVARIANCE

The DFT can be represented by the phase and power spectra Φ, P:

$$\Phi(m,n) = \mathrm{atan}\frac{\mathrm{Im}[F(m,n)]}{\mathrm{Re}[F(m,n)]} \tag{2}$$

$$P(m,n) = \sqrt{\mathrm{Re}[F(m,n)]^2 + \mathrm{Im}[F(m,n)]^2}. \tag{3}$$

Since the image I is real, $\mathrm{Re}[F]$ is even and $\mathrm{Im}[F]$ is odd symmetric around the origin. In other words, $F(-m,-n) = F^*(m,n)$. If the original image I is circularly translated into $\tilde{I}(x,y) = I(x+x_x, y+y_0)$, its DFT will be given by $\tilde{F}(m,n) = F(m,n) \cdot \mathrm{e}^{-2\pi i(\frac{x_0}{M}+\frac{y_0}{N})}$. Therefore the power spectrum of the translated image will be $\tilde{P} = P$, while the phase spectrum will be altered. In order to build a translation-invariant representation, we retain the power spectrum, and discard the phase spectrum.

By virtue of the power spectrum's even symmetry, one only needs to retain the coefficients $P(m,n)$ in the two quadrants $\{(m,n), m \geq 0, n = \frac{-N}{2}, \ldots, \frac{N}{2}-1\}$, thereby reducing storage needs by half. It should be noted that, albeit shift invariant, this initial representation involves a loss of information and is therefore not invertible.

3.2 ROTATION AND SCALE INVARIANCE

In order to achieve invariance to rotation and scaling transformations also, we perform further non-linear and non-invertible operations on the power spectrum P. The first is a coordinate transformation, from cartesian (m,n) to logarithmic-polar (ρ,ϑ), where $\rho = \log\sqrt{(m-m_0)^2+(n-n_0)^2}$, $\vartheta = \mathrm{atan}\frac{n-n_0}{m-m_0}$, and (m_0,n_0) is the origin of the new coordinate frame with respect to the cartesian one.

Although such a coordinate change is bijective in the continuous domain, it is not so for the discrete case, where the coordinates of the transformed domain only assume the following values:

$$\rho = \frac{1}{5}r, \frac{2}{5}r^2, \ldots, \frac{64}{5}r^{R-1} \quad (4)$$

$$\vartheta = 0, \frac{\pi}{S}, \frac{2\pi}{S}, \ldots, \frac{(S-1)\pi}{S} \quad (5)$$

In our implementation, we used $r = 10^{\frac{1}{64}}$ for the 64 magnitude bins, and $S = 32$ orientation bins, appropriate for images of size 256×256. The difference between this mapping and the traditional polar one is in the use of the logarithm, which results in an exponential sampling frequency as a function of the distance from the origin (m_0, n_0). In order to remove the aliasing effect introduced by this subsampling process, a low-pass filtering operation is performed. Since the distance between samples varies, the cut-off frequency of the filter should decrease with the distance from the origin. We do so by filtering the domain P with a Gaussian filter before the log-polar mapping, according to the following space-varying convolution:

$$P'(m,n) = \frac{1}{2\pi\sigma^2(m',n')} \sum_{k,l} P(k,l) \cdot \exp[-\frac{(m'-k)^2 + (n'-l)^2}{2\sigma^2(m',n')}] \quad (6)$$

