HUMAN-INSPIRED SHAPE-BASED IMAGE RETRIEVAL USING WEIGHTED CITY-BLOCK DISTANCE AND FOURIER DESCRIPTORS

Emir Sokic, Delila Halac, Samim Konjicija

Faculty of Electrical Engineering, University of Sarajevo
Bosnia and Herzegovina
esokic@etf.unsa.ba

ABSTRACT

Contour-based Fourier descriptors are very simple and effective shape description method used for content-based image retrieval. Similarity between Fourier descriptors is usually computed using measures such as City-block or Euclidean distance. These similarity measures consider all harmonics to be equally important, therefore harmonics with larger magnitude tend to have larger significance during the computation of shape similarity. In order to increase the importance of harmonics with lower magnitude, we propose to use weighted City-block distance for computing shape similarity. The proposed weighting coefficients are inspired by the contrast sensitivity of the human visual system to different spatial frequencies, known as the Contrast Sensitivity Function (CSF). Although weighted distances generally do not improve the retrieval performance, experimental results clearly demonstrate that human observers favour the retrieval system based on the weighted distances, and find it more accurate and relevant.

Index Terms— Content based image retrieval, Fourier descriptors, weighted distance, contrast sensitivity function, frequency

1. INTRODUCTION

Shape is widely used as a discriminative element in the field of content-based image retrieval (CBIR). In many applications, shape captures the most of the perceptual information of the observed objects on images. A variety of shape description techniques have been developed over the years [1, 2].

Fourier descriptors (FD) are established as compact shape descriptors, well known for their low computational complexity and relatively high retrieval performance. Although Fourier descriptors may be used as a region-based method [3], they are quite more often used as a global contour-based shape description technique [4–6]. Fourier descriptors are computed by first applying the Discrete Fourier transform over the shape signature function (such as Perimeter area function [4], Farthest distance function [6], Complex coordinates [5] etc.), and subsequently applying different procedures in order to the achieve invariance under translation, rotation, change of scale and/or starting point of the contour. Fourier descriptors have a hierarchical representation in frequency domain, where low frequency harmonics contribute to coarse description of the shape, while high frequency harmonics contain details and/or noise.

The similarity between shapes is commonly measured using City-block or Euclidean distance between their corresponding Fourier descriptors [4–6], therefore harmonics with larger magnitude tend to have larger significance during the computation of shape similarity. Since the low frequency harmonics usually have larger magnitude, middle and high frequency components hardly contribute in shape matching. It may seem technically correct to reduce the importance of high frequency components, since they are more susceptible to noise and carry (probably irrelevant) details. However, the magnitude and frequency of the harmonics are not linearly correlated to their semantical contribution, as it will be shown in the paper. This is in fact in accordance with the human visual perception, since the human visual system is most sensitive in detecting contrast differences occurring at "middle" frequencies. Therefore, a computer-based shape retrieval system performance may prosper if the shape difference analysis is focused around middle frequencies instead of the low frequencies. In order to exploit this effect, we propose to use a weighted City-block distance. Experimental results do not show significant improvement of retrieval performance, but human users assess the weighted distance-based retrieval system to be more accurate, and more suitable to human shape perception.

The paper is organized as follows. Section 2 gives a brief introduction to contour-based Fourier descriptors. Contrast sensitivity function is explained in Section 3. Section 4 introduces three proposed weighting schemes. Methodology and experimental results are given in Section 5 and 6, while the conclusion and guidelines for future work are given in the last section.

2. CONTOUR-BASED FOURIER DESCRIPTORS

The shapes that are analyzed in this paper can be described as single plane closed (discrete) curves. In preprocessing stage, the coordinates of the shape boundary are extracted from the image, and re-sampled with the fixed number of points N using equal arc-length sampling. The re-sampled points of the contour $P_n = (x_n, y_n)$ $n = 0, 1, ..., N-1$ may be represented using Complex coordinates shape signature [5]:

$$Z_n = x_n + jy_n. \tag{1}$$

The Discrete Fourier Transform is computed using:

$$a_k = \frac{1}{N} \sum_{n=0}^{N-1} Z_n e^{-j2\pi nk/N}, \tag{2}$$

where $k = 0, 1, ..., N-1$. Fourier coefficients a_k are used to compute FD, therefore they must be additionally transformed in order to be invariant under translation, rotation, scale and starting point change. Invariance under rotation and starting point change is easily obtained using only the magnitude of the Fourier descriptors, while invariance under translation is achieved by disregarding the coefficient a_0. In order to introduce scale invariance authors in [5] proposed to use an effective scaling coefficient $Sc = \sum_{i=1}^{N-1} |a_i|$. Therefore, a Fourier descriptor is given with:

$$\mathbf{F} = \left\{ \frac{|a_{-M/2}|}{Sc}, ..., \frac{|a_{-1}|}{Sc}, \frac{|a_1|}{Sc}, \frac{|a_2|}{Sc}, ..., \frac{|a_{M/2}|}{Sc} \right\}, \tag{3}$$

where M is the chosen number of Fourier coefficients (M is smaller than N). It is important to note that the discrete Fourier transform is a periodic discrete sequence (with period N), which explains the notation $a_j = a_{j+N}$ (for $j = -M/2, ..., -1$) in equation (3). In order to use a simpler notation, a substitution $f_i = |a_i|/Sc$ ($i = 1, 2, ..., M/2, N - M/2, ..., N-1$) is used, while f_0 is introduced and set to zero. Also, we introduce a measure of the "energy" of the shape reconstructed with M coefficients as:

$$E_M = \sum_{i=-M/2}^{M/2} |f_i|. \tag{4}$$

It is important to note that the nominal energy of the shape is equal to one ($E_{(M=N)} = 1$).

During matching stage, two Fourier descriptors $\mathbf{F^I} = \{f_i^I\}$ and $\mathbf{F^{II}} = \{f_i^{II}\}$ based on Complex coordinates shape signature are compared using City-block distance (Euclidean distance is used for other shape signatures):

$$d(\mathbf{F^I}, \mathbf{F^{II}}) = \sum_{i=-M/2}^{M/2} |f_i^I - f_i^{II}|. \tag{5}$$

The problem with this simplistic approach is that low frequency components f_i ($|i| \leq 5$) have considerably larger magnitude than the high frequency components, therefore they tend to be more important for shape discrimination. To support this fact, boxplot analysis of the first 20 coefficients f_i of 1400 shapes from MPEG-7 CE-1 Set B [7, 8] [1] is given in Figure 1. Low frequency components contain the coarse description of the shape, while the high frequency components append details and eventually noise. This can be seen in Figure 2a)-e), where a shape is reconstructed from M lowest frequency coefficients ($f_{-M/2}, ..., f_{M/2}$) for $M = 4, 8, 16, 32, 64$. It is interesting to note that the shape given in Figure 2a) has almost 70% of the energy of the initial shape, but its resemblance with the original shape is rather low. This means that the magnitude of the harmonic is not directly correllated with the semantical contribution of that component. Therefore we propose to use the weighted City-block distance:

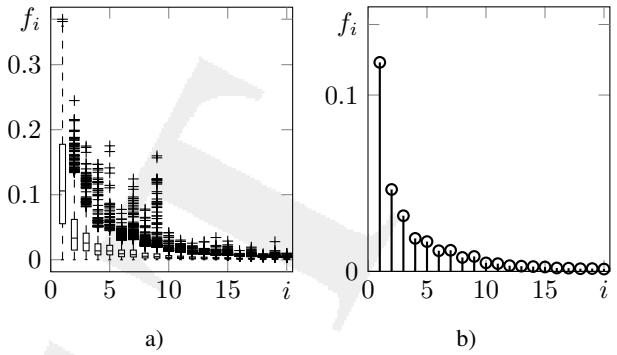

Fig. 1. Magnitude analysis of Fourier coefficients f_i of shapes from MPEG-7 CE-1 Set B: a) boxplot - on each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme datapoints considered not to be outliers, and the outliers are plotted individually, b) the average values of magnitudes.

