

SAARLAND UNIVERSITY

FACULTY OF MEDICINE

Visual Computing for Motion Processing and Multimodal Imaging: Addressing Cross-Domain Challenges in Neuroscience Applications

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften der
Medizinischen Fakultät der Universität des
Saarlandes

vorgelegt von:

Philipp Flotho, M.Sc.

erstmals eingereicht im November 2024



UNIVERSITÄT
DES
SAARLANDES

SAARLAND UNIVERSITY

FACULTY OF MEDICINE

Tag der Promotion:

19.05.2026

Dekan der Medizinischen Fakultät:

Prof. Dr. Matthias Hannig

Berichterstattende:

Prof. Dr. Dr. Daniel J. Strauss

Prof. Dr. Andreas Keller

Prof. Dr. Andrés Bruhn

Abstract

Applying computer vision methods developed for natural RGB images to biomedical imaging, particularly in neuroscience, presents unique challenges due to data characteristics and domain-specific requirements. Cross-domain adaptation of visual computing methods from natural images to medicine, life sciences, and, particularly, neuroscience therefore shapes how images are used in biomedical research. Most methods in computer vision are optimized for specific domains and innovations often happen for natural, RGB images first. The challenge in applying these methods in systems neuroscience lies in understanding the unique statistics of the data, the limitations of the methods and the specific requirements of the target domain and application.

This thesis addresses these challenges for applications in dynamic face analysis and neuroimaging with new methods for motion analysis, motion visualization, motion correction and multimodal image acquisition. The key contributions of this work include the design of a fully calibrated, multimodal camera array consisting of RGB, thermal and *near infrared* (NIR) cameras that exploits 3D camera geometry for multimodal mapping. This setup enables the projection of landmarks detected in RGB images onto thermal images for functional analysis. It has been used to record large-scale datasets to investigate remote screening for respiratory diseases. Furthermore, building on the insights from that study, this thesis introduces a new dense and multimodal RGB-thermal landmark detection approach that achieves state-of-the-art performance on challenging thermal images. High framerate RGB videos acquired with the multi-camera setup are a rich source to study microexpressions and micromovements in faces. One approach towards face analysis is the visualization of subtle motion by means of motion magnification. Building on identified limitations of existing motion magnification methods, this thesis introduces new Lagrangian approaches for facial motion magnification by fine tuning deep learning-based *optical flow* (OF) estimation specifically for faces and using OF field processing. These methods enable high-quality magnification of micromovements even with head motion. They allow for the extraction of distinct spatiotemporal movement components, as well as the decomposition into global and local facial movements.

While the goal of motion magnification is the visualization of movement, estimated motion can also be removed for motion correction. The second part of the thesis introduces innovative approaches for motion correction of *2-photon* (2P) microscopy videos and linescans. To this end, variational optical flow estimation is revisited, and recent advances and data insights are integrated into a model that is specifically tailored to the unique statistics of 2P neuroimaging videos. The resulting motion correction toolbox, *Flow-Registration*, achieves state-of-the-art performance on challenging neuroimaging videos and novel synthetic benchmarks. By addressing limitations of existing methods for handling non-elastic deformations, such as those occurring during drug injections in tissue, and resolving them within a well-studied mathematical framework, Flow-Registration significantly advances the analysis of high-speed 2P videos. Insights into 2P motion statistics further lay the groundwork for a deep learning-based approach to 2P motion correction using synthetic label generation. Preliminary results suggest that this method achieves performance similar to Flow-Registration, particularly on synthetic benchmarks.

Zusammenfassung

Die Nutzung von Bilddaten im Bereich der Neuro- und Biowissenschaften hängt eng mit der Entwicklung neuer Methoden in verwandten Wissenschaftsbereichen zusammen. Insbesondere Methoden des maschinellen Sehens werden häufig zuerst für RGB-Bilder entwickelt und sind für diese optimiert. Die Anwendung von Methoden wie beispielsweise dem optischen Fluss und tiefe neuronale Netze in den systemischen Neurowissenschaften und der biomedizinischen Bildgebung erfordert daher die Entwicklung von Methoden, die an die spezifischen Eigenschaften der biologischen Daten angepasst sind.

Diese Dissertation löst diese Herausforderungen für Anwendungen im Bereich der Gesichtsanalyse und der Neurobildgebung und entwickelt neue Methoden zur Bewegungsanalyse, Bewegungsvisualisierung, Bewegungskorrektur und der multimodalen Bildaufnahme. Ein wichtiger Beitrag dieser Arbeit stellt die Entwicklung eines vollständig kalibrierten, multimodalen Kamerasystems dar. Das System besteht aus RGB-, Wärmebild- und Nahinfrarotkameras (NIR). Mithilfe dreidimensionaler Kamerageometrie erlaubt es die präzise Zuordnung von Merkmalen zwischen den einzelnen Modalitäten. So können beispielsweise im RGB-Bild erkannte Bereiche leicht im Wärmebild identifiziert werden. Im Rahmen der Erforschung von Methoden zur kontaktlosen Erkennung von Atemwegsinfektionen wurde das Kamerasystem für die Aufnahme großer Datensätze eingesetzt und es wurden Ansätze zur Infektionserkennung damit untersucht. Während die Eignung des Systems für Aufnahmen außerhalb des Labors nachgewiesen wurde, hat diese Studie auch Limitierungen existierender Methoden für die automatische Identifizierung von Gesichtsmerkmalen (engl. landmarks) im Wärmebild hervorgehoben. Darauf aufbauend wird in dieser Dissertation ein neuer, multimodaler Ansatz für die Bestimmung von Gesichtsmerkmalen im RGB und Wärmebild entworfen. Auf einem anspruchsvollen Datensatz mit Wärmebildern erreicht die neue Methode dabei bessere Ergebnisse als der bisherige Stand der Wissenschaft. Ein weiterer Schwerpunkt dieser Arbeit ist die Entwicklung neuer Methoden zur Bewegungsverstärkung von Mikrobewegungen im Gesicht. Existierende Methoden haben Schwierigkeiten mit der Verstärkung solcher Mikrobewegungen, die während Kopfbewegungen auftreten können. Zu diesem Zweck werden neue Lagrange-Ansätze entwickelt und eine auf neuronalen Netzen basierende Methode zur Bestimmung des optischen Flusses für Gesichter mit einem neuen Ansatz optimiert. Die neuen Methoden zerlegen hierfür die berechneten Bewegungsfelder in getrennte Bewegungskomponenten, sowie in globale und lokale Bewegungen, die daraufhin unabhängig voneinander für die Visualisierung verarbeitet werden können und erreichen dabei eine sehr hohe, visuelle Qualität.

Während das Ziel der Bewegungsverstärkung die Visualisierung von Bewegungen ist, lässt sich der optische Fluss auch für die Korrektur von Bewegungen verwenden. Diese Dissertation stellt innovative Ansätze der Bewegungsstabilisierung für die 2-Photonen (2P) Mikroskopie in 1D und 2D vor. Dazu werden die spezifischen Eigenschaften von 2P Videos untersucht um damit ein variationelles Modell für die Bestimmung des optischen Flusses für 2P Mikroskopie zu entwickeln. Die so entwickelte Methode *Flow-Registration* übertrifft den bisherigen Stand der Wissenschaft bei der Bewegungskorrektur anspruchsvoller Videos. Flow-Registration stellt ein wichtiges Werkzeug für die Neurowissenschaften dar und darauf aufbauend wird in dieser Dissertation bereits ein neuer Ansatz zur Generierung synthetischer Label für das Training tiefer, neuronaler Netze für die Bewegungskompensation von 2P Aufnahmen vorgeschlagen. Vorläufige Ergebnisse zeigen dabei, dass diese Methode insbesondere auf synthetischen Benchmarks bereits eine ähnliche Leistung wie Flow-Registration erreicht.

“Everything flows, nothing stands still.”

— Heraclitus of Ephesus

Contents

Abstract	v
Zusammenfassung	vii
Contents	xi
1 Introduction	1
1.1 Motivation	1
1.2 Overview and background	4
1.2.1 Multi-camera alignment	5
1.2.2 Motion magnification for faces	8
1.2.3 Motion correction	10
1.3 Original contributions	12
1.3.1 Scientific publications	14
1.3.2 Organization	17
2 Image acquisition and motion analysis	19
2.1 Multimodal data acquisition at SARS-CoV-2 drive through screening centers	20
2.1.1 Introduction	20
2.1.2 Materials and methods	21
2.1.3 Results	22
2.1.4 Conclusion	26
2.2 Dense landmarks for visible-thermal face analysis	27
2.2.1 Introduction	28
2.2.2 Materials and methods	32
2.2.3 Results	36
2.2.4 Conclusion	44
2.3 Lagrangian motion magnification revisited	46
2.3.1 Introduction	46
2.3.2 Materials and methods	48
2.3.3 Results	51
2.4 Lagrangian motion magnification with landmark-prior and double sparse OF decomposition for microexpressions and facial micromovements	54
2.4.1 Introduction	54
2.4.2 Materials and methods	57

2.4.3	Results	64
2.4.4	Conclusion	69
2.5	Markerless XROMM using DeepLabCut and marker masking	72
2.5.1	Introduction	72
2.5.2	Materials and methods	73
2.5.3	Results	76
2.5.4	Conclusion	78
2.6	Broader impact	79
2.6.1	Quantitative motion analysis in a study for auditory attention . .	79
2.6.2	Core temperature estimation	80
2.6.3	Influence of video compression on OF estimation	82
2.7	Summary	84
3	Motion correction	87
3.1	Variational 1D alignment	89
3.1.1	Introduction	89
3.1.2	Materials and methods	92
3.1.3	Results	96
3.1.4	Discussion	107
3.2	Flow-Registration: Non-parametric image registration of 2-photon imaging data	109
3.2.1	Introduction	109
3.2.2	Materials and methods	112
3.2.3	Results	118
3.2.4	Discussion	122
3.2.5	Conclusion	125
3.3	RAFT-2P meets Flow-Registration: Synthetic label generation for 2P motion compensation	127
3.3.1	Introduction	128
3.3.2	Materials and methods	129
3.3.3	Results	134
3.3.4	Conclusion	135
3.4	Broader impact	139
3.4.1	Flow-Registration	139
3.4.2	Motion compensation for remote EDA from RGB videos of the human palm	141
3.5	Summary	146
4	Outlook, discussion and limitations	147
4.1	Outlook and limitations	149
4.1.1	2P motion compensation and motion analysis	149
4.1.2	Lagrangian motion magnification and motion decomposition . . .	150
4.1.3	Facial landmarks on thermal images	151
	Abbreviations	153

A	Comparison of forward warping approaches	155
B	Motion-induced dependence of signal and noise	161
C	Motion magnification error	163
D	Parameter selection and data presentation for Flow-Registration	167
D.1	Parameter selection	167
D.2	False color representation	168
E	OF estimation	169
F	Variational OF optimization	173
G	Additional DTSC results	175
 Personal contributions		 178
Bibliography		183
Acknowledgements		203

Chapter 1

Introduction

1.1 Motivation

Image processing, computer vision and scientific visualization are important areas of computer science that deal with image data. They are collectively referred to as *visual computing*. In recent years, advances in visual computing have had a huge impact on many scientific areas such as medicine, the life sciences, and, in particular, neuroscience. Neuroscience is the study of the nervous system at all scales from molecular mechanisms, cellular processes and neural circuits up to higher-order functions such as behavior, network level functions and social interactions. Systems neuroscience, as a holistic approach, relies on different imaging modalities to study these processes across different scales. This includes functional optical imaging and *2-photon* (2P) microscopy for studying brain activity in animal models up to the analysis of facial features in humans with RGB and thermal images. In comparison to natural images, neuroimaging modalities are very specialized. As a result, advances in imaging are often driven by solving problems for natural images such as faces or natural scenes. These data are highly available and the methods find broad application — from smartphones to self-driving cars. Applying those methods to out-of-domain data such as animals in lab settings is non-trivial, even when using the same modality such as *red-green-blue* (RGB) color images.

One such example of successful transfer and domain adaptation with a significant impact on the biomedical communities is *DeepLabCut* (DLC) [1, 2, 3] which is a framework for pose estimation. It took two years from the publication of the DeeperCut paper for human pose estimation [4] to its adaptation for tracking of animals by means of few and zero shot learning in DLC.