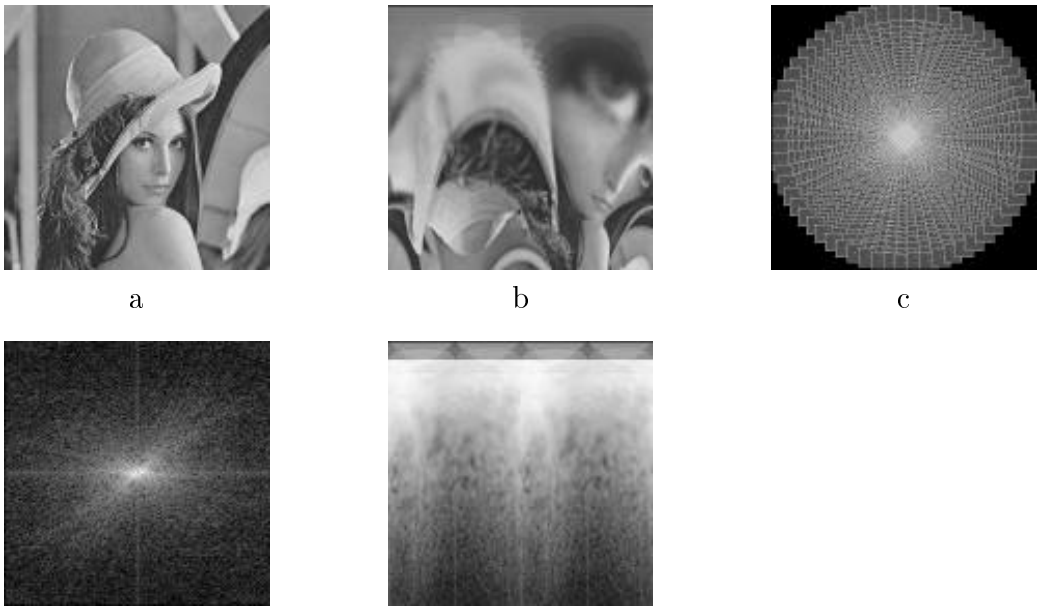
$$\sigma(m',n') = \max(a, b \cdot \log\sqrt{(m')^2 + (n')^2}) \quad (7)$$

$$(m',n') = (m - m_0, n - n_0) \quad (8)$$

where $a = 2.0$ is a constant guaranteeing a finite upper bound for the cut-off frequency to the low-pass filter for points close to the origin (m_0, n_0), and $b = 0.07$ is a scaling coefficient. For efficiency reasons, the convolution is computed by truncating the Gaussian coefficients outside a square window, whose size is a function of $\sigma(m', n')$. Figure 1 shows the log-polar representation, together with the underlying sampling grid, applied to both an input image (for illustration purposes) and its power spectrum. In both cases the origin of the mapping is located at the center of the images.

An important parameter of this transformation is the origin (m_0, n_0), since different values generally produce very different representations. When it is applied to the power spectrum of natural images, the most appropriate choice is to set $(m_0, n_0) = (0, 0)$, since the magnitude of the Fourier coefficients is known to follow an approximately negative exponential distribution, around the DC component. Furthermore, as pointed out above, P is even symmetric for real input images, and this choice of the origin also leads to savings in memory space.

Let $L(\rho, \vartheta)$ denote the log-polar representation of the power spectrum. The advantage of using L instead of P resides in some degree of invariance to rotation and scaling transformations of the input image I, on the top of the translation invariance, already achieved by P. Indeed, it is easy to show that a rotation of the image by an angle ϕ results in a rotation of the Fourier coefficients by the same angle. Therefore, the rotation produces a shift in the ϑ coordinate of L by ϕ. Similarly, a scaling transformation on the input image, i.e. the division of the coordinates (x, y) by a constant α, has the effect of multiplying the coordinates (m, n) of P by the same constant. Due to the logarithmic sampling of L, this multiplication reduces to a shift by $\log\alpha$ in the ρ coordinate of L.



a b c

d e

Figure 1: Illustration of the logarithmic-polar mapping. First row: (a) input image; (b) transformed image; (c) log-polar sampling grid. Second row: (d) Fourier power spectrum of image a; (e) transformed image.

To summarize, the proposed log-polar representation of the power spectrum is inherently invariant to image translations, and also provides invariance to rotation and scaling, up to a translation factor.

3.3 APPLICATION TO IMAGE ARCHIVAL AND RETRIEVAL

The representation L introduced above is invariant to translation and, up to a coordinate shift $(\log \alpha, \phi)$, also to scaling and rotation. In order to get rid of the latter shift, one may compute, once again, the power spectrum $\mathcal{L}$ of L, obtaining what is known as the Fourier-Mellin transform. This would make $\mathcal{L}$ completely invariant to the three transformations, and make it a candidate representation for comparing images in a database. The effective size of this representation, due to the symmetry properties described above, would be one half of L's size. For the sampling scheme described above (32×64 matrix L, appropriate for input images of size 256×256), this would produce an invariant signature vector of 1'024 coefficients. One may think of any distance function in this multidimensional feature space, such as the Euclidean distance, in order to assess the visual similarity between two images.

