# Subpixel precision in registration of multimodal datasets

IOP Conf. Series: Materials Science and Engineering **949** (2020) 012007 doi:10.1088/1757-899X/949/1/012007

# Subpixel precision in registration of multimodal datasets

**Matěj Lébl[1,2], Jan Blažek[1], Jana Striová[3], Raffaella Fontana[3], Barbara Zitová[1]**

(1) The Czech Academy of Sciences, Institute of Information Theory and Automation
(2) Faculty of Mathematics and Physics, Charles University, Czech Rep.
(3) INO National Research Council (CNR)-National Institute of Optics (INO)

E-mail: blazek@utia.cas.cz

**Abstract.** The motivation for our research is the huge demand for registration of multimodal datasets in restorers practice. With an increasing number of various screening modalities, each analysis built on the acquired dataset starts with the registration of images acquired from different scanners and with varying levels of mutual correspondence. There is currently no well-suited state of the art method for this task. There are many existing approaches, i.e. based on control points or mutual information, but they do not provide satisfying (subpixel) precision, thus the registration is very often realized manually in Adobe Photoshop™or any similar tool. Another popular option is to use scanners able to produce registered datasets by design. During the last 10 years, datasets from these devices have extended available analytical techniques the most.

In our research, we focus on solving the mentioned registration task. In [1] we concluded that the work with misregistered modalities is possible but limited. Now we present results of our experiments challenging these limits and conditions under which we can precisely register data from different modalities. The achieved results are promising and allow usage of more complex artificial neural networks (ANN) for dataset analysis e.g. [2]. We describe the construction of registration layers for estimation of shift, rotation and scale and a useful strategy and parametrization for ANN optimizer.

## 1. Introduction

The registration of artwork images is a crucial part of any higher level analysis. The usefulness of the registered dataset was demonstrated e.g. in [3, 4, 5, 6] where spectral reflectance of each pixel was used for pigment identification. Another demonstration, working with registered data, is focused on layer identification e.g. in [7, 8, 9, 10]. There exist scanners producing already registered datasets (macro X-ray fluorescence (MA-XRF) [11], visible (VIS) and near infrared (NIR) [12], XRF + VIS - NIR [13]) however their set of acquired modalities is limited. Except these few scanners input datasets from other devices must be registered using some software based approach.

### *1.1. Modality*

Registration of a multimodal dataset is more complex than photograph stitching because the information content of the images from different modalities varies. Sometimes less (VIS and NIR), sometimes more (ultraviolet fluorescence (UVF) and VIS) but in some cases, the

information content is completely different (some MA-XRF channels and VIS). The current state of the art uses either a manual registration in Photoshop$^{TM}$, approaches based on control point selection [14] or information based similarity measures (like e.g. mutual information [15]), which are especially useful when information content is similar.

As far as we know there is no publication quantifying the similarity level needed for registration convergence as well as there is a lack of studies comparing the usefulness of particular information measures for registration of XRF, UVF, VIS, NIR, terahertz imaging (THz), optical coherent tomography (OCT) and other. The correlation [14] and the mutual information [16] are the most common thanks to their available implementations in gradient descent methods.

In our approach where neural networks will be used for transformation estimation, we assume that:

(i) The transformation parameters are limited (to prevent local optima convergence). This means that images are roughly pre-registered.

(ii) Registered images have non-trivial pixel intensity gradients (to be able to use gradient descent methods). An image pair should contain edges, corners, various noise distribution in different parts.

(iii) The global minimum of square error function corresponds to the correct transformation parameters.

The first two of our assumptions are achievable, the third we will discuss further.

### *1.2. Typical distortions*

Before we start to build our model it is appropriate to restrict the space of possible transformations according to the physical properties of acquisition devices. By this restriction we reduce the number of optimized parameters and in this way we improve the convergence of gradient descent methods. As well we should define the required precision of the registration according to the following data application noise robustness. We establish two levels of necessary precision i.e.:

- rough - for studying large areas - up to $1px$ misplacement.
- fine - for studying pixel spectral responses - less than $1px$ misplacement error.

The physical behavior of acquisition devices produces two types of transformations we are interested in: perspective, caused by the positioning of acquisition device and spherical, caused by lens distortion.

Perspective transformation is well described by 4 control points and we can describe it by 8 parameters $[a, b, c, d, e, f, u, v]$. Coordinates $[x, y]$ are transformed to $[x_T, y_T]$ as follows [17]:

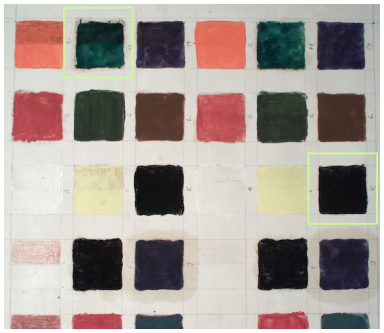
$$x_T = \frac{ax + by + c}{ux + vy + 1} \text{ and } y_T = \frac{dx + ey + f}{ux + vy + 1} \tag{1}$$

Very often we are also able to keep the orientation of the painting between two captures and therefore rotation can be eliminated from the list of allowed transformations.

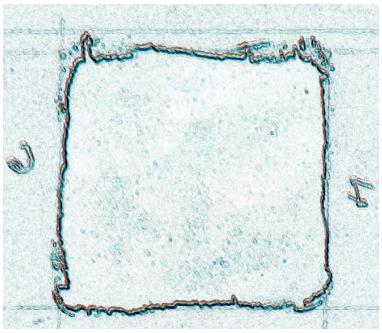
Radial distortion is present when a lens is used for capturing (the whole painting is captured at once). This effect is nicely demonstrated on VIS+NIR screening made by the same camera see Figure 1.

Radial distortion is not mathematically precisely defined and we approximate it by [18]:

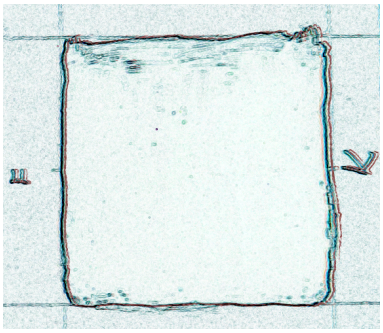
$$x_T = x + \frac{x - c_x}{1 + k_1 r^2 + k_2 r^4 + \ldots} \text{ and } y_T = y + \frac{y - c_y}{1 + k_1 r^2 + k_2 r^4 + \ldots}, \tag{2}$$



(a) Origin of cuts (original size was $3156 \times 4752px$). Green rectangle indicates zoomed areas.



(b) **Left top part**. Misplacement of NIR and VIS contour is approx. 6 pixels to the right



(c) **Mid right part.** Misplacement of NIR and VIS contour is approx. 6 pixels to the top.

Figure 1: **Misplacement of VIS and NIR images.** Comparison of two images which were taken by a Canon EOS 500D from the same position. Blue (VIS) was taken with UV/NIR blocking filter, red (NIR) with NIR $> 850nm$ passing filter. Images were reduced by edge detector and alpha blended.

where $[c_x, c_y]$ denotes center of radial distortion coordinates, $\forall i \in N, k_i$ are estimated parameters and

$$r = \sqrt{(x - c_x)^2 + (y - c_y)^2}. \tag{3}$$

Expected parameters are then $[c_x, c_y]$ and $k_i$ $(\forall i \in N)$. Higher number of $k_i$s (bigger $i$) provides higher precision of the approximation.

## 2. Method

Our target algorithm for registration has to take into account both types of transformations (see Equations (1),(2)) but it should be modular as well. Modularity is useful in the cases when one can eliminate some of the transformations (e.g. rotation, scale, sheer) which leads to a lower number of parameters to be optimized and therefore better convergence. We try to construct such a modular system and test how it performs with real data.

### *2.1. Multilayer ANN with backpropagation algorithm*

The core of our construction is a multilayer ANN (see Figure 2). This ANN has two types of inputs. First, a pixel coordinates $[x, y]$ which are transformed by very first layers to $[x_T, y_T]$. And secondly, an image from source modality which is used for transformation of $[x_T, y_T]$ to pixel intensity $I_{VIS}(x_T, y_T)$. The following ANN layers are pretty much the same as in [2] section 3.3. which transform source modality to target modality by minimizing:

$$\min_{f_T} \sum_{x_T, y_T} I_{NIR}(x, y) - f_T(I_{VIS}(x_T, y_T)) \tag{4}$$

This construction allows requested modularity as well as universality because we can setup layers according to expected transformation. We can establish trained parameters corresponding with:

- Rotation angle $[r_{ang}]$
- Scale $[s_x, s_y]$
- Shift $[t_x, t_y]$
- (Radial distortion parameters $[c_x, c_y, k_1, k_2, \dots]$)

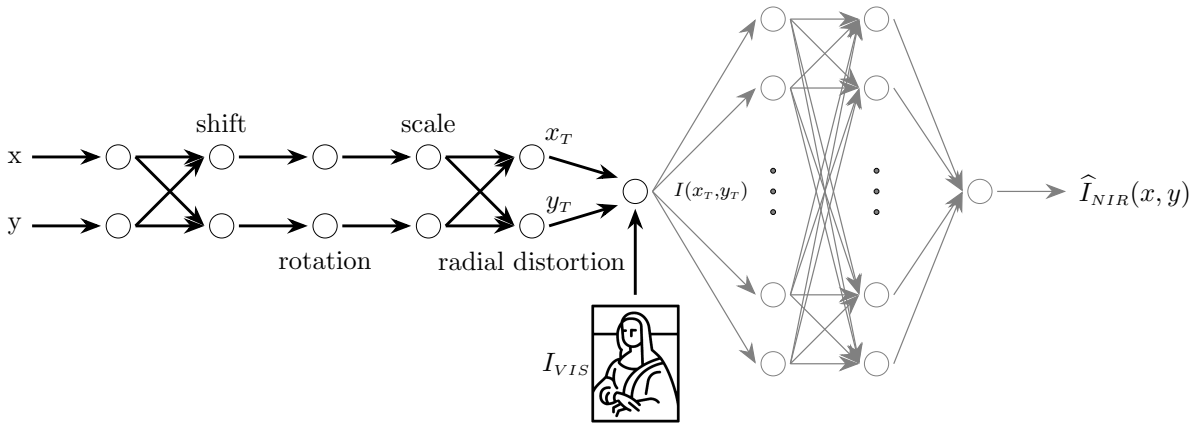



Figure 2: ANN schema. Left part (black) symbolically describes registration layer of ANN. Here we have four layers with various transformations. Transformed pixel coordinates are then used for pixel intensity interpolation from source modality which is then (gray) transformed, according to [2], to target modality. As an error is used mean square error (MSE) of target intensity with the predicted.

*2.1.1. Shear* Adding shear is more complicated and it will be part of further research. The problem is caused by the fact that to get an affine transformation, we have to apply shear twice (for x and y-axis). Alternatively, we can achieve shear with two rotations and an anisotropic scale. However, in both cases, we have to apply the same parameter in two different layers of ANN. In this way, we lose the transparency and interpretability of ANN layers.

*2.1.2. Normalization and optimization* Because shift, rotation, and scale all attaining different values (tens of pixels for the shift, tenths of radians for the rotation) they have a different impact on error function. This impact also depends on the dataset and level of pre-registration. For this reason, we suggest normalization of parameters as well as recommend the usage of the ADAM optimizer. For $t_x, t_y, r_{ang}, s_x, s_y$ we estimate reasonable ranges. Then, the parameters are scaled i.e. estimated ranges were stretched to $[-1, 1]$ interval. In this way, all parameters have approximately the same effect on the error function. Finally, the ADAM optimizer changes the learning rate for each parameter separately and thus compensates for the variances in the input data.

During testing, we encountered a problem with ADAM's convergence to the subpixel level. Near the true transformation parameters, ADAM fails to converge and oscillates around the exact solution. This could happen as it was described in [19]. To solve this we used the scaled gradient descent (SGD) optimizer to refine the ADAM's solution. Without the SDG's refinement, we obtain the error in the order of pixels while with SDG refining step we achieved sub-pixel precision ($2^{\text{nd}}$ decimal place).

## 3. Implementation and testing

For the implementation, we have used Python library TensorFlow where custom layers were written. Hyperparameters were then found in large scale testing in Metacentrum - the Czech national computation grid. Our code can be found at *https://github.com/gimlidc/igre*

*3.1. Tensorflow*

From the TensorFlow custom layers we have derived 4 custom layers (see Figure 2) which estimate shift $[t_x, t_y]$, rotation $r_{ang}$ and scale $[s_x, s_y]$. The fourth layer does bilinear interpolation

and converts $[x_T, y_T]$ to $I_{VIS}(x_T, y_T)$. Layers for radial distortion estimation are in progress.

(i) **Shift layer** - It is a custom layer with two parameters representing translation in $x$ and $y$ axis. Parameter values are constrained with the hyperbolic tangent. The expected shift is set to $[-50, 50]$. For the shift to be learned, we had to accelerate the learning rate by a factor of 2000. The default value is zero.

(ii) **Scale layer** - It is a custom layer with two parameters representing the anisotropic scale in $x$ and $y$ axis. Parameters in this layer do not need to be constrained, bound for scale was set to be 10%, i.e. values in the interval $[0.9, 1.1]$. The default value of the parameter is one (in scaled range it is zero in the expected range $[-1, 1]$), the parameter represents percentage deviation from the default scale which is one.

(iii) **Rotation layer** - It is a custom layer with one parameter representing rotation angle in radians. The learning rate for this layer was scaled down by a factor of 10. No additional constraints were used, expected rotation is within range $[-4°, 4°]$. Default value is $0°$.

(iv) **Interpolating layer** - It is a custom layer that takes image coordinates transformed by the previous three layers (similarity transformation) and assigns a pixel value to each coordinate using linear interpolation on $2 \times 2$ neighborhood. The custom gradient was also implemented as a standard image gradient but over the larger neighborhood ($4 \times 4$ patch without corners = 12 pixels)

*3.1.1. Learning rate issues* As mentioned above, we encountered great obstacles while setting up the optimization. First, let's take into account only the shift layer. With default/recommended parameters for the ADAM optimizer, the translation never moved far from the initial values, resulting in a detected shift being under one pixel in any situation. Increasing the learning rate solved this problem however finding the correct parameters proved to be difficult. The effect of increasing the learning rate is not linear. For a large range of values, the optimizer gets stuck in some local minimum near the initial position. Increasing learning rate too much leads to parameter explosion i.e. unconstrained growth.

Adding another layer further complicates the convergence. Keeping the settings from shift-layer-only case leads to convergence for purely shifted images however it failed to detect rotation/scale. Interestingly for certain settings rotational parameter was correctly estimated only in the cases where translation was also present. Setting the correct learning rates for each layer thus cannot be done separately and the influence of presence/absence of each transformation must be explored.

We set our ANN to detect transformation with translation up to $50px$, rotation up to $4°$ and scale up to 10%. Convergence was tested for all possible combinations of these transforms (e.g. just translation, scale + rotation, etc.). These bounds should be more than sufficient in the case where input images are (roughly) preregistered.

*3.1.2. Local optimum and Data preprocessing* One of the main reasons for failed registration is the convergence into a local minimum. This is more frequent in the cases of multimodal data processing. To prevent this phenomenon and to help the optimizer we prepared stages of training with blurred images. In stage one we apply heavy Gaussian blur. This leads to alignment based on larger more uniform regions rather than details. We repeat this process in stages two and three with moderate and small blur. Finally, in the fourth stage, we use original, clean data. After this stage, the error is around 1 pixel. Finally, in the fifth stage, we switch optimizer for SGD. This helps us further improve the error to subpixel precision. Note that using the ADAM optimizer for the early stages as well as applying the blur is necessary. The SGD optimizer alone was not able to handle even small transformations. Even in the case of non-blurred data, the optimization diverged or got stuck in a local minimum. The last stage with SGD optimizer

fixed ADAM oscillations near the correct solution. Our last experiments shown that SGD is necessary especially for $t_x$ and $t_y$ estimation. In the last stage it is therefore possible to disable *trainable* parameter for $r_{ang}$, $s_x$ and $s_y$ and refine the translation alone.

### *3.2. Metacentrum*

ANN convergence to the true transformation highly depends on many parameters. Moreover, without extra assumptions about the data, the convergence is hard to achieve. For this reason, we set up experiments instead of an analytical approach to direct method evaluation.

We have started with shift estimation. As an input for the experiment data from INO-CNR captured by 32 band VIS-NIR scanner [12] precisely registered by design was used. We simulated the transformation by shifting VIS bands and then tried to find out shifting parameters by ANN. This first experiment also helps us with setup for the learning rates $[lr, beta1, beta2]$ and training stages (a blur of the input image in the first three stages, SGD usage for final refinement).

In this first experiment shift range was set in the range $[-50, 50]$px which we consider as sufficient for nowadays available 20Mpx cameras. We have processed 43 image samples of size $400 \times 400$px of different kinds (detailed structures and line sketches as well as gradual color transitions). Every experiment was repeated with the same parametrization, but with different gradient descent seed, 20 times to obtain statistically relevant data.

In the second experiment, we have enabled similarity transformation (rotation and scale layers were added). The ranges for transformation were set to $[-50, 50]px, [-4°, +4°], [0.9, 1.1]$. Again, we expect rough pre-registration by humans or process of screening where these limits seem achievable for nowadays available cameras. In this test case, 350 different transformations were tested in two repeats per each sample (30k runs altogether).

In the last experiment, we tried to combine registration layers: scale, rotation, and shift, with extrapolation layers from our previous paper [2]. In this case, we focus on the performance of registration in sense of combining different modalities (first two tests works mainly with the same modality in the input layer as well as in the error computation).

For such big scale computation, we used the National computation grid Metacentrum.

## 4. Results

The test scenarios described in the previous subsection were evaluated and we illustrate them by following graphs.

The first test (shift estimation) evaluates if our approach converges at all. We have expected good convergence for the same modality but we also try to estimate modality distance, where convergence occurred.

This test shows that custom layers written in TensorFlow work as expected. Moreover, we show (see Figure 3) that for shift convergence slight shift in a spectral band is not an issue. We have used as an input modality sub-band with middle wavelength $700nm$ and we were able to find out correct $[s_x, s_y]$ for input modality from $620nm$ up to $1150nm$. The wider modality distance for input and output ($450nm$ versus $80nm$ in NIR versus VIS) is caused by lower variability of reflectance signal between NIR subbands.

In the second test we have tried more complex transformations and optimization of 5 parameters $[t_x, t_y, r_{ang}, s_x, s_y]$. We also tested all transformation combinations on 43 different datasets to evaluate, how relevant are the data itself in the sense of registration convergence. Results are shown in the Figure 3. The worst results have samples 26-34 which contain drawings (other samples were paintings).

Finally, the last experiment with approximation layers should demonstrate, that the composition of ANN by registration and approximation layers works as expected. The expected output is a good registration convergence for distant bands. This last experiment was not so successful especially for $t_x, t_y$ where the error was about a $1px$ misplacement. However, we were

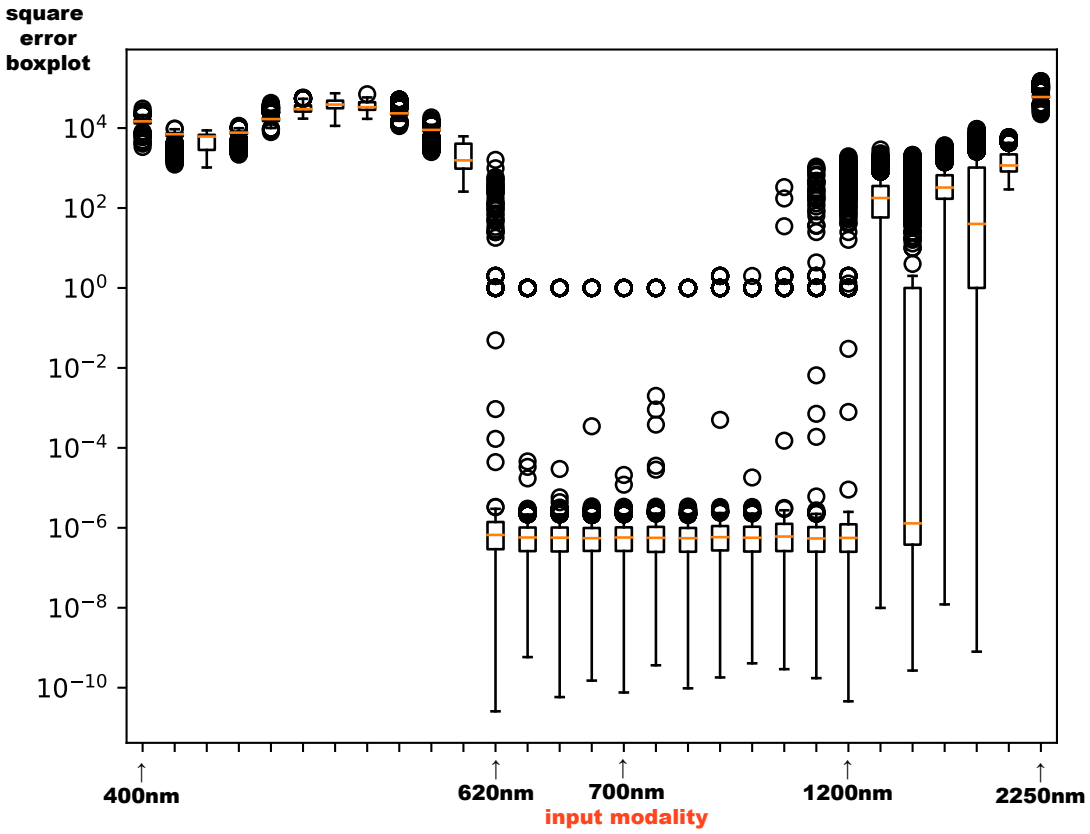



Figure 3: Convergence of shift estimation for various modality distances

able to demonstrate, that spectral distance between input and output modality can be bigger without any negative effect on registration convergence. The Figure 5 shows that we were able to extend convergence of the ANN to the correct registration parameters from $80nm$ spectral distance between input and output in VIS range up to $220nm$ and for NIR range from $350nm$ up to more than $900nm$ which is sufficient output for standard DLSR camera with removed NIR filter.

## 5. Conclusion

Half of our goal, to register multimodal images of artwork, was met. We suggested new architecture of ANN for combining registration layers with modality transformation layers. We tested this architecture for a shift, rotation and scale on real datasets at a huge scale and set up limits for successful registration layers application. We suggested the configuration of ANN and algorithm for ANN training suitable for estimation of registration parameters and we evaluated it on more than 2M experiments.

Our approach to registration of multimodal data is very promising. We see here a significant unlock which will enable the application of algorithms for pigment or layer identification to low-costly acquired data. Moreover, the usefulness of pigment databases now available will be proven in the short term in everyday restorers and conservators' practice.

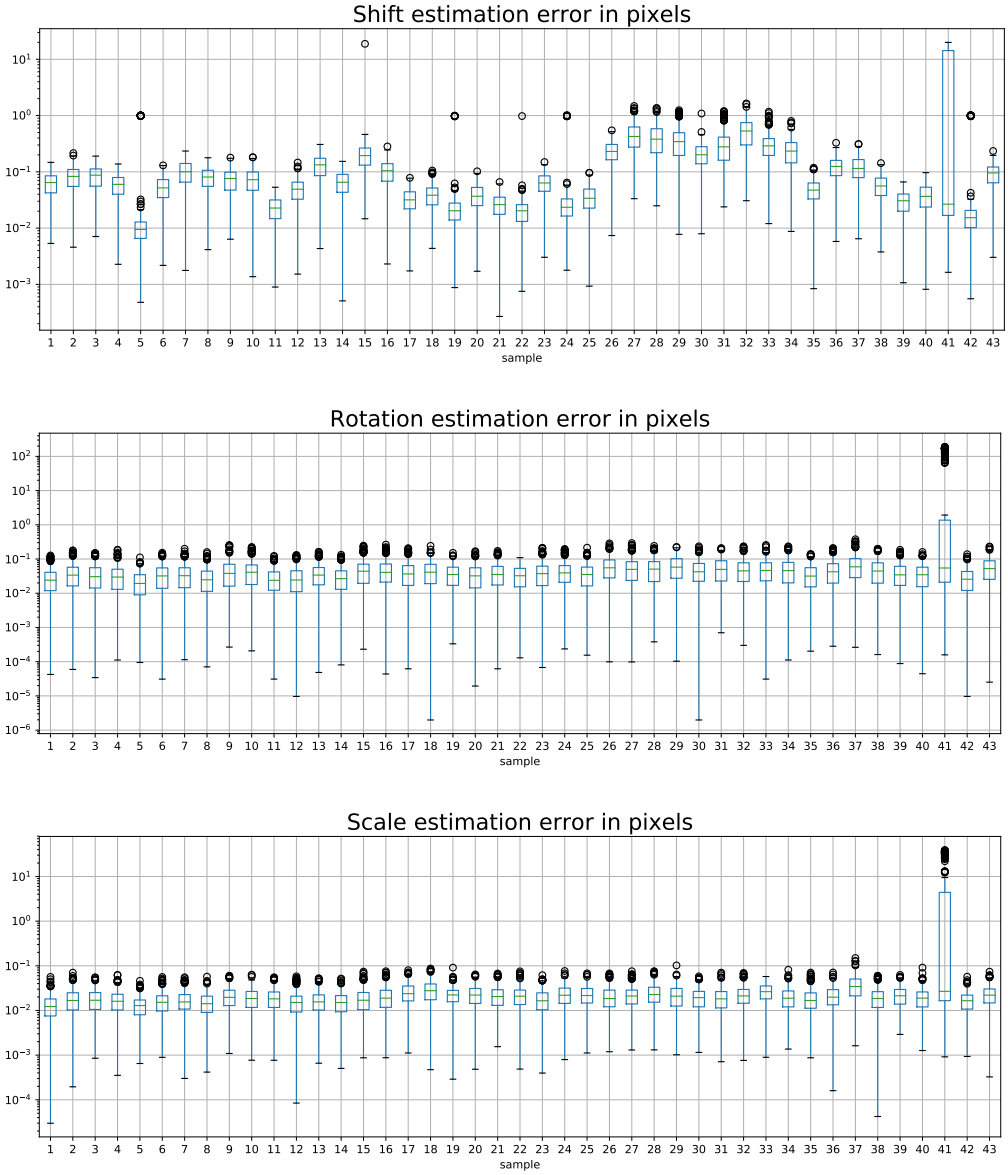



Figure 4: Convergence overview for testing samples: shift, rotation and scale. The boxplots demonstrate that data strongly influence the result. Samples 26-34 had problem to converge $t_x, t_y$. We do not currently know if this is caused by optimizer configuration which must be adapted for such data (expected) or by the type of the data itself.

## 6. Future work

There is still a large amount of work necessary to fully cover artwork dataset registration needs. In the next few months, we would like to develop and test an ANN layer for radial distortion. In parallel, we still need to improve ANN optimizers for better convergence, especially for a complex scenario.

Our current plan for pushing our work forward is as follows:

- Optimizer setup for the shift, rotation, and scale (before Heritech 2020)
- Spherical distortion software compensation (Jun 2020)

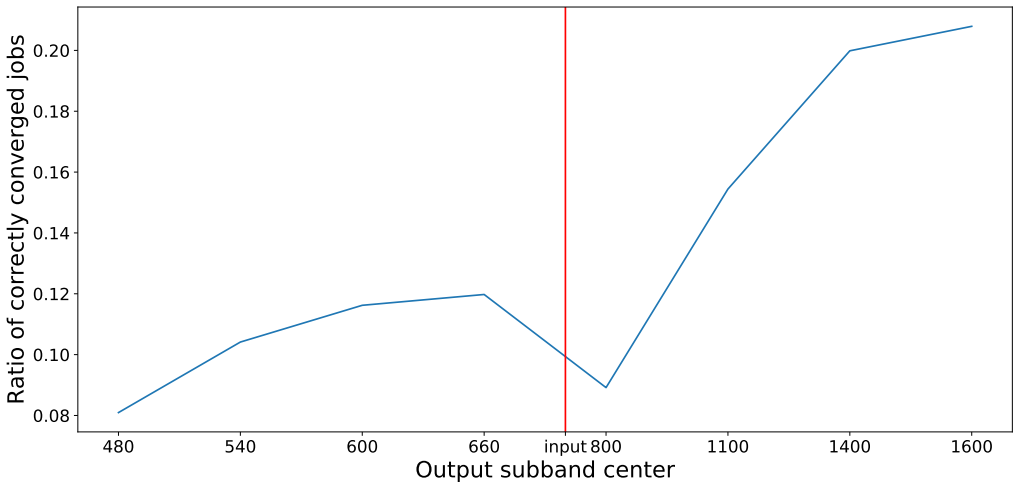



Figure 5: Last experiment takes as an input image in 700nm and output wavelength goes from 480 up to 1600, we have tested how many of test runs on 44 different samples converge to the correct transformation (error less than $1px$). ANN were constructed from registration layers (shift, rotation and scale) and two layer extrapolating intensity from input modality to output modality.

- Extension of TensorFlow python libraries with designed registration layers (Mid-Late 2021)

## 7. Acknowledgement

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

## 8. References

[1] Jan Blažek, Oldřich Vlašic, and Barbara Zitová. Improvement of the visibility of concealed features in misregistered NIR reflectograms by deep learning. *IOP Conference Series: Materials Science and Engineering*, 364:012058, jun 2018.

[2] Jan Blažek, Jaa Striová, Raffaella Fontana, and Barbara Zitová. Improvement of the visibility of concealed features in artwork NIR reflectograms by information separation. *Digital Signal Processing*, 60:140–151, jan 2017.

[3] Mauro Bacci, Marcello Picollo, Giorgio Trumpy, Masahiko Tsukada, and Diane Kunzelman. Non-invasive identification of white pigments on 20th-century oil paintings by using fiber optic reflectance spectroscopy. *Journal of the American Institute for Conservation*, 46(1):27–37, 2007.

[4] Antonino Cosentino. Identification of pigments by multispectral imaging; a flowchart method. *Heritage Science*, 2(1):8, 2014.

[5] Bartosz Grabowski, Wojciech Masarczyk, Przemysław Głomb, and Agata Mendys. Automatic pigment identification from hyperspectral data. *Journal of Cultural Heritage*, 2017:1–12, 2018.

[6] J. Striova, C. Ruberto, M. Barucci, J. Blažek, D. Kunzelman, A. DalFovo, E. Pampaloni, and R. Fontana. Spectral Imaging and Archival Data in Analysing Madonna of the Rabbit Paintings by Manet and Titian. *Angewandte Chemie - International Edition*, 2018.

[7] Aurèle J L Aurele J L Adam, Paul C M Planken, Sabrina Meloni, and Joris Dik. Terahertz imaging of hidden paint layers on canvas. In *34th International Conference on Infrared, Millimeter, and Terahertz Waves, IRMMW-THz 2009*, volume 17, pages 1–2. IEEE, sep 2009.

[8] Matthias Alfeld, Wout De Nolf, Simone Cagno, Karen Appel, D. Peter Siddons, Anthony Kuczewski, Koen

Janssens, Joris Dik, Karen Trentelman, Marc Walton, and Andrea Sartorius. Revealing hidden paint layers in oil paintings by means of scanning macro-XRF: a mock-up study based on Rembrandt's "An old man in military costume". *J. Anal. At. Spectrom.*, 28(1):40–51, 2013.

[9] Piotr Targowski and Magdalena Iwanicka. Optical coherence tomography: Its role in the non-invasive structural examination and conservation of cultural heritage objects-A review. *Applied Physics A: Materials Science and Processing*, 106(2):265–277, 2012.

[10] David Thurrowgood, David Paterson, Martin D de Jonge, Robin Kirkham, Saul Thurrowgood, and Daryl L Howard. A Hidden Portrait by Edgar Degas. *Scientific Reports*, 6:29594, 2016.

[11] Matthias Alfeld, Koen Janssens, Joris Dik, Wout de Nolf, and Geert van der Snickt. Optimization of mobile scanning macro-XRF systems for the in situ investigation of historical paintings. *Journal of Analytical Atomic Spectrometry*, 26(5):899, 2011.

[12] C Bonifazzi, P Carcagnì, R Fontana, M Greco, M Mastroianni, M Materazzi, E Pampaloni, L Pezzati, and D Bencini. A scanning device for VIS–NIR multispectral imaging of paintings. *Journal of Optics A: Pure and Applied Optics*, 10(6):064011, jun 2008.

[13] John K. Delaney, Damon M. Conover, Kathryn A. Dooley, Lisha Glinsman, Koen Janssens, and Murray Loew. Integrated X-ray fluorescence and diffuse visible-to-near-infrared reflectance scanner for standoff elemental and molecular spectroscopic imaging of paints and works on paper. *Heritage Science*, 6(1), 2018.

[14] Ernestine Zolda Paul Kammerer, Allan Hanbury. a Visualization Tool for Comparing Paintings and Their Underdrawings *. pages 148–153, May 2004.

[15] Paul A Viola. Alignment by Maximization of Mutual Information. Technical Report 1548, Massachusetts Institute of Technology, jun 1995.

[16] Anila Anitha, Andrei Brasoveanu, Marco Duarte, Shannon Hughes, Ingrid Daubechies, Joris Dik, Koen Janssens, and Matthias Alfeld. Restoration of X-ray fluorescence images of hidden paintings. *Signal Processing*, 93(3):592–604, 2013.

[17] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, oct 2003.

[18] Andrew W Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR*, pages 125–132, 2001.

[19] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *CoRR*, abs/1904.09237, 2019.