Visual and Quantitative Comparison of Real and Simulated Biomedical Image Data

Tereza Nečasová and David Svoboda

Centre for Biomedical Image Analysis
Masaryk University, Brno, Czech Republic
necasovat@fi.muni.cz

Abstract. The simulations in biomedical image analysis provide a solution when the real image data are difficult to be annotated or if they are available only in small quantities. The progress in simulations rapidly grows in the recent years. Nevertheless, the comparative techniques for the assessment of the plausibility of generated data are still unsatisfactory or none. This paper aims to point out the problem of insufficient comparison of real and synthetic data, which is done in many cases only by visual inspection or based on subjective measurements. The selected texture features are first compared in a univariate manner by quantile-quantile plots and Kolmogorov-Smirnov test. The evaluation is then extended into multivariate assessment using the PCA for a visualization and furthermore for a quantitative measure of similarity by Jaccard index. Two different image datasets were used to show the results and the importance of the validation of simulated data in many aspects.

Keywords: Feature comparison · Validation of simulation · Statistical evaluation · Similarity visualisation.

1 Introduction

The research in the last decades showed the power of technical progress besides other things in biomedical imaging, especially in the generation of artificial image data, which should resemble the real images. From the images of spots and particles, over nuclei and subcellular components, also the images of multiple-target, cell populations and tissues are possible to be synthesized as summarized in [16].

Despite the methods for simulations differ, the main objectives for generating artificial image data are the same: 1) to use the simulated data for validating the segmentation algorithms with unhidden ground truth; 2) to perceive the biological processes and understand the cell behavior; and last but not least: 3) to reduce the time and inconsistency among manual annotators, which is even higher in three-dimensional image data and time-lapse sequences.

Nevertheless, one can legitimately ask for a plausibility of the artificial data. There are many characteristics that can be computed and subsequently used for the comparison of real and synthetic data, such as shape, number of cells in the

image, number of various elements etc. The comparison made on these features could support the expectation that the simulated data are of sufficient quality. This paper is focused on methods for validating texture, as we found this feature to be variously interpreted and independent of the work of annotator. This feature is also given by every single image, which makes this feature applicable to all available images.

The first step of quality assessment is the visual inspection, which is mentioned almost every time. While some of the works ends with the visual evaluation of experts, other outspread with quantitative measurements. In [1], for example, the quality of generated images is assessed by measuring the largest magnification in which the image data look realistic. In [8], the authors suggested to compute the sensitivity and specificity over 6 differently experienced human test subjects, who had to classify, whether the image is synthetic or real. This decision was made in limited time, which should imitate the real conditions. In contrast to expert-based evaluation, there also exist the approaches based on quantitative evaluation. In particular, the Q-Q plots of texture descriptors were used in [8, 12]. Furthermore, [13] showed the histograms of real and synthetic data accompanied by means, standard deviations and p-values of Kolmogorov-Smirnov tests. Note, that all of these validation techniques were assessed univariately. This paper extends these evaluations into multiple dimensions to compare the mutual information of the data points. The aim of this paper is also to visualize the data and their mutual comparison. We suggest this by a reduction of a multidimensional feature space using principal component analysis.

2 Datasets

For the demonstration of the proposed approach two different datasets of images are presented. The first dataset comes from Uppsala university [8]. In this paper, a simulation framework for generating images used for Pap smear analysis in cervix cancer screening was developed. The total number of 25 monochromatic 2D histology images of real data and 5 batches of synthetic image data (each consisting of 5 images, 25 images in total) were included in the reference dataset provided by the authors.

The images in the second dataset consisted of 180 images of lung cancer cells with filopodial protrusions from 3D time-lapse acquired by fluorescence microscope [11]. The two subtypes of lung cancer cells were analyzed in this paper – the cells with overexpressing phenotype (90 real images and 90 synthetic images) and the cells with phospho-defective phenotype (also 90 real images and 90 synthetic images).