The availability of a new, state-of-the-art pose estimation method that can be retrained with only a few examples for the marker-less tracking of any animal and without computer vision expertise was transformative for behavioral research. It enabled big-data analysis in the context of animal studies which required manual annotations and analyses previously. Similarly, the neural network architecture *U-net* [5] was originally developed for the segmentation of biomedical images but evolved into a foundational building block for various types of modern deep learning-based approaches, even guided diffusion image generation [6] which is still the backbone of modern image generation models.

The emergence of other, generically applicable algorithms such as foundation models for computer vision and their high availability in easy-to-use programming frameworks as well as the existence of new medical datasets rendered the communities much more closely connected. Today, even conferences dedicated purely to the advances of computer vision such as the *Institute of Electrical and Electronics Engineers* (IEEE) conference on *computer vision and pattern recognition* (CVPR) have regular workshops on biomedical topics¹. At the same time, conferences on biomedical engineering such as the *IEEE engineering in medicine and biology conference* (EMBC) contain specific sections for different aspects of image processing and computer vision applications².

However, to a large part those are about modalities with broad availability, while niche problems and data can lack cross-disciplinary adoption. For example, *optical flow* (OF) methods, which are methods to compute motion in videos and which have a rich theoretical base, have long been considered not ideal for motion correction of 2P neuroimaging. Instead, researchers often had to reinvent the wheel and develop new methods from scratch. Similarly, as we will later see in this thesis, studies on thermal facial landmarks often use weak baseline methods and benchmarks making it difficult to judge the real world performance of new methods. Applying computer vision methods that have been developed for RGB images of natural images to specialized biomedical imaging modalities is challenging. This domain adaptation requires domain-specific knowledge as well as technical expertise and theoretic understanding to adapt and develop methods. It requires knowledge of the statistics in the data and access to large datasets. This can be difficult to achieve with niche sensors such as thermal imaging or multispectral recordings. With their unique, spectral properties as well as niche imaging areas with unique statistics, such as the surface of the hand or the human ear, they can be important sources of information to study new effects in neuroimaging. Consequently, it can also take many years for advances from visual computing to find applications in areas of biomedical research.

¹The CVPR 2022 offered the 9th CVF/CVPR Medical Computer Vision Workshop.

²The EMBC 2022 technical program contained seven sessions on image analysis and computer vision.

When considering thermal and optical imaging or microscopy, the statistics in the images are so different such that vanilla RGB methods rarely work out-of-the-box. At the same time, they have enough similarities for successful cross-domain adaptation of existing methods.

We established the *Visual Computing for Neuroscience* lab at the *Systems Neuroscience and Neurotechnology Unit* (SNNU) in 2018 after having identified areas where imaging modalities can complement the experimental setups during publication of results from my master's thesis and preliminary experiments with multi-camera setups at the unit. As lab head, I was given the opportunity to coordinate efforts to fill the gap between the results from the theoretical neurosciences and real-world applications and to develop frameworks for new methods and modular, complementary computer vision setups. The core contribution of the visual computing lab as a coordinated effort of multiple researchers is an image acquisition framework which allows to record multispectral, big-data recordings in multiple lab and out-of-lab scenarios alongside traditional, contact-based gold standard measurements such as measurements of respiration or pulse. Together with cross-modal calibration and dense, camera geometry-based mapping, this allows to generate unique annotations in video datasets to develop new, camera-based methods for physiological and psychophysiological assessment. Our approaches have proven flexible and scalable, for example, we were able to acquire recordings from over 400 participants in an out-of-lab setting during their visit to SARS-CoV-2 drive through screening centers when the COVID-19 pandemic put all other, ongoing experiments on hold. This formed the foundation for the joint research project VI-Screen which investigates remote methods for the assessment of respiratory diseases. With the acquisition of large datasets, we were able to identify new challenges related to the non-contact, psychophysiological assessment from faces.

Most methods in this thesis fall under the category of domain adaptation. They fill gaps between existing algorithms and the practical requirements of systems neuroscience. Similarly to how DLC transformed animal pose estimation by adapting human pose estimation models, my work focuses on the adaptation of different methods such as OF to address the unique challenges in neuroimaging, thermal imaging and other domains.

This thesis presents advances in methods for image acquisition, motion correction, motion analysis and multimodal imaging that have been the focus of my work. It consists largely of my first-authored publications, but also includes preliminary results from follow-up work and contributions from co-authored publications. This chapter continues with an overview and background for related methods. Finally, it summarizes the contributions of this thesis with a detailed description of my personal contributions for each of the projects. The rest of the thesis is split into two main chapters followed by a conclusion and future work chapter.

1.2 Overview and background

Existing methods in computer vision are often optimized for specific applications and data and innovations often happen for natural, RGB images first. The contributions in this thesis revolve around cross-domain adaptation and transfer of methods and concepts from visual computing to systems neuroscience and neuroimaging. This requires the analysis of existing methods, to identify their limitations and to adapt them to the unique statistics of the target neuroimaging data. Furthermore, it is crucial to identify strong baselines and define benchmarks that allow meaningful comparisons with existing methods and realistic assessment of their performance.

Systems neuroscience has a holistic and interdisciplinary approach to understanding the brain which happens on many different scales and modalities. Methods from computer vision find applications at each scale such as from the understanding of faces and human pose down to the segmentation and analysis of individual neurons in functional microscopy videos.

The core contributions of this thesis include methodological advancements in **motion analysis**, **motion visualization**, **motion correction**, and **multimodal image acquisition and alignment** for sensor fusion. Some of the approaches are dense in the sense that motion information exists for each pixel of the input image. The dense motion approaches rely on the calculation and adaptation of OF and the adaptation of OF models to different modalities. Here, we retrain deep learning-based OF for faces and adapt variational OF estimation to the statistics of 1D neuroimaging data as well as 2D 2P videos. Sparse approaches estimate motion or spatial information for a sparse set of image points. In the context of this thesis, we are interested in facial landmarks. Facial landmarks are distinct points on the face such as the nose or the eyes that can be used in downstream tasks. For facial landmarks, we will use the term *dense* also when we depict landmark conventions that define more than 70 points.

This section provides an introduction to multi-camera alignment, motion magnification, and motion correction to give context for this thesis.

1.2.1 Multi-camera alignment

Multi-camera alignment plays a crucial role for the multimodal camera arrangements in this thesis (see Figure 1.2). In those camera systems, we want to fuse the information from different image sensors. A core problem is the localization of the same image points in different cameras. Two complementary approaches to obtain this mapping are followed in this work. First, geometric camera calibration, which enables depth estimation and projection of image points from one camera to their corresponding locations in the other cameras. Second, pose estimation networks that are fine tuned to the domain of each camera. The background for the first approach is briefly introduced below. A more in depth introduction of the related concepts can be found in [7].

An important concept from optics is the pinhole camera model: This is a simplified camera model, where the projection of 3D world points onto the image plane is expressed as a perspective projection. The pinhole camera model is based on the *camera obscura* and assumes a pinhole camera with infinitely small aperture [8]. The relationship between 3D world coordinates and pinhole camera coordinates is defined by the rotation and translation between the world coordinate system and the camera. The translation vector and rotation matrix make up the so called extrinsic matrix and the intrinsic camera matrix is made up of focal length, effective image resolution and principal point. Extrinsic and intrinsic camera matrices describe the full camera model and the estimation of those matrices is called camera calibration.

Figure 1.1 visualizes the epipolar constraint that underlies stereo matching in calibrated camera pairs. For a 3D point M , its projections m_i must lie on epipolar lines obtained as the intersection of the epipolar plane with the respective image plane. This reduces correspondence search from a 2D area to a 1D line. In our multi-camera rigs we calibrate the pairs (C_1, C_2) and (C_2, C_3) , triangulate using a calibrated pair, and then project the recovered 3D point into further calibrated views such as C_3 .

Successful camera calibration requires pairs of 3D object points together with their corresponding 2D image points. This is implemented with the help of geometric calibration patterns that consist of structures that can be detected by a combination of operators from traditional image processing and low level computer vision. A commonly used pattern is the checkerboard pattern [8] which consists of a checkerboard pattern of known spacing. Different applications can also require a circleboard pattern with equally placed, solid color circles (see Figure 1.3).

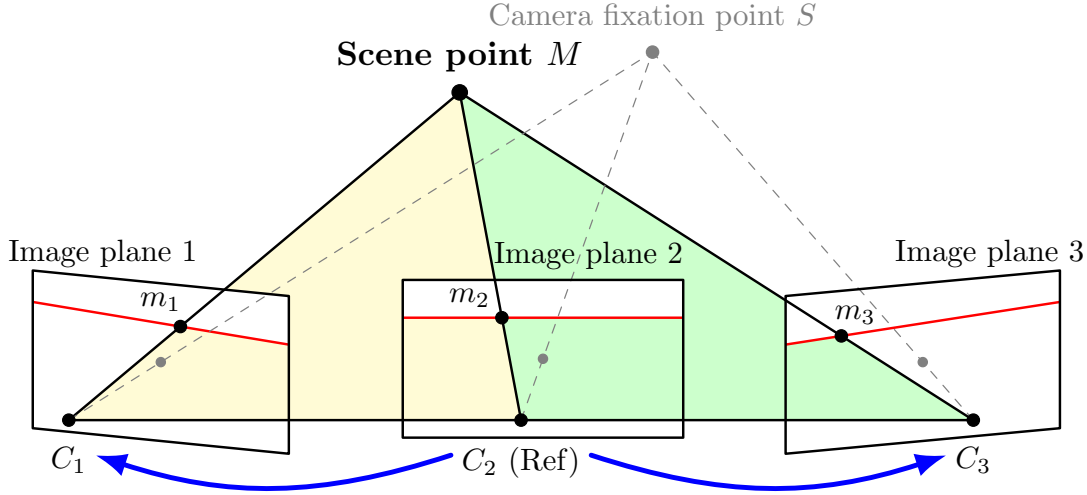


Figure 1.1: Schematic of epipolar geometry in a convergent three-camera system with reference camera C_2 . All cameras are pointed towards a common fixation point S . A 3D scene point M projects to image points m_1, m_2, m_3 on the image planes of cameras C_1, C_2, C_3 . For each calibrated stereo pair (C_1, C_2) and (C_2, C_3) , the corresponding camera centers together with M span an epipolar plane (yellow for (C_1, C_2) and green for (C_2, C_3)). Its intersections with the image planes yield the associated epipolar lines (red lines). Given a point m_2 in the reference view, the matching points in views 1 and 3 must lie on the corresponding epipolar lines in each view. Conversely, m_1 and m_3 induce epipolar lines in the reference image plane, which coincide in this example because the cameras are collinear and the optical axis of C_2 is orthogonal to this baseline. Pairwise calibration enables triangulation from corresponding image points and subsequent projection of reconstructed 3D points into any other calibrated view.

The methods in this thesis assume a pinhole camera model with radial and tangential lens distortion, usually modeled as the barrel or pincushion transform [8]. The proposed multi-camera system is calibrated using pairwise stereo calibration. The camera system consists of a depth component alongside cameras of other modalities, such as thermal or hyperspectral cameras. The depth component may include two cameras of the same modality for stereo-based depth estimation or other types of depth camera. Calibration allows us to compute 3D world coordinates in absolute units.

Stereo estimation uses two different camera views to estimate depth by means of triangulation. The traditional approach to stereo estimation exploits the fact that the intrinsic and extrinsic camera parameters constrain a point in one camera to lie on a line in the other view, the so called epipolar line. This relationship is covered by the so called fundamental and essential matrices. For stereo estimation, an epipolar constraint can be incorporated into the correspondence problem for joint OF and stereo estimations [9]. It is also used to rectify images from stereo cameras to bring the images into a canonical form.

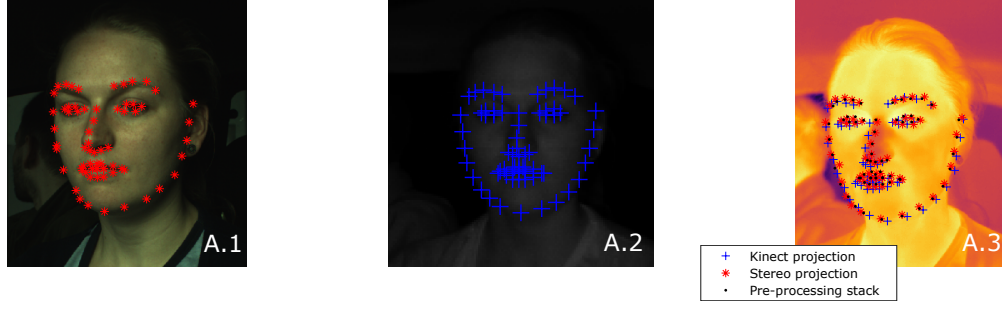


Figure 1.2: Landmark detection in the different modalities. For RGB (A.1) and NIR (A.2) facial landmarks have been computed with the *deep alignment network* (DAN) [10], for thermal (A.3) pre-processing functions [F1] and cross-modal mapping from RGB or NIR are used. Chapter 2 later discusses the suitability of pre-processing functions from [F1] and DAN for thermal images.

