EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2022)
R. Pintus and F. Ponchio (Editors)

# A New Baseline for Feature Description on Multimodal Imaging of Paintings

Jules van der Toorn[1], Ruben Wiersma[1], Abbie Vandivere[2], Ricardo Marroquim[1], and Elmar Eisemann[1]

[1]Delft University of Technology
[2] Mauritshuis

**Abstract**

*Multimodal imaging is used by conservators and scientists to study the composition of paintings. To aid the combined analysis of these digitisations, such images must first be aligned. Rather than proposing a new domain-specific descriptor, we explore and evaluate how existing feature descriptors from related fields can improve the performance of feature-based painting digitisation registration. We benchmark these descriptors on pixel-precise, manually aligned digitisations of "Girl with a Pearl Earring" by Johannes Vermeer (c. 1665, Mauritshuis) and of "18th-Century Portrait of a Woman". As a baseline we compare against the well-established classical SIFT descriptor. We consider two recent descriptors: the handcrafted multimodal MFD descriptor, and the learned unimodal SuperPoint descriptor. Experiments show that SuperPoint starkly increases description matching accuracy by 40% for modalities with little modality-specific artefacts. Further, performing craquelure segmentation and using the MFD descriptor results in significant description matching accuracy improvements for modalities with many modality-specific artefacts.*

**CCS Concepts**

• ***Computing methodologies*** → ***Image processing;*** • ***Applied computing*** → ***Fine arts;***

## 1. Introduction

Painting conservators and scientists make extensive use of non-invasive imaging technologies to analyse and visualise the composition of historic paintings. Typical digitisations include visual light photography, infrared reflectography, ultraviolet fluorescence photography, and X-radiography. Detailed comparisons of painting regions within- and across modalities can reveal information that would otherwise remain hidden, such as the composition of pigments, the presence of underdrawings and evidence of changes to the painting over time [VWvdBvL19, APvEH*13, GDE*21, vLNdM*20].

To facilitate the direct comparison of specific painting regions across modalities, the digitisations need to be aligned as close as possible. While a simple alignment could be done with manual tools, often hundreds of high-detail patch digitisations need to be mosaicked together, which is inhibitively time-consuming for conservators. Hence, an automatic image registration algorithm is desired with high accuracy.

Current approaches for automatic image registration can be broadly classified into two classes. First of all, in *area-based* image registration, a sliding window is used whose alignment is optimised using pixel similarity metrics such as cross-correlation. While this enables sub-pixel accurate alignment, the low degree of freedom of the sliding window fundamentally limits registration flexibility. Secondly, in *feature-based* image registration, multiple keypoints are detected in both images, which are then matched based on characteristic details around the keypoints. While this approach requires no assumptions on the transformation type, the resulting registrations are often less precise than area-based methods [Bro92, ZF03].

Given the different strengths of area- and feature-based image registration, a popular approach is to combine the two techniques. In such a setup, feature-based image registration is used for obtaining the flexible transformation for a rough alignment, after which area-based optimization is used for further aligning the images with sub-pixel accuracy. This is also done in the state-of-the-art painting digitisation registration algorithm developed by Conover et al. [CDL15]. Here, shearlet wavelets are used for feature selection, which are subsequently matched using a cross-correlation sliding window.

While such an approach works well in theory, the classical feature-based SIFT image registration algorithm [Low04] has been used previously for multimodal painting digitisation registration with varying success. Zacharopoulos et al. [ZHK*17] found that they could only get sufficient image registration performance when aligning digitisations from adjacent spectral bands. Further, Mirhashemi [Mir19] had to include a manual pre-crop stage and added a custom iterative feature match filtering algorithm.

Previous work in the realm of image registration for paintings has often ignored recent advancements in multi- and unimodal image registration, often comparing to SIFT as the baseline approach. In this work, we explore how more recent, existing feature descriptors can be used to improve the performance of feature-based painting digitisation registration. We selected two recent descriptors based on their applicability to the painting domain and overall image registration performance: the handcrafted multimodal MFD descriptor [NP17], and the learned unimodal SuperPoint descriptor [DMR18]. We evaluate the descriptors on two paintings, across four modalities: *Girl with a Pearl Earring* by Johannes Vermeer (c. 1665, Mauritshuis) and an anonymous *18th-Century Portrait of a Woman*, using visual light (VIS), X-radiography (XR), ultra violet (UV) and infrared reflectography (IRR). Furthermore, we describe a preprocessing step using craquelure segmentation to improve performance and provide insights that can be built on by practitioners in the field as well as future research improving image registration for multimodal digitisations of paintings.

Summarising, our main contributions are:

- A survey of multi- and unimodal feature descriptors and their applicability to registration of multimodal digitisations of paintings.
- A thorough and objective evaluation of the most suitable and recent feature description algorithm performances.
- The proposal of a novel craquelure segmentation preprocessing step for increasing description matching accuracy for painting digitisation modalities with many modality-specific artefacts.

## 2. Related Work

**Multimodal Registration for Historic Painting Digitisations** Zacharopoulos et al. [ZHK*17] made use of classical SIFT matching for the registration of an unaligned spectral cube. They modified the descriptor to make use of all 16-bit colour information and consecutively matched images from adjacent spectral bands.