The texture of histopathology images for Pap smear analysis was evaluated over the whole 2D image as it was described by the authors of [8]. Unlike this dataset, in case of lung cancer cells, only the texture of the central interior regions ($31 \times 31 \times 3$ voxels) was compared in real and synthetic data following the same procedure as proposed by the authors of [11].

These two image datasets have been separately used as the input samples for testing the proposed validation method.

3 Methods

Haralick texture features [3] belong to very popular and widespread [2, 15, 7, 13] image descriptors. In this paper, we evaluated 14 Haralick descriptors. In the following text, methods of statistical comparison will be presented. Although these methods are well known and widely used, they have not been applied together in this context.

3.1 Univariate comparison of feature distributions

After achieving 14 values for each image in both, real and synthetic data, a comparison of distributions in these groups was performed using the quantile-quantile plot (Q-Q plot) [17]. When comparing two samples, the empirical quantiles are plotted in the x-y figure against each other for every descriptor. This method reveals the identical distribution in both samples if the points of the quantiles lie along the straight line with the slope of 1, while diversion of the points from this line indicates differences in distribution. This method has no assumptions put on the input data, and furthermore, it is not necessary to identify the distributions that are compared. In case of image descriptors, the Q-Q plot can help to compare the distribution of both groups of data in each descriptor separately, i.e. univariately.

Since the Q-Q plot is only a visual technique for data comparison, this procedure can be accompanied by the statistical test of Kolmogorov-Smirnov [9]. The tested null hypothesis is that the cumulative distribution functions is the same for the both samples $A = \{a_i | i = 1, \ldots, n_A; a_i \in \mathbb{R}\}$ and $B = \{b_i | i = 1, \ldots, n_B; b_i \in \mathbb{R}\}$. The empirical cumulative distribution for sample A is defined as $F_A(x) = \frac{\#\{a \in A | a \leq x\}}{n_A}$ and in the same manner for sample B. In our case, A stands for a sample of real data and B for a sample of synthetic data. The test statistic for the two samples is based on the largest distances between the two empirical distribution functions, which is $KS(A, B) = \sup_{x \in \mathbb{R}} |F_A(x) - F_B(x)|$. The value of Kolmogorov-Smirnov statistic is then compared to the critical value and the null hypothesis is then rejected if $KS(A, B) > \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}\kappa_\alpha$, where κ_α is chosen according to the α level of significance. Note that the Kolmogorov-Smirnov test is a non-parametric technique, which means it has no assumptions put on the given data. Therefore, it is easy to apply Kolmogorov-Smirnov test also to the Haralick descriptors with no limitation and with a clear decision if the data reveals the similarity in a particular descriptor.

3.2 Reduction of the feature space

As one could be interested in the assessment of all 14 Haralick features simultaneously, we suggest the multivariate assessment. This approach describes every

image by a vector in a 14-dimensional space. It is not possible to see visually the positions of the real and synthetic data points in such a high-dimensional vector space. Therefore the reduction with the retention of most of the total variability in three features is convenient. Since some of the Haralick descriptors are correlated, the reduction of the feature space is possible to perform and was done by applying the Principal Component Analysis (PCA), originally proposed by [10] and then extended by [4]. PCA is a reduction procedure with transformation to another uncorrelated feature space. The aim of the technique is to preserve the most of the variability in the new first features (components) by extracting the principal pattern of the linear system of descriptors. Afterwards, only some of the components can be selected to represent the new reduced data with the particular proportion of the variability. Other redundant features can be discarded.

