SHAPE RETRIEVAL WITH ROBUSTNESS AGAINST PARTIAL OCCLUSION

Michael Höynck, Jens-Rainer Ohm

{hoeynck, ohm}@ient.rwth-aachen.de
Aachen University (RWTH), Institute of Communications Engineering (IENT)
Melatener Strasse 23, 52072 Aachen, GERMANY

ABSTRACT

Object shape features are powerful when used in similarity search-&-retrieval and object recognition because object shape is usually strongly linked to object functionality and identity. Many applications, including those concerned with visual objects retrieval or indexing, are likely to use shape features. Those systems have to cope with scaling, rotation, deformation and partial occlusion of the objects to be described. The ISO standard MPEG-7 contains different shape descriptors, where we focus especially on the region-shape descriptor. Since we found the region-shape descriptor not being very robust against partial occlusion, we propose a slightly changed feature extraction method, which is based on central-moments. Further, we compare our method with the original region-shape implementation and show that, applying the proposed changes, the robustness of the region-shape descriptor against partial occlusions can be significantly increased.

1. INTRODUCTION & RECENT WORK

The amount of worldwide available audio-visual information has been growing rapidly during the past years. In this context, the strong need arises to develop automatic systems to categorize, search for, link and organize this information. Over the last decade research activities for multimedia content description systems have grown considerably. Recently, research activities are forced towards automatic content understanding techniques based on features like color, texture, shape or motion. Automatic content understanding systems could for instance guide and support a user by providing functionalities for search-&-retrieval, management, storage and selection of multimedia content.

Object shape features are very powerful when used in similarity search-&-retrieval and object recognition. This is because the object's shape is usually strongly linked to its functionality and identity, where this property distinguishes shape from other elementary visual features, such as color or texture [1]. Many applications, including those concerned with visual objects retrieval or indexing, are likely to use shape features.

The extraction of a *shape* can be basically performed manually or automatically. Manual segmentation requires a user to specify a certain region of interest, where the region can be, subsequently, described with respect to its color, texture or shape. Although manually guided segmentation is very time consuming, it still leads to the most robust results. Automatic shape extraction methods require some additional information for object region extraction. This can be done by applying a robust background-thresholding technique [2], where sophisticated algorithms regard more complex features like motion [3], disparity information delivered by a stereo camera system, radar or laser beams [4]. However, the problem of automatic object segmentation is not solved yet. Practical problems like partial occlusion of objects as well as non perfect segmentation results often lead to unsatisfying results.

After having determined object regions, certain distinctive shape features need to be extracted for shape description. This can be useful in search-&-retrieval applications, where a user draws the contour of an object based on which a query-by-sketch is performed. Obviously, additional features like color or texture must be regarded to eliminate false retrieval results, see Fig.1.

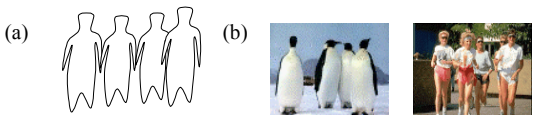


Fig.1. Example: user supplied contour (a) and s-&-r results (b).

Recently, the MPEG-group has developed the multimedia content description interface MPEG-7. Significant research activity for shape description has been carried out in the standardization process for the visual part. The standard contains several shape descriptors: the *region-shape*, which covers the shape information of binary masks, the *contour-shape*, for closed contour based shape description and the *shape3D* for 3-D shape based object matching.

We aim on a shape description, where the object can also be composed of a set of regions, e.g. caused by partial occlusions. We will focus on the region-shape descriptor throughout this paper, since it regards all pixels (being connected or not) constituting an object's shape. The

properties of the descriptor have been extensively tested and evaluated during the standardization process. Its invariance against scaling and rotation and its robustness against perspective- and non-rigid deformation were tested in well defined *Core Experiments (CEs)* with excellent retrieval results. The applied *test-sets* contain shape masks from various object-classes with several examples, where different scaling, rotation and deformation had been applied to [1].

Besides this, an application of a shape description for search-&-retrieval requires the applied descriptor to be robust against partial occlusion of the object under investigation. In practice, occlusions can be caused by foreground objects overlapping the object under investigation, or by segmentation artifacts that are not avoidable during the image segmentation procedure.

In this paper, we study the robustness of the region-shape descriptor against partial occlusions. We further propose a change in the descriptor extraction method, which increases its robustness against partial occlusion, while retaining its excellent invariance against rotation and scale, and its robustness against perspective and non-rigid deformation at the same time. In section 2, we describe the MPEG-7 region-shape descriptor and its extraction method. The proposed changes, which allow more robustness of the descriptor in case of partial occlusion, are explained in Section 3. We present our simulation results in section 4 and summarize this paper in section 5.