SIFT was also used for registration of unrelated image modalities in the work of Mirhashemi [Mir19]. Here, regular SIFT detection and description was used, but a custom iterative filtering and matching algorithm was proposed for better matching performance.

Conover et al. [CDRL11,CDL15] proposed a custom registration technique, which is a hybrid approach between feature-based and area-based image matching. They assume an initial rough alignment of the reference and template images and use phase correlation to optimise its translative component. Feature patches in the template image are selected using the magnitudes from a wavelet transform. Subsequently, an area-based matching using normalised cross-correlation is done. Lastly, feature match candidates are filtered by iteratively refitting a bilinear function.

Finally, Sindel et al. [SMC21] developed a machine-learned image registration pipeline. They propose CraquelureNet, a fully-convolutional neural network that jointly learns keypoint detection and description. They focus on detecting and describing craquelure, the fine pattern of dense cracking which can form on the surface of ageing paintings. Features are matched using the brute forced mutual nearest-neighbour algorithm.

**Feature Description for Multimodal Registration** Hasan et al. [HPJ12] looked into optimising SIFT for multimodal feature matching. Among other things, they preserve keypoints with low contrast, change the criterion for calculating the principal keypoint orientation, use a larger descriptor window with more subwindows, and propose a three layer matching method as opposed to the original nearest-neighbour algorithm.

A descriptor for matching images with nonlinear intensity variations based on Log-Gabor (LG) filters is proposed by Aguilera et al. [ACST15]. Features are selected using the FAST detector [RPD10]. Subsequently, pixels in subwindows are binned based on the magnitude response of LG filters at different orientations and different scales. Lastly, features are matched using the nearest-neighbour algorithm.

Li et al. [LHA20] propose a descriptor based on phase congruency (PC) and maximum index maps (MIM). Corner and edge feature points are detected using a generated PC map. Subsequently, pixels in subwindows of the feature patches are binned using the MIM response. Features are matched on Euclidean distance, where outliers are removed using a normalised barycentric coordinate system.

Finally, Xie et al. [XJC21] propose a descriptor that uses shearlet-based orientation maps (SOM). Features are selected using a PC-based detector. For each feature patch, a shearlet decomposition at different scales is calculated. The resulting SOM is then flattened and used as feature description. The nearest neighbour algorithm with ratio check is used for feature matching.

While various custom multimodal descriptors have been proposed, they all originate from the remote-sensing or medical domain, and no earlier work has been done in applying them to the domain of painting digitisation registration. Next to this, no off-the-shelf learned feature descriptors were previously used for painting digitisation registration in the literature. To this end, we will investigate if multimodal or learned feature descriptors can improve classical painting digitisation registration performance.

## 3. Background

For our experiments, we consider three different feature description algorithms: SIFT, MFD, and SuperPoint. This section will give a brief overview of their implementation details.

**Scale Invariant Feature Transform (SIFT)** The SIFT descriptor was developed by Lowe in 2004 [Low04]. It uses a handcrafted feature description algorithm, and is intended for unimodal image registration. First, the algorithm subdivides a feature patch in 16 subwindows. Then, for each subwindow, the 8 principal pixel gradient orientations are binned into a histogram. Finally, the 16 histograms are concatenated and normalized to form a 128-dimensional feature description vector.

The SIFT description algorithm is well known in the literature, and has previously been used for registering painting digitisations. To that end, we consider this descriptor as our baseline for painting digitisation registration performance.

**Multispectral Feature Descriptor (MFD)** Developed by Nunes et al. in 2017, MFD is a multimodal handcrafted feature description algorithm [NP17]. Similar to SIFT, the algorithm subdivides a feature patch in 16 subwindows. Then, for each subwindow, the 5 principal edge orientations at two different scales are calculated using various Log-Gabor filters. The highest response orientations are put in a maximum index map (MIM), which is then normalized to form a 160-dimensional feature description vector.

**Self-Supervised Interest Point Description (SuperPoint)** SuperPoint is a learned feature descriptor proposed by DeTone et al. [DMR18] in 2018, and uses a convolutional neural network to infer feature description vectors. It is based on the VGG network architecture [SZ15], and applies self-supervision by transforming training data using randomly sampled homographies. Different to the handcrafted description algorithms, this architecture outputs a dense grid of feature descriptions for each pixel in the input image. Specifically, each pixel is described using a 256-dimensional feature description vector.

## 4. Dataset

To evaluate the performance of feature descriptors across different modalities, a ground truth registered multimodal dataset is required. Specifically, for two images from different modalities, the ground truth transformation matrix from the *template* image to the *reference* image has to be known in advance. Given two matched features, it can then be verified if the match is in line with the ground truth transformation.

### 4.1. Digitisation Collection and Registration

As our ground truth dataset source, we assembled various high resolution images (digitisations) from the famous historic painting *Girl with a Pearl Earring* by Johannes Vermeer (c. 1665, Mauritshuis, Figure 3, top) and from *18th-Century Portrait of a Woman* (Figure 3, bottom). *Girl with a Pearl Earring* was examined systematically in various modalities by conservators and scientists at the Mauritshuis, as part of the 'Girl in the Spotlight' project in 2018 [VWvdBvL19]. The modalities include ultraviolet-induced fluorescence, infrared reflectography, X-radiographs, and hyperspectral image cubes. As some digitisations were only made for certain regions in the paintings, and others were visually very similar to the visual spectrum, we chose 3 distinct full-painting digitisations: X-radiography (XR), ultra violet (UV), and infrared reflectography (IRR). More details on the acquisition of each of the digitisations and what can be learned from them is given by Vandivere et al. [VWvdBvL19, VvLD*19]. For a second set of digitisations, we selected the same modalities from *18th-Century Portrait of a Woman*.