This means that images are transformed into a representation where world points are projected onto the same horizontal lines in both images, which can also be achieved by physical alignment of the cameras in a parallel stereo setup [8]. This can also be exploited for uncalibrated stereo, where F and E are estimated from corresponding points in the cameras via iterative model estimation methods such as *random sample consensus* (RANSAC), however, uncalibrated stereo can usually not be used to compute metric depth in absolute units.

With high accuracy, large displacement, dense OF approaches, we can achieve unconstrained correspondence matching and directly identify corresponding points. For two corresponding points in two camera views, homogeneous 3D world coordinates $\tilde{X}$ of a scene point M can be computed via triangulation by solving the linear system $A\tilde{X} = 0$ where A is constructed from the two projection matrices P_1 and P_2 and the corresponding image point coordinates $\tilde{x}_1$ and $\tilde{x}_2$ of image points m_1 and m_2 . The solution is obtained using *singular value decomposition* (SVD) of A .

Having computed $\tilde{X}$ with camera C_1 and camera C_2 , we can project any point to a camera C_3 view with the corresponding 3×4 projection matrix P_3 such that $\tilde{x}_3 = P_3\tilde{X}$ yields the homogeneous coordinates of image point m_3 up to scale (see Figure 1.1). This allows us to project any image point from a stereo camera pair or depth camera into any other calibrated camera.

1.2.2 Motion magnification for faces

Motion magnification methods are a group of visualization techniques that magnify movements in videos: Videos can contain small motions with no more than subpixel magnitude that are not visible to the human eye. Traditional methods fall into two groups, namely, Lagrangian and Eulerian motion magnification. This naming scheme has its motivation in fluid simulations, where Eulerian methods work on a fixed grid, while Lagrangian methods track individual particles. Accordingly, Eulerian motion magnification works on a fixed pixel grid [11], while Lagrangian methods manipulate point trajectories in images [12]. One of the first methods for motion magnification is the Lagrangian method by Liu et al. [12]. The authors use a combination of image segmentation and point tracking approaches to generate displacements for different image regions which then are amplified each by a different factor.

Traditionally, motion estimation has been regarded as a computationally expensive and challenging problem. As a result, the performance of Lagrangian methods is often constrained by the efficiency and accuracy of the underlying motion estimation approach [13]. Therefore, the primary strength of the Eulerian approach of Wu et al. [11] is that it does not require explicit motion estimation. Instead, it exploits the linearization assumption [14], where motion in a video is assumed to be locally linear changes in pixel intensity. This renders the motion magnification problem as a manipulation of local image gradients. Those gradients can be implicitly represented using Laplacian image pyramids. This allows for computationally very efficient implementations that use different amplification factors in different spatial bands. However, this intensity-based approach is only effective when the linearized motion assumption holds true. In scenarios with significant noise or large motions, it can lead to artifacts. Phase-based motion magnification has been proposed to extend those capabilities. It amplifies local phase shifts in the video [15]. While this method generally outperforms intensity-based motion magnification visually, it can introduce blurring artifacts around image edges. Without explicitly estimated motion, both Eulerian methods rely on spatiotemporal filtering and assume periodic motion in the videos.

In this thesis, we are interested in amplifying subtle motion in faces. Examples of such movements are microexpressions, which are fast and involuntary expression changes. Facial micromovements are usually irregular, fast and small. At the same time, they can be masked by large head motion, which introduces non-linear displacements. Furthermore, while motion magnification can be a suitable tool for the manual analysis of subtle motion in the face, real-world applications require motion information.

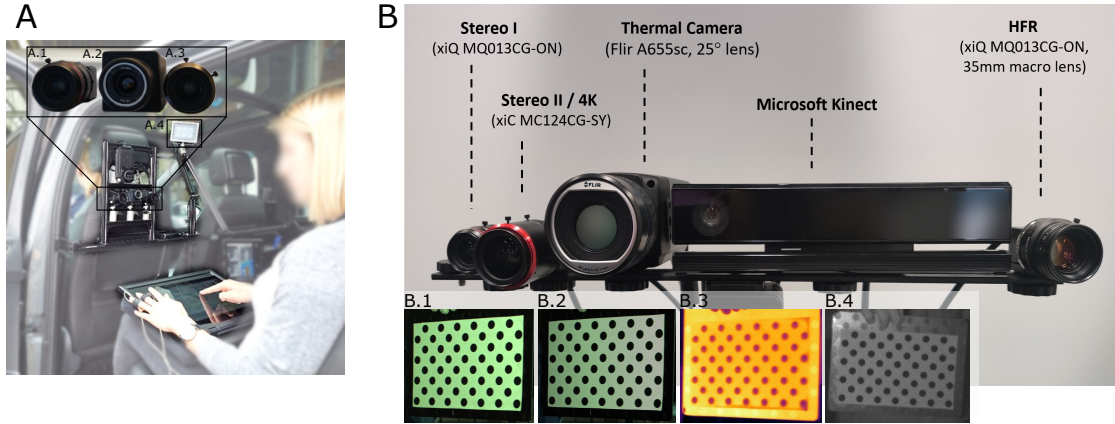


Figure 1.3: Different alignments for RGB/NIR/Thermal cameras. Inside of experimental vehicles such as in [F2], the participants sit on the left backseat during the experiment and perform tasks on a tablet (A). Here, cameras A.3 and A.1 as well as A.3 and A.2 are pairwise calibrated. The setup is illuminated by diffuse NIR illumination at 850 nm (A.4). For a focus on mapping across cameras, another possibility is to use a RGB-RGB-Thermal-NIR configuration (B) from [F1]. Here, a fast camera (B.1) and a thermal camera (B.3) are calibrated with a slower high resolution camera (B.2). In this setup, the Kinect sensor (B.4) produces RGB, NIR and depth information and is calibrated with the thermal camera so that points detected in the Kinect streams can be projected into the thermal camera image plane via the depth information. Calibration is performed with a circular calibration board that is visible in all imaged bands.

The motion information within Eulerian as well as many deep learning-based motion magnification approaches cannot be extracted in a straightforward way. Advances in dense OF estimation in recent years have led us to revisit Lagrangian motion magnification with dense displacement fields for facial motion visualization.

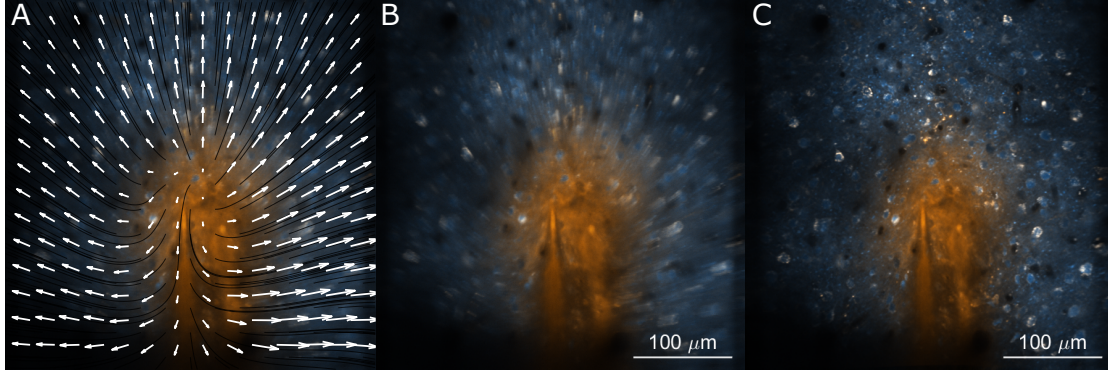


Figure 1.4: Frames from a video recorded during injection into the brain with arrow visualization of the motion (A) where the tissue expands outward from the injection point [F3]. Averaging without motion correction introduces motion blur (B) while the average of stabilized videos preserves crisp image details (C).

1.2.3 Motion correction

Microscopy videos of the brain in awake and behaving animals contain various sources of movement. This makes the analysis of signals of interest in such videos difficult: An example of such signals is the intensity change of voltage or metabolic indicators attached to individual neurons, which can be very subtle. In fact, very small signals can be substantially distorted by tiny motion, and motion can hinder automated signal extraction (see Appendix B and C). Hence, the removal of motion and stabilization of videos is an important part of many neuroimaging pipelines.

In the visual computing literature, removing motion from videos has different naming conventions that depend on the community. *Motion compensation* is often used to compensate motion effects on image intensity for applications in compression [16] whereas *motion correction* is often used for biomedical applications [17].

The term *warping* describes resampling an image under a transformation, such as a displacement field using interpolation. *Image registration* computes such displacements or a transformation between two images, and warping is then used to align the moving image to the reference. In the context of *medical image registration*, the images can be from different modalities to combine information or contain a template against which a new image is compared [18]. In the OF literature, the term *warping* is also used to compensate flow increments, for example across different scales of the images [19]. One goal of this thesis is to develop methods that motion correct different functional neuroimaging modalities. In motion corrected videos, intensity timecourses can be easily extracted for analysis. In the same way as for motion magnification, we will use dense OF for motion correction of difficult motion in videos with very low *signal-to-noise ratio* (SNR).

This thesis contains methods that fall into the groups of variational OF and deep learning-based OF estimation with the goal to remove non-elastic deformations (e.g. see Figure 1.4). For variational OF, the challenge is to find a suitable mathematical model that describes the expected data. For deep learning-based OF, on the other hand, architectural innovations are secondary in this thesis and we focus on the data generation for motion correction.

Both require us to estimate a displacement field where each vector points from a location in the reference image to the corresponding location in a moving image, ideally with subpixel precision. A dense displacement field is a vector field $\mathbf{w} := (u, v)^\top : \Omega \rightarrow \mathbb{R}^2$ defined on an image domain $\Omega \subset \mathbb{R}^2$ that maps from a reference image f_{ref} to an image f such that $f_{\text{ref}}(\mathbf{x})$ depicts the same part of the image as $f(\mathbf{x} + \mathbf{w}(\mathbf{x}))$. In this thesis, we often define $\mathbf{x} := (x, y)^\top$ to simplify notation. With the knowledge of $\mathbf{w}$, f can be warped into the coordinate system of the reference via backward warping such that we can define a stabilized image as $\hat{f}(\mathbf{x}) := f(\mathbf{x} + \mathbf{w}(\mathbf{x}))$. The goal of motion correction or registration is to estimate $\mathbf{w}$ such that $\hat{f}$ and f_{ref} depict the same structures. In the idealized noise-free case with brightness constancy, this means $\hat{f} = f_{\text{ref}}$. In practice, this corresponds to $f_{\text{ref}}(\mathbf{x}) \approx f(\mathbf{x} + \mathbf{w}(\mathbf{x})) = \hat{f}(\mathbf{x})$.

1.3 Original contributions

This thesis contributes to biomedical image processing of neuroimaging data, imaging of faces under difficult conditions and multimodal, multi-camera image acquisition. This is achieved by addressing cross-domain challenges: The approaches build on methods and strategies from computer vision and image processing and adapt them to the unique requirements and statistics of neuroscience applications and data.

The contributions are linked to the establishment of a Visual Computing for Neuroscience Lab at SNNU. The lab focuses on the development of methods from the field of computer vision and image processing as well as high-performance computer vision image acquisition systems for biomedical applications. The core contributions of this thesis are the following:

- Development of the Flow-Registration framework and toolbox for dense OF-based motion compensation of 2-photon microscopy videos in a collaboration with the optical neuroimaging unit at the Okinawa Institute of Science and Technology, Graduate University (OIST). Flow-Registration outperforms the current state-of-the-art and interfaces with popular 2-photon imaging suites and is being actively applied and adapted in high-impact studies such as [20, 21, 22]. The insights from Flow-Registration laid the foundation for a deep learning-based method for 2P motion correction with synthetic label generation.
- A new and fast method for the alignment of 1D signals that showed higher robustness and accuracy than established approaches such as dynamic time warping or cross correlation-based alignment. Furthermore, it allows the alignment of non-flat displacements in various neuroimaging data.
- Design of fully calibrated multi-camera setups for multimodal and functional thermal imaging that are suitable for out-of-lab data acquisition, particularly, in the context of remote respiratory disease screening. This work represents an important proof-of-concept for out-of-lab contactless data acquisition and evaluation in the joint projects VI-Screen BMBF-FZ 13N15753 (2021-2024) and VI-Screen-PRO BMBF-FZ 13N17308 (2025-2027) — coordinator neurotechnology is Prof. Dr. Dr. Strauss and involved research institutions are Saarland University, Technical University Berlin and Saarland University Hospital.