However, this straightforward approach presents some weaknesses. On one hand, the dimensionality of this feature space is clearly too large for efficient indexing. On the other hand, the feature vector is still too close to a global "photographic" representation of the input image, in the transformed space. Ideally, one should be able to assess the perceptual similarity between two images in a more "local" fashion, e.g. by distinguishing between the overall background of an outdoor picture, and its foreground objects. Since the basis functions of the DFT are not spatially localized, and all spatial information has been dropped with the phase spectrum, it is not possible to isolate coefficients which describe spatially-distinct components of an image. However, "foreground" objects are often defined over a limited frequency band (as opposed to the "background"). Therefore

one may envisage the use of some robust distance function which discounts occasional band-limited differences between two vectors. Still, the design of such a distance function for the representation $\mathcal{L}$ would be quite complex, since frequency-band and orientation information are merged together.

In order to overcome the above problems, rather than computing the power spectrum of L, we compute two 1-D signature vectors by projecting $L(\cdot,\cdot)$ onto the two axes ρ and ϑ. In other words, we compute the marginals

$$L_\rho(s) = \sum_\vartheta L(s,\vartheta) \tag{9}$$

$$L_\vartheta(t) = \sum_\rho L(\rho,t) \tag{10}$$

where ρ and ϑ vary as defined in eq. 5. However, the representation L is still RTS-invariant up to a shift, and so are the vectors L_ρ, L_ϑ. In fact, L_ρ is already invariant to rotation (thanks to the projection) but not to scaling, while the opposite holds true for L_ϑ. Both can be made fully RTS-invariant by computing their power spectra $\mathcal{L}_\rho, \mathcal{L}_\vartheta$. Figure 2 shows the signature vectors $\mathcal{L}_\rho, \mathcal{L}_\vartheta$ computed from two images, which have been RTS-transformed with respect to one another. It can be seen that the two vectors are indeed invariant, despite all subsampling and numerical approximation errors.