The original feature space given by d descriptors ($d = 14$) and n_A real images and n_B synthetic images ($n_A + n_B = n$) can be represented as a matrix

$$X_{n,d} = \begin{pmatrix} x^A_{11} & \dots & x^A_{1d} \\ \vdots & \ddots & \vdots \\ x^A_{n_A 1} & \dots & x^A_{n_A d} \\ x^B_{11} & \dots & x^B_{1d} \\ \vdots & \ddots & \vdots \\ x^B_{n_B 1} & \dots & x^B_{n_B d} \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix} \tag{1}$$

where x_{ij} is value of $j-$th descriptor measured in $i-$th image. The matrix X can be centered by subtracting the sample multivariate mean $\overline{x}$ and scaled/standardized before further operations. In case of centered PCA, the distances among objects are equal to the distances in the original space, but the central point of the axis is shifted to the centroid of objects. This is used especially, when the scales of the variables are similar. In case of standardized PCA the variables are transformed to the variables with unit variance. It can be used if the scales of the variables are measured in different units, such as the case of Haralick descriptors. The transformation of PCA is then given by

$$Y_{n,d} = (X_{n,d} - 1_n \overline{x}^T)\hat{\Gamma}, \tag{2}$$

where $\hat{\Gamma}$ contains the eigenvectors of the sample covariance or correlation matrix $\hat{\Sigma}$ of the input data so that, $\hat{\Gamma}^T \hat{\Sigma} \hat{\Gamma} = \hat{\Lambda} = diag(\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_d)$, where the eigenvalues are in the order from the highest to the lowest $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_d \geq 0$. The new feature space is formed by principal components Y_i (the $i-$th column of the matrix $Y_{n,d}$).

PCA has an assumption of quantitative variables (but there exist some modifications), independence of objects and multivariate normality, which is sometimes difficult to achieve. However, some works [6] state that PCA is sufficiently robust to overcome this.

Showing only the first two or three principal components in an x-y plot or x-y-z plot, respectively, one can have an idea about mutual position and clusters

of the real and synthetic data in the uncorrelated reduced feature space. Note that the PCA is independent of the group assignment in this case and it is used only for visualization.

3.3 Overlap of samples

Now we have an analogue situation as we had in a univariate approach. The visual inspection gives us a subjective information about the similarity of the groups, but the objective assessment is missing. Let us compare the overlap of the data points of each sample in the reduced feature space. We will do that by comparing the ellipsoids that envelope these data points (see Fig. 2). The construction of ellipsoids is based on the sample covariance matrices and sample means to fit the data points on a level of 95% joint confidence interval. This enables to omit 5% of potential multidimensional outliers. Finally, the intersection volume and the volume of union are compared by the Jaccard similarity index [5]:

$$J(A, B) = \frac{V_A \cap V_B}{V_A \cup V_B}, \tag{3}$$

where the V_A and V_B stand for the volume of the ellipsoid constructed in reduced feature space over sample A or B, respectively. The index ranges values between $0 \leq J(A, B) \leq 1$, in case of total similarity it yields 1, in case of no intersection the index is equal to 0.

4 Results

First, the datasets were assessed univariately using the Q-Q plots for all of the 14 descriptors. To see the differences and similarities, the datasets were also compared in real image data only, divided randomly into two groups with an expectation of good results of homogeneous groups. The randomization into two groups was performed three times with similar results. In case of reference Pap smear real data divided into two groups (see Fig. 1), one can see the high level of similarity. Even if the points diverse from the straight line in some of the cases, the p-values of Kolmogorov-Smirnov test retain the null hypothesis stating to have the same distribution in both groups of real data.

To assess the mutual relationships in a multidimensional feature space, PCA was applied to all 14 Haralick features. The data were pre-processed by Tukey's ladder of powers in case of a violation of normality, centred and normalized, because the individual descriptors are measured in different scales. The three principal components with the largest contribution on the explained variability were used for a visualization. Analyzing the results of PCA, many input variables (descriptors) entering the transform were found to be highly correlated. This supports the idea that reduced space for visualization is needed.

To see as much as possible of the variability and the mutual positions of the samples, the points were depicted in the transformed 3D feature space given by principal components as axes. Inspecting the reference phospho-defective lung

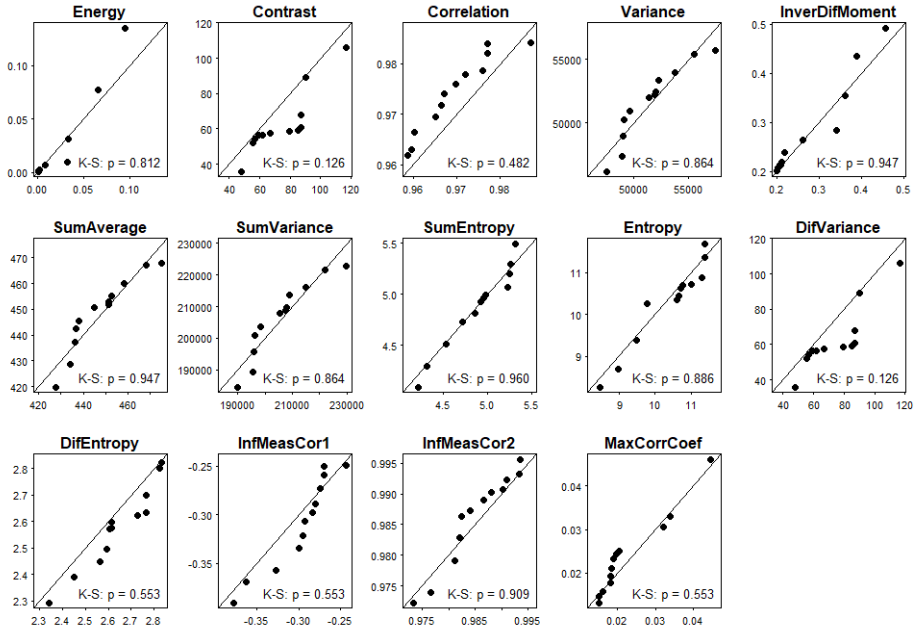

Fig. 1. Q-Q plots showing the similarity in distributions of 14 Haralick descriptors in comparison of two random samples of real image data from Pap smear test. The each plot is enriched by p-values of Kolmogorov-Smirnov test (K-S) supporting the difference between sample distributions when the value is ≤ 0.05 (in case of level of significance $\alpha = 0.05$).

cancer cells dataset (see Fig. 2), one can see only a partial overlap of the ellipsoids enveloping the synthetic data (blue) with those for real image data (yellow)(a-c). On the contrary, the visualisation of the two subsets of real data reveals a high overlap of the ellipsoids in all views (d-f). For a detailed description of all results, see the Table 1. The values of Jaccard similarity indices computed over volumes of ellipsoids correspond to the visualization done by PCA. When comparing the real and synthetic data of Pap smear images, the intersection volume covered only 16.1% of the volume of union of the two ellipsoids ($J = 0.161$). The best results were observed when comparing real *vs.* real image datasets ($J \geq 0.652$). The visual comparison of real overexpressing lung cancer cells with synthetic revealed visually a good overlap, however the Jaccard index was smaller according to the fact, that ellipsoid enveloping synthetic data was smaller than the ellipsoid enveloping real data.