- Development of new Lagrangian motion magnification methods with magnitude driven, non-linear motion scaling and approaches for motion decomposition in the context of faces. In contrast to most Eulerian motion magnification approaches, the presented methods allow the magnification of micromovements in faces during head movement, enabling imaging in less constrained setups.
- Development and description of a multi-camera setup for scene-flow analysis of ear motion in a psychophysiological experiment for auditory attention.
- DLC-based tracking approach for *x-ray reconstruction of moving morphology* (XROMM) recordings. The new approach enables marker-less tracking in x-ray recordings (i.e. without the need for implanted, radiopaque beads) with high accuracy that allows 3D reconstruction in stereo setups.
- New deep learning-based approach for the detection of dense, facial landmarks in thermal and RGB images together with a corresponding synthetic dataset. The multimodal method outperforms the state of the art on thermal images.
- **Preliminary results:**
 - RAFT-2P: Synthetic label generation for 2P motion correction with RAFT optical flow.
 - Thermal image analysis and alignment on the VI-Screen dataset for the development of deep learning-based, RGB-driven correction of the body core temperature.

1.3.1 Scientific publications

Most of this thesis has been peer-reviewed and published in journals as well as in the proceedings of scientific conferences. This thesis builds on those publications and contains direct excerpts. References [F4, F1, F3, F5]³ **are first author journal publications**. Reference [F6] **is a first author full paper in the main proceedings** of the IEEE/CVF *Conference on Computer Vision and Pattern Recognition* (CVPR). References [F7, F2, F8] **are co-authored journal publications** and references [F9, F10, F11, F12, F13] **are preliminary works** that have been published in MEDLINE and IEEE indexed and [F14] in IEEE indexed conference proceedings. References [F15, F16, F17, F18, F19] **are unpublished 1-page abstracts** presented at conferences.

The unaltered, signed, and published author contributions for each of the journal publications that are used within this thesis are as follows.

- *Vestigial auriculomotor activity indicates the direction of auditory attention in humans* [F7]: Daniel J Strauss: Conceptualization, Resources, Formal analysis, Supervision, Validation, Investigation, Methodology, Writing - original draft, Project administration, Writing - review and editing, Farah I Corona-Strauss: Conceptualization, Software, Formal analysis, Supervision, Validation, Investigation, Visualization, Methodology, Andreas Schroeer: Software, Formal analysis, Validation, Investigation, Visualization, **Philipp Flotho**: Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Ronny Hannemann: Conceptualization, Methodology, Steven A Hackley: Conceptualization, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review and editing.
- *Fast variational alignment of non-flat 1D displacements for applications in neuroimaging* [F4]: **Philipp Flotho**: Conceptualization, Methodology, Software, Visualization, Writing – Original Draft, Investigation, Formal analysis, Data Curation David Thinnies: Software, Data Curation, Visualization, Writing – Review & Editing Bernd Kuhn: Writing – Review & Editing, Conceptualization, Supervision, Resources Christopher J. Roome: Conceptualization, Data Curation Jonas F. Vibell: Writing – Review & Editing, Conceptualization, Data Curation Daniel J. Strauss: Supervision, Conceptualization, Writing – Review & Editing, Project Administration, Resources, Funding acquisition.

³Throughout this thesis, my own contributions follow the citation convention [F#], whereas general literature is cited as [#].

- *Multimodal data acquisition at SARS-CoV-2 drive through screening centers: Setup description and experiences in Saarland, Germany* [F1]: Daniel J. Strauss & **Philipp Flotho** conceived and designed the study. Mayur J. Bhamborae, Tobias Grün, David Thinnes, and Dominik Limbach conducted experiments. Mayur J. Bhamborae and Tobias Grün implemented the acquisition system. The data were processed, analyzed, and curated by **Philipp Flotho**, Tobias Grün, Carlos Trenado, Mayur J. Bhamborae, David Thinnes, and Dominik Limbach. The manuscript was written by **Philipp Flotho** and Daniel J. Strauss.
- *Software for non-parametric image registration of 2-photon imaging data* [F3]: **Philipp Flotho** contributed to software, methodology, conceptualization, data curation, formal analysis, visualization, and writing-original draft. Shinobu Nomura contributed to data curation, investigation, writing-review & editing, and validation. Bernd Kuhn contributed to conceptualization, writing-review & editing, supervision, and resources. Daniel J. Strauss contributed to conceptualization, writing-review & editing, and supervision.

Author contributions of original works that have been published without credit statements are as follows.

- *T-FAKE: Synthesizing thermal images for facial landmarking* [F6]: **Philipp Flotho** and Moritz Piening share equal author contribution. **Philipp Flotho** implemented and developed the landmarker training and ablation and contributed to writing-original draft, data curation, analysis, investigation and conceptualization. Moritz Piening implemented and developed the thermalization model and contributed to writing original draft, data curation, analysis, investigation and conceptualization. Anna Kukleva contributed to writing-review & editing. Gabriele Steidl contributed to writing-review & editing and conceptualization.
- *Rodent tongue kinematics with Markerless XROMM Using DeepLabCut and Marker Masking* [F13]: **Philipp Flotho** and Mohamed Eltabbal share equal author contribution. **Philipp Flotho** developed and implemented the DLC / xmalab interfaces, the marker masking and training, contributed to writing the original draft, data curation, investigation and analysis. Mohamed Eltabbal implemented the tongue kinematics analysis pipeline, contributed to writing the original draft, data curation, investigation and analysis. Teresa E. Lever contributed to writing-review & editing. Kazutaka Takahashi contributed to conceptualization, data curation, writing-review & editing, and supervision.

- *Lagrangian motion magnification with double sparse optical flow decomposition* [F5]: **Philipp Flotho** developed and implemented fine tuning of the deep learning-based optical flow method for faces and contributed to conceptualization, writing the original draft, data curation, investigation and analysis. Cosmas Heiss developed and implemented the double sparse PCA decomposition as well as the OpenGL-based fast forward warping and contributed to writing the original draft, investigation and analysis. Gabriele Steidl contributed to writing-review & editing and supervision. Daniel J. Strauss contributed to conceptualization, writing-review & editing, and supervision.
- *Lagrangian motion magnification with landmark-prior and sparse PCA for facial microexpressions and micromovements* [F12]: **Philipp Flotho** and Cosmas Heiss share equal author contribution. **Philipp Flotho** developed and implemented fine tuning of the deep learning-based optical flow method for faces, as well as the landmark-based decomposition and contributed to writing the original draft and conceptualization. Cosmas Heiss developed and implemented the sparse PCA decomposition and contributed to writing the original draft and conceptualization. Gabriele Steidl contributed to writing-review & editing and supervision. Daniel J. Strauss contributed to conceptualization, writing-review & editing, and supervision.
- *Lagrangian motion magnification revisited: Continuous, magnitude driven motion scaling for psychophysiological experiments* [F9]: **Philipp Flotho** developed and implemented the motion magnification approaches and contributed to writing the original draft, data curation, investigation and analysis. Mayur Bhamborae contributed to writing-review & editing and analysis. Lars Haab contributed to writing-review & editing and supervision. Daniel J. Strauss contributed to conceptualization, writing-review & editing, and supervision.

1.3.2 Organization

Following this introduction, this thesis is split into two main chapters which are followed by an outlook and discussion chapter. Chapter 2, *Image acquisition and motion analysis*, introduces a multimodal, multi-camera system that is evaluated in the context of remote respiratory disease screening. The chapter discusses different use-cases and solutions for sparse and dense motion analysis, visualization (i.e. motion magnification) and calibrated multi-camera systems. This chapter follows the progression of efforts to analyze videos from faces in out-of-lab environments and the extension of those concepts to out-of-domain data. Starting with the multi-camera system, limitations of existing methods as well as suitable baselines are identified. For facial landmarking in high resolution thermal imaging, Chapter 2 considers RGB methods as baseline alongside methods tailored for thermal images. Based on those insights, a novel, deep learning-based, multimodal RGB-thermal method for facial landmarks is proposed. In the context of motion magnification, limitations of Eulerian motion magnification for the visualization of facial micromovements are identified and new approaches for Lagrangian motion magnification of facial micromovements in the presence of head movements are proposed. For that purpose, a strategy for OF fine tuning on faces is developed.

While Chapter 2 focuses on analyzing motion in dense and sparse motion data, the approaches in Chapter 3, *Motion correction*, aim at removing any motion from videos. This chapter describes the development of new dense motion correction methods for various types of neuroimaging data. It revisits OF-based methods for motion correction of 2-photon neuroimaging videos. The main contribution is the development of suitable models that capture the unique statistics of 2P microscopy images and incorporating them with recent advances in variational and deep learning-based OF computation on natural images.

Finally, Chapter 4, *Outlook, discussion and limitations*, discusses the presented methods and their limitations and gives an outlook on future work. The results are revisited in the broader context of cross-domain adaptation of methods from image processing and computer vision.

Chapter 2 and Chapter 3 consist of multiple publications. The introduction of each chapter describes the relation of each section to the covered publications.

Chapter 2

Image acquisition and motion analysis

This chapter explores the implementation and applications of multi-camera systems in biomedical and psychophysiological research, and presents methodological advancements for facial landmark detection in thermal images and motion magnification for faces. For that purpose, first, a modular, pairwise calibrated multi-camera system consisting of RGB, thermal, and *near infrared* (NIR) cameras is presented. This system forms the basis for large-scale data acquisition during psychophysiological and medical studies. With hundreds of monitored patients recorded from SARS-CoV-2 screening stations alone, this dataset ranks among the largest paired RGB-thermal datasets in biomedical research available today. The large-scale data acquisition and evaluation conducted in the context of respiratory disease screening provided unique insights. These insights informed the development of methods for robust landmark detection and motion magnification. These methods are discussed in later sections of this chapter.

This chapter contains results and direct excerpts from the publications [F1, F7, F2, F9, F12, F5, F13, F6] and from poster abstracts [F16, F17]. See 1.3.1 for the unaltered, signed, and published author contributions for each of the journal publications.

In particular, section 2.1 is taken from [F1]. Section 2.2 describes a thermal landmark detection method and contains direct excerpts from preprint [F6]. Section 2.3, based on [F9], and section 2.4, partially taken from [F12, F5], discuss new approaches for motion magnification. Section 2.5 is an excerpt from [F13].

Section 2.6 discusses the broader impact of this chapter. It includes preliminary results from poster presentations [F16, F17], with direct excerpts from the submitted abstracts.

2.1 Multimodal data acquisition at SARS-CoV-2 drive through screening centers

In 2020, SARS-CoV-2 *drive through screening centers* (DTSC) had been implemented worldwide as a fast and secure way of mass screening. During this study, DTSCs were used as a platform for the acquisition of multimodal datasets that are needed for the development of remote screening methods. The acquisition setup consists of an array of thermal, infrared and RGB cameras as well as microphones and this section applies methods from computer vision and computer audition for the contactless estimation of physiological parameters. At the DTSC, a multimodal dataset of DTSC participants in Germany for the development of remote screening methods was recorded. The results from this section demonstrate that the acquisition in the early stages of a pandemic and in regions with high infection rates can facilitate and speed up the identification of infection specific symptoms and large-scale data acquisition at DTSC is possible without disturbing the flow of operation.

This section contains direct excerpts from [F1]¹.

2.1.1 Introduction

Research on digital technologies to combat the COVID-19 pandemic includes the computational analysis of video and audio data [23, 24, 25]. Due to their contactless nature, such methods are particularly promising and needed for mass screening purposes [25] and besides fever, there are other atypical and non-severe symptoms (e.g. [26, 27, 28, 29]) which allow for non-contact medical assessment. The value of such screening systems would of course be directly related to the achievable sensitivity and specificity for detecting SARS-CoV-2 infections. However, it is challenging to acquire homogeneous data sets for the development and assessment of such remote systems without interfering with medical services due to the pressure of the ongoing pandemic.