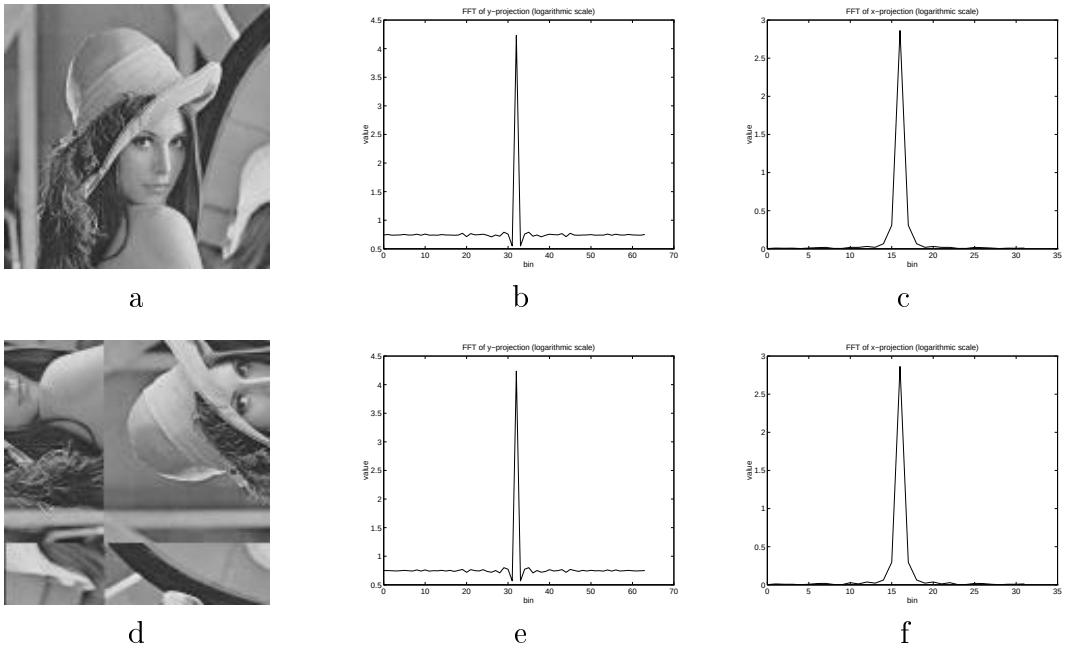


Figure 2: Example of invariant signature vector pairs extracted before (first row) and after (second row) RTS transformation. (a) Input image I; (b) vector $\mathcal{L}_\rho$; (c) vector $\mathcal{L}_\vartheta$; (d) transformed image I'; (e) transformed vector $\mathcal{L'}_\rho$; (f) transformed vector $\mathcal{L'}_\vartheta$.

On one hand, using these vectors as signatures dramatically reduces the dimensionality of the data, thereby facilitating indexing. Also, thanks to the symmetry of the power spectra of the two real-valued vectors, only half of the power spectra coefficients $\mathcal{L}_\rho, \mathcal{L}_\vartheta$ need to be retained. For the sizes used in our current implementation, this amounts to $\frac{1}{2}(32+64)+2 = 50$ coefficients. On the other hand, by separating frequency band magnitude from orientation, it is easier to design a robust distance function on the $\mathcal{L}_\rho$ vector that discounts (or identifies) band-limited objects.

It should be noted that the above signature vectors represent a gray–level image. For color images, three different pairs of vectors should be computed (one for each color plane), which we align into $3 \times 50 = 150$ dimensional vectors. It is appropriate in these cases to first perform a color mapping from RGB to another space, such as YIQ, where the luminance information (Y) is separated from the two chrominance channels.

4 EXPERIMENTAL VALIDATION

In order to validate any image representation for content-based retrieval, it is important to employ an experimental dataset for which a ground-thruth classification is available. However, it is very difficult (except for a few classes, such as faces, buildings, and landscapes) to create a collection of still pictures that can be clustered into disjoint classes without any subjective evaluation.

In order to overcome this problem, we decided to employ still pictures estracted from videoclips. We assumed that the action of uninterrupted recording from a video camera naturally implies a continuity in its visual content, with gradual changes from frame to frame due to camera operations and subject motion, including object appearance and disappearance.

In particular, we employed video clips of various news programs, all compressed in the MPEG-1 format, provided by the Greek television MegaChannel and by the Swiss television TSR. For each clip, we run a program that automatically segments a clip into *shots* by detecting scene cuts [10], and extracts a keyframe right after each cut. In addition to the keyframe, we also manually extracted from some shots a number of other frames in the middle and at the end of the corresponding shot, whenever object or camera motion was causing major changes between their visual appearance. The images classes were thus objectively defined by shot membership. In total, 1405 shots were extracted, providing a total of 2082 still images.

In order to determine the similarity between any two images I^i, I^j in the database, the simple Euclidean distance function between their signatures was used, and their average value was retained as the overall distance:

$$d(i,j) = \frac{1}{2}\left(\sqrt{\sum_s (\mathcal{L}^i_\rho(s) - \mathcal{L}^j_\rho(s))^2} + \sqrt{\sum_s (\mathcal{L}^i_\vartheta(s) - \mathcal{L}^j_\vartheta(s))^2}\right) \ . \tag{11}$$

In this way, all images from the database could be compared with any query image (also selected from the database), and ranked by the value of d. A fixed number W of top-ranked images could then be displayed to the user, enabling browsing through the database. Figure 3 shows sample sets of top-ranked images retrieved from two query images belonging to shots that present considerable camera and subject motion (query images correspond to the highest-ranked hits, reported in hte top-left corners). It can be seen that, even though camera and subject motion clearly violate the RTS transformation model underlying the image representation, the target images are always retrieved with high similarity.

