In Eighth International Conference on Information Visualisation.
Los Alamitos: IEEE Computer Society Press, 2004, p. 513-518.
ISBN 0-7695-2177-0.

Evaluation of Two Principal Approaches to Objective Image Quality Assessment

Martin Čadík, Pavel Slavík
Department of Computer Science and Engineering
Faculty of Electrical Engineering, Czech Technical University in Prague
Karlovo nám. 13, 121 35 Prague, Czech Republic
cadikm@fel.cvut.cz, slavik@fel.cvut.cz

Abstract

Nowadays, it is evident that we must consider human perceptual properties to visualize information clearly and efficiently. We may utilize computational models of human visual systems to consider human perception well. Image quality assessment is a challenging task that is traditionally approached by such computational models. Recently, a new assessment methodology based on structural similarity has been proposed. In this paper we select two representative models of each group, the Visible Differences Predictor and the Structural SIMilarity index, for evaluation. We begin with the description of these two approaches and models. We then depict the subjective tests that we have conducted to obtain mean opinion scores. Inputs to these tests included uniformly compressed images and images compressed non-uniformly with regions of interest. Then, we discuss the performance of the two models, and the similarities and differences between the two models. We end with a summary of the important advantages of each approach.

1. Introduction

Image quality assessment and comparison metrics play an important role in various graphics oriented applications. They can be used to monitor image quality for quality control systems, they can be employed to benchmark image processing algorithms, and they can be embedded into the rendering algorithms to optimize their performances and parameter settings. It is well known [8], that classical comparison metrics like Root Mean Square error are not sufficient when applied to the comparison of images, because they poorly predict the differences between the images as perceived by the human observer. This fact has led to the development of more advanced perceptual quality assessment techniques.

Traditional perceptual image quality assessment approaches are based on measuring the errors (signal differences) between the distorted and the reference images, and attempt to quantify the errors in a way that simulates human visual error sensitivity features.

Different from the traditional error-sensitivity-based approach, *structural similarity based* image quality assessment has been recently outlined. This approach is based on the following philosophy: the main function of the human visual system is to extract structural information from the viewing field, and the human visual system is highly adapted for this purpose. Therefore, a measurement of structural information loss can provide a good approximation to the perceived image distortion [10].

In this paper we evaluate the two mentioned principal approaches to image quality assessment. We judge the responses of representative models of each group using the subjective opinion scores. We have conducted subjective tests to obtain our own values of the scores. The test input images included uniformly compressed images and images compressed non-uniformly with regions of interests. Region of interest (ROI) image compression allows less degradation for the ROIs than for the other parts of the image.

The paper is organized as follows. In Section 2, we summarize the two representative models – the visible differences predictor and the structural similarity index. In Section 3, we describe subjective and objective testing that was performed. Finally, in Section 4, we examine and discuss the results of the tests and we conclude with a summary of the important advantages of each approach.

2. Background

Recently, several studies on performance of traditional perceptual image quality models have been published [5, 4, 2, 11]. However, as far as we know, no independent evaluation of traditional and structural similarity approaches have

been carried out, apart from the article where the SSIM model was introduced [10].

For comparison of the traditional error-sensitivity and structural similarity based approaches we have chosen a representative from each group. The Visible Differences Predictor (VDP) is a typical example of an image quality metric based on *error sensitivity,* whereas the Structural SIMilarity index (SSIM) is a specific example of a *structural similarity* quality measure. We have included the responses of the Universal Quality Index (UQI) into the evaluation as well. The UQI is a special case of the SSIM.

The input to the models consists of two images and (in the VDP case) parameters for viewing conditions, whereas the output is a map describing the visible differences between them. The output map defines the probability of detecting the differences between the two images as a function of their location in the images.

2.1. Visible Differences Predictor

The VDP model [1] interprets early vision behaviour, from retinal contrast sensitivity to spatial masking. The use of the VDP consists of three main stages: components for *calibration* of the input images, a *human visual system (HVS) model* and a method for *displaying* the HVS visible *differences*.

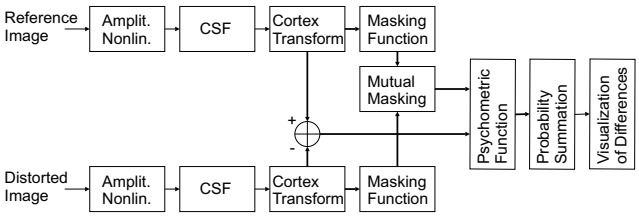

Figure 1. Block diagram of the Visible Differences Predictor (heavy lines indicate parallel processing).

The key element of the VDP is the *human visual system model*, see Figure 1. It concentrates on the lower-order processing of the visual system, such as the optics, retina, lateral geniculate nucleus, and striate cortex. The HVS model consists of a number of processes that limit visual sensitivity. Three main sensitivity variations are accounted for, namely, as a function of *light level*, *spatial frequency*, and *signal content*. Sensitivity S is defined as the inverse of the contrast C_T required to produce a threshold response, $S = 1/C_T$, where contrast is defined as $C = (L_{max} - L_{mean})/L_{mean}$, where L_{max} and L_{mean} refer to the maximum and mean luminances.

The variations in sensitivity as a function of light level are simulated by *amplitude nonlinearity*. Each input luminance L_{ij} is transformed by a simplified version of the retinal response to an "amplitude non-linearity value" b_{ij} defined as: $b_{ij} = L_{ij}/(L_{ij} + 12.6L_{ij}^{0.63})$, where the constants 12.6 and 0.63 apply when luminance is expressed in cd/m^2. For this model the adaptation level for an image pixel is solely determined from that pixel.

The variations as a function of spatial frequency are modeled by the *contrast sensitivity function* (CSF), implemented as a filtering process. A Fast Fourier transform is applied to the values b_{ij}. The resulting magnitudes, $f_{uv}(b)$ are filtered by a CSF which is a function of the image size in degrees and light adaptation level L_m.

The variations in sensitivity due to a signal content are referred to as *masking*. Masking effects are modeled by the *detection mechanism*, which is the most complicated element of the VDP. It consists of four subcomponents: image channeling, spatial masking, psychometric function, and probability summation.

During the image channeling stage, the input image is fanned out from one channel to 31 channels or bands as follows. Each channel is associated with one cortex filter which consists of a *radial filter* (*dom*, difference of mesa filter) and an *orientational filter* (*fan* filter). The total number of radial filters is six resulting in five frequency bands and one base band. Each of these bands except for the base band is further fanned out into six channels of different orientation. Thus five frequency bands times six orientations per bands plus one base band results in 31 channels.

2.2. The Structural SIMilarity Index

The Structural SIMilarity Index (SSIM) [10] is a specific example of a *structural similarity* quality measure. The structural information in an image is defined as those attributes that represent the structure of objects in the scene, independent of the average luminance and contrast. The diagram of the quality assessment using the SSIM is shown in Figure 2. (Note: the SSIM is a generalization of the Universal Quality Index [9]).

The SSIM separates the task of similarity measurement into three comparisons. First, the *luminance* of each signal x and y is compared. The luminance comparison function $l(x, y)$ is a function of μ_x and μ_y: $l(x, y) = \frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1}$, where the $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$ (the mean intensity) is the estimate of luminance.

Second, the mean intensity is removed from the signal and the *contrast comparison* function is evaluated as follows: $c(x, y) = \frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2}$, where the

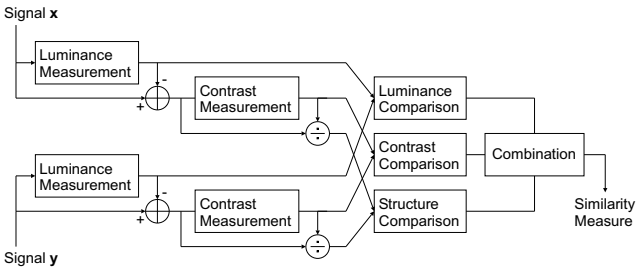

Figure 2. Diagram of the structural similarity (SSIM) measurement system.

$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)^2\right)^{1/2}$ (the square root of variance) is the estimate of the signal contrast.

Third, the signal is normalized by its own standard deviation, so that the two signals being compared have unit standard deviation. The *structure comparison* is defined as follows: $s(x,y) = \frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}$, using the correlation (inner product) σ_{xy} between the normalized signals, $\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$. C_1, C_2, and C_3 are constants included to avoid instabilities.