Given the three unaligned high resolution digitisations, the exact transform to the high resolution visual image had to be determined. It was assumed that the transformation was projective, having 8 degrees of freedom. To solve the transformation matrix equations, the exact offset of 4 keypoints in the visual image had to be matched in the multimodal digitisations. To that end, 4 distinct craquelure patterns were manually selected in the visual image,

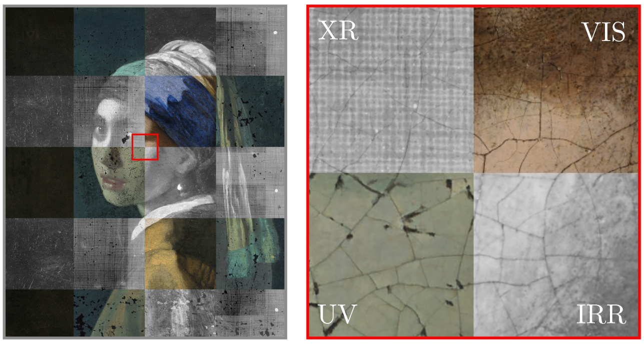



**Figure 1:** *Mosaic image of the ground truth aligned digitisations. On the right a close-up is displayed, which shows the pixel-precise continuity of the craquelure in the painting over the four different digitisation modalities. XR and IRR by René Gerritsen Art & Research Photography, VIS by Hirox Europe/Jyfel.*

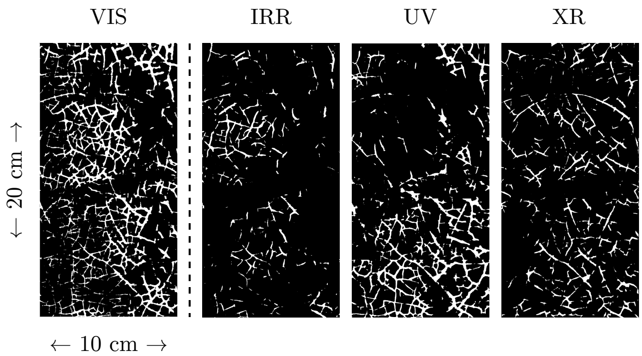



**Figure 2:** *Close-up of the craquelure segmented painting digitisations of Girl with a Pearl Earring for the four different digitisation modalities.*

striving to select regions with enough distance in between for registration robustness. Within the craquelure regions, all four digitisations were laid next to each other, and an intersection in the craquelure that was clearly visible in all digitisations was chosen as the keypoint. Subsequently, using the 4 keypoints, a pixel-precise homography was calculated for each multimodal digitisation.

Given the ground truth homographies, each multimodal digitisation was registered onto the visual digitisation. The resulting dataset for each painting is a collection of 4 high resolution digitisations in different modalities where any 2 digitisations physically align at each pixel coordinate. In turn, the correctness of a feature match can be verified by simply checking if their keypoint coordinates are equal. A mosaic image of these digitisations is shown in Figure 1.

### 4.2. Digitisation Craquelure Segmentation

When we created the ground truth dataset, we found that the most reliable way to select corresponding keypoints was by looking at the craquelure patterns in the painting digitisations. Because craquelure impacts paintings on a structural level, it is clearly visible in all digitisation modalities. This insight was also mentioned in earlier literature. Conover et al. state that "In the case of a painting, the regions likely to match are the painting's texture features, such as cracks, brushstrokes, bubbles, and blisters" [CDL15, p. 3]. Furthermore, Sindel et al. [SMC21] exploit this idea with the CraquelureNet descriptor they developed, which extracts feature descriptions from craquelure patterns in historic paintings.

To investigate the possible benefit of exploiting craquelure, we also experiment with feature description on craquelure segmented masks of the ground truth digitisations. For this, we made use of the VGG16 segmentation network [SZ15] trained on a crack segmentation dataset [ZYZZ16, YZY*20, ESS*17, SCQ*16, ACIB16, ZCL*12], of which an implementation was made by Github user Khanhha[†]. A preview of the generated masks is shown in Figure 2.

## 5. Method

We created a controllable image registration pipeline to evaluate feature description performance in isolation. This section will go over the decisions that were made for this, and some of its important implementation details.

### 5.1. Libraries and Tools

The image registration pipeline was implemented in a Python 3 Jupyter notebook to allow for quick iteration and experimenting. The main dependencies of the pipeline are OpenCV for image processing, and NumPy for matrix operations. For the feature description algorithms, open-source implementations were used[‡].

### 5.2. Feature Detection

In a regular image registration pipeline, distinctive regions in the input image, referred to as features, are selected by a dedicated feature detector. Examples of distinctive features are edges and corners, present at different sizes and orientations. For our experiments we require the ability to precisely specify the desired number of selected features, their orientations and their sizes. To that end, we implemented a simple disjunct random feature detector. Given an input image, random locations are sampled in the painting, referred to as keypoints. Each descriptor is evaluated on a patch around a keypoint. Keypoints that cause patches to overlap are filtered out, until a fixed number of keypoints are obtained.