Two statistical measures were computed to assess system performance. One is called *recall*, and consists of the proportion of target images (i.e. images from the same class as the query) that have been retrieved among the top W hits. This measure is clearly monotonically increasing with W and attains 100% when W includes the whole dataset. For simplicity we shall only report the value of recall for the number of images that could fit into a display window of size $W = 12$. The other measure is *precision*, and consists of

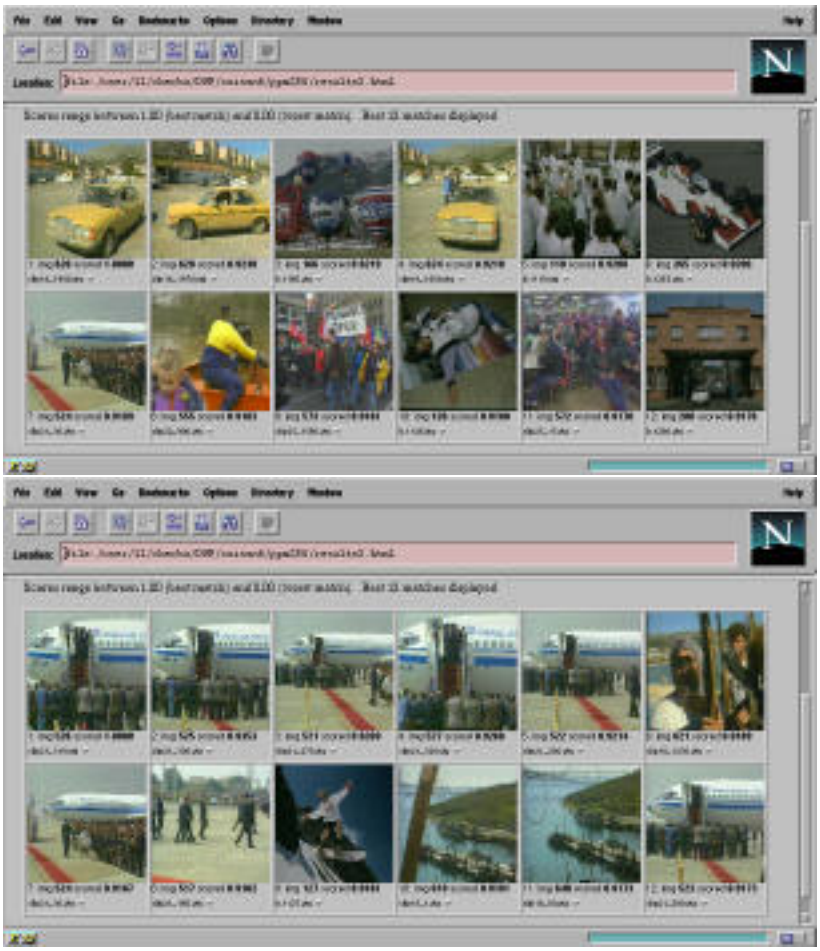


Figure 3: Results of database retrieval using the query image shown in the top-left corner of each window. Images are ranked from left to right, top to bottom, by decreasing similarity (increasing distance measure d).

the number of target images that are retrieved among the top W hits. However, since the image classes are very small (up to 7 images per class), we employed a slightly different definition of precision, namely the proportion of target images that are retrieved up to the last correct one. A high value of precision therefore indicates that the top-ranked hits *all* contain target images.

The average recall measure computed from 16 queries of this kind, matched against the whole database of 2082 images partitioned into 1405 classes was 66.7%. The average precision at 100% recall was 54.6%. These results confirm the robustness of the invariant signature well beyond rigid 2D transformation, and its applicability to the indexing of real images, captured with a moving camera. Indeed, the shots from which all query images where selected present complex tracking and dollying motion (in addition to pan/tilt/zoom producing RTS transformations) that would in principle require more complex and computationally time-consuming geometric models for signature extraction (cf. [3]). Furthermore, subject motion causes objects to enter and leave the camera field of view, thereby altering the image content.

As far as storage and CPU requirements are concerned, the proposed method refers to color images of size 256×256, for which it computes two 1D feature vectors, for a total of 192 coefficients. At archival time, the extraction of the signature vectors for a new image requires approximately 6 s, including file I/O, FFT calculations and the log-polar

mapping. At query time, the search through the database is performed linearly, which requires on average 399 μs per image comparison on a Sun Ultra-2 workstation. The full database of 2082 images is searched in 0.83 s.