For the rapid collection of samples for polymerase chain reaction (PCR) based screening, *drive through screening centers* (DTSCs), e.g., Kwon et al. [30], are already in use in several countries. There are advantages of DTSCs for the acquisition of contactless recordings of the patients: They simulate a scenario, where contactless screening of infectious diseases might be implemented one day, such as the entrance of an employee parking area.

¹The code for this section is available at https://github.com/phf1ot/multimodal_cam_calib.

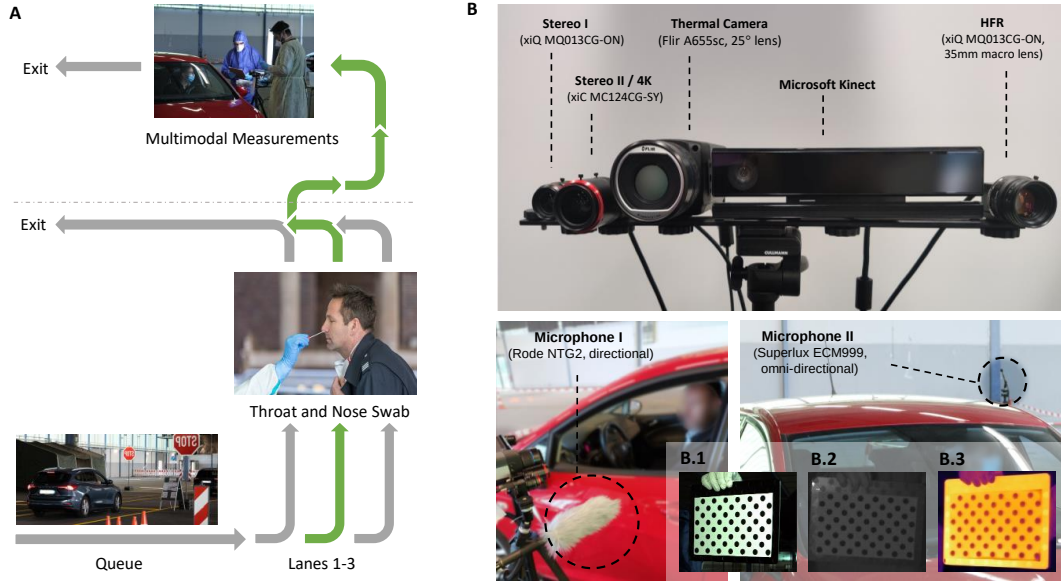


Figure 2.1: Schematic depiction of the flow at the SARS-CoV-2 DTSC (A) and of the multimodal acquisition system (B). Measurements were done through the opened window of the car with RGB, NIR, depth and thermal cameras as well as with microphones. We recorded from the thermal camera at 50 fps (B, top center), the *high framerate* (HFR) camera with macro lens for face closeups at 120 fps (B, top right) and the high resolution camera in a stereo setup at 10 fps (B, top left). The stereo cameras covered the same field of view as the thermal camera. The camera setup was calibrated with a circular calibration pattern glued to a metal plate with pattern cut out as seen in the RGB (B.1), near infrared (B.2), and thermal (B.3) image. Of the three possible lanes for the swab tests, the middle one was used during the time of our measurements.

The exposure of equipment and personnel to patients is minimized and at the same time the exposure of healthy participants to contaminated air or equipment is fully controlled and can be completely avoided as the patients stay seated in their own car.

We propose an acquisition system along with a processing pipeline for rapidly acquiring such data at DTSC without disturbing their flow of medical operation and present a multimodal dataset of DTSC users as well as preliminary evaluations.

2.1.2 Materials and methods

We recorded our dataset between May and July 2020 at the SARS-CoV-2 DTSC located at the former fairground area in Saarbrücken, State of Saarland, Germany. The study was approved by the responsible ethics committee (ethics commission at the Ärztekammer des Saarlandes, ID No 90/20) and after a detailed explanation of the procedure, all included participants signed a consent form.

Admission to and recommendation for the tests was given by the participants' general practitioner if a patient had a potential SARS-CoV-2 infection based on the Robert Koch Institute's guidelines [31]. The PCR-test result for SARS-CoV-2 from the individual nose and throat swab was accessible for us at the responsible public health office. The recordings were done through an opened window with the participants sitting in their car. Our multimodal setup consisted of RGB, NIR, depth and thermal cameras as well as microphones (see Figure 2.1). We recorded at 120 fps (face closeups, RGB), 50 fps (thermal camera), 30 fps (NIR) and 10 fps (high resolution RGB, stereo) and used custom acquisition routines and frame grabbers to minimize the user interaction with the recording systems. The investigators followed the same guidelines for the personal protective equipment (PPE) as the physicians taking the swab samples. Participants waiting for the experiment were regularly informed about the estimated waiting period and had the option to quit the experiment between two recordings, to reduce the contamination with ambient sound during audio recordings. The experiment began with a set of yes/no/unknown questions with the goal to generate uniform voice samples that can be compared between subjects. These also provided additional medical history. The second segment was a free speech sample, where the participants were asked to tell the circumstances that led to them visiting the DTSC. Subsequently, we asked the participants to briefly present their hands from both sides towards the camera, to catch potential cues to skin rashes. In the final segment, participants were asked to take 10 deep breaths and to breathe normally for 30s afterwards. The entire study/data acquisition took around 6 minutes. A total of 436 participants with signed consent form participated in our study, aged 19–86 (mean age 45.6 ± 15.2 , 215 males, 221 females, 7 participants did not provide their age). 34 subjects reported chronic or acute respiratory diseases or symptoms. Despite a relatively high participation of 36% of the DTSC users in our study, our dataset contained only two subjects with SARS-CoV-2 positive PCR results from the swab tests at the DTSC, see discussion.

2.1.3 Results

We recorded up to 6 minutes of data per subject using the HFR or high-resolution multimodal cameras and microphones. We applied already available procedures from computer vision and computer audition for assessment of the data quality. The evaluations and respective methodology used as proof-of-concept are described below.

Appendix G, Figure G.2 provides an overview of the modality-wise evaluations on the DTSC dataset, in particular how the contactless physiological parameters compare to gender and age specific norm values and groups them by a presence/absence of symptoms. Due to ethics / privacy conditions, evaluations on the full dataset are limited to the local computing infrastructure of Saarland University.

However, we have recordings of individual participants that agreed to have their recordings made publicly available and offer those for download on our project website.

2.1.3.1 System calibration and camera mapping

The camera setup was calibrated with a 4×13 circular calibration pattern glued to a cut-out metal plate. The pattern was heated with two heating blankets prior to acquisition and therefore was visible in the thermal, NIR, and RGB cameras (see Figure 2.1 B.1-B.3). Since the macro lens of the high-speed camera required manual adjustment prior to each recording and had a drastically different field of view, this camera was not calibrated into the system. We applied pairwise stereo calibration with the goal of mapping from the Kinect and from the stereo pair into the thermal camera. For stereo-thermal calibration, the left camera (see Figure 2.1, Stereo II) was used as the origin and for Kinect-thermal, the Kinect was the origin. The 4K camera (Stereo II) was spatially downsampled to the resolution of the low resolution right stereo camera (Stereo I). Mapping into the thermal camera was implemented via the Kinect intrinsic matrix and the depth channel, as well as triangulation and projection from the stereo cameras.

2.1.3.2 Functional thermography

From the thermal recordings, we analyzed static temperatures from the orbital, periorbital, maxillary, and nose region (see Figure G.2 (C) in Appendix G). Vanilla landmark detectors did not perform consistently for all participants on the thermal camera, so we developed a stack of pre-processing methods such as image inversion and unsharp masking to a set of 10 randomly sampled frames for each subject, applied a facial landmark detector [10] to each of the images with different pre-processing and averaged the results per frame. Failed detections were manually annotated.

For the two demo recordings (two different recording days and system calibrations), we compared this approach with projecting landmarks from the Kinect NIR image via the Kinect depth channel as well as with triangulation from landmarks on the stereo RGB frames into the thermal camera frame.

The root mean square error (RMSE) between stereo and thermal was 2.4 and 5.3 pixels and between Kinect and thermal / stereo was a few orders of magnitude larger. The low performance of the Kinect can partially be explained by outliers due to missing depth values, which could be solved by depth map filtering. Hence, the median error is much lower with 2.7 and 4.6 for Kinect and thermal as compared to 1.8 and 2.5 for stereo and thermal. Due to the generally low RMSE between stereo and thermal and good qualitative performance of the thermal detection approach, we used our method in the thermal domain for the static evaluations.

This has the advantage that it does not depend on the calibration of the recording day and is easy to deploy. Our results showed a significant difference in temperature for subjects reporting fever vs. no fever in the maxillary, periorbital and nose region, see [F1] for qualitative results.

2.1.3.3 Discussion

A major limitation of our study is the marginal number of participants with positive PCR test result for SARS-CoV-2. The reason for this was the generally very low incidence rate in the study period in the region where the DTSC was located. In fact, the positive rate was below 0.4% at the DTSC Saarbrücken in the respective period. Thus, the specificity and sensitivity of the described approach with respect to SARS-CoV-2 infections cannot be assessed. However, our proof of concept shows that a remote data acquisition of SARS-CoV-2 infection related symptoms at DTSCs is possible. Our experiences and results enable the installation of similar approaches in regions that do massive DTSC testing.

Additionally, we have recorded an unprecedented, multimodal dataset with a high number of subjects that can be used for the development and refinement of computer vision methods. Many state-of-the-art computer vision studies with isolated modalities have to resort to smaller, publicly available datasets: Considering the development of algorithms for the detection and classification of microexpressions, Li et al. report between 80-210 subjects for common microexpression datasets with evoked microexpressions and they present a dataset with 20 subjects for spontaneous microexpressions, recorded at 100 fps (RGB) and 25 fps (NIR) [32]. Davison et al. record 32 subjects at 200 fps for spontaneous microexpressions [33]. Our dataset contains recordings of 436 participants at 120 fps (RGB), 50 fps (thermal) and 30 fps (NIR) over 6 minutes.

While some subjects moved out of frame during various parts of the experiment for the HFR face close-up recordings, we expect a similar percentage of successful recordings that allow for microexpression annotation as for our HR evaluations and we can additionally report HFR thermal recordings with wider field of view. See Appendix G for additional results for remote physiology.

For a facial landmarking task in the context of functional thermal imaging, datasets of around 2935 frames from 90 subjects with full manual annotation can be considered among the state-of-the-art [34]. Our 436 recordings at 50hz of up to 6 minutes potentially enable the generation of a dataset with 7-8 million frames which would be multiple orders of magnitude larger. Our preliminary results suggest the possibility of partial, automatic annotation with either an appropriate stack of pre-processing methods in the thermal image, projection from Kinect or stereo cameras or a combination of both.

Unlike approaches that rely on homographies for crossmodal mapping (e.g. Palmero et al. [35]), our geometric projection approach guarantees distance invariant performance without parallax errors. Together with tracking approaches in the thermal image, any RGB or gray scale based landmarking or detection algorithm could be used for a full annotation of the dataset in the future.

Computer vision algorithms have different requirements for environmental parameters. In the context of *remote photoplethysmography* rPPG methods, Wang et al. require illumination with constant spectrum to reconstruct *photoplethysmography* PPG signals from videos in talking and static scenarios and of various skin types and then achieve high signal-to-noise ratio of the spectrograms [36]. The employed studio illumination in our setup at the DTSC fulfills those requirements and allowed us to record high-quality data for rPPG measurements.

The redundant camera setup allows for investigations of optimal and minimal sensor configurations of similar setups: With the minimal sensor configuration of one high speed camera with macro lens and a thermal camera, the results in this section can be reproduced. However, the macro lens required manual adjustments for each subject which aggravates calibration. Additionally, the narrow field of view makes it more likely for a participant's face to move out of the frame. Employing a Kinect that is calibrated with the thermal camera is an affordable way to project results from tracking or detection algorithms from the NIR domain to the thermal image. In terms of available algorithms and constancy assumptions, an additional pair of RGB stereo cameras forms the better solution. Using one or two high resolution cameras in the stereo setup could allow for additional analysis of skin and facial parameters.

2.1.4 Conclusion

We have proposed a setup to record multimodal data and have recorded a unique dataset of DTSC users across all age groups. To our knowledge (May 2021), this is the first time that a multimodal video and audio dataset has been recorded at a SARS-CoV-2 DTSC. We have shown in our preliminary evaluations that the data quality is sufficient for the application of already available procedures from computer vision and computer audition. This could allow for SARS-CoV-2 infection related symptoms assessment from such data in the future.