2. REGION-SHAPE DESCRIPTOR

In general, the shape of an arbitrary object may consist of either one single region or a set of regions; as well some holes might be contained in the object under investigation. A description of an object by its contour only may not be sufficient in any case.

2.1. The Angular Radial Transform

The *Angular Radial Transform (ART)* is utilized to extract the region-shape feature, where only a low order of computational complexity is required [5]. The two dimensional complex transform is defined on a unit disk in polar coordinates:

$$F_{nm} = \int_0^{2\pi} \int_0^1 V_{nm}^*(\rho,\theta) \cdot f(\rho,\theta)\, \rho d\rho d\theta \ , \qquad (1)$$

where F_{nm} is an *ART* coefficient of order n and m, $f(\rho,\theta)$ is a binary object mask in polar coordinates, and $V_{nm}(\rho,\theta)$ is an *ART* basis function. The *ART* basis functions are separable along the angular and radial directions and defined as follows [5]:

$$V_{nm}(\rho,\theta) = A_m(\theta) \cdot R_n(\rho) \ , \qquad (2)$$

where

$$A_m(\theta) = \frac{1}{2\pi} \exp(jm\theta) \ , \qquad (3)$$

$$R_n(\rho) = \begin{cases} 1 & n = 0 \\ 2 \cdot \cos(\pi \cdot n \cdot \rho) & n \neq 0 \end{cases} . \qquad (4)$$

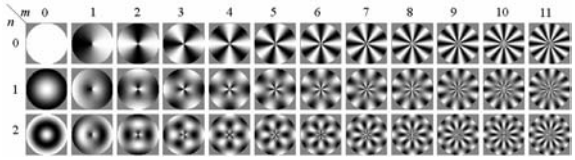


Fig. 2. Real parts of the ART basis functions.

2.2. Extraction Procedure

Due to reasons of simplicity and a fast implementation, the calculation is directly performed in Cartesian coordinates rather than converting the image and basis functions into polar coordinates. This can be done, since the basis functions are separable, as shown in equation (2).

First, the set of *ART* basis functions $V_{nm}(x,y)$ is computed and stored in a lookup-table *(LUT)*. In the next step, the binary object-mask is normalized. The center of mass *(centroid)* of the shape-mask is aligned to coincide with that of the *LUT*. If the sizes of the mask and *LUT* are different, linear interpolation is applied to map the mask onto the corresponding *LUT*. Here, the size of the object is defined as twice the maximum distance from the centroid of the object, which corresponds to the radius of the minimum circle around the object, centered in the centroid of the object. The real and imaginary parts of the *ART* coefficients are computed by summing up the multiplication of a pixel in an image to each corresponding pixel in the *LUT*, in raster scan order. After normalization and quantization, an arbitrary shape is characterized by 35 coefficients in 4 bit resolution, which results in 17.5 bytes for the description.

3. PROPOSED EXTRACTION METHOD

We carried out extensive experiments concerning the robustness of the region-shape description against partial occlusion. During our experiments we discovered that the *region-shape feature (RSF)* is not robust against occlusions. We found one of the reasons for this in the extraction procedure (see subsec. 2.2), where the object area is mapped onto the *LUT,* which contains the *ART* basis functions. The final mapping of the object pixels depends on the minimum radius around the shape-centroid, where the selection of the radius is essential for the shape description; see an example in Fig. 3.

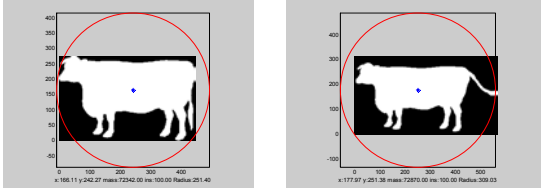


Fig.3. Original extraction method: radius of minimum circle around centroid (*); left: 251 pixels; right: 309 pixels.

The original shape-mask (*cattle-14.tif, CE1-B)* on the left is modified by lifting the tail. Although only small changes were made, the chosen radii around the shapes differ significantly (23 %). A matching of the resulting region-shape feature vectors (see Tab.1) may not work reliably in practice.

RSF	1	2	3	4	5	6	7	8	9	10	11	12	13	…
left	10	15	2	4	2	15	14	15	8	7	9	11	11	…
right	15	15	2	2	4	15	6	15	9	6	5	9	8	…

Tab.1. *RSF*-vectors of left/right shape in Fig.3, *sum of absolute differences (SAD)* between them is 93 (inv. quant. 1.0003).

We study the application of central-moments for the selection of appropriate radii in order to obtain a more robust region-shape feature.

