Scale invariant image registration in the domain of art conservation

Mihail Spasov
Supervisors: Ruben Wiersma, Ricardo Marroquim
EEMCS, Delft University of Technology, The Netherlands

23-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Image registration is the process that overlays two or more images from different sources taken at different times and angles. Art conservators take various scans of paintings and then register them against the original in order to learn more about the working style of the artist, materials used and physical changes throughout time. This paper describes how scale space theory could be applied with a variant of cross-correlation to find the right scale at which to register the input images. From there, the standard state-of-the-art approach is used to register the images. This added step allows automatically registering a wider range of images.

1 Introduction

Art conservators use various imaging techniques to improve the condition, find insights into the working process of an artist and analyze changes throughout time. Some of these insights are only apparent when overlaying the scanned image on top of the regular image of the painting. Manual inspection and alignment is costly and error-prone, which raises the question: *Can this be done automatically?*

The task of registering images is not particularly new, but in the multimodal context, solutions are quite recent and often lack a common feature, namely scale invariance. This is the main focus of this research. This invariance would enable automatically registering partial scans of images, which might then be used for mosaicking a fully registered image. The scale invariance has been achieved in the unimodal context - SIFT being a prominent algorithm [4]. However, these scale invariant approaches do not extend well in situations where intensity, features and information highly vary across modalities. A new robust approach is outlined which solves the aforementioned concerns.

Manual scaling might become inefficient and error-prone when dealing with a lot of image patches. Using well established algorithms for automatic scaling, such as SIFT, do not always produce an optimal result. This approach allows automatic scaling to take place and then register images with an approach developed by Conover et al. [1].

Phase correlation has proven to be very effective in stitching and mosaicking images across various modalities [1,3,5]. Our method extends the idea of cross-correlation of images to the three dimensional domain. This is done through building a scale space of the phase image for both images and then computing cross-correlation of the 3D signals as shown in Figure 1. Peak selection is conducted to find the largest corresponding scale of each image. Our algorithm estimates the optimal scale of the template image. The method relies on the premise that a good correlation between two image sections would be apparent at multiple scales and the overall peak would be achieved where the input images are at the same size.

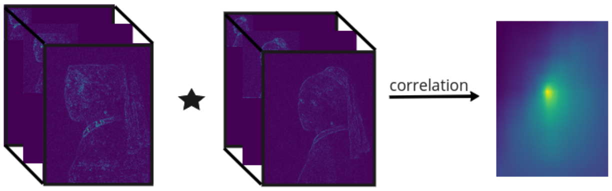

Figure 1: Cross-correlation of three dimensional signals

2 Related Work

Conover et al., propose an algorithm for registering multimodal images of paintings based on phase correlation [1]. It first computes the wavelet transform of the template image and uses local maximums to distinguish feature points. Based on a variant of phase correlation, it computes the locations of the feature points in the reference image to form pairs of matches. After that, the points are filtered to remove false matches. The remaining pairs are used to define a spatial transform for aligning the two images. Conover's approach relies on the premise that the input images have more or less the same size and orientation. The scale variance is what the proposed algorithm aims to remove.

Lowe proposed robust scale and rotation invariant approach for registering images called SIFT [4]. Based on scale-space theory [2], it first constructs a Gaussian scale space in order to calculate a difference of Gaussian (DoG) pyramid. Key points are then found by searching for local extreme points in the DoG pyramid. The description of the keypoints is done via statistical gradient histograms. Keypoints are extracted for both images and then matches are computed. One of the biggest drawbacks of SIFT is that it was designed for regular RGB and generally does not extend well to the multimodal domain, i.e. images taken from different sensors. The proposed algorithm takes the idea of scale space, but unlike SIFT, uses phase transformation and cross-correlation to compute the transformation function.

Phase correlation is an efficient way to estimate the translation change between a pair of similar images. A limitation to the phase correlation is that the rotation and scale differences affect accuracy tremendously. A log-polar transform could be used to recover the differences in scale and rotation because of its properties - scaling in Cartesian space is equivalent to translation along the radial coordinate, rotation in Cartesian space is equivalent to translation along the angular coordinate of log-polar space, translation differences in the spatial domain do not impact the magnitude spectrum in the frequency domain. Two algorithms build on top of this idea [6,7]. The scale factor up to which they reliably register images is 2.0, which is sufficient in the domain of medical imagery. In the context of art conservation - where an increasing effort is made to capture paintings at larger and larger scale with an immense amount of detail - there is no guarantee that the input images will not differ by a scale factor larger than 2.0.