While we captured only a marginal number of PCR positive subjects, our dataset can help with the development and refinement of computer vision algorithms beyond the COVID-19 pandemic: After full annotation for various computer vision tasks such as landmarking or microexpression analysis, it has the potential to rank among the state-of-the-art in terms of the number of participants and age statistics. It has been recorded in an out-of-lab setting with realistic participant interaction, which for example could be encountered at the entrance of an employee parking area.

2.2 Dense landmarks for visible-thermal face analysis

The previous section highlights the importance and diverse applications of functional thermal imaging. Thermal imaging is an important image modality with applications in imaging areas such as life sciences, medicine and biometrics. Facial landmark analysis and detection is essential for extracting facial temperatures and enabling a wide range of applications in security, autonomous driving, entertainment, and healthcare. Despite the availability of various facial RGB datasets, the thermal modality has been largely overlooked. While there exist tailored approaches for the estimation of facial landmarks from thermal images, the experiences from working with the DTSC data have shown that RGB methods can perform reasonably well with the right pre-processing approach. This motivates further exploration of the relationship between RGB and thermal training data for facial landmark detection.

To this end, the RGB and thermal images from the DTSC dataset from the previous section are aligned using camera geometry-based dense alignment with the goal to train an RGB-to-thermal image translator. With different thermalization approaches, the publicly available, large-scale, Face Synthetics dataset (FAKE) is thermalized. We refer to the resulting thermalized FAKE variants as T-FAKE-P1 and T-FAKE. T-FAKE-P1 is the prototype dataset that, while not containing realistic thermal images, demonstrates that training RGB landmarking models with low quality gray scale images boosts thermal landmark performance beyond baselines and informed the development of the RGB2Thermal loss. T-FAKE-P1 is generated with an initial supervised Pix2Pix-style thermalizer trained on paired RGB-thermal data including data from the previous section and T-FAKE is the final dataset with simulated warm/cold conditions generated with our domain-adaptive thermalization model trained on SEJONG using the RGB2Thermal loss, see section 2.2.2.1 and Figure 2.3.

Leveraging the thermalized datasets, probabilistic landmark prediction, and label adaptation networks, we demonstrate significant improvements in landmark detection methods on thermal images across different landmark conventions on a difficult benchmark. The models show excellent performance with both sparse 70-point landmarks and dense 478-point landmark annotations.

An in-depth description of the thermalization approach can be found in [F6]. This section contains direct excerpts from [F6]².

²The latest version of the code and models is available at <https://github.com/phflot/tfake> and as the `thermal-face-alignment` package on pip.



Figure 2.2: Predictions of 70 (middle) and 478 (bottom) point facial landmarks on the challenging CHARLOTTE dataset [37] which are estimated accurately by training a model with our synthetic T-FAKE dataset. © 2025 IEEE.

2.2.1 Introduction

Functional thermal imaging is an acquisition modality with growing importance in many areas from life sciences over medicine to biometrics. Due to the close relation of body temperature with physiological states, thermal imaging of faces emerged as an important tool for screening infectious diseases [25, 38, 39] and for the analysis of neurocognitive and affective states [40]. As a consequence, there exists a growing need for thermal images with annotated facial landmarks. However, in contrast to RGB images, annotated data for thermal images of faces is still rarely available. For example, the largest thermal datasets with facial landmark annotations contain only tens to hundreds of individuals [34, F1, 41, 37], while the largest datasets for face detection for RGB images reached the mark of a million faces (of more than 500k individuals) with the MegaFace benchmark already in 2016 [42]. Moreover, large synthetic RGB datasets, like FAKE [43], appeared to be valuable for sparse and dense landmark detection training.

Generally, the detection of facial landmarks is an important pre-processing task for further facial analysis. These 2D coordinates of corresponding points across faces are part of methods, e.g., for alignment of faces to extract regions of interest [44], the animation of 3D avatars [45], motion magnification [F12, F5] or deepfakes [46]. The most common convention for facial landmark annotations uses 68 points across the face and face boundary, see, e.g. [47]. Even though the 68-point convention remains challenging for thermal images [48, 49], there is additional demand for dense landmarkers with more than 300 points [43, 44, 45].

Faces in thermal images share structural similarities with RGB images. However, temporally varying environmental factors such as temperature and rain droplets, and physiological factors such as blood perfusion changes and vascular details lead to much higher variability in the appearance of faces within individuals [50]. Sweating, flushing, or outdoor rain can result in time-dependent texture changes. Features that are very stable in RGB images such as the contrast of the nose and the eyes can become inverted when the nose or the face is cooled down. Due to the structural similarities, landmarks trained on RGB datasets were also used under specific thermal imaging conditions [F1]. However, they show high failure rates and produce imprecise landmark predictions on challenging thermal images. Up to now, there exist only a few sparse thermal landmark detection methods [34, 49, 41, 51, 52], and we are unaware of dense thermal landmarking approaches.

2.2.1.1 Related work

Facial RGB landmark prediction. Different annotation conventions exist for labeled landmarking datasets. An important distinction is between 2D annotations and 3DA-2D annotations which project 3D landmarks onto images [53]. 2D is closer to human perception, but 3DA-2D enables annotation of occluded face regions. The most widely used annotation convention defines 68-point 2D landmarks available for many facial alignment datasets [54, 55, 56, 10]. Sometimes points on the eye or other details are additionally added [43, 37]. While an in-depth review of facial RGB landmarking is beyond the scope of this work, we introduce those that are relevant to thermal landmarking methods. Kazemi et al. [57] introduced a popular traditional computer vision method for landmark prediction via the dlib library [58]. Kowalski et al. [59] proposed the deep alignment network (DAN) which uses convolutional neural networks and iteratively refines predictions. Around the same time, Bulat et al. [10] proposed the facial alignment network (FA). Both methods predict heat maps of landmark positions. Explicit integration of uncertainty quantification into a landmark prediction model via the LUVLi loss has been proposed by Kumar et al. [60]. Recently, image annotation quality became a limiting factor which led to the emergence of synthetic data with high-quality ground truth [61]. Wood et al. [43, 44] trained 3DA-2D landmarks with excellent performance using their proposed synthetic FAKE dataset with 100k frames of at least 2k individuals.

Dataset name	Year	DS avail.	Model avail.	Landmarks	Size (F/S)	Distance (cm)
Aachen[34]	2018	✓	✓	68	2935/90	90
ARL-MMFD [62]	2019	(request)	✗	6	888/111	250, 500, 750
DTSC (in-house) [F1]	2021	✗	✗	Mapping	~7-8M/436	~80-150
ARL-VTF [41]	2021	✗	✗	68	500k/395	210
Car (in-house) [F2]	2022	✗	✗	Mapping	?/24	~80-150
TFW [52]	2022	✓	✓	5	9982/147	100-?
SF-TL54 [51]	2022	✓	✓	54	2556/142	100
CHARLOTTE [37]	2022	✓	✗	73 (43 profile)	7684*/10	100-660
VI-Screen (in-house, unpublished)	2024	✗	✗	Mapping	?/ 191	~80-150
Hand (in-house) [F8]	2024	✗	✗	Mapping	?/25	~80-150
T-FAKE-P1	2024	✓	✓	70 and 478	100k/>2k	~80-150
T-FAKE	2024	✓	✓	70 and 478	200k/>2k	~80-150

Table 2.1: Comparison of thermal landmark datasets. In-house datasets recorded with fully calibrated camera setups similar to the one described in the previous section are marked in bold. For those datasets, geometric mapping is available and RGB landmarks can be projected onto the thermal images. Number of frames (F) and number of subjects (S) are indicated as well as distance to the camera if known. *Note that for CHARLOTTE, we report the total number of available 16bit image - annotation pairs in the final dataset rather than the reported number of acquired frames.

Thermal landmark prediction. Many tailored approaches for thermal landmarking use similar architectures as RGB landmarkers and are mainly developed alongside the publication of new datasets, see Table 2.1. Kuzdeuov et al. [52] released the TFW dataset with 5 facial landmarks and bounding boxes and trained YOLOv5 networks for landmark prediction. The SF-TL54 dataset [51] has been used for 2D landmark prediction with a dlib and a U-net model. Kopaczka et al. [34] also use a dlib model for their AACHEN dataset with 68-point 2D landmarks. A highly structured 2D landmarking dataset is the CHARLOTTE thermal face dataset with controlled temperature changes by Ashrafi et al. [37]. It contains around 10k images in total and 7684 images with annotations. In contrast to these publicly available datasets, datasets with limited or no public availability include the ARL-VTF dataset [41] and the in-house dataset from the previous section recorded in SARS-CoV-2 drive-in stations (here abbreviated with DTSC) [F1] which contains six minutes recordings at 50Hz of 436 subjects. In the previous section, we found that a stack of pre-processing functions already leads to a good performance of existing RGB landmark models on thermal images. Similarly, Poster et al. [63] proposed to use transfer learning on a siamese DAN to improve thermal landmark prediction with RGB facial landmarks. Mallat et al. [49] used RGB2Thermal image translation (‘thermalization’), active appearance models, and deep alignment networks of RGB datasets for thermal landmark prediction. In contrast to the previous work, our dataset contains thermal images of at least 2k individuals that are annotated with sparse 70 3DA-2D as well as dense 478 2D point landmarks.

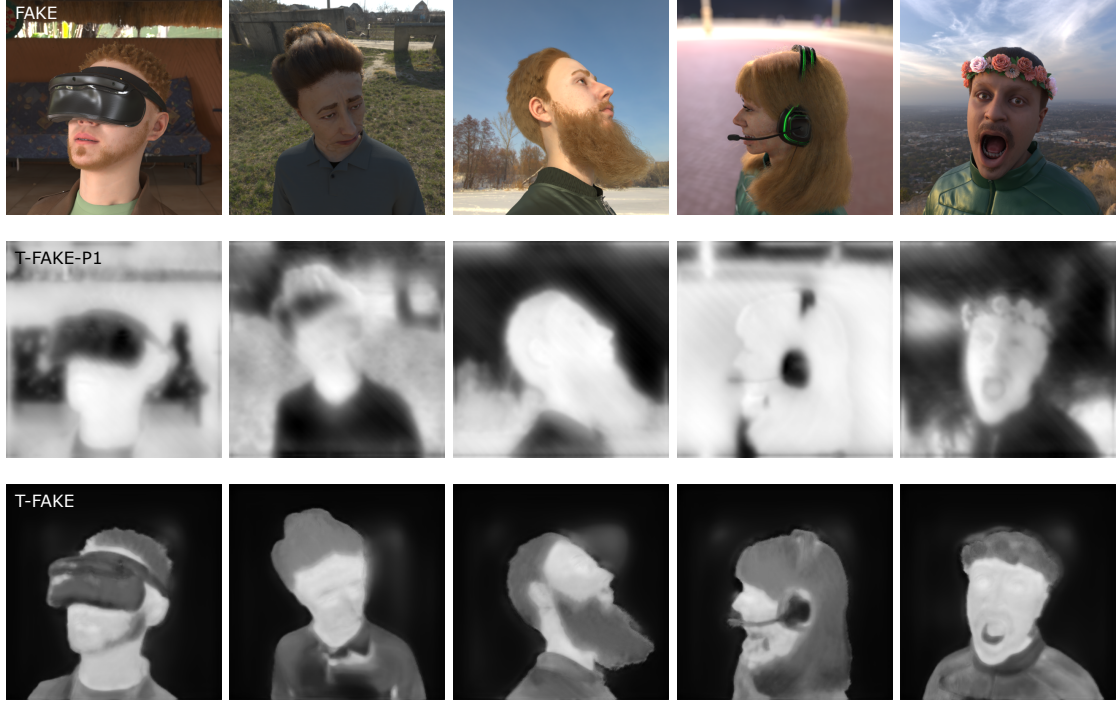


Figure 2.3: Frames from the FAKE dataset [43] (top), early prototype T-FAKE-P1 (middle) and T-FAKE (bottom). The high variability in the FAKE dataset resulted in poor generalization of the supervised approach (middle). To address this, T-FAKE-P1 images are created by averaging multiple predictions to mitigate artifacts. T-FAKE employs a novel regularization technique to enforce thermal image statistics on FAKE images [F6].

Thermalization. In addition to thermal landmark datasets, various datasets of paired and aligned thermal and RGB images without annotations exist. These can be employed for the training of thermalization models. In contrast to the rich literature on Thermal2RGB synthesis [64, 65, 66], the literature on RGB2Thermal image synthesis using such datasets is more limited. Mallat et al. thermalized facial images by minimizing a perceptual loss [49] using the VIS-TH dataset [67]. Similarly, generative adversarial networks (GANs) have been used for Thermal2RGB synthesis [62, 50] using the NCKU-VTF [50] dataset and the ARL-MMFD [62] dataset containing facial image data. In contrast to the limited public availability of these two datasets, a publicly available dataset with paired facial images is the multimodal disguise *Sejong face database* (SEJONG) with 100 subjects, 15 disguise variations and 4849 RGB+Thermal image pairs by Cheema and Moon [68]. Note that all these datasets and resulting models are mostly restricted to frontal views of faces with controlled lighting. Thermalization of general scenes has been achieved by Kniaz et al. [69] with a modified Pix2Pix GAN, where they used the publicly available ThermalWorld dataset with 5098 triplets of RGB images, thermal images, and segmentation masks.

2.2.2 Materials and methods

2.2.2.1 T-FAKE-P1 and T-FAKE datasets

To generate a synthetic, large-scale dataset for thermal landmark prediction, we translate the FAKE dataset [43] to the thermal domain (T-FAKE). The FAKE dataset contains 100k synthetic RGB images of simulated human faces with ground truth for facial landmarks and segmentation. In contrast to thermal landmarking datasets with a fixed distance between the subjects and the camera [34, 41, 51], the distance varies within an estimated range of 80cm to 150cm for the T-FAKE dataset. Moreover, the synthesized images are based on facial scans with a wide range of age and ethnicity and a balanced number of people identifying as male or female [43]. In addition to the original ground truth sparse 3DA-2D landmark annotations, we add dense 2D annotations by transferring MediaPipe [45] predictions for the FAKE to the T-FAKE dataset. This makes our dataset the first densely annotated thermal landmarking dataset. While we found many large thermal landmarking datasets to be non-public [62, F1, 41], our complete dataset with both thermal conditions and annotations is available for download (see Table 2.1).

The first prototype of a thermalized FAKE dataset is obtained with a pixel-to-pixel translation *generative adversarial network* (GAN) [70] that is trained on the DTSC multimodal dataset from the previous section and the SEJONG dataset. For that purpose, the RGB and thermal images from the DTSC dataset were aligned through i.) calibration of each recording with the corresponding calibration recording, ii.) disparity estimation and iii.) projection of the RGB frames onto the thermal image frames via projective camera geometry. For disparity estimation, we used dense inverse search optimization (DIS) OF [71]. While more recent correspondence methods may improve the disparity estimates, in our DTSC recordings the dominant limitation was calibration quality due to the frequent need for re-calibration. Moreover, our downstream use is restricted to the facial area and DIS OF produced accurate RGB-thermal alignment on faces. Of the 436 subjects, the calibration of 294 recordings proved sufficient for dense mapping between RGB and thermal domain. From those recordings, we constructed a dataset with 248707 frames with dense mapping. Both the SEJONG and the DTSC dataset contain only limited poses and limited variation in scene composition and backgrounds. First experiments of the trained network revealed poor generalization to the target domain. To mitigate those effects, the first prototype T-FAKE-P1 was obtained by averaging multiple predictions which, however, results in blurry images with poor thermal image features (see Figure 2.3, middle).

In [F6], we follow two approaches for better generalization on the FAKE dataset. Firstly, we use the segmentation masks from the FAKE dataset during training. We enforce average temperatures on segmented image areas to lie in the range reported in the literature [37]. Here, we use both a warm and a cold condition following the intuition that different temperatures in the face have huge impact on thermal facial representation. Secondly, we introduce a patch-based loss function using the Wasserstein distance [72, 73]. This patch-based loss is similar to the discriminator in pixel-to-pixel translation network, but it is applied to images in the target domain and ensures that the patch distributions in thermalized FAKE images and real thermal images are more similar.

For the final thermalization model, we use a U-Net T_θ with a Resnet-34 encoder pre-trained on ImageNet [74] and train it for 10 FAKE epochs with SEJONG batches of size 64 and FAKE batches of size 64. This corresponds to approximately 100 SEJONG epochs. Further, we use an Adam optimizer with an initial learning rate of 0.001 which we reduce to 0.0001 after 4 epochs. Based on a random split, we use 80% of the SEJONG data for training and all available FAKE data. The final T-FAKE dataset is obtained from applying T_θ to the complete FAKE dataset. For each thermal condition, the dataset contains 100k images featuring at least 2k unique subjects. Each image is available in a simulated ‘cold’ and a ‘warm’ environment adding up to 200k images. This number of frames and subjects surpasses the numbers in other thermal landmarking datasets [34, 62, F1, 41, 52, 51, 37] (see Table 2.1).

We train an RGB2Thermal network T_θ by minimizing

$$\mathcal{L}(\theta) := \frac{1}{N} \sum_{n=1}^N \|T_\theta(X_{\text{RGB}}^n) - X_{\text{T}}^n\|_2^2 \quad (2.1)$$

$$+ \lambda_W W(T_\theta(\mathbf{X}_{\text{FAKE}}), \mathbf{X}_{\text{T}}) \quad (2.2)$$

$$+ \lambda_R \frac{1}{M} \sum_{m=1}^M R(T_\theta(X_{\text{FAKE}}^m)), \quad \lambda_W, \lambda_R > 0. \quad (2.3)$$

Here, W matches multiscale patch statistics of generated and real thermal images using an entropy-regularized optimal transport (Sinkhorn) distance, and R enforces region-wise mean temperatures using the available FAKE segmentations (warm/cold priors) [F6]. We denote by $(X_{\text{RGB}}^n, X_{\text{T}}^n)_{n=1}^N$ the aligned RGB/thermal training pairs (with thermal ground truth X_{T}^n) and by $(X_{\text{FAKE}}^m)_{m=1}^M$ the unpaired FAKE RGB images (with their provided segmentation masks), and write $\mathbf{X}_{\text{RGB}}, \mathbf{X}_{\text{T}}, \mathbf{X}_{\text{FAKE}}$ for the corresponding collections. The weights $\lambda_W, \lambda_R > 0$ control the influence of the patch-matching term W and the segmentation-based temperature regularizer R , respectively.

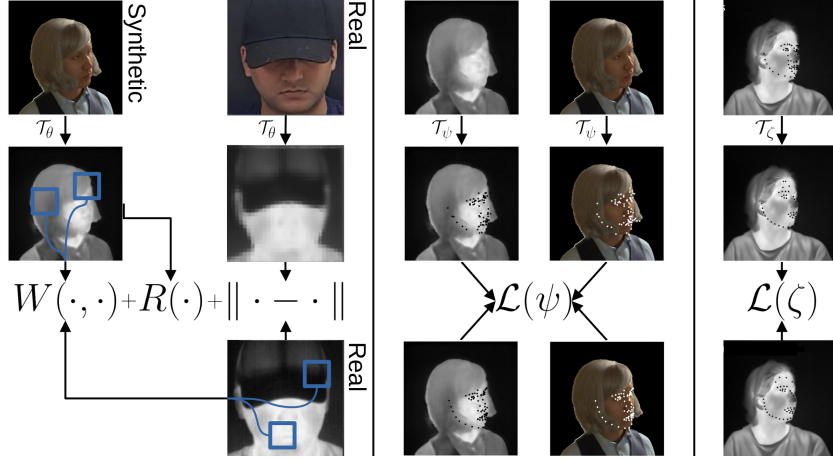


Figure 2.4: Three steps of our training pipeline (models and their loss functions): 1. thermalization network $\mathcal{T}_\theta$ with loss (2.1) - (2.3). 2. landmark prediction network $\mathcal{T}_\psi$ with (2.5). 3. landmark adaptation network $\mathcal{T}_\zeta$ with (2.6). We show black dots for thermal landmarks and white dots for RGB landmarks. Note the landmark adaptation on the forehead in the last step. Figure originally published in [F6]. © 2025 IEEE.

Conceptually, W measures how well the generated thermal images match the patch statistics of real thermal images across scales. We refer to [F6] for the full definition of W and implementation details.

Having trained $\mathcal{T}_\theta$, we do inference on the FAKE to create the final T-FAKE dataset. For each thermal condition, the dataset contains 100k images featuring at least 2k unique subjects. Each image is available in a simulated ‘cold’ and a ‘warm’ environment adding up to 200k images. This number of frames and subjects surpasses the numbers in other thermal landmarking datasets [34, 62, F1, 41, 52, 51, 37] known to the authors.

In contrast to thermal landmarking datasets with a fixed distance between the subjects and the camera [34, 41, 51], the distance varies within an estimated range of 80cm to 150cm for the T-FAKE dataset. Moreover, the synthesized images are based on facial scans with a wide range of age and ethnicity and a balanced number of people identifying as male or female [43]. In addition to the original ground truth sparse 3D-2D landmark annotations, we add dense 2D annotations by transferring MediaPipe [45] predictions for the FAKE to the T-FAKE dataset. This makes our dataset the first densely annotated thermal landmarking dataset. While we found many large thermal landmarking datasets to be non-public [62, F1, 41], our complete dataset with both thermal conditions and annotations is available for download. In combination with the original FAKE dataset, we can train multimodal facial landmarks jointly on RGB and thermal images as described in the next section. Table 2.1 displays a comparison between thermal landmarking datasets.

2.2.2.2 Facial landmarking

Learning with synthetic images has proven to be very useful in facial landmarking [62, 43, 44]. Based on the proposed T-FAKE and the original FAKE dataset, we can train a supervised thermal or multimodal landmark prediction model for thermal and RGB images using synthetic data only. Moreover, we adapt a model to new annotation conventions by training a small-scale label adaptation model.

Landmark Prediction. We apply the construction in [44] and learn a landmark prediction model through probabilistic landmark regression from given paired data $(X^j, \mathbf{y}_j)_{j=1}^J$, where $X^j \in \text{FAKE} \cup \text{T-FAKE}$ and $\mathbf{y}_j := (y_l^j)_{l=1}^L$, $y_l^j \in \mathbb{R}^2$ are the corresponding L landmarks. We learn a network T_ψ mapping from the space of images to the space of L -fold two-dimensional Gaussian distributions $(\mathcal{N}(\mu_l, \sigma_l I_2))_{l=1}^L$. Here we employ a MobileNet V2 architecture for landmark prediction [75]. Recalling that the Gaussian distribution $\mathcal{N}(\mu, \sigma I_2)$ has the negative log-likelihood function

$$-\log p(y|\mu, \sigma) = \log(2\pi\sigma^2) + \frac{\|\mu - y\|_2^2}{2\sigma^2}, \quad (2.4)$$

the authors of [44] proposed to learn the network $T_\psi := (T_\psi^1, T_\psi^2)$ with $X \mapsto (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu} = (\mu_l)_{l=1}^L \in (\mathbb{R}^2)^L$ and $\boldsymbol{\sigma}^2 = (\sigma_l^2)_{l=1}^L \in \mathbb{R}^L$ by minimizing the loss function

$$\mathcal{L}(\psi) := \sum_{l=1}^L \left(\sum_{j=1}^J \log T_{\psi,l}^2(X^j) + \frac{\|T_{\psi,l}^1(X^j) - y_l^j\|_2^2}{2T_{\psi,l}^2(X^j)} \right). \quad (2.5)$$

During training, we clip $T_{\psi,l}^2 \geq \varepsilon > 0$ to avoid numerical instability due to too small variances. The first sum controls the location accuracy and the second one the uncertainty prediction. Having learned the network T_ψ , it produces not only the expected values $\boldsymbol{\mu}$ of the corresponding landmarks, but also a measure for its uncertainty $\boldsymbol{\sigma}^2$.

Label Adaptation. A common approach that is especially crucial for synthetic trained data and comparison of 2D and 3DA-2D landmarks is label adaptation [43, 44]. It allows us to adapt to different landmark conventions between the training dataset and an evaluation dataset. Given two sequences $(\hat{\mathbf{y}}_k, \mathbf{y}_k)_{k=1}^K$ with predicted landmarks $\hat{\mathbf{y}}_k := (\hat{y}_k^l)_{l=1}^{\hat{L}} \in \mathbb{R}^{2\hat{L}}$, and with ground truth landmarks $\mathbf{y}_k := (y_k^l)_{l=1}^L \in \mathbb{R}^{2L}$ with possibly $\hat{L} \neq L$, we aim to train a third label adaptation model $T_\zeta : \mathbb{R}^{2\hat{L}} \rightarrow \mathbb{R}^{2L}$. Here, we use the loss

$$\mathcal{L}(\zeta) = \frac{1}{K} \sum_{k=1}^K \|T_\zeta(\hat{\mathbf{y}}_k) - \mathbf{y}_k\|_1. \quad (2.6)$$

Since we deal with low-dimensional data, we use a simple multi-layer fully connected perceptron with 5 layers for this task.

2.2.3 Results

In this section, we present details of our final implementation and our benchmarking pipeline. Note that we clamp the temperature of all thermal images between 20 °C and 40 °C for all experiments and visualizations.

2.2.3.1 Implementation details

Landmark prediction. We simultaneously train our landmark prediction model using a sparse 70-point 3DA-2D and dense 478-point 2D landmark convention with an image size of 224×224 on our new T-FAKE thermal dataset and the original FAKE RGB dataset. Due to the limited, public availability of the dense 3DA-2D ground truth in the FAKE dataset [43], we use the 478-point 2D predictions of the MediaPipe model for the synthetic RGB images as our dense ground truth annotations.

On thermal images, we fill in the background with random textures from the texture dataset [76] with a probability of 0.25. Furthermore, we include random landmark positions on images from the same dataset as negative examples into the dataset to increase the learned uncertainty σ^2 on images without faces. During inference, we use a multiscale sliding window evaluation to generalize our model to varying image sizes and face scales. We downsample iteratively with a factor of 0.75 until the height or the width reaches 224. For each image scale, we run our model on sliding windows of size 224×224 with a stride of 20. For our final landmark prediction, we pool all predictions and use the landmark with the smallest predicted standard deviation across all scales and all sliding windows. For training the landmarker T_ψ , we fine tune a model that has been pre-trained on the original FAKE dataset for the sparse landmarker for the results in Table 2.4. During refinement, thermal images are used with a probability of 0.4 split with equal probability for ‘cold’ and ‘warm’ conditions. We fine tune for 100 epochs with a learning rate of 0.0004, Adam optimizer with weight decay, batch size of 512 and OneCycleLR scheduler. Our final model is trained for 4000 epochs with a learning rate of 0.001 on FAKE and T-FAKE (p=0.4).

Augmentation details. We make use of spatial augmentations, including random shear, rotations, resizing, and cropping to allow the landmarker to learn a large variation of face orientations without the need for a dedicated face detector.

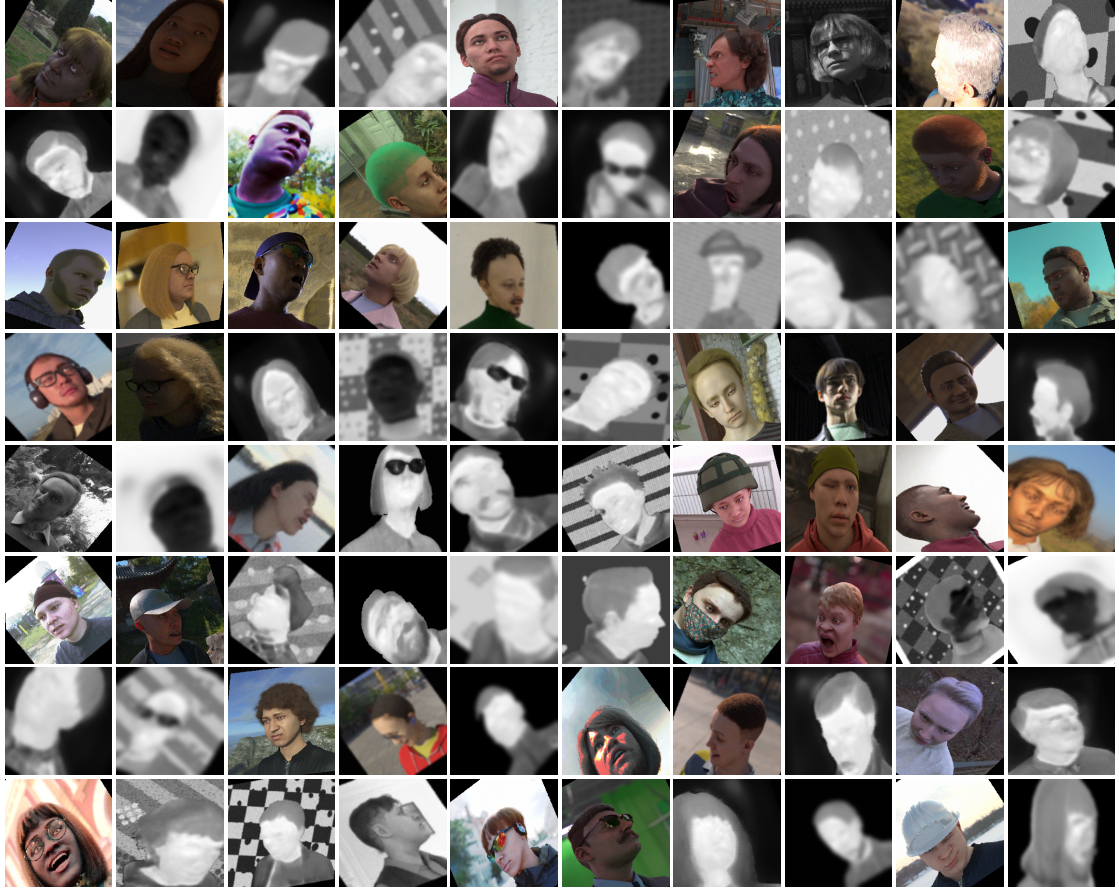


Figure 2.5: Example images from FAKE and T-FAKE with the augmentations applied during training. © 2025 IEEE.

Specifically, we use geometric augmentations which apply random rotations (up to $\pm 45^\circ$) and random shear ($\pm 7^\circ$) to both the image and landmarks, preserving geometric consistency. In addition, a random resized cropping operation is performed, with the cropped region size scaling between 40% and 200% of the original image and an aspect ratio ranging between $\frac{3}{4}$ and $\frac{4}{3}$. Furthermore, we use photometric transformations and random Gaussian smoothing. The thermal images are randomly jittered to simulate temperature variations and thermal images are randomly inverted with a probability of 0.1. Additionally, random noise is applied with a probability of 0.2 to simulate sensor noise. See Figure 2.5 for image examples with augmentations applied.

RGB Model Inference. We use two different pre-processing approaches to include landmarks solely developed for RGB images into the evaluation. Firstly, we use gray-value images, where the temperature is normalized between 20° and 45° and, secondly, we apply the pre-processing stack proposed in the previous section [F1].

The pre-processing stack consists of temperature clamping between 20°C and 45°C, unsharp masking with two sets of parameters with and without temperature inversion. The reported landmarks are the averages over all detected faces. This simple pre-processing stack is a simple method for boosting RGB landmarker performance for thermal images [F1]. As a result, we can include RGB landmarks as a baseline for thermal landmarking models.

Label Adaptation. Our label adaptation network is trained solely on the two-dimensional landmark predictions and ground truth annotations of our random training split of a benchmark dataset and not on the images. This dataset is later excluded from the final evaluation. We do not use the predicted standard deviation at this stage. Again, we use random rotations and shear transformations during training. Label Adaptation was trained for each method on the predictions on all detected faces on a random 1000 image CHARLOTTE split. We train a model T_ζ for 2000 epochs with a learning rate of 0.002, OneCycleLR and Adam optimizer. As landmark augmentations, we apply random rotation up to 45° as well as random shearing during training. The label adaptation network is a five-layer perceptron with fully connected layers that takes the predicted landmarks together with the resize factor as input. The latter accounts for varying degrees of quantization at different image sizes in the CHARLOTTE ground truth. The label adaptation generally handles even outlier predictions but can also contain fail cases (see Figure 2.6).

Evaluation Metrics for Landmarking. We report the *normalized mean error* (NME) [43, 47, 77] for evaluation, i.e., the mean absolute error (MAE) normalized with an image-dependent constant to account for varying image sizes. Often, normalization is performed by the distance between the outermost eye landmarks (NME IO) [43, 47]. However, the CHARLOTTE benchmark dataset contains many examples with side profile views where only one of the eyes is visible. Therefore, we use the average height and width of the bounding box around the face for normalization [77] (NME W/H). Apart from the NME, we report the failure rates for each of the reference methods. The failure rate is the percentage of frames where no faces could be found. Frames without recognized faces are not included nor weighted in the NMEs. The method of rejection varies for each landmarker. We use vanilla parameters except for MediaPipe [45], where we set the minimal detection confidence to 0.01. For our method, we evaluate it i) on all frames without uncertainty rejection and ii) on all frames where the average uncertainty score σ_l of the predicted landmarks l , $1 \leq l \leq L$, is less than a fixed value $\bar{\sigma}$.



Figure 2.6: Label adaptation examples. For side profile views in the CHARLOTTE dataset only landmarks for visible parts of the face exist and additional positions on the forehead are marked (top). Label adaptation translates three of the different landmark conventions we use for evaluation (bottom) into the CHARLOTTE convention (middle). © 2025 IEEE.

Method	Year	Landmarks	Thermal Images
FA [10]	2017	68	✗
Aachen [78]	2018	68	✓
3FabRec [79]	2020	68	✗
Pre-processing stack [F1]	2021	—	✓
MediaPipe [45]	2020	478	✗
TFW [52]	2022	5	✓
SF-TL54 [51]	2022	54	✓
FAKE Baseline [44]	2022	70	✓
Star [81]	2023	68	✓

Table 2.2: Overview of the landmarking methods we include in the benchmark sorted by year.

Metric	Method	Training Dataset	High	Low	Side	Front	Full
NME W/H ↓	3FabRec [79]	300W [†]	0.0931	0.1084	0.0675	0.1344	0.1009
	3FabRec*	300W [†]	0.0814	0.1558	0.0791	0.1602	0.1196
	Star [81]	300W [†]	0.0707	0.1021	0.0673	0.1064	0.0868
	Star*	300W [†]	0.0639	0.1273	0.0703	0.1227	0.0965
	MediaPipe [45]	<i>undisclosed</i> [‡]	0.1900	0.2745	0.2879	0.1865	0.2262
	MediaPipe*	<i>undisclosed</i> [‡]	0.1400	0.2621	0.1838	0.2164	0.2023
	FA [10]	LS3D-W [†]	0.1070	0.2790	0.1222	0.2508	0.1875
	FA*	LS3D-W [†]	0.1007	0.2207	0.1075	0.2123	0.1593
	FAKE[43, 44]	FAKE [†]	0.1055	0.2675	0.1241	0.2534	0.1887
	FAKE*	FAKE [†]	0.0933	0.2682	0.1312	0.2348	0.1824
	Aachen [78]	AACHEN	0.1677	0.2405	0.2942	0.1731	0.2054
	TFW [52]	TFW	0.0787	0.1249	0.0701	0.1311	0.1012
	SF-TL54 U-Net [51]	SF-TL54	0.2476	0.3134	0.1757	0.3596	0.2829
	SF-TL54 Dlib [51]	SF-TL54	0.2159	0.2174	0.1365	0.2741	0.2167
	P1 dense ($\bar{\sigma} < \infty$)	FAKE [†] + T-FAKE-P1	0.0916	0.1658	0.0974	0.1621	0.1297
	P1 ($\bar{\sigma} < \infty$)	FAKE [†] + T-FAKE-P1	0.0759	0.1479	0.0731	0.1527	0.1129
	P1 ($\bar{\sigma} < 15 \times 10^{-4}$)	FAKE [†] + T-FAKE-P1	0.0751	0.1152	0.0648	0.1220	0.0939
	Ours dense ($\bar{\sigma} < \infty$) [F6]	FAKE [†] + T-FAKE	0.0854	0.1395	0.0764	0.1500	0.1132
	Ours ($\bar{\sigma} < \infty$) [F6]	FAKE [†] + T-FAKE	0.0683	0.1239	0.0645	0.1294	0.0969
	Ours ($\bar{\sigma} < 6 \times 10^{-4}$) [F6]	FAKE [†] + T-FAKE	0.0671	0.1118	0.0593	0.1179	0.0889
	Ours ($\bar{\sigma} < 2.2 \times 10^{-4}$) [F6]	FAKE [†] + T-FAKE	0.0661	0.1018	0.0551	0.1059	0.0815
Failure rate (%) ↓	Aachen [34]	AACHEN	49.32	48.72	72.85	25.15	49.01
	MediaPipe [45]	<i>undