IMAGE SIMILARITY MEASUREMENT BY KULLBACK-LEIBLER DIVERGENCES BETWEEN COMPLEX WAVELET SUBBAND STATISTICS FOR TEXTURE RETRIEVAL

Roland Kwitt and Andreas Uhl

Department of Computer Sciences
University of Salzburg, Jakob-Haringer-Str. 2, A–5020 Salzburg, Austria

ABSTRACT

In this work, we present a texture-image retrieval approach, which is based on the idea of measuring the Kullback-Leibler divergence between the marginal distributions of complex wavelet coefficient magnitudes. We employ Kingsbury's Dual-Tree Complex Wavelet Transform for image decomposition and propose to model the detail subband coefficient magnitudes by either two-parameter Weibull or Gamma distributions for which we provide closed-form solutions to the Kullback-Leibler divergence. The experimental results indicate that our approach can achieve higher retrieval rates than the classical approach of using the pyramidal Discrete Wavelet Transform together with the Generalized Gaussian model for detail subband coefficients.

Index Terms— Image texture analysis, Wavelet transforms, Statistics, Texture Retrieval

1. INTRODUCTION

A reasonable similarity measure between images is one of the essential parts of every image retrieval and image classification system. We focus on the problem of texture-image retrieval, where we have an arbitrary query image and we want to obtain the K most similar images from a given image database, according to some similarity criterion. Considering the fact that the database can contain a large number of images, we favor a computationally inexpensive way to measure similarity. Many classical approaches in the field of texture-image retrieval leave the spatial domain and work in a transform domain instead. Commonly used image transformations include the Gabor Wavelet Transform or the classical pyramidal Discrete Wavelet Transform (DWT). Our work is based on the contributions of [1], where the authors propose a statistical framework in which the feature extraction and similarity measurement step are closely related to each other. They propose to measure image similarity by computing the Kullback-Leibler divergence (KL-divergence) between the marginal distributions of DWT detail subband coefficients. In our approach, we work with the Dual-Tree Complex Wavelet Transform (DT-CWT), originally proposed by Kingsbury [2], which eliminates the DWT disadvantages of shift-dependency and lack of directional selectivity at the cost of very limited redundancy. We note, that the DT-CWT has already been successfully used in the context of image retrieval [3], however by using the traditional mean and standard deviation features of [4]. Our contribution is to propose a reasonable model for the marginal distributions of DT-CWT detail subband coefficients and to integrate these models into the statistical texture image retrieval framework of [1].

This work has been partially supported by the Austrian Science Fund (FWF) under Project No. L366-N15

The remainder of this paper is structured as follows: in Section 2 we briefly review the connection between the KL-divergence and the principle of maximum-likelihood selection. Section 3 then introduces the distributional models for the marginal distributions and provides closed-form solutions for the corresponding KL-divergences. Finally, Section 4 presents a comparison of the experimental retrieval results and Section 5 concludes the paper with a brief summary and an outlook on further research.

2. PROBABILISTIC IMAGE RETRIEVAL

First of all, we establish the formal framework of probabilistic image retrieval by merging the illustrations of [1] and [5]. Lets assume that we have N images $\mathcal{I}_i, 1 \leq i \leq N$ in our database. Each image is represented by a data vector $\mathbf{x}_i = \{x_{i1}, \ldots, x_{in}\}$, which is an element of some feature space $\mathcal{X} \subset \mathbb{R}^n$ and is obtained by feature extraction (FE). The retrieval task is to search the K most similar images to a given query image $\mathcal{I}_q$, according to some similarity criterion. We further assume that all images have equal prior probability and the query image is represented by data vector $\mathbf{x}_q$. From the probabilistic point of view, each data vector contains n realizations of i.i.d. random variables $X_1, \ldots, X_n$, which follow a parametric distribution with probability density function (PDF) $p(x|\theta), \theta \in \mathbb{R}^d$. Given that we have a consistent estimator $\hat{\theta}$ for the parameter vector θ, we can use $\hat{\theta}$ without limitations. Under these premises, it is natural to select the most similar image $\mathcal{I}_r$ to $\mathcal{I}_q$ as the one, whose parameter vector θ_r leads to a maximization of the likelihood/log-likelihood function, i.e.

$$r = \operatorname*{argmax}_j \frac{1}{n} \sum_{i=1}^{n} \log\left(p(x_{ji}|\theta_j)\right). \tag{1}$$