$$d_w(\mathbf{F^I}, \mathbf{F^{II}}) = \sum_{i=-M/2}^{M/2} w_i |f_i^I - f_i^{II}|. \tag{6}$$

where w_i ($i = -M/2, ..., M/2$) are the weighting coefficients and $\sum_{i=-N/2}^{N/2-1} w_i = 1$. To the best of the authors' knowledge, this is the first time that weighted distance is proposed to be used in conjunction with Fourier descriptors. In the following sections, three different sets of weighting coefficients w_i are proposed.

3. MOTIVATION - CONTRAST SENSITIVITY FUNCTION (CSF)

Since there could be literally unlimited possibilities for the weighting functions, the proposed analytical forms are pri-

[1] http://www.cis.temple.edu/~latecki/TestData/mpeg7shapeB.tar.gz

Fig. 2. Shape reconstructed with $M = 4$, $M = 8$, $M = 16$, $M = 32$ and $M = 64$ Fourier coefficients (shape energies are 0.6899, 0.8047, 0.8707, 0.9313, 0.9655 respectively)

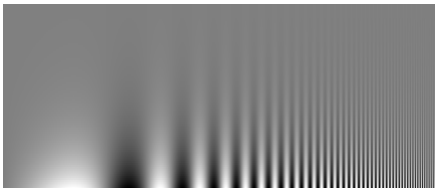

Fig. 3. Illustration of the CS. Although the contrast on the image increases linearly from the bottom, and all vertical lines are the same length, the lines in the middle seem longer. (Note that the illustration may not be completely understandable on certain zoom levels on a computer monitor, because of aliasing)

marily inspired by the sensitivity of the human visual system to certain spatial frequencies.

The level of contrast necessary to elicit a perceived response by the human visual system is known as the contrast threshold [9]. The inverse of the threshold is known as the contrast sensitivity (CS). The contrast sensitivity function (CSF) describes the contrast levels at a given spatial frequency necessary to elicit a perceptual response for a given spatial pattern, luminance level, and temporal frequency [9, 10]. Another description, as given by [11], is to refer to the contrast sensitivity function as a threshold modulation function that normalizes all frequencies such that they have equal contrast thresholds. It is generally well understood that achromatic contrast sensitivity can be described with a band pass spatial filter peaking around 3-4 cycles per degree of visual angle [11].

The illustration of CS is given on Figure 3. The contrast on the image increases linearly from the bottom. Although all vertical lines are the same length, it seems that the lines in the middle of the image are longer. This happens because the human visual system is most sensitive to these "middle" frequencies. Therefore, all proposed weighted distances in the paper favour "middle" frequencies and reduce the importance of low frequency components.

4. WEIGHTED FUNCTION MODELS

Three different forms of weighting coefficients are proposed, based on the following functions: CSF [11], Rayleigh distribution [12], and Log-normal distribution [13].

CSF model - This model is inspired by the most common and popular analytical form of the contrast sensitivity function proposed by Barten [11] (which is very similar to the model proposed by Mohshon and Kiorpes [14]):

$$\text{csf}(f) = a \cdot f \cdot e^{-bf}\sqrt{1 + ce^{bf}}, \tag{7}$$

where parameters $a = 540(1 + 0.7/L)^{-0.2}/[1 + 12/X/(1 + f/3)^2]$, $b = 0.3(1 + 100/L)^{0.15}$ and $c = 0.06$ are determined as functions of illumination L and the size of the pattern X in degrees, and f is the frequency in *cpd*. For typical usages $X = 45^\circ$ and $L = 500[cd/m^2]$.

The CSF model given by (7) represents the actual CSF function of the human visual system. However, using direct search to find the optimal parameters is slightly complicated, because these parameters do not independently affect the shape of CSF function. That is why we propose two more, different but simpler models: Rayleigh and Log-normal distribution models. Shapes of these functions are very similar to the CSF function (as shown in Figure 6), but they are considerably simpler for interpretation and analysis.

Rayleigh distribution model - It is given by [12]:

$$\text{csf}_R(f) = \frac{f}{\sigma^2} e^{-\frac{f^2}{2\sigma^2}}. \tag{8}$$

Unlike the model proposed with relation (7), this model is more convenient for optimization since it has only one parameter (so called scale parameter σ), and the statistical parameters such as the mean, maximum, variance etc. are given by simpler relations.

Log-normal distribution model - Unlike the Rayleigh distribution model, Log-normal distribution [13] has two degrees of freedom - standard deviation σ and expected value μ:

$$\text{csf}_{LN}(f) = \frac{1}{f\sigma\sqrt{2\pi}} e^{-\frac{(lnf-\mu)^2}{2\sigma^2}}. \tag{9}$$

In contrast to the Barten CSF model (7), function parameters (σ and μ) have certain physical meaning and facilitate optimization procedures.

5. METHODOLOGY

In order determine the optimal weighted distance, a commonly adopted Bulls-Eye retrieval performance score was chosen for evaluation [15]. Bulls-Eye score is defined as the

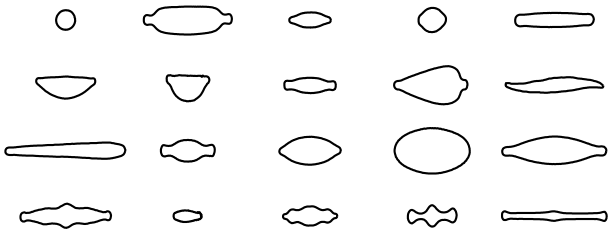

Fig. 4. Representative shapes of the modified Diatoms dataset (total of 425 shapes distributed in 20 classes).

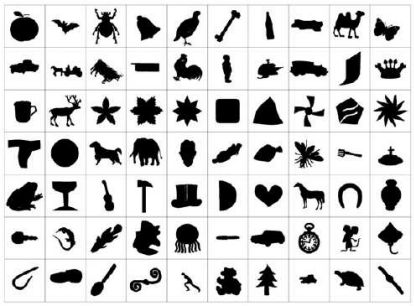

Fig. 5. MPEG-7 CE-1 Set B representative shapes (70 classes with 20 variations per class).

percentage of relevant results in the first $2 \cdot K$ retrieved results of a query, where K is the number of elements in the shape class which the shape belongs to [15]. Average Bulls-Eye score is computed after all elements in the dataset have been used as a query.

Prior to conducting optimization procedures, a suitable dataset had to be chosen. The proposed dataset is based on the original Diatoms dataset initially presented in [15], but shape classes that have extremely similar contours or differ only by scale, are removed. Therefore, the initial set is reduced to the total of 425 elements distributed into 20 classes (each class has at least 20 shapes). Representative shapes for each class are presented in Figure 4. The modified Diatoms dataset is essentially curve-based, and does not contain projective transformations, non-rigid transformations, articulation, noise etc., all of which may affect retrieval performance. Thus, the elements in the proposed dataset differ only by their shape i.e. by the magnitudes of the harmonics.

The unknown parameters for Rayleigh and Log-normal function model were estimated by direct search using Bulls-Eye score as the cost function, while the CSF model parameters are found using least-squares curve fitting of Rayleigh and Log-normal model.