3.1. Central-Moments

As we have shown above, even if only small differences exist between the shapes, choosing the image area with respect to the minimum circle around the shape-centroid can cause significant differences in the region-shape feature. We propose to apply central-moments for the selection of the mapping radius, which can be applied to characterize the geometrical shape of a visual signal in the space domain. The coordinate related central-moment of order *k+l* of a discrete 2-D signal *x(m,n)* is defined as:

$$\mu^{(k,l)} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} m^k n^l \cdot x(m,n) \quad . \tag{5}$$

The shape-centroid's coordinates are calculated as:

$$\overline{r} = \frac{\mu^{(1,0)}}{\mu^{(0,0)}} \quad ; \quad \overline{s} = \frac{\mu^{(0,1)}}{\mu^{(0,0)}} \quad . \tag{6}$$

The centroid-centered central-moments of order *k+l* are:

$$\rho^{(k,l)} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} (m-\overline{r})^k (n-\overline{s})^l \cdot x(m,n) \cdot \tag{7}$$

The *Eigenvalues* derived from central-moments of order 2 are rotation-invariant and express the variance, i.e. the spread of the shape along its principal axes [6]. Central-moments of order 2 can be interpreted as ellipsoids having axes-lengths λ_1 and λ_2, positioned around the shape centroid, where λ_1 and λ_2 can be derived as follows:

$$\lambda_{1,2} = \frac{1}{2}\left[\rho^{(0,2)} + \rho^{(2,0)} \pm \sqrt{\left(\rho^{(0,2)} - \rho^{(2,0)}\right)^2 + 4 \cdot \left(\rho^{(1,1)}\right)^2}\right] \cdot \tag{8}$$

The first Eigenvalue λ_1 describes the length of the taller axis of the ellipsoid (8). Since the subsequent mapping step regards a circular image area, we propose to select the radius of the circle in dependence of λ_1. We use a scaling factor *f* to ensure, that the whole shape is regarded in the feature extraction process, where we determined the optimal *f,* with respect to all shapes contained within the test-sets.

3.2. Proposed changes

The change in the extraction method (see subsec. 2.2) influences the selection of the radius for mapping the image area onto the *ART* basis functions, where the other parts of the region-shape feature extraction method remain unchanged. Fig. 4 shows the same example like Fig. 3, but in contrast we perform the selection of the mapping radius for feature-extraction based on central-moments. The small changes applied to the shape cause a difference in the selected radius of only 1.8 %, compared to the unchanged shape. This is in fact less than one twelfth of the difference that can be achieved using the original implementation, where the complexity for the calculation of central-moments is comparable to that of the original method.

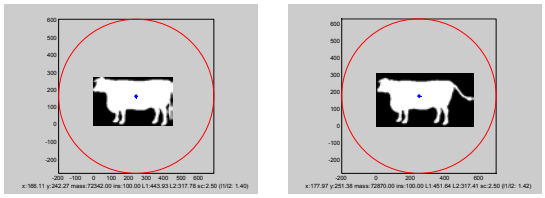


Fig.4. Proposed method: radius chosen based on λ_1 , *(f=2.5)*; left: 444 pixels; right: 452 pixels (* centroid)

RSF	1	2	3	4	5	6	7	8	9	10	11	12	13	…
left	15	15	2	3	5	15	12	15	8	5	7	11	5	…
right	15	15	2	3	4	15	12	15	9	4	8	10	6	…

Tab.2. RSF-vectors of left/right shape in Fig.4, where SAD between them is 44 (inv. quant. 0.4342)

4. SIMULATION RESULTS

4.1. Testing Procedures

For the evaluation of the proposed extraction methods, we use the test-sets that were defined during the standardization process of MPEG-7, based on which we test our proposed extraction methods in terms of robustness of the resulting features against scale, rotation, deformations and non-rigid transformations of the objects.

For testing the robustness against partial occlusion we generated another test-set, based on test-set *CE1-B*, which contains 1400 objects from 70 different object-classes. We occluded the shapes deterministically applying different percentages of occlusion (see Fig.5) from several directions (see Fig.6) to them, in raster-scan order.

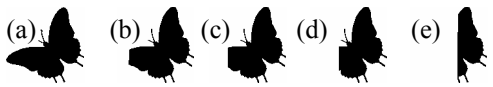


Fig.5. From left hand side occluded objects; (a) original *butterfly-3.tif*; by (b) 5, (c) 10, (d) 20 and (e) 50 % occluded shape.

Fig.6. By 10 % occluded object; (a) original *(butterfly-3.tif)*, occluded from (b) left, (c) right, (d) top, (e) and bottom.

4.2. Performance Measure

As shown above, partial occlusion causes a distortion of the resulting *RSF*. In practice, the classification of shapes is usually performed by comparing the *RSF*-vector of an object under investigation with feature vectors that have been already (manually) classified. The closer two feature vectors are in feature space (e.g. regarding *SAD*, [5]), the more similar the shapes are supposed to be. Since we do not have any information about the specific shape-feature database to be used for matching, we are reasoning: *the smaller the distortion D caused by the occlusion of the shape feature - the more robust a classification of the object can be performed in practice*, and define the following distortion-measure:

$$D = \sum_{i=0}^{34} \left| M_{orig}(i) - M_{occl}(i) \right| , \qquad (9)$$

where D is the *SAD* between original (M_{orig}) and occluded RSF-vector (M_{occl}). As described above, this measure is further determined for each shape within the test-set.

Finally, we determine minimum, maximum and mean distortion D for each of the applied occlusion percentages, where we average the four different directions. This results in three values (min, max and mean) expressing the distortion caused in feature space for a specific percentage of occlusion.

4.3. Results

Using the test-sets described in subsection 4.1, we evaluated the performance of our *central-moment (CM)* based method, where we conducted the same experiments that were carried out during the standardization process of the MPEG-7 region-shape descriptor [1]. The results prove (see Tab.3) that using CM-based scaling, the shape-description retains, in overall, its robustness against scaling, rotation and diverse deformations of the shape.

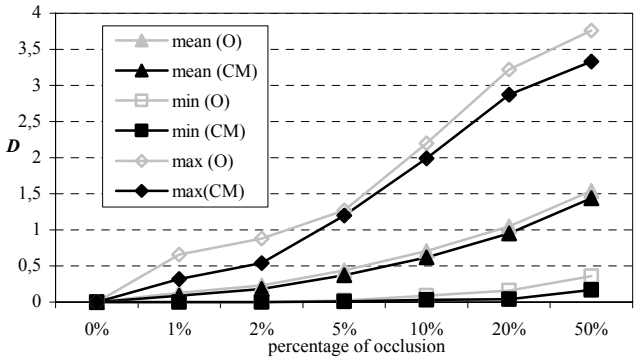
CE		original	CM-based scaling	Difference
1	A-1	97.70 %	97.82 %	+0.12 %
	A-2	100.0 %	100.0 %	0.00 %
	B	68.34 %	67.59 %	-0.75 %
	C	92.00 %	91.50 %	-0.50 %
2	A-1	86.20 %	86.80 %	+0.60 %
	A-2	98.06 %	99.39 %	+1.33 %
	A-3	98.54 %	97.44 %	-1.10 %
	A-4	62.65 %	63.25 %	+0.60 %
	B	66.13 %	63.04 %	-3.09 %

Tab.3. CE retrieval results *(f = 2,5)*, for details see [1].

At the same time, applying the CM-based region-shape-feature extraction method significantly increases the robustness of the descriptor against partial occlusions (see Tab.4). The mean distortion (as well as the minimum and maximum distortion) decrease for all given percentages of occlusion (Tab.5).

Occlusion	original scaling			CM-based scaling		
[%]	mean	min	max	mean	min	Max
0	0,00	0,00	0,00	0,00	0,00	0,00
1	0,13	0,00	0,66	0,09	0,00	0,32
2	0,23	0,00	0,88	0,18	0,00	0,54
5	0,44	0,02	1,27	0,37	0,01	1,20
10	0,71	0,09	2,20	0,62	0,03	1,99
20	1,05	0,16	3,22	0,95	0,04	2,87
50	1,54	0,36	3,76	1,44	0,17	3,33

Tab.4. Results: distortion on test-set *(f=2,5)*, for details see 4.1.



Tab.5. Distortion Results; comparison of original (O) and central-moment (CM) based shape feature extraction approach.

5. CONCLUSION AND OUTLOOK

In this paper, we have studied the robustness of the MPEG-7 region-shape descriptor, where we especially focused on robustness against partial occlusion. Since we found the descriptor not being very robust against occlusions of the shape, we propose a modification in the feature extraction method of the descriptor. Our results prove, that due to these changes, the robustness of the description against partial occlusions can be significantly increased, while retaining the excellent robustness of the original implementation at the same time.

6. REFERENCES

[1] B. S. Manjunath, P. Salembier, T. Sikora. *Introduction to MPEG-7*. J. Wiley & Sons, Inc., 2002.

[2] L. G. Shapiro; G. C. Stockman. *Computer Vision.* Prentice Hall, 2001.

[3] C. Kim; J. Hwang. Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications. IEEE Transactions on CSVT, Vol.12 No.2, Feb. 2002.

[4] L. Zhao; C. E. Thorpe. Stereo- and Neural Network-Based Pedestrian Detection. IEEE Transactions on ITS, Vol.1, No.3, Sep. 2000.

[5] ISO/IEC JTC1/SC29/WG11. 15938-3 FDIS. Information-Technology - Multimedia content description interface - Part 3 Visual. Doc. N4358. Sydney. Jul. 2001.

[6] W. K. Pratt. *Digital Image Processing*. J. Wiley & Sons, Inc., 2nd edition, 1991.