Note, that the additional factor $1/n$ does not affect the maximization result. By applying the weak law of large numbers to Eq. (1) as $n \to \infty$ (asymptotic case), we obtain

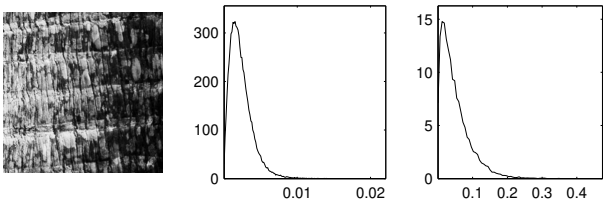
$$r \overset{n\to\infty}{=} \operatorname*{argmax}_i \mathbb{E}_{p(x|\theta_q)}(\log(p(x|\theta_i))) \tag{2}$$

$$= \operatorname*{argmax}_i \int_D p(x|\theta_q) \log(p(x|\theta_p)) dx, \tag{3}$$

where the term $\mathbb{E}_{p(x|\theta_q)}(\cdot)$ denotes the expectation with respect to $p(x|\theta_q)$ and D denotes the domain of $p(x|\cdot)$. By observing that $p(x|\theta_q)$ is an independent term for maximization, we can rewrite Eq. (3) as the following minimization problem:

$$r = \operatorname*{argmin}_i \left\{ -\int_D p(x|\theta_q) \log(p(x|\theta_i)) dx \right\} \tag{4}$$

$$\equiv \operatorname*{argmin}_i \int_D p(x|\theta_q) \log\left(\frac{p(x|\theta_q)}{p(x|\theta_i)}\right) dx \tag{5}$$



(a) Bark.00 (b) B_{12} (c) B_{25}

Fig. 1. Example texture and two histograms for the DT-CWT detail subbands B_{12} and B_{25} (The notation B_{sk} denotes the subband at scale s and orientation $k \in \{1, \ldots, 6\}$).

However, the last term in Eq. (5) is the KL-divergence between $p(x|\theta_q)$ and $p(x|\theta_i)$, which we will denote as $\mathrm{KL}(p_q||p_i)$ using the abbreviation $p_i := p(x|\theta_i)$. Hence, we have established the connection, that in the asymptotic case ($n \to \infty$) maximum-likelihood selection is equivalent to minimization of the KL-divergence. To obtain the *second most-similar* image to $\mathcal{I}_q$, we simply repeat the selection procedure on the $N-1$ remaining image samples. After N iterations we have an ordering of the image database induced by the KL-divergence, with the order-relation defined as $\mathcal{I}_i \leq \mathcal{I}_j :\Leftrightarrow \mathrm{KL}(p_q||p_i) \leq \mathrm{KL}(p_q||p_j)$ w.r.t. the query image $\mathcal{I}_q$.

3. COMPUTING IMAGE FEATURES

We employ Kingsbury's Dual-Tree Complex Wavelet Transform (DT-CWT) [2] to compute a redundant image representation with six oriented complex detail subbands at each decomposition level. The advantages of this complex wavelet transform variant are its approximate shift-invariance, its directional selectivity and the very efficient implementation scheme by four parallel 2-D DWTs. All of these properties come at the very low cost of four-times redundancy in 2-D.

To model the marginal density of the complex coefficient magnitudes of the detail subbands we consider three distributional models: the Rayleigh distribution, the two-parameter Weibull distribution and the two-parameter Gamma distribution. In Section 5 we will use a chi-square goodness-of-fit (GoF) test to quantify the suitability of these models. We like to remind that a visual comparison based on histogram fits alone can be elusive due to the dependency on histogram binning. To get a first impression of the principal shape of the marginal coefficient distributions, Figures 1(b) and 1(c) show two histograms of coefficient magnitudes from two detail subbands obtained by decomposing the example texture in Figure 1(a) by a three-scale DT-CWT.