The obtained weighting coefficients are validated on the popular MPEG-7 CE-1 Set B [8]. Representative elements of MPEG-7 CE-1 Set B are depicted in Figure 5. MPEG-7 CE-1 Set B consists of 1400 shapes representing real life objects, classified into 70 classes with 20 similar shapes for each

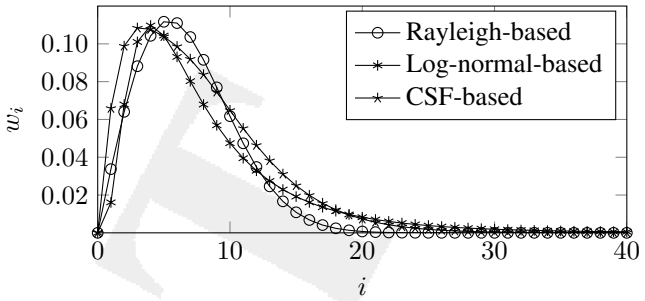

Fig. 6. Comparison of different weighting functions: CSF, Rayleigh and Log-normal based functions.

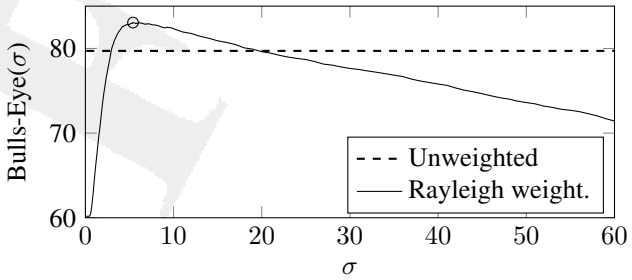

Fig. 7. Bulls-Eye score vs. scale parameter of the Rayleigh distribution σ, achieved on the modified Diatom dataset. A maximal score of 83.06 is achieved for $\sigma = 5.4$ (denoted with mark o).

class. This database is more difficult for shape-based image retrieval, since it includes rotation, scaling, skew, stretching, defection, indentation and articulation of shapes.

Finally, the retrieval performance is analyzed using an online survey, filled in by human users. Different retrieval results with similar performance scores, obtained with weighted and unweighted distances, were presented to human observers. The users voted for retrieval lists that they assessed as more relevant and accurate. The main conclusions of this survey are presented in the following section.

6. EXPERIMENTAL RESULTS

The analytical forms of the obtained weighting coefficients w_i on the modified Diatom set for each of the three models presented in the Section 4 are given in Table 1, and depicted in Figure 6. In order to further illustrate the effects of the weighted distances to the performance of the retrieval, the Bulls-Eye score vs. Rayleigh distribution scale parameter σ is given in Figure 7. Highest Bulls-Eye scores are achieved when the expected value of the distribution is located closer to the "middle" frequencies. This leads to a conclusion that "middle" frequencies are more important than low and high frequencies for shape retrieval, and clearly carry more information. Moreover, retrieval performance on this dataset can

Table 1. The proposed weighted City-block distance coefficients w_i ($i = -M/2, ...M/2$).

Weighted distance	Coefficients	Optimal parameters	Bulls-Eye
Unweighted	$w_i = 1$	-	79.71
Rayleigh	$w_i = \frac{\|i\|}{\sigma^2} e^{-\frac{i^2}{2\sigma^2}}$	$\sigma = 5.4$	83.06
Log-normal	$w_i = \frac{1}{\|i\|\sigma\sqrt{2\pi}} e^{-\frac{(ln(i)-\mu)^2}{2\sigma^2}}$	$\sigma = 0.7077$, $\mu = 1.8871$	83.63
CSF	$w_i = a \cdot \|i\| \cdot e^{-b\|i\|}\sqrt{1 + ce^{b\|i\|}}$	$a = 675(1 + 0.7/L)^{-0.2}/[1 + 12/X/(1 + f/3)^2]$ $b = 0.8016$, $c = 0.06$, $L = 0.1621$, $X = 0.001927$	81.55

a) b)

Fig. 8. a) Original classic car shape, b) shape reconstructed using weighting coefficients (contains only 4.52% of the initial energy).

be increased by 3 to 4%, just by emphasizing these "middle" frequencies using a weighted distance.

Another interesting fact is illustrated in Figure 8. The classic car shape is reconstructed with and without multiplication with the weighting coefficients. On first glance it seems that the cars in Figure 8a) and Figure 8b) are not extremely similar. However, the shapes in Figure 8a) and Figure 8b) are clearly more similar than the shapes in Figure 8a) and Figure 2a), although the shape depicted in 8b) has the energy equal to 0.0452, while the shape in Figure 2a) has the energy equal to 0.6899. This leads to the conclusions that low frequency components are not the most important for shape discrimination, and that the magnitude of the harmonics is not directly correlated with the respective semantical contribution.

The proposed weighted distances are validated on the MPEG-7 CE-1 Set B. The achieved Bulls-Eye scores are presented in Table 2. The results indicate small improvements of the Bulls-Eye score, moreover Rayleigh and Log-normal weighted coefficients seem to sligthly underperform. From the retrieval performance point of view, it appears that weighted distances introduce a rather small technical contribution. However, the usage of weighted distances yield another very interesting result. They allow reducing the average energy of the 1400 descriptors from the MPEG-7 CE-1 Set B from 1.0 to 0.049 without significantly altering retrieval performance. This indicates a large redundancy in FD shape description, and points out the importance of middle frequency components over low and high frequency components.

Finally, the effects of weighted distances on human perception were investigated. Human users were asked to assess

Table 2. Bulls-Eye scores for different weighted City-block distances on MPEG-7 CE-1 Set B.

Weighted distance	Bulls-Eye score
Unweighted	75.75
Rayleigh	75.51
Log-normal	74.09
CSF	76.44

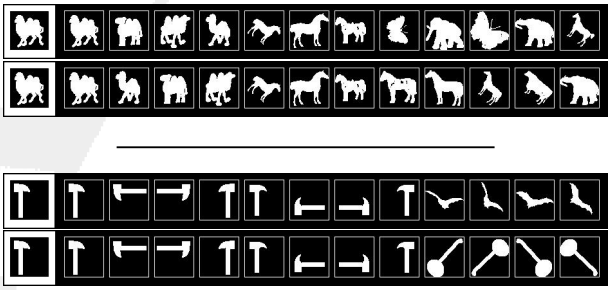

Fig. 9. Retrieval results for different query shapes on MPEG-7 CE-1 Set B. First image on the left is the query image. For both queries, the first row presents the results with unweighted distance, while the second row presents the results for the weighted distances (CSF for the "camel" shape, Log-normal for the "hammer" shape).

the retrieval results achieved with unweighted and weighted distances. Users had to fill in an online survey with 21 questions. Two questions were repeated, so that the non-consistent users are excluded from the statistics. Every question contained two retrieval result lists, and users had to choose the one which they found more accurate and relevant. In order to obtain objective results, the users were not fully acquainted with the exact purpose of the research. Two example questions are presented in Figure 9. The survey was completed by 46 contestants, aged from 20 to 52, both males and females. In order to evaluate the results of the survey a measure given by $k = n_w/n_{tot}$ is introduced, where n_w is the number of users that voted for weighted distance results and n_{tot} is the number of all participants. Hence $k = 100\%$ is the best result (meaning all users voted for the weighted distance), and $k = 0\%$ is the worst result (none of the users voted for the

Table 3. Survey results - the percentage of votes for weighted distances against unweighted distance.

Weight.distance	$k(\%)$
Rayleigh	69.25
Log-normal	76.08
CSF	78.26

weighted distance). In the end, the average value of k is computed. The final results are presented in Table 3.

It may be concluded from the results given in Table 3 that users would rather see the retrieval results obtained with weighted distance, especially CSF-based weighted distance, regardless of the fact that retrieval performance (Bulls-Eye score) is almost the same. This is justified by the fact that the non-relevant retrieval results obtained with the weighted distances are more closer to human visual understanding. As illustrated in Figure 9, in the first query, the "camel" shape is more similar to "horse", "cow" or "elefant" than the "butterfly" shape, and in the second query, the "hammer" shape is more similar to the "spoon" shape than it is to the "bat" shape. Although both queries have similar retrieval performance (since the "spoon" and "bat" are in fact dissimilar to the "hammer" shape), the users find weighted distance based retrieval system to be more accurate and results to be more relevant.

7. CONCLUSION

Using weighted distances inspired by the human visual system may enhance users' experience of the shape retrieval system, although the quantitative retrieval performance remains the same. The weighted distances did not exhibit significant retrieval performance improvement on MPEG-7 CE-1 Set B, mostly due to the Fourier descriptors limitations, and sensitivity to non-rigid transformations. Moreover, experiments demonstrate that the energy/magnitude/frequency of the harmonics is not directly related to their semantical contribution in the shape description, and that in fact the "middle" frequencies carry most of the important information about the shape.

The work presented in this paper should give new insights to human-based shape understanding. As part of future work, the weighted coefficients will be analyzed in conjunction with other shape description methods.

8. REFERENCES

[1] A Amanatiadis, VG Kaburlasos, A Gasteratos, and SE Papadakis, "Evaluation of shape descriptors for shape-based image retrieval," *Image Processing, IET*, vol. 5, no. 5, pp. 493–499, 2011.

[2] Johan W. H. Tangelder and Remco C. Veltkamp, "A survey of content based 3D shape retrieval methods," *Multimedia Tools and Applications*, vol. 39, no. 3, pp. 441–471, Dec. 2007.

[3] Dengsheng Zhang and Guojun Lu, "Shape-based image retrieval using generic Fourier descriptor," *Signal Processing: Image Communication*, vol. 17, no. 10, pp. 825–848, 2002.

[4] Bin Wang, "Shape retrieval using combined Fourier features," *Optics Communications*, vol. 284, no. 14, pp. 3504–3508, July 2011.

[5] Emir Sokic and Samim Konjicija, "Novel fourier descriptor based on complex coordinates shape signature," in *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on*. IEEE, 2014, pp. 1–4.

[6] Akrem El-ghazal, Otman Basir, and Saeid Belkasim, "Farthest point distance: A new shape signature for Fourier descriptors," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 572–586, Aug. 2009.

[7] *ISO/IEC 15938-3, Information Technology - Multimedia content description interface - Part 3: Visual*, 2002.

[8] L.J. Latecki, R. Lakamper, and T. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*. 2000, vol. 1, pp. 424–429, IEEE.

[9] Brian A Wandell, *Foundations of vision*, vol. 8, Sinauer Associates Sunderland, MA, 1995.

[10] Garrett M Johnson and Mark D Fairchild, "On contrast sensitivity in an image difference model," in *IS and TS pics conference*. Society for imaging science & technology, 2002, pp. 18–23.

[11] Peter GJ Barten, *Contrast sensitivity of the human eye and its effects on image quality*, SPIE press, 1999.

[12] Athanasios Papoulis and S Unnikrishna Pillai, *Probability, random variables, and stochastic processes*, Tata McGraw-Hill Education, 2002.

[13] Norman L Johnson, Samuel Kotz, and N Balakrishnan, "Continuous univariate distributions, vol. 2 of wiley series in probability and mathematical statistics: applied probability and statistics," 1995.

[14] J Anthony Movshon and Lynne Kiorpes, "Analysis of the development of spatial contrast sensitivity in monkey and human infants," *JOSA A*, vol. 5, no. 12, pp. 2166–2172, 1988.

[15] AC Jalba and MHF Wilkinson, "Shape representation and recognition through morphological curvature scale spaces," *Image Processing, IEEE*, vol. 15, no. 2, pp. 331–341, 2006.