3 Background

This section elaborates on some of the terminology used, going formally into the mathematical equations that the ideas are based on and some assumptions that were made during the crafting of the algorithm.

3.1 Cross correlation

Cross correlation is a similarity measure of two signals as a function of the translation of one relative to the other. Formally, when dealing with discrete finite signals $f, g \in \mathbb{C}^N$, such as images, the cross-correlation is defined through the following equation:

$$(f \star g)[n] \triangleq \sum_{m=1}^{N-1} \overline{f[m]} g[(m+n)_{\text{mod } N}]$$

where $f[n]$ is the value of f at location n. The bar operator, the vertical line above $f[m]$, denotes the complex conjugate of that number. The complex conjugate is the number that has the same real value, the complex value has the same magnitude, but an opposite sign. Because images comprise of only real valued numbers, in our case $\overline{f[m]} = f[m]$.

In practice, the cross-correlation does not use modulus when the index is out of bounds, but relies on padding. If the padding is circular, then effectively the result would be the same. Padding could also be constant, where the same number is used to pad everywhere, and if the signals are normalized a zero padding would be a reasonable choice. We avoid using padding by ensuring that the size of the template is smaller than the reference at all times.

3.2 Phase correlation

Phase correlation is a variant of the cross-correlation technique. It operates in the frequency domain, usually by doing fast Fourier transforms of the images. Phase correlation is used to estimate the translative offset of two similar images. It is widely used in the context of image registration.

3.3 Scale space

Scale space theory has advanced the notion of object representation, more specifically allowing an object to be present at multiple scales to deal with the fact that *objects are only meaningful entities over certain ranges of scale* [2].

The typical structure for constructing scale space is represented by a pyramid where two terms play an important role - layers and octaves. Different layers are achieved through repeatedly applying a Gaussian filter on the image, i.e., blurring it. The octaves are produced by repeatedly downsampling the image by a given factor. Such a Gaussian pyramid is depicted in Figure 2.

The algorithm deviates from the standard approach in two major ways. Firstly, the three dimensional geometrical object that represents the images is a rectangular cuboid, hereinafter referred to as just cuboid, in contrast with the pyramid. Secondly, the approach combines layers and octaves into one variable - $levels$. The levels are built by repeatedly applying the following: 1) downscale by a given factor, 2) apply a Gaussian filter with $sigma$ value of 2 * downscale factor / 6.0, which covers more than 99% of the Gaussian distribution.

3.4 Assumptions

In order to perform cross-correlation without relying on padding, the size of the template image must be less than or equal to the size of the reference image. This imposes an upper bound on the estimated size of the template image. More specifically, the estimated size will always be less than or equal to the size of the reference image. This essentially means that the template image should not depict a bigger section of the painting than what the reference has captured. If that is not the case, then the maximum size of the template image is equal to a subsection of the image itself, resulting in a mismatch between the optimal and the estimated value.

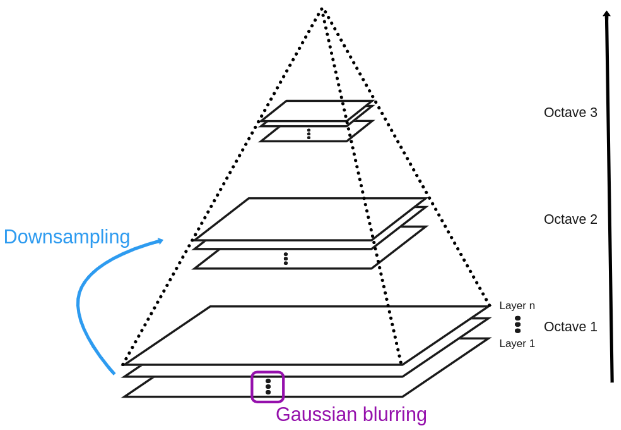

Figure 2: Gaussian pyramid where each octave represents a different scale and each layer - a different resolution

4 Method

The goal of the algorithm is to rescale the template image to the size of the reference image, to within 5 % margin of error in order to register images using Conover's algorithm [1].