A downside of this approach is that the selected features are not guaranteed to represent distinct regions, which makes the features more difficult to describe uniquely. However, by running experiments between descriptors on the same set of random features, and repeating these with multiple new sets of random features, a fair comparison between descriptors can still be made.

[†] Available at github.com/khanhha/crack_segmentation

[‡] Used implementations: github.com/ducha-aiki/numpy-sift, github.com/cfgnunes/mfd, and github.com/rpautrat/superpoint

It should be noted, however, that the presented overall accuracy scores could be further improved by using a dedicated feature detector algorithm, but this is not the focus of the current study.

### 5.3. Feature Description

The descriptors are then evaluated on patches around each keypoint for each modality, yielding a feature vector for each keypoint and for each modality. To get a fair comparison between the three feature description algorithms considered, some modifications to the original implementations had to be made.

SIFT was developed as a rotation invariant descriptor. This is realized by calculating a global orientation for a given feature patch, which the feature description vector then is normalized to. However, in our research we purely want to focus on feature description performance, and not take rotation into account. To that end, we modified the SIFT description algorithm to always assume a global orientation of zero degrees.

For MFD, no alterations were necessary. This descriptor does not claim rotation invariance, and can generate a description vector for any input feature patch size.

Lastly, the SuperPoint descriptor had to be adapted to support describing feature patches of any size. This learned descriptor originally generates a dense grid of descriptors for each pixel in the input image. As we want to experiment with different feature patch sizes, we added an additional rescaling stage to the descriptor. This stage downscales the input image such that each pixel has the same physical size as the evaluated feature patch size.

### 5.4. Matching and Evaluation

Finally, we match keypoints across modalities by comparing their feature vectors. For feature matching, we implemented a simple nearest neighbor matching strategy in the feature space. Each keypoint in the reference image is linked to the keypoint in the template image whose feature description vector has the shortest Euclidean distance to the feature description vector of the reference keypoint. We can then compute the *accuracy* of a descriptor as the ratio of keypoints that is matched correctly.

Classical feature matching often applies an additional match filtering stage. A common strategy is to discard a feature match if the ratio of distances of the nearest and second-nearest feature descriptions is above a certain threshold. We decided not to apply any match filtering for the following reasons: First and foremost, initial experiments showed that match filtering gave a similar trade-off between precision and recall regardless of the used descriptor. Second, adding match filtering complicates objective evaluation, as there is no unique optimal trade-off between precision and recall. The trade-off depends on the choice of algorithm for homography estimation. Given a feature matching in this setup, each keypoint in the reference image is matched, either correctly or incorrectly.

## 6. Experiments

In our experiments, we investigate the descriptor performance of different registration pipelines. For this, two overarching


**Figure 3:** *Examples of feature patches in the four different modalities at three different physical sizes. The red bounding boxes represent physical patch sizes of 1 mm, 20 mm, and 40 mm.*

experiments were conducted. Our initial experiment investigates the descriptor performance on original painting digitisation images, and a follow-up experiment investigates the descriptor performance on painting digitisation images after going through craquelure segmentation.

To give an overview of the performance profile of a feature descriptor, we look at its description matching accuracy as function of feature patch size. This gives two main insights. First of all, the optimal matching accuracy shows how descriptive the descriptor can be at ideal conditions, and is used to compare its overall performance against other descriptors. Secondly, the specific performance curve motivates how the descriptor can best be applied in an image registration pipeline. If performance plateaus, and is consistent over a large range of patch sizes, the descriptor is stable and could be applied for both scale and transform registration in a single pass. However, if performance has a clear peak, it makes sense to decouple an overall registration into separate scale and transform stages. After a rough scale has been determined, the descriptor can then be run on its optimal feature size.

Additionally, the robustness of a feature descriptor to keypoint translational and rotational noise could have been investigated. If a descriptor stays descriptive under such noise, keypoint selection can be more lenient, and rough registrations can still be found. However, the aim of registering high resolution painting digitisations is for conservators to analyze details at sub-millimeter precision. Because of this, accurate registration is desired over transform flexibility, which motivated us to solely focus on description matching performance on precisely selected keypoints.

In each experiment, different permutations of feature patch sizes, descriptor algorithms, and digitisation modalities are considered. For feature patch sizes we consider a range of 0.5 to 40 millimeters, with a step size of 0.5 millimeters. Subsequently, the three descriptor algorithms that were introduced earlier are evaluated for each feature patch size on the three non-visual digitisation modalities. Figure 3 illustrates the level of detail in feature patches at different sizes in the different modalities.

The number of selected keypoints was held constant at 100 keypoints. This quantity has the same order of magnitude as the 536 keypoints selected in the literature example of SIFT [Low04], while still allowing experimentation with non-overlapping feature patches of large sizes.

Because a random keypoint selector is used in the simulated registration pipeline, descriptor performance is correlated to the distinctiveness of the randomly drawn keypoints. To that end, each experiment is repeated 10 times, where each run uses a new seed for random keypoint selection. The presented performance measures are an average of all repeated runs.

### 6.1. Descriptor Performance on Original Digitisations

In this experiment, the descriptor performance on the original digitisation images is investigated. The resulting description matching accuracy for different feature sizes and digitisation modalities are shown in Figure 4.

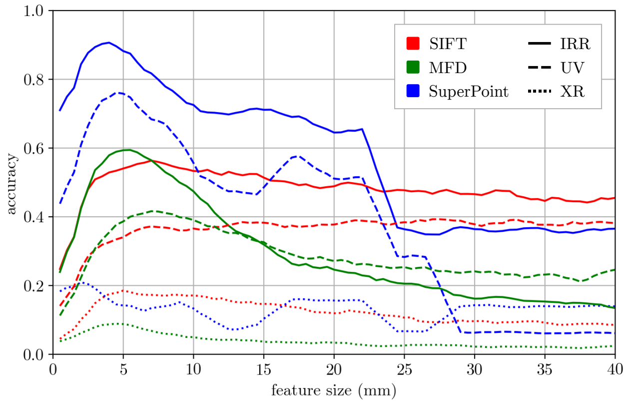
An overall insight is that matching accuracy for the IRR and UV modalities reach between 40% and 90%, but that XR significantly lacks behind with an optimal matching accuracy of 20%. Because the performance of XR registration is so low, its matching accuracy as a function of feature size does not show significant patterns, and seems to be mainly based on contextual noise. To that end, only the results of the IRR and UV registrations are taken into account for conclusions on optimal feature sizes in this experiment.

The matching accuracy of the SIFT descriptor quickly grows as the feature patch size increases to 5 millimeters, but then plateaus. Given that larger feature patches are relatively easier to describe distinctively, and because smaller feature patches are favorable for accurate registration, we conclude that a feature patch size of 5 millimeters is optimal for this descriptor.

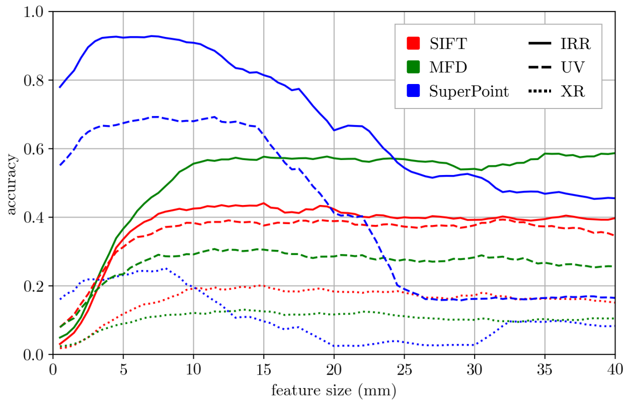
For the MFD descriptor, we see significantly distinct performance behavior for the two paintings. While description matching accuracy peaks at around 6 millimeters in Figure 4a, it stays stable after patch sizes larger than 10 millimeters in Figure 4b. The performance peak of the former could be explained by the small-scale craquelure that is more clearly visible in this painting. Visually inspecting the maximum index maps generated by this descriptor also show that the patches of different modalities look similar at different sizes, while still having high detail.

Lastly, the SuperPoint descriptor has optimal performance at a feature patch size of 4 millimeters. This could be explained by the fact that this scales the image down to a resolution of around 100x100 pixels, which is close to the resolution that the descriptor network was trained on (280x320 pixels) [DMR18].

To compare the relative performance of the descriptors, we look at the registration accuracy at corresponding optimal feature




**(a)** *Description matching accuracies on Girl with a Pearl Earring.*




**(b)** *Description matching accuracies on 18th-Century Portrait of a Woman.*

**Figure 4:** *Feature description matching accuracy (y-axis) as function of feature patch size (x-axis, in millimeters) on the original painting digitisations.*

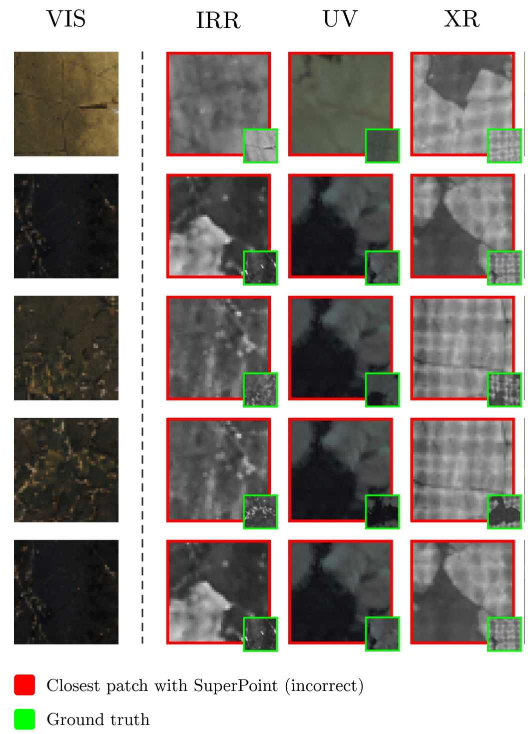



**Figure 5:** *Feature patches that were incorrectly matched by SuperPoint in all modalities from Girl with a Pearl Earring. Rows represent different keypoint samples, columns represent different digitisation modalities.*

sizes. In this setting, SIFT obtains an accuracy between 30% and 50%, MFD an accuracy between 30% and 60%, and SuperPoint an accuracy between 70% and 90%. From this, we can conclude that each descriptor has a similar performance variance and is not severely sensitive to the specific structure of a different modality. Registration for IRR consistently performs around 20% better than for UV, which can be explained by the dark blobs that are visible in UV, but not in IRR and VIS.

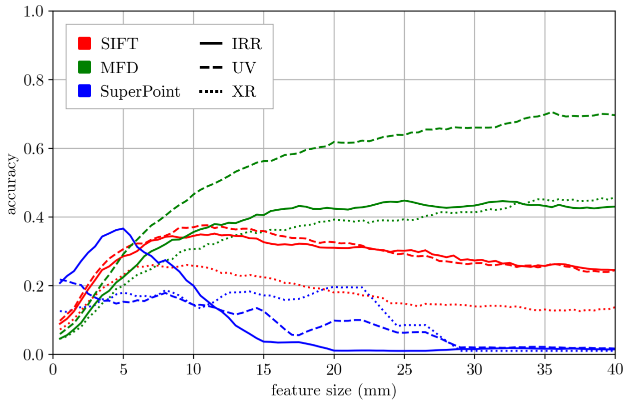
The overall conclusion from this experiment is that the multimodal MFD descriptor performs only slightly better than the classical SIFT descriptor, while the learned unimodal SuperPoint descriptor achieves almost double the performance of SIFT. A reason for the limited performance of MFD could be explained by the fact that the descriptor was mainly developed and evaluated for the domain of aerial images, which might accentuate different characteristics for different modalities. On the other hand, the SuperPoint descriptor performs very well, even though it was originally developed as a unimodal descriptor. Its high overall performance can be attributed to the fact that it uses a convolutional neural network, which was trained and automatically optimized for thousands of images. During training, synthetic data augmentation techniques such as Gaussian noise and motion blur were used as well, which can explain the retention of high performance under multimodal registration.

While SuperPoint achieves impressive performance, it still incorrectly matches a fraction of the keypoints. Given a random sample of 100 keypoints at a patch size of 4 millimeters, we found that 10 keypoints were incorrectly matched in *all* modalities. Upon closer inspection, those all originate from the homogeneous background of the painting. The feature patches of 5 of the incorrectly matched keypoints are shown in Figure 5.

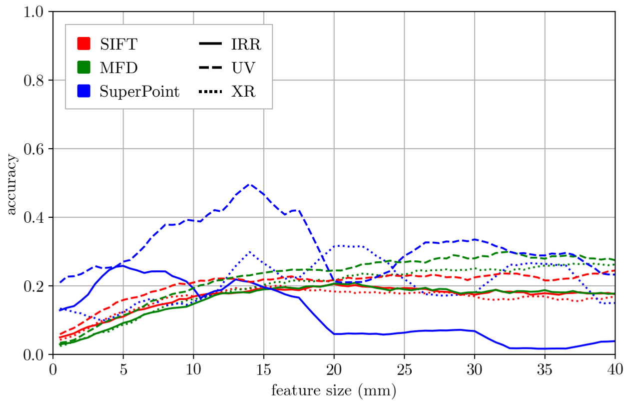
### 6.2. Descriptor Performance on Craquelure Segmented Digitisations

As became clear in the previous experiment, no descriptor was able to achieve sufficient performance at registering the XR digitisations. This is not unexpected, as the XR digitisation has significantly more modality-specific artifacts than IRR and UV. In the XR digitisation, the canvas structure behind the painting is visible, which makes it difficult to manually recognize higher level features when inspecting patches at centimeter granularity.

Aiming to improve modality invariance, we investigate the description performance of registering craquelure segmented masks of the multimodal painting digitisations. Besides this




**(a)** *Description matching accuracies on Girl with a Pearl Earring.*




**(b)** *Description matching accuracies on 18th-Century Portrait of a Woman.*

**Figure 6:** *Feature description matching accuracy (y-axis) as function of feature patch size (x-axis, in millimeters) on the craquelure segmented painting digitisations.*

additional preprocessing step, all other variables were held constant with regards to the previously conducted experiment. The resulting description matching accuracy for different feature sizes and digitisation modalities are shown in Figure 6.

Comparing the overall registration accuracy of the segmented images with respect to the original digitisations, both SIFT and SuperPoint perform worse under all modalities, however, MFD shows improved performance under certain conditions. Because of this, only the performance results of MFD are discussed in this experiment.

While the matching accuracy of MFD started plateauing after 10 millimeters in the previous experiment, its matching accuracy of craquelure masks strictly increases with patch size, and only starts to stabilize at around 25 millimeters. This can be explained by the lower spatial resolution of the segmented masks, which still show high detail at higher patch sizes.

The highest description matching accuracy obtained by MFD is 70% for the UV digitisation of *Girl with a Pearl Earring*, a stark increase from the 40% obtained on the non-segmented digitisation. While this is an impressive result, it is still lower than the UV matching accuracy of 75% obtained by SuperPoint on the original digitisation image. Another promising result from this experiment is the performance of MFD on the XR digitisation of *Girl with a Pearl Earring*. It manages to achieve a matching accuracy of 45%, which is more than two times as high as the highest accuracy achieved on registering the original digitisation image of this painting. This shows that it could be beneficial to perform craquelure segmentation on painting digitisations with a lot of modality-specific artifacts before running image registration.

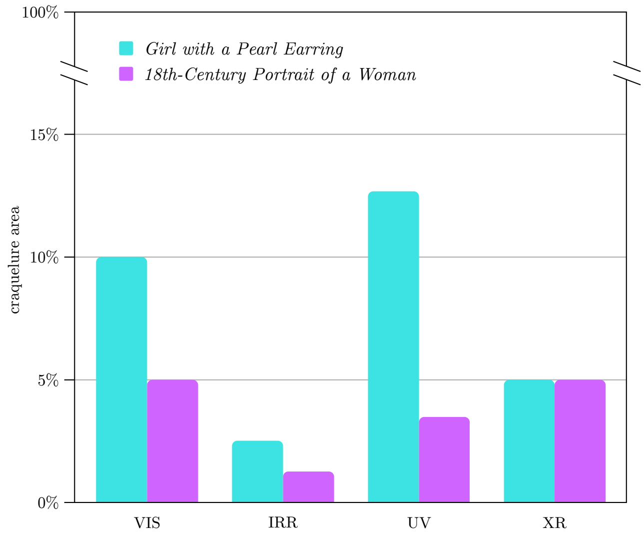



**Figure 7:** *The relative area of craquelure (y-axis, percentage) that was segmented in painting digitisations of the four modalities (x-axis).*

The performance of MFD on the craquelure segmented digitisations varies quite severely across modalities and the different paintings. The reason for this difference becomes apparent when it is compared to the relative area of craquelure in each segmented digitisation, which is shown in Figure 7. First of all, twice as much craquelure could be detected in the visual digitisation of *Girl with a Pearl Earring* compared to *18th-Century Portrait of a Woman*, which explains the overall better performance of MFD on the former painting. Secondly, the big relative area of craquelure in the UV digitisation of *Girl with a Pearl Earring* explains why MFD has the best overall performance on this modality.

In general, it seems that the description matching accuracy of MFD scales linearly with the relative area of detected craquelure. Visually inspecting the different digitisation images, much of the craquelure that is visible was not properly segmented by the segmentation network. To that end, it would be valuable to develop a more robust crack segmentation algorithm, which could result in significantly higher registration accuracy for all modalities.

## 7. Conclusion

We present a thorough evaluation of different feature descriptors for multimodal historic painting digitisations, striving to improve the robustness of image registration algorithms used by art conservators. The classical SIFT feature descriptor, which is used in most literature on feature-based painting digitisation registration, is compared to more recent feature description algorithms. We consider MFD, a handcrafted descriptor developed for multimodal aerial image registration, and SuperPoint, a popular deep-learned descriptor for unimodal image registration.

From our experiments we conclude the following points. First of all, SuperPoint achieves an impressive performance improvement over SIFT for registering multimodal digitisations with little modality-specific artifacts, increasing description matching accuracy by more than 40% for the IRR and UV modalities. Second, when many modality-specific artifacts are present in digitisations, description matching performance can be improved by preprocessing digitisations with craquelure segmentation. Description matching accuracy of MFD for the XR digitisation of *Girl with a Pearl Earring* increased by 20% after craquelure segmentation, doubling the accuracy obtained by SIFT.

Given these insights, it is proposed to combine both descriptors for a robust image registration pipeline. In an initial iteration, running the SuperPoint descriptor on original painting digitisations provides high description-matching accuracy for most modalities. However, when it is detected that features are matched with low certainty, a second iteration could perform craquelure segmentation and fall back on registering the painting digitisations with MFD, which often increases matching accuracy for noisy modalities.

## 8. Acknowledgements

The authors would like to acknowledge the conservators and scientists at the Mauritshuis and Rijksmuseum for sharing their insights and knowledge, among them Francesca Gabrieli. We thank John Delaney and Damon Conover for generously sharing the implementation of their image registration pipeline. Thanks to Matthias Alfeld and Joris Dik for insightful discussions and sharing data.

The IRR and X-radiography of *Girl with a Pearl Earring* by Johannes Vermeer (c. 1665, Mauritshuis) were captured by René Gerritsen Art & Research Photography; the visible light photograph was created from 3D digital microphotographs taken by Emilien Leonhardt, HIROX/JYFEL.

The MA-XRF and RIS data of *18th-Century Portrait of a Woman* were acquired at the Rijksmuseum Amsterdam by A. van Loon and F. Gabrieli. X-ray radiography and technical photography was done by René Gerritsen Art & Research Photography. The dataset of *18th-Century Portrait of a Woman* was provided by J. Dik and M. Alfeld of the Department of Materials Science and Engineering of the TU Delft.

## References

[ACIB16] AMHAZ R., CHAMBON S., IDIER J., BALTAZART V.: Automatic Crack Detection on Two-Dimensional Pavement Images: An Algorithm Based on Minimal Path Selection. *IEEE Transactions on Intelligent Transportation Systems 17*, 10 (2016), 4

[ACST15] AGUILERA-CARRASCO C. A., SAPPA A. D., TOLEDO R.: LGHD: A feature descriptor for matching across non-linear intensity variations. *2015 IEEE International Conference on Image Processing (ICIP)* (2015), 2

[APvEH*13] ALFELD M., PEDROSO J., VAN EIKEMA HOMMES M., VAN DER SNICKT G., TAUBER G., BLAAS J., HASCHKE M., ERLER K., DIK J., JANSSENS K.: A mobile instrument for in situ scanning macro-XRF investigation of historical paintings. *Journal of Analytical Atomic Spectrometry 28* (2013), 1