Finally, the three components are combined to yield an overall similarity measure (structure similarity index):

$$SSIM(x,y) = [l(x,y)]^{\alpha}\cdot[c(x,y)]^{\beta}\cdot[s(x,y)]^{\gamma},$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters used to adjust the relative importance of the three components. This definition satisfies the conditions of symmetry: $SSIM(x,y) = SSIM(y,x)$, boundedness: $SSIM(x,y) \leq 1$, and unique maximum: $SSIM(x,y) = 1$ iff $x = y$. The Universal Quality Index (UQI) corresponds to the special case that $C_1 = C_2 = 0$, which produces unstable results when either $(\mu_x^2+\mu_y^2)$ or $(\sigma_x^2+\sigma_y^2)$ is very close to zero.

For *image quality assessment*, it is useful to apply the SSIM index locally. Localized quality measurement can provide a spatially varying quality map of the image, which delivers more information about the quality degradation of the image. The local statistics μ_x, σ_x, and σ_{xy} are computed within a local window, which moves pixel-by-pixel over the entire image. When a single overall quality measure of the entire image is required, we use a mean SSIM (MSSIM) index to evaluate the overall image quality: $MSSIM(X,Y) = \frac{1}{M}\sum_{j=1}^{M} SSIM(x_j,y_j)$, where X and Y are the input images, x_j and y_j are the image contents at the $j-$th local window, and M is the number of samples in the quality map.

3. Measurements

First, we have conducted subjective testing to obtain subjective opinion scores. Then, we have executed the objective testing using the VDP and the SSIM using the same group of inputs. The results of these two tests were then compared to evaluate the performance of the inquired models.

3.1. Subjective Testing

As visual stimuli we have used the 33 JPEG 2000-compressed photos of an urban construction site. We have conducted two tests: in the first test all the input images were compressed uniformly. In the second test the images contained manually-specified regions of interest (ROI) compressed with better quality than the other areas of the image. The total number of 32 subjects were asked to express the difference between the original and compressed images by ratings. Subjects had normal or corrected-to normal vision and were non-experts in the field of image comparison. Mean opinion scores (MOSs) were computed for each image pair from raw scores of each subject.

3.2. Objective Testing

The pairs of images used in the subjective tests were used as input stimuli for both the VDP and the SSIM models. The parameters of the VDP model were set properly to correspond with the subjective observers configurations.

Both the inspected models are relative because they do not describe an absolute value of image quality but instead they address the problem of differences between two images. The output is a map describing the visible differences between them, see Figure 4 as an example. The output map defines the probability of detecting the differences between the two images as a function of their location in the images. An advantage is that we can see the nature of the difference and we can use this information for further improvement of the design.

However, since we need a single overall quality measure as well, we use a mean SSIM index in the case of the SSIM model. For the VDP model we use the approach as follows [6]: the difference between images $D_{0.75}$ is the percentage of pixels for which the probability of difference detection is greater than 0.75. It is assumed, that the difference can be perceived for a given pixel when the probability value is greater than 0.75 (75%), which is the standard threshold value for discrimination tasks.

3.3. Test Results

As a visual illustration of the relationship between subjective data and model predictions, scatter plots of MOS

versus the VDP and the SSIM predictions are shown in Figure 3. Each point in a graph represents one test image, with its vertical and horizontal coordinates representing its subjective MOS and the model prediction, respectively. As we can see, the SSIM results exhibit better consistency with the subjective data than the results of the VDP.

For the numerical evaluation we use both the standard (Pearson) and the non-parametric (Spearman) correlations [3]. We use the 3^{rd} order polynomial fit function prior to computation of the correlation coefficients, because the mapping of the objective model outputs to the subjective MOS is generally non-linear. These non-linear regression functions are used to transform the set of model outputs to a set of predicted MOS values. Correlation coefficients are then computed between these values and the subjective MOS.

4. Discussion

In this section we discuss the performances of the models. First, we discuss the performances for uniformly compressed input images, then we consider the results for ROI-compressed images. Finally, we point out the advantages of each model.