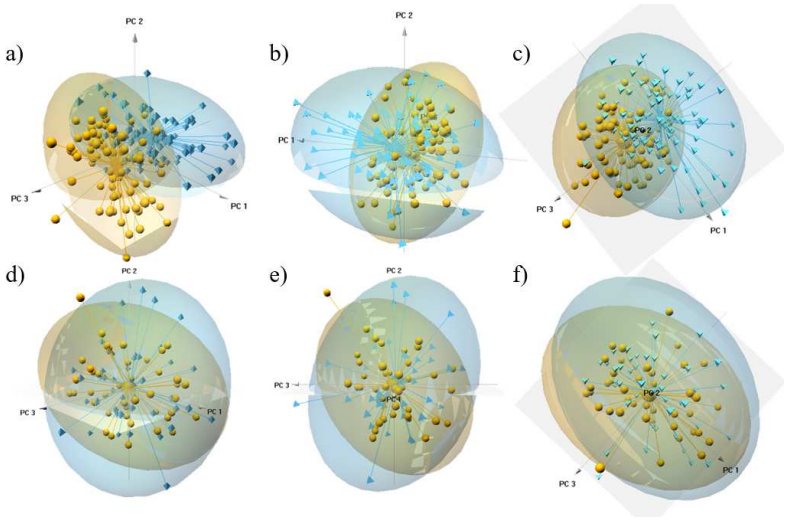

Fig. 2. Visual comparison of data in 3D feature space reduced by PCA into the three main components (axes PC1, PC2, PC3). Frontal (a, d), horizontal (b, e) and vertical (c, f) view on real (yellow) and synthetic (blue) data (a-c) and two groups of real data of lung cancer cells (d-f). The interactive tool of `pca3d` library in R software [14] enabled to rotate the set of points in all directions with the possibility of showing id of images, distances to the centroids and ellipsoids around each subset. The greater the intersection of each pair of ellipsoids, the higher the similarity of the corresponding evaluated samples can be expected.

5 Discussion

The real and synthetic data were compared to each other, at first univariately in Q-Q plots, which gave us some information about distributions. However, it was not possible to conclude from Figure 1 that the synthetic images covered only a limited part of the variability as it is shown in Figure 2(a). The PCA helped to reduce the high-dimensional feature space and therefore to visualize the important part of the variability and the position of the groups of images in a feature space. It was possible to visually assess the overlap of the ellipsoids constructed over 95% joint confidence interval representing a given group. To express this overlap numerically, the Jaccard index was suggested.

Exploring the subset of real images, the Jaccard values was never higher than 0.8. This is according to the variability given by each subset of real data and it should be kept in mind also when comparing real and synthetic data. The value of Jaccard $J = 1$ is impossible to achieve and if so, this result would refer to the overfitted simulator to the particular real dataset.

The low performance of the comparison of cervix data could result from evaluation of the Haralick features over the whole image consisting of many cells,

Dataset		Q-Q plots	K-S p-value	Visualization by PCA	Jaccard
Pap smear	**R×S**	Diverging a lot from the straight line in 4 descriptors (++)	Significant difference in 6 cases (+)	Synthetic data lie in the plane crossing the ellipsoid of real data (+)	0.161
	R×R	Good (++)	All right, no rejection (+ + +)	Ellipsoids have large overlap (+ + +)	0.652
Lung cells					
Over-expressing	**R×S**	Diverging a lot from the straight line in 1 descriptor (++)	Significant difference in 4 cases (++)	The ellipsoid of the synthetic data is smaller, but fitting a majority (++)	0.583
	R×R	Perfect (+ + +)	All right, no rejection (+ + +)	Ellipsoids have large overlap (+ + +)	0.796
Phospho-defective	**R×S**	Diverging a lot from the straight line in 5 descriptors (++)	Significant difference in 6 cases (+)	Ellipsoids overlap only in part of the data (++)	0.284
	R×R	Perfect (+ + +)	All right, no rejection (+ + +)	Ellipsoids have large overlap (+ + +)	0.659

Table 1. Results of methods for comparison on datasets ($R \times S$ – real vs. synthetic; $R \times R$ – real vs. real). The performance of the methods is described in a sentence (+++ suggest no difference in the samples, + stands for a low similarity).

in contrast to the lung cancer cells. The inner regions of single cells provide more homogeneous part of the image suitable for the evaluation.

Note that the similarity evaluation of cervix data was done only on 25+25 images. Despite the small sample size, the limitations of the synthesized variability were conspicuous.

In this paper, only the texture feature was assessed, however the PCA could be applied also to other characteristics of an image, such as shape. The descriptors are nevertheless based on some pre-processing and could be affected by the chosen method.

6 Conclusion

In the last years, the methods of cell image synthesis have been rapidly improving. However, the evaluation of the similarity of such generated data compared to real data, is still not sufficient.

This paper aimed to show new possibilities for texture comparison in 1) viewing data in context of multiple dimensions given by texture descriptors, 2) observe the mutual position of real and synthetic image data points in feature space reduced by PCA, and 3) quantitative measurement of similarity by comparing the volumes of ellipsoids enveloping given group of data in a feature space.

The described methods were applied to two datasets showing differences between groups of images. In two randomly chosen subsets of real images there was observed only a small difference between the groups with Jaccard index ≥ 0.652. The experiments comparing real and synthetic data showed different levels of

similarity. The Jaccard indices revealed corresponding results to the visual inspection in reduced feature space and univariate statistical comparisons.

In the future work we plan to extend our method for time-lapse image data, where the texture is time varying. The vision is also to give a constructive feedback to the designer of the synthesizing algorithm in identifying, why the images differ in a chosen descriptor.

Acknowledgement.
This work was supported by Czech Science Foundation, grant No. GA17-05048S.

References

1. Apou, G., Feuerhake, F., Forestier, G., Naegel, B., Wemmert, C.: Synthesizing whole slide images. In: 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA). pp. 154–159 (Sept 2015)
2. Boland, M.V., Murphy, R.F.: A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells. Bioinformatics **17**(12), 1213–1223 (2001)
3. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **SMC-3**(6), 610–621 (Nov 1973)
4. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psych. **24** (1933)
5. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Société Vaudoise des Sciences Naturelles **37**, 547–579 (1901)
6. Jolliffe, I.: Principal component analysis. Springer Verlag (2002)
7. Kovacheva, V.N., Snead, D., Rajpoot, N.M.: A model of the spatial tumour heterogeneity in colorectal adenocarcinoma tissue. BMC bioinformatics **17**(1), 255 (2016)
8. Malm, P., Brun, A., Bengtsson, E.: Simulation of bright-field microscopy images depicting pap-smear specimen. Cytometry, Part A **87**, 212–226 (2015)
9. Massey, F.J.: The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association **46**(253), 68–78 (1951)
10. Pearson, K.: LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901)
11. Sorokin, D.V., Peterlík, I., Ulman, V., Svoboda, D., Maška, M.: Model-based generation of synthetic 3D time-lapse sequences of motile cells with growing filopodia. In: IEEE Internationl Symposium on Biomedical Imaging. pp. 822–826 (2017)
12. Sorokin, D.V., Peterlík, I., Ulman, V., Svoboda, D., Nečasová, T., Morgaenko, K., Eiselleová, L., Tesařová, L., Maška, M.: Filogen: A model-based generator of synthetic 3d time-lapse sequences of single motile cells with growing and branching filopodia. IEEE Transactions on Medical Imaging (2018), *in press*
13. Svoboda, D., Ulman, V.: Mitogen: A framework for generating 3d synthetic time-lapse sequences of cell populations in fluorescence microscopy. IEEE Transactions on Medical Imaging **36**(1), 310–321 (Jan 2017)
14. Team, R.D.C.: R: A language and environment for statistical computing. R Foundation for Statistical Computing (2010), http://www.r-project.org

15. Tesar, L., Smutek, D., Shimizu, A., Kobatake, H.: 3d extension of haralick texture features for medical image analysis. In: Proceedings of the Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications. pp. 350–355. SPPRA '07, ACTA Press, Anaheim, CA, USA (2007)
16. Ulman, V., Svoboda, D., Nykter, M., Kozubek, M., Ruusuvuori, P.: Virtual cell imaging: A review on simulation methods employed in image cytometry. Cytometry Part A **89**(12), 1057–1072 (2016)
17. Wilk, M.B., Gnanadesikan, R.: Probability plotting methods for the analysis for the analysis of data. Biometrika **55**(1), 1–17 (1968)