5 CONCLUSIONS

In this paper we presented a method for constructing an image signature invariant to 2D rigid transformations, such as rotation, translation, and scaling (RTS), in order to provide a representation suitable for content-based image retrieval. This method requires the computation of the image power spectrum, performs a coordinate mapping from cartesian to logarithmic-polar, projects a 2-D representation onto two 1-D feature vectors, and computes their power spectra.

This signature has been shown to withstand more complex transformations than RTS, such as affine transformations induced by motion around the camera axis, as well as unrestriced camera and subject motion. The proposed representation has been shown to capture the perceptually relevant information of an image, since it provides good retrieval performance for a database of 2082 images extracted from a variety of news video clips. In comparison to methods extracting different representations and matching tools for color, texture, and shape features, it employs a holistic approach, and does not require the user to set appropriate weights to combine their results. With respect to other holistic approaches, such as those based on the wavelet transform, its invariant properties are a clear advantage.

Future work will improve the computational performance at query time, by indexing the feature vector and by using hierarchical search strategies. By virtue of the underlying holistic representation, it is also possible to introduce weights for fine-tuning the matching function, without compromising the interpretability of the search process by the user. These weights may, for instance, describe the relevance of color, or indicate the most relevant frequency band corresponding to an object or an image region selected by the user.

ACKNOWLEDGEMENTS

We would like to thank A. Jacot-Descombes for useful comments and F. Deguillaume for providing the shot segmentation programs used for the experimental validation. D. Squire proof-read the article and helped in evaluating system performance. Images and videos comprising the experimental dataset were kindly provided by the television studios Télévision Suisse Romande (Switzerland) and MegaChannel (Greece).

REFERENCES

[1] F. Idris and S. Panchanathan, Review of image and video indexing techniques. J. Visual Comm. and Image Repres. 1997, 8(2), pp. 146-166.

[2] M. De Marsico, L. Cinque, and S. Levialdi, Indexing pictorial documents by their content: a survey of current techniques. Image and Vision Computing 1997, 15, pp. 119-141.

[3] S. Startchik, R. Milanese and T. Pun, Projective and photometric invariant representation of planar disjoint shapes. Image and Vision Computing Journal, accepted for publication, 1998.

[4] B. Furht, S.W. Smoliar, H. Zhang, Video and image processing in multimedia systems. Kluwer Academic Publishers, 1995.

[5] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele and P. Yanker, Query by image and video content: the QBIC system. IEEE Computer, Sept. 1995, pp. 23-32.

[6] A. Gupta and R. Jain, Visual Information Retrieval. Communications of the ACM, Vol. 40, 5, May 1997.

[7] F. Idris and S. Panchanathan, Image indexing using wavelet vector quantization. Proc. SPIE Storage and retrieval of image and video databases III, Vol. 2420, Febr. 1995, pp. 373-380.

[8] C.E. Jacobs, A. Finkelstein, and D.H. Salesin, Fast Multiresolution Image Querying. Proc. Siggraph '95, ACM, New York, 1995.

[9] W. Y. Ma, Y. Deng and B.S. Manjunath, Tools for texture/color based search of images. Proc. SPIE Conference on Human Vision and Electronic Imaging II, 1997, Vol. 3106, San Jose, CA, pp. 496-507.

[10] R. Milanese, F. Deguillaume, and A. Jacot-Descombes, Video segmentation and camera motion characterization using compressed data. Proc. SPIE Conf. on Multimedia Storage and Archiving Systems II, Dallas (TX), Nov. 2-7, 1997.

[11] V.E. Ogle and M. Stonebraker, Chabot: retrieval from a relational database of images. IEEE Computer, Vol. 28, No. 9, Sept. 1995.

[12] A. Pentland, R.W. Picard and S. Sclaroff (1994). Photobook: tools for content-based manipulation of image databases. Proc. SPIE Storage and Retrieval for Image and Video Databases II, San Jose, CA, Feb. 1994. Vol. 2185, pp. 34-37.

[13] S. Sclaroff, L. Taycher and M. La Cascia, ImageRover: a Content-based Browser for the World Wide Web. Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, June 1997, San Juan, Puerto Rico.

[14] J.R. Smith and S.F. Chang, VisualSEEk: a fully automated content-based image query system. Proc. Fourth ACM International Multimedia Conference, Nov. 1996, Boston, MA, USA.