4.1. Quality Assessment Performances

Calculated values of the Pearson (CC) and Spearman (SROCC) correlation coefficients are presented in Table 1. The values in the first column (overall performances) show that the correlation to the MOS is much better for the SSIM model than for the the VDP. The absolute values of the CCs are generally lower than the published results [10]. This slight discrepancy is probably caused by the selection of input stimuli. A more comprehensive set of input images would be valuable to draw more general conclusions.

As one may see in the second and third column of Table 1, the performances of models in the ROI task were comparable and quite poor. This has led us to experiments with the ROI functions, as described in the following section.

Model	CC	CC-ROI	SROCC-ROI
VDP	0.22	0.44	0.39
SSIM	0.72	0.51	0.45
UQI	0.55	0.20	0.40

Table 1. Quality assessment performances of the VDP, the SSIM and the UQI models.

4.2. ROI Quality Assessment

Half of the input stimuli in the subjective tests were images compressed with manually-defined ROIs. Since we had the explicit information about the ROI structure for each of the input images we were able to incorporate it into the computation of the VDP and the SSIM responses. Based on the information about the ROI structure (spatial arrangement and compression ratios) we have constructed the "ROI functions" that we have used to scale the output probability maps. We have experimented with various ROI functions including the absolute values of compressions, smoothed and reduced range function, approximation by Gaussians, and the inversion of the absolute values. See Figure 5.

For all of the tested ROI functions the values of correlation coefficients between the scaled results and MOSs were less than in the non-ROI-aware case, both for the VDP and for the SSIM. We suppose that the ROI information must be involved directly during the computation of an actual model taking into consideration such issues as the foveation, etc. This offers a wide area for future research and thorough psychophysical testing.

On the other side, the SSIM model *detects* the regions of interest in the image quite well, as we can see for example on the bottom right image in Figure 4. This feature of the SSIM would be used to assess the locations of regions of interest automatically. However, we suppose that visual attention-aware models [7] would even outperform the SSIM in this task.

4.3. Advantages of Evaluated Models

The *VDP model* is based on generally-held human visual system assumptions and is able to handle various phenomena such as the visual masking. The consistency of the VDP model has been previously verified [6] and the model was successfully applied to various issues not just in the area of image quality assessment, but also for the computer graphics algorithms [6].

The correlation coefficients between our subjective data set and the *SSIM model* responses exhibit that the SSIM model shows better consistency with subjective data than the VDP does. As we have seen, the SSIM model is able to detect regions of interest in the image. This feature is promising for future research on ROI issues. The SSIM model is simple to implement in comparison to the VDP model, and the MatLab code is publicly available. The computation of the SSIM does not require time consuming Fourier transformations (as the VDP does) and it is certainly faster than the computation of the VDP model.

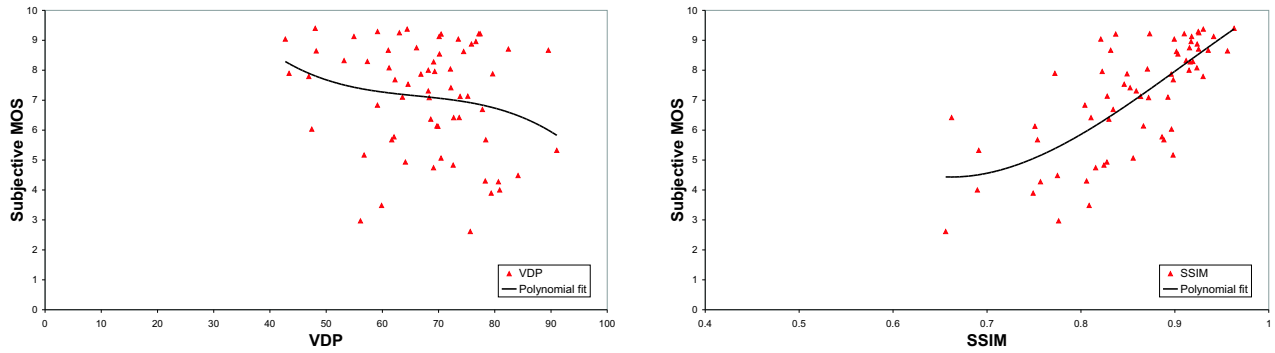

Figure 3. Quality predictions compared to subjective mean opinion scores (MOS) for the Visible Differences Predictor (left) and for the Structural SIMilarity index (right).

5. Conclusion

Image quality assessment models are relevant for various computer graphics applications. In this paper, we have presented the independent comparison of two image quality assessment approaches. We have evaluated the VDP model and the SSIM model as the representatives of the traditional perceptual approach and the structural similarity based approach, respectively. We have described the subjective tests that we have conducted to obtain mean opinion scores both for uniformly and ROI compressed images.

The evaluation of subjective results and predictions of the models shows that the structural based approach outperforms the traditional approach for involved input stimuli. As the implementation of the SSIM model is more straightforward than the implementation of the VDP, we propose that the SSIM model is a significant alternative to the thoroughly verified VDP model. The SSIM model is able to detect the ROIs in the image. However, both models perform poorly in the ROI image assessment task. Moreover, the input stimuli set and the group of observers should be more comprehensive to obtain more general results.

Acknowledgements

This project has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic under research program No. Y04/98: 212300014 (Research in the area of information technologies and communications), and by the Czech Technical University in Prague – grant No. CTU0408813.

References

[1] S. Daly. The visible differences predictor: An algorithm for the assessment of image fi delity. In A. B. Watson, editor, *Digital Images and Human Vision*, pages 179– 206, Cambridge, MA, 1993. MIT Press.

[2] R. Eriksson, B. Andren, and K. Brunstrom. Modelling the perception of digital images: A performance study. In *IS&T/SPIE Conference on Human Vision and Electronic Imaging III, SPIE volume 3299*, January 1998.

[3] M. J. Gardner. *Statistics with Confi dence*. Altman D. G., 1989.

[4] W. B. Jackson, M. R. Said, D. A. Jared, J. O. Larimer, J. L. Gille, and J. Lubin. Evaluation of human vision models for predicting human observer performance. In *Proc. SPIE Vol. 3036, p. 64-73, Medical Imaging 1997: Image Perception, Harold L. Kundel; Ed.*, pages 64–73, Apr. 1997.

[5] B. Li, G. Meyer, and R. Klassen. A comparison of two image quality models. In *SPIE Conf. on Human Vision and Electronic Imaging III*, volume 3299, 1998.

[6] K. Myszkowski. The visible differences predictor: applications to global illumination problems. In *Proceedings of the Ninth Eurographics Workshop on Rendering*, pages 223–236, Wienna, Austria, 1998.

[7] W. Osberger, N. Bergmann, and A. Maeder. An automatic image quality assessment technique incorporating higher level perceptual factors. In *Proceedings ICIP-98*, USA, 1998.

[8] P. C. Teo and D. J. Heeger. Perceptual image distortion. In *Human Vision, Visual Processing and Digital Display*, volume 2179, 1994.

[9] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81– 84, 2002.

[10] Z. Wang, A. C. Bovik, H. R. Seikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. In *IEEE Transactions on Image Processing*, volume 13, 2004.

[11] B. Watson, A. Friedman, and A. McGaffey. Using naming time to evaluate quality predictors for model simplifi cation. In *CHI Letters*, volume 2, pages 113– 120, 2000.