Since the frequency responses of the DT-CWT closely resemble frequency responses of the Gabor wavelet transform [3], we first consider the Rayleigh distribution as a reasonable model [6]. The PDF of the Rayleigh distribution is given by

$$p(x|b) = \frac{x}{b^2} \exp\left(-\frac{x^2}{2b^2}\right), x > 0, \tag{6}$$

with scale-parameter $b > 0$. By inserting Eq. (6) into Eq. (5) we can derive a closed-form solution for the KL-divergence between two Rayleigh distributions with parameters b_i and b_j as

$$\underset{\text{Rayleigh}}{\mathrm{KL}}(p_i||p_j) = \frac{b_i^2}{b_j^2} - 2\log(b_i) + 2\log(b_j) - 1. \tag{7}$$

The expression in Eq. (7) only involves the distribution parameters, which allows a fast computation in case we have estimators for b_i and b_j. Throughout this work, we will use maximum-likelihood estimators (MLE) for the distribution parameters. The MLE of b can be found in [7] for instance. In the absence of a solution like Eq. (7), we would have to resort to the discrete version of the KL-divergence, which can be quite problematic to compute in view of the already-mentioned problem of histogram binning. Interestingly, assuming a Rayleigh distribution has some theoretical background too. If we assume that the real and imaginary part of the complex coefficients are just realizations of two random variables A and B following a normal distribution with zero mean and equal variance σ^2, then the magnitude $C = \sqrt{(A^2 + B^2)}$ follows a Rayleigh distribution with parameter σ. However, in the context of a medical image classification problem, it was shown in [8] that the assumptions of the Rayleigh model are very strong and are often violated. The authors propose the two-parameter Weibull distribution as a reasonable alternative, since it includes the Rayleigh distribution as a special case and allows for more freedom in shape, due to one additional parameter. The Weibull PDF is given by

$$p(x|c,b) = \frac{c}{b}\left(\frac{x}{b}\right)^{c-1} \exp\left\{-\left(\frac{x}{b}\right)^{c}\right\}, x > 0, \tag{8}$$

with shape parameter $c > 0$ and scale parameter $b > 0$. For $c = 2$ and $b = \sqrt{2}b$ it is trivial to see that Eq. (8) reduces to the Rayleigh distribution. Solutions for the MLEs of b and c are again given in [7]. Unfortunately, the MLE for c has no explicit form and can only be computed as the solution to a non-linear equation. By inserting Eq. (8) into Eq. (5), we can derive a closed-form solution for the KL-divergence between two Weibull distributions with parameters c_i, b_i and c_j, b_j as

$$\begin{aligned}\underset{\text{Weibull}}{\mathrm{KL}}(p_i||p_j) = \Gamma\left(\frac{c_j}{c_i}+1\right)\left(\frac{b_i}{b_j}\right)^{c_j} + \log\left(b_i^{-c_i}c_i\right) - \\ \log\left(b_j^{-c_j}c_j\right) + \log(b_i)\,c_i - \log(b_i)\,c_j + \frac{\gamma c_j}{c_i} - \gamma - 1,\end{aligned} \tag{9}$$

where γ denotes the negative of the digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$ at $x = 1$ ($\gamma \approx 0.577$).

The third alternative model we consider is the Gamma distribution, which has already been proposed as an alternative to the Rayleigh distribution for modeling the magnitudes of Gabor filter outputs [9]. By looking at Figures 1(b), 1(c) and recalling the principal shape of the Gamma distribution, we conjecture that this model might also be suitable for the DT-CWT detail subband coefficient magnitudes. The Gamma PDF is given by

$$p(x|a,b) = \frac{b^{-a}x^{a-1}}{\Gamma(a)} \exp\left(-\frac{x}{b}\right) \tag{10}$$

According to [9], a closed-form solution for the KL-divergence between two Gamma distributions with parameters a_i, b_i and a_j, b_j exists and can be computed by

$$\begin{aligned}\underset{\text{Gamma}}{\mathrm{KL}}(p_i||p_j) = \psi(a_i)(a_i - a_j) - a_i + \log\left(\frac{\Gamma(a_j)}{\Gamma(a_i)}\right) + \\ a_j \log\left(\frac{b_j}{b_i}\right) + \frac{a_i b_i}{b_j},\end{aligned} \tag{11}$$

with ML-estimates for the shape and scale parameter given in [7]. Like in the Weibull case, ML parameter estimation involves finding the root of a non-linear equation. Until now, we can only compute a similarity measure between the marginal distributions of two subbands. Assuming independency of the subband data allows to derive a simple similarity measure between two images in this framework

[1]. Since the KL-divergence can be expressed in terms of entropy and cross-entropy, we can apply the chain rule of entropy [10] and obtain the result that the overall KL-divergence between all subbands is simply the sum over the individual KL-divergences. Given that $\hat{p}_{sk}^{\mathcal{I}} := p(x|\hat{\theta}_{sk}^{\mathcal{I}})$ and $\hat{p}_{sk}^{\mathcal{J}} := p(x|\hat{\theta}_{sk}^{\mathcal{J}})$ denote the fitted distributions for each DT-CWT detail subband of images $\mathcal{I}$ and $\mathcal{J}$, the final similarity measure can be written as

$$\mathcal{S}(\mathcal{I},\mathcal{J}) = \sum_{s=1}^{J}\sum_{k=1}^{6} \mathrm{KL}(\hat{p}_{sk}^{\mathcal{I}}||\hat{p}_{sk}^{\mathcal{J}}) \tag{12}$$

where J denotes the decomposition depth of the DT-CWT. Eq. (12) is actually very simple to compute, since our expressions for the KL-divergences only involve the parameter estimates.

4. EXPERIMENTAL RESULTS

In order to compare our image retrieval approach to the one presented in [1], we use almost the same experimental setup. We work with the same 40 texture images from the MIT Vision Texture (VisTeX) database [11] (512×512 pixel) and split each image into 16 non-overlapping subimages (128×128 pixel). In addition to normalizing the subimages by subtracting the mean and dividing by the standard deviation, we conduct a contrast enhancement step using contrast-limited adaptive histogram equalization (CLAHE) [12]. Regarding the filter sets for the wavelet transforms, we use 8-tap Daubechies filters for the DWT and Kingsbury's Q-Shift (14,14)-tap filters (levels ≥ 2) in combination with (13,19)-tap near-orthogonal filters (level 1) for the DT-CWT [13].

As we mentioned in Section 3, we check the suitability of our distributional models by employing a chi-square GoF tests on the coefficient data. The chi-square test is applied on all subbands of the DT-CWT decomposed subimages (using a 3-scale DT-CWT). In all tests, the null-hypothesis is that the subband coefficients follow the assumed distribution and the significance level is set to $\alpha = 0.01$ (i.e. the probability of falsely rejecting the null-hypothesis although it is true, is less than 1%). The number of bins to compute the chi-square statistic is set to 0.3σ, where σ denotes the sample standard deviation (same setting as in DATAPLOT [14]). In case we observe bins containing less than 5 observations in the tail regions, the corresponding adjacent bins are scaled to include these observations. Table 1 lists the percentage of rejected null-hypothesis for our three distributional assumptions per scale and on average.

Scale	Rayleigh	Gamma	Weibull
1	93.98	64.79	69.17
2	76.02	39.12	33.48
3	44.64	12.63	9.51
Avg.	71.55	38.85	37.39

Table 1. Percentage of rejected null-hypothesis ($\alpha = 0.01$) per scale and on average.

As we can see, the Weibull and Gamma distribution are better models than the Rayleigh distribution, which seems obvious due to the higher degree of freedom in adapting to the underlying data. We further observe that higher decomposition depths lead to smaller numbers of rejected null-hypothesis. Even in case of the Rayleigh assumption, the rejection rate drops to approximately 45% for level three. For the purpose of comparision, the rejection rate in case of assuming a Laplace distribution for the marginal detail subband coefficients of a three-scale DWT is 63% on average and 14.39% when assuming a Generalized Gaussian distribution (GGD).

In order to evaluate the performance of the retrieval system, we have to establish some evaluation criteria. First, we check the number of correctly retrieved images among the top K matches. By *correct* we understand the correct texture class (the parent of the subimage). For each query image (i.e. each subimage) we know the correct membership index set $Q = \{r_1, \ldots, r_B\}$, where B denotes the number of subimages. Given that the index set for the top K matches is denoted by $\{q_1, \ldots, q_K\}$, we calculate

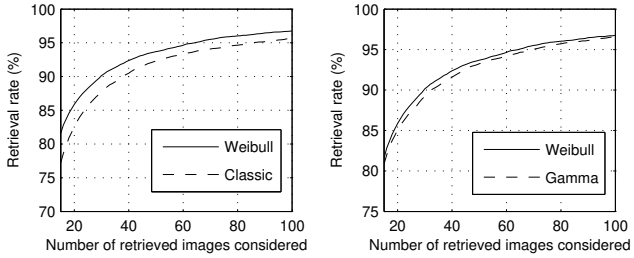
$$s_k = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}_Q(q_i), \quad \mathbf{1}_Q(x) := \begin{cases} 1, \text{ if } x \in Q, \\ 0, \text{ else} \end{cases} \tag{13}$$

which gives us the percentage of correctly retrieved images. Since each image is split into 16 subimages, we set $K = B = 16$. Second, we evaluate the retrieval performance according to the number of considered retrieved images. This means, that we calculate Eq. (13) for varying values of K. This gives us some kind of operating characteristic curve. Table 2 lists the percentage of correctly retrieved images among the top $K = 16$ matches with decomposition depths less than three.

Scale	Laplace	GGD	Rayl.	Wbl.	Gamma	Classic
1	41.00	64.35	63.52	**71.81**	71.17	68.12
2	52.07	70.91	69.57	**76.81**	76.06	72.06
3	57.27	74.24	75.07	**81.58**	80.86	77.23

Table 2. Top $K = 16$ retrieval results.

As we can see, the highest rates are obtained by relying on the Weibull assumption, although the Gamma model performs at a comparable level. However, an interesting observation is that by using the classical mean and standard deviation features of [4] (abbreviated by *Classic*) together with the unweighted Euclidean distance as a similarity measure, we obtain even better retrieval rates than the DWT & GGD (abbreviated by *GGD*) approach. Disregarding the fact that the Weibull model leads to the highest rates, these results are particularly noteworthy, since the computational effort to compute the empirical mean and empirical standard deviation is by far smaller than computing the MLEs of the GGD by finding the roots of a transcendental equation [1]. In Figure 2 we show a comparison of the retrieval performances for different DT-CWT detail subband features in case of a 3-scale DT-CWT.

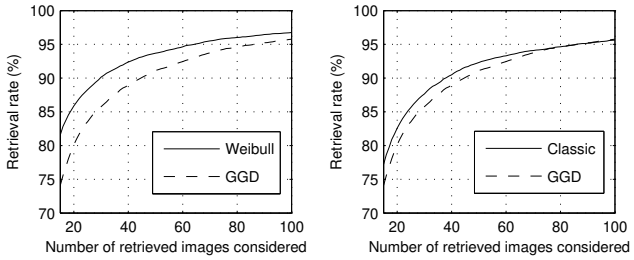


(a) Wbl. vs. Classic (b) Wbl. vs. Gamma

Fig. 2. Comparison between various DT-CWT features.

The Weibull model leads to the consistently highest retrieval rates and is slightly better than the Gamma model. Further, we see the effect mentioned in [3], that as the number of considered images grows, the slope of the graph becomes more shallow. This implicates that in order to obtain retrieval rates equal to the Weibull approach, we need to consider many more images in the Classic and

Gamma case. In Figure 3 we compare the best results of the DWT & GGD approach to DT-CWT & Classic and DT-CWT & Weibull approaches.



(a) GGD vs. Wbl. (b) GGD vs. Classic

Fig. 3. Comparison between DWT and DT-CWT features.

On average, the combination DT-CWT & Weibull is superior to the combination DWT & GGD by 2.8% for $15 \leq K \leq 100$. In the GGD vs. Classic case, the traditional approach of using mean and standard deviation achieves better rates than the GGD approach for $K \leq 80$. To see the differences in retrieval rates for the 40 textures of our experiments, we additionally provide the texture-specific rates in Table 3. For 90% of the textures, the Weibull approach achieves

Texture	GGD	Wbl.	Texture	GGD	Wbl.
Bark.00	64.45	70.70	Food.08	99.22	100.00
Bark.06	42.19	60.55	Grass.01	69.14	64.06
Bark.08	64.84	77.34	Leaves.08	63.28	71.48
Bark.09	51.17	77.34	Leaves.10	47.66	42.19
Brick.01	82.03	86.33	Leaves.11	72.66	88.28
Brick.04	61.33	76.95	Leaves.12	76.56	79.69
Brick.05	71.88	86.72	Leaves.16	74.22	80.08
Buildings.09	96.09	99.61	Metal.00	69.92	72.27
Fabric.00	80.47	91.41	Metal.02	100.00	100.00
Fabric.04	62.11	65.23	Misc.02	82.42	83.59
Fabric.07	96.88	100.00	Sand.00	76.17	92.19
Fabric.09	67.97	97.27	Stone.01	58.59	60.55
Fabric.11	55.86	76.17	Stone.04	69.14	73.05
Fabric.14	100.00	100.00	Terrain.10	39.84	51.17
Fabric.15	85.16	95.70	Tile.01	51.17	48.44
Fabric.17	94.53	100.00	Tile.04	96.48	99.22
Fabric.18	94.92	99.22	Tile.07	100.00	100.00
Flowers.05	70.31	81.64	Water.05	96.09	98.83
Food.00	75.39	90.63	Wood.01	31.64	31.25
Food.05	88.28	96.48	Wood.02	84.77	97.66

Table 3. Texture-specific retrieval results.

higher retrieval rates. In case of the Gamma assumption, the results look similar, although we cannot provide a full listing of the results here, due to space limitations.

5. CONCLUSION

In this work we have taken up the idea of measuring image similarity by KL-divergences between the marginal distributions of wavelet detail subband coefficients and extended it to the complex wavelet domain. We have shown that the magnitudes of the complex coefficients can be modeled by either two-parameter Gamma or Weibull distributions. We have further validated the results of previous work [8], that the Rayleigh distribution is not flexible enough to model these subband coefficients. An interesting result of our experiments is, that the mean and standard deviation based features from DT-CWT detail subbands together with the Euclidean distance outperform the combination DWT & GGD. Especially, from the computational point of view, this result is actually very appealing. We conclude that the advantages of the DT-CWT over the classical DWT are reflected in the good results. The additional computational load that comes along with the DT-CWT is neglectable since the DT-CWT allows a parallel implementation and is still of linear complexity. Future work includes the evaluation of the Generalized Gamma distribution as another alternative distribution, since it includes the Rayleigh, Weibull and Gamma model as special cases, however at the cost of complex parameter estimation.

REFERENCES

[1] M. Do and M. Vetterli, "Texture similarity measurement using kullback-leibler distance on wavelet subbands," in *Proceedings of the IEEE International Conference on Image Processing (ICIP'00)*, Vancouver, Canada, Sept. 2000, vol. 3, pp. 730–733.

[2] Nick G. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters," in *Proceedings of the IEEE Digital Signal Processing Workshop, DSP '98*, Bryce Canyon, USA, Aug. 1998, pp. 9–12.

[3] P. Rivaz and N. Kingsbury, "Complex wavelet features for fast texture image retrieval," in *Proceedings of the IEEE International Conference on Image Processing (ICIP'99)*, Kobe, Japan, 1999, pp. 109–113.

[4] W.Y. Ma B.S. Manjunath, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[5] N. Vasconcelos and A. Lippman, "A unifying view of image similarity," in *Proceedings of the International Conference on Pattern Recognition (ICPR'00)*, Barcelona, Spain, 2000, pp. 38–41.

[6] S. Bhagavathy, J. Tesic, and B.S. Manjunath, "On the rayleigh nature of gabor filter outputs," in *Proceedings of the IEEE International Conference on Image Processing (ICIP'03)*, Barcelona, Spain, 2003, vol. 3, pp. 745–748.

[7] K. Krishnamoorthy, *Handbook of Statistical Distributions with Applications*, Chapman & Hall, 2006.

[8] R. Kwitt and A. Uhl, "Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images," in *Proceedings of the IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'07)*, Rio de Janeiro, Brasil, 2007, pp. 1–8.

[9] J.R. Mathiassen, A. Skavhaug, and K. Bo, "Texture similarity measure using kullback-leibler divergence between gamma distributions," in *Proceedings of the 7th European Conference on Computer Vision (ECCV'02)*, Copenhagen, Denmark, 2002, pp. 133–147, Springer.

[10] Thomas C. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[11] MIT Vistion Texture Database, "MIT vision and modeling group," [Online], Available from: `http://vismod.media.mit.edu/vismod/`.

[12] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, Paul S. Heckbert, Ed., pp. 474–485. Morgan Kaufmann, 1994.

[13] Nick G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 234–253, May 2001.

[14] N.A. Heckert and J.J. Filliben, *NIST Handbook 148: DATA-*

PLOT Reference Manual, vol. 1, National Institute of Standards and Technology Handbook Series, 2003.