[Bro92] BROWN L. G.: A Survey of Image Registration Techniques. *ACM Computing Surveys 24*, 4 (1992), 1

[CDL15] CONOVER D. M., DELANEY J. K., LOEW M. H.: Automatic registration and mosaicking of technical images of Old Master paintings. *Applied Physics A 119*, 4 (2015), 1, 2, 4

[CDRL11] CONOVER D. M., DELANEY J. K., RICCIARDI P., LOEW M. H.: Towards automatic registration of technical images of works of art. In *Electronic Imaging* (2011), 2

[DMR18] DETONE D., MALISIEWICZ T., RABINOVICH A.: SuperPoint: Self-Supervised Interest Point Detection and Description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018), 2, 3, 5

[ESS*17] EISENBACH M., STRICKER R., SEICHTER D., AMENDE K., DEBES K., SESSELMANN M., EBERSBACH D., STOECKERT U., GROSS H.-M.: How to get pavement distress detection ready for deep learning? A systematic approach. *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), 4

[GDE*21] GABRIELI F., DELANEY J. K., ERDMANN R. G., GONZALEZ V., VAN LOON A., SMULDERS P., BERKEVELD R., VAN LANGH R., KEUNE K.: Reflectance Imaging Spectroscopy (RIS) for Operation Night Watch: Challenges and Achievements of Imaging Rembrandt's Masterpiece in the Glass Chamber at the Rijksmuseum. *Sensors (Basel, Switzerland) 21* (2021), 1

[HPJ12] HASAN M., PICKERING M. R., JIA X.: Modified SIFT for multi-modal remote sensing image registration. *2012 IEEE International Geoscience and Remote Sensing Symposium* (2012), 2

[LHA20] LI J., HU Q., AI M.: RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform. *IEEE Transactions on Image Processing 29* (2020), 2

[Low04] LOWE D. G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* (2004), 1, 2, 5

[Mir19] MIRHASHEMI A.: Configuration and Registration of Multi-Camera Spectral Image Database of Icon Paintings. *Computation 7* (2019), 1, 2

[NP17] NUNES C. F. G., PÁDUA F. L. C.: A Local Feature Descriptor Based on Log-Gabor Filters for Keypoint Matching in Multispectral Images. *IEEE Geoscience and Remote Sensing Letters 14* (2017), 2, 3

[RPD10] ROSTEN E., PORTER R. B., DRUMMOND T.: Faster and Better: A Machine Learning Approach to Corner Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32* (2010), 2

[SCQ*16] SHI Y., CUI L., QI Z., MENG F., CHEN Z.: Automatic Road Crack Detection Using Random Structured Forests. *IEEE Transactions on Intelligent Transportation Systems 17* (2016), 4

[SMC21] SINDEL A., MAIER A. K., CHRISTLEIN V.: Craquelurenet: Matching The Crack Structure In Historical Paintings For Multi-Modal Image Registration. *2021 IEEE International Conference on Image Processing (ICIP)* (2021), 2, 4

[SZ15] SIMONYAN K., ZISSERMAN A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. *2015 International Conference on Learning Representations (ICLR)* (2015), 3, 4 [vLNdM*20] van Loon A., Noble P., de Man D., Alfeld M., Callewaert T., der Snickt G. V., Janssens K., Dik J.: The role of smalt in complex pigment mixtures in Rembrandt's Homer 1663: combining MA-XRF imaging, microanalysis, paint reconstructions and OCT. 1

[VvLD*19] Vandivere A., van Loon A., Dooley K. A., Haswell R., Erdmann R. G., Leonhardt E., Delaney J. K.: Revealing the painterly technique beneath the surface of Vermeer's Girl with a Pearl Earring using macro- and microscale imaging. 3

[VWvdBvL19] Vandivere A., Wadum J., van den Berg K. J., van Loon A.: From 'Vermeer Illuminated' to 'The Girl in the Spotlight': approaches and methodologies for the scientific (re-)examination of Vermeer's Girl with a Pearl Earring. *Heritage Science 7* (2019), 1, 3

[XJC21] Xie J., Jin X., Cao H.: SMRD: A Local Feature Descriptor for Multi-modal Image Registration. *2021 International Conference on Visual Communications and Image Processing (VCIP)* (2021), 2

[YZY*20] Yang F., Zhang L., Yu S., Prokhorov D. V., Mei X., Ling H.: Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. *IEEE Transactions on Intelligent Transportation Systems 21* (2020), 4

[ZCL*12] Zou Q., Cao Y., Li Q., Mao Q., Wang S.: CrackTree: Automatic crack detection from pavement images. *Pattern Recognition Letters 33* (2012), 4

[ZF03] Zitová B., Flusser J.: Image registration methods: a survey. *Image and Vision Computing 21* (2003), 1

[ZHK*17] Zacharopoulos A., Hatzigiannakis K., Karamaoynas P., Papadakis V. M., Andrianakis M., Melessanaki K., Zabulis X.: A method for the registration of spectral images of paintings and its evaluation. *Journal of Cultural Heritage 29* (2017), 1, 2

[ZYZZ16] Zhang L., Yang F., Zhang Y. D., Zhu Y. J.: Road crack detection using deep convolutional neural network. *2016 IEEE International Conference on Image Processing (ICIP)* (2016), 4