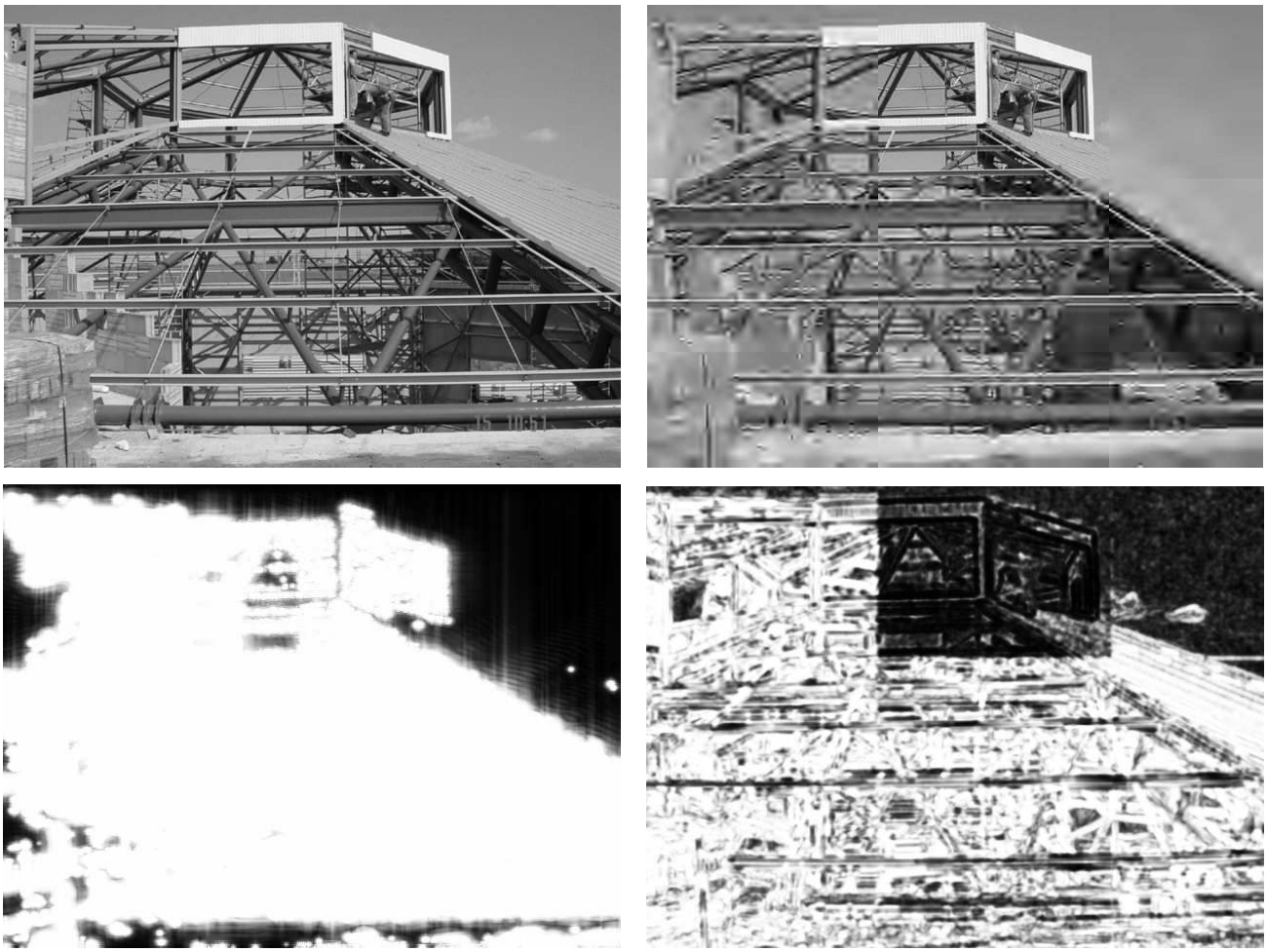

Figure 4. Image quality assessment using the VDP and the SSIM. Original image (top left), ROI compressed image (top right), VDP detection probability map (bottom left), SSIM detection probability map (bottom right).

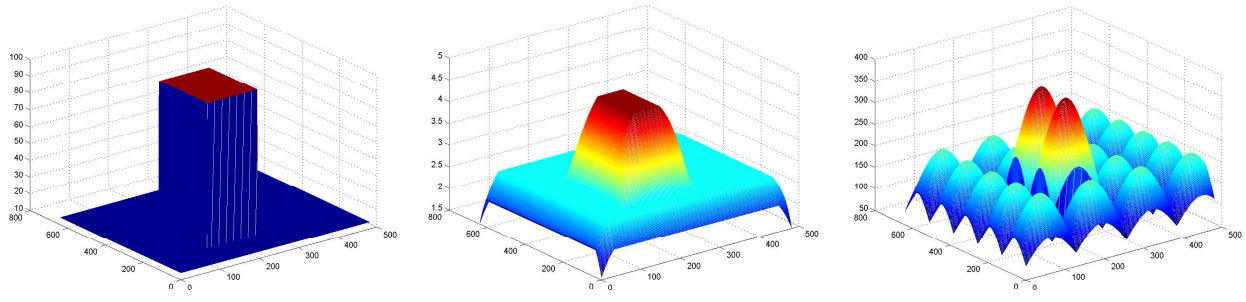

Figure 5. ROI functions. Absolte values (left), smoothed and reduced range (middle), approximated by Gaussians (right).