Initial transform of template image

The template image is repeatedly downscaled by a factor of 2 until the size of the template is smaller than the size of the reference. The argumentation for this step is that cross-correlation that does not rely on padding requires the template signal to be smaller or equal to the reference signal. The choice of factor 2 does not affect the end result as the later stages of the algorithm correct any misalignment.

Core part of algorithm : greedy optimization

Given a downscale scale factor df_n and upsale factor uf_n where $1 \leq n \leq factors$:

- build a cuboid from the phase image of the reference image with the downscale factor ds_n : c_{ref}
- build a cuboid from the phase image of the unchanged template image with a downscale factor ds_n : c_{temp}
- downscale template image with downscale factor ds_n and calculate phase image, build cuboid from it with ds_n : c_{dtemp}
- If upscaling template by a factor of uf_n does not make the width and height bigger than the ones of the reference, upscale template by a scale factor of uf_n. Build

scale space from the phase image of the upscaled template with downscale factor of ds_n : c_{utemp}

After the 3D signals have been obtained, cross-correlation is performed between each of c_{temp}, c_{dtemp}, c_{utemp} and c_{ref}. The peaks of the cross correlations are computed and their respective scores are yielded - s_{temp}, s_{dtemp}, s_{utemp}. The one with the highest value corresponds to the best cross-correlation so far. The highest score is picked and the template is updated with the corresponding version of the template: s_{temp} corresponds to unchanged, s_{dtemp} corresponds to downscaled template, s_{utemp} corresponds to upscaled template. The process is repeated a predefined number of times - $steps$, until the scale factors are changed to the next from their respective list of scale factors - $factors$. As the scale factors are becoming closer and closer to one, the change is becoming increasingly smaller, eventually converging to the optimal scale.

Peak selection

The peak selection is based on the following formula:

$$peak = \frac{max(xs) - \mu(xs)}{\sigma(xs)}$$

where xs is the input array, μ is the mean of xs after the maximum has been removed and σ is the standard deviation of xs after the maximum has been removed.

Normalizing the value is necessary to ensure that we find a unique peak. For example, if cross-correlation is performed on two images, the smaller one being completely white, the peak would be the maximum possible peak achievable, but it does not necessarily mean that the two images are very similar. Subtracting the mean and dividing by the standard deviation addresses this problem. When computing the mean and standard deviation of the input array, the peak is removed as it is presumed to be an outlier.

Informal algorithm overview

In order to perform cross correlation over cuboids, downscaled images need to be consistently placed at the same place, in our case top-left. For example, the eye region from the template image is always compared to the same region in the reference image, regardless of the level of the cuboid. The different levels emphasize features at different frequencies - the lower levels of the cuboid, where the image has been less altered, contain more detail in contrast to the higher levels of the cuboid, where downsampling and blurring are applied repeatedly. Two distinct cuboids with the same number of levels are built for comparison, one for the template and one for the reference image. The two cuboids are then cross-correlated. Essentially, 2D cross-correlation is performed on all levels and then the results are summed per pixel. By taking into account all levels of the cuboid, the algorithm is more robust to small changes in scans and noise - apparent in the high frequency domain, while not being too coarse - comparing practically the same thing in the low frequency domain. Figure 4 depicts a cuboid with three levels.

Figure 3 depicts the first few iterations of the algorithm. The dashed lines show when the scale factors are changed and what values they are changed to. The levels of the tree represent each time the algorithm estimates the scale of the template image. On each level, excluding the start node, three nodes are present. The left one denotes the downscaled template image, the middle node represents the unchanged template image, the right node represents the upscaled template image. The numbers inside the node show the normalized peak obtained when cross-correlating the cuboids of the given template and reference images. The highlighted in green nodes are the highest values per level and they denote which direction the algorithm takes, hence the tree expands under the highlighted nodes. A little optimization was used for performance reasons, which is to immediately update the scale factors if the highest cross-correlation was achieved by the unchanged template image because the computations are deterministic and the results would repeat if the scale factors remain unchanged. Consequently, only one level corresponds to the scale factors of 2.

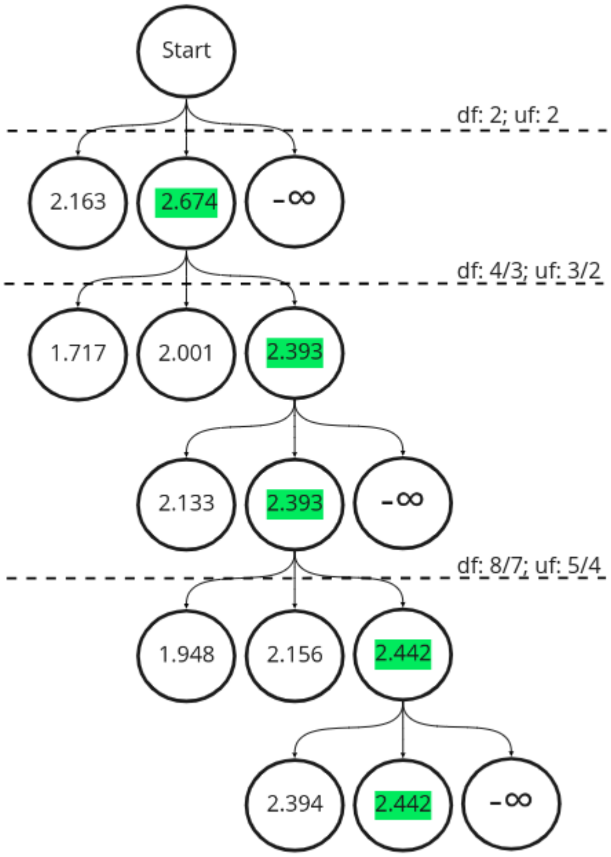

Figure 3: Tree-like structure of an instance of the algorithm

Comparison to algorithmic paradigms

The algorithm follows the Greedy algorithm paradigm. Fundamentally, at each step the algorithm makes the best local choice.

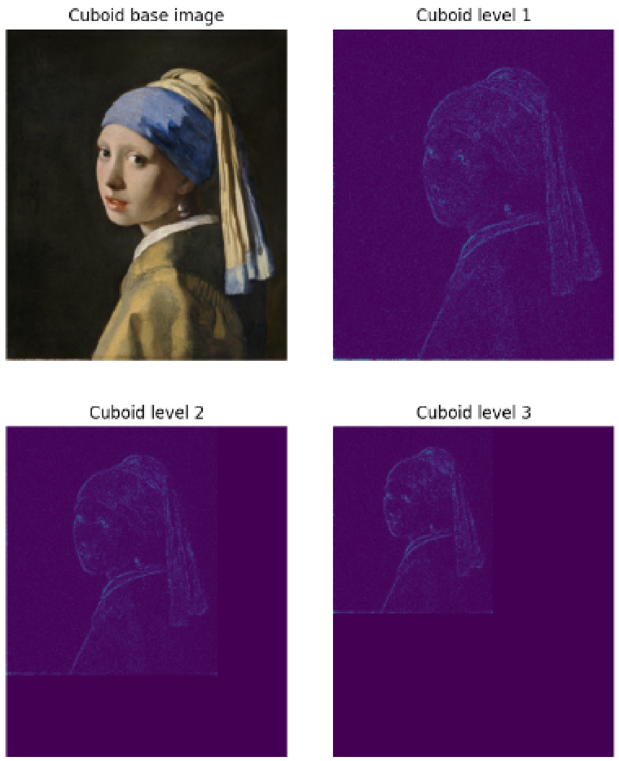

Figure 4: A three level cuboid with downscale factor of 1.33 of *Girl with a pearl earring* by Johannes Vermeer (c. 1665, Mauritshuis). Photography by René Gerritsen Technical Art & Research Photography.

5 Experimental Setup and Results

The goal of the experiments is to show that the method reliably finds a scale of the template image that could be used to register the rescaled template image against the reference through the use of Conover's algorithm [1]. Although, the desirable margin of error in that paper is described as *within 5-10% size difference*, we stick to a more strict bound of 5% in order to ensure that registering works flawlessly. If the error is within 5-10%, the output is labeled as 'maybe' as in these scenarios the algorithm works on a per case basis. If the size difference is more than 10% , the output is labeled as 'reject' and our algorithm has failed. The function that labels the result is as follows:

$$f(x, opt) = \begin{cases} Accept, & \text{if } 0.95 * opt \leq x \leq 1.05 * opt \\ Reject, & \text{if } x < 0.9 * opt \vee x > 1.1 * opt \\ Maybe, & \text{otherwise} \end{cases}$$

where x is the estimated scale relative to the original size of the template and *opt* is the optimal value.

We define the downscaling operation as dividing the template image by a given factor. In contrast, our definition of upscaling is multiplying the template image by a factor. As long as the use is consistent, the different types of operation do not yield a different result, i.e., we could have equivalently defined both operations with multiplication only.

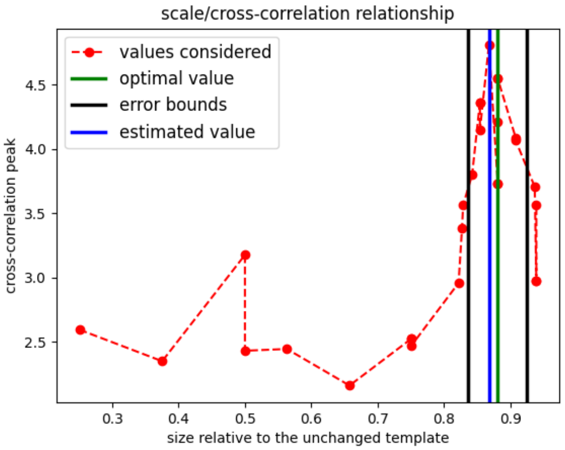

Figure 5: Instance of our method, margin of error is 1.44%

We use two sequences to define these factors, for downscaling the sequence of elements is defined by the following equation:

$$df_n = \frac{2^n}{2^n - 1}$$

and the upscaling factors are defined by:

$$uf_n = \frac{2^{n-1} + 1}{2^{n-1}}$$

Both sequences are monotonically decreasing and converging to one. This is desirable because we want to limit the search space of scales with with the advancement of the algorithm.

Figure 5 shows a single instance of our method in a good case where the result is accepted. On the x-axis the scale relative to the original size of the template image is shown. The y-axis is the corresponding cross-correlation peak of the two cuboids. It is worth noting that the start of the algorithm is at scale of 0.5 as the template image is larger than the reference. Our method explores values and makes a decision based on the peaks of the cross correlation. In an ideal case, the number of values considered will increase as they get closer to the optimal value because the scaling factor decreases, as is the depicted case. The bounds are set to 5% deviation from the optimal value and the estimated value is well within the bounds and is considered as success.

Independent parameters

A small empirical study is done to see how the choice of independent variables - $steps, dims, factors$, affected the end result. The study is conducted based on the same template and reference image. The two images remain the same in order to focus only on how the different choice of parameters affects the end result. By rescaling the template image from the ground truth size, the best result is known beforehand. The algorithm estimates the size of the template and outputs a number. From this number two metrics are computed - the margin of error in percentages, and the status of the result,

Exp	Est	MoE	status	dims	steps	factors
1.901	0.437	77%	R	3	2	6
1.901	1.216	36%	R	3	3	6
1.901	1.216	36%	R	3	3	8
1.901	1.874	1.42%	A	4	2	6
1.901	1.899	0.11%	A	4	2	8
1.901	1.874	1.42%	A	5	2	6
1.416	1.406	0.71%	A	3	2	6
1.416	1.406	0.71%	A	4	2	6
1.416	1.406	0.71%	A	5	2	6
0.556	0.545	1.98%	A	3	2	6
0.556	0.545	1.98%	A	4	2	6
0.556	0.486	12.6%	R	5	2	6
0.556	0.503	9.53%	M	5	2	8
0.556	0.507	8.81%	M	5	3	8

Table 1: Independent variable analysis

which can either be accept, reject or maybe. Table 1 shows how the different choices of independent variables affect the end result. Three different scales are approximated - 1.901, 1.416 and 0.556. The choice for these three cases is random and aims to be representative.

The $dims$ parameter corresponds to the number of levels the cuboids have. The higher the number, the more the algorithm values well matching features in the low frequency domain. It is a thin balance between coarseness and strictness. In the first case the 3 levels of the cuboid are not enough to calculate a good estimate of the scale. Contrary, in the third case the 5 levels are too many and the algorithm is emphasizing the low-frequency features too much. The value of 4 empirically shows the best results of balancing out the low and high frequency feature matching. This intricate balance point is not uniform across all pairs of images. Consequently, individual cases might require adjusting the $dims$ parameter.

The $factors$ variable expresses how many elements we take from the aforementioned infinite sequences ds_n, us_n. The changes that might occur from the given scale factors are increasingly smaller. If our bounds were stricter, we would increase the value to a larger one, but in the experiments the value of 6 sufficed.

The $steps$ parameter defines how many times the algorithm is going to estimate the size based on a given downscale and upscale factor. This variable is affected by the choice of scaling factors, which motivates us to use the value of 2. The experiment does not show a significant difference between the value of 2 and 3.

6 Discussion

The algorithm is designed for estimating the size of the template image based on a reference image in the multimodal context. It works best if the template and reference image depict similar sections of the painting.

Naturally, as a greedy algorithm our approach has certain limitations. It heavily relies on the premise that the best local choice would lead to the global optimal. We have not proven that this strong assumption always holds. Our assumption could be expressed by the following sentence: *When cross-correlating two cuboids of differently sized template image against the cuboid of the reference image, the better peak of the two would be achieved through the cuboid of the template image that has a size closer to the optimal value.* This assumes that for a given scale factor the function describing the template size/peak of cross-correlation relationship is first monotonically increasing until it reaches the absolute peak, afterwards it is monotonically decreasing. This is quite similar to the behaviour of a quadratic function, but no analysis is done to show if it is a good approximation of the real function. We note that our algorithm does not analyze the whole spectrum of the function, but does two estimations per scale factor, which is at most 4 distinct points of that function because some will get repeated.

From the tested scenarios, the assumption does not hold when trying to match a subsection of a painting to the entire painting. The differences in modalities also play a significant role in how well the scale is going to be approximated. Our approach relies on cross-correlation of phase images at different scales, but these phase images might vary significantly across certain modalities and there our approach comes short.

7 Responsible Research

Ethical implications of this research area are estimated to be relatively low. All computations are run against images of paintings, targeting the art conservators community. The proposed algorithm does not make any decisions that impose the need of morality.

Reproducibility of the results is high given the right access of resources, namely code and images. Although the code is not publicly available, the steps we took are documented in this paper and replication of the results is possible. Small deviations in performance are expected because of implementation details and/or hardware differences.

The major problem for accurate reproducibility is the access to all imagery. The access to various scans of paintings is limited for many reasons. Usually, the scans are conducted by researchers in museums and their active work involves these images. Moreover, these images are property of the museum, which makes distribution an intricate task.

8 Conclusions and Future Work

We have proposed an alternative algorithm that could be used to approximate the size of an image relative to a reference image through the use of scale space theory and cross-correlation. The algorithm has been tested in multiple modality pairs and has proven its effectiveness.

A limitation of the current approach is that it is highly dependent on the assumption that the input images have the same orientation. This might or might not be desirable and future work could be done to remove this assumption, possibly through the use of histogram of orientated gradients as a feature descriptor.

Further work could be done to analyze the behaviour of the relationship between template size and peak of cross-correlation of the cuboids. More specifically, this paper assumed a quadratic behaviour, but this is not always a good approximation of the real function.

References

[1] Damon M. Conover, John K. Delaney, and Murray H. Loew. Automatic registration and mosaicking of technical images of old master paintings. *Applied Physics A*, 119(4):1567–1575, Jun 2015.

[2] Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.

[3] Changying Liu, Hongliang Liu, Yang Liu, Tongtong Li, and Tianhao Wang. Normalized cross correlation image stitching algorithm based on minimum spanning tree. *Optik*, 179:610–616, 2019.

[4] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004.

[5] Vladan Rankov, Rosalind J. Locke, Richard J. Edens, Paul R. Barber, and Borivoj Vojnovic. An Algorithm for image stitching and blending. In Jose-Angel Conchello, Carol J. Cogswell, and Tony Wilson, editors, *Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XII*, volume 5701, pages 190 – 199. International Society for Optics and Photonics, SPIE, 2005.

[6] B.S. Reddy and B.N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996.

[7] Jignesh N Sarvaiya, Suprava Patnaik, and Salman Bombaywala. Image registration using log-polar transform and phase correlation. In *TENCON 2009 - 2009 IEEE Region 10 Conference*, pages 1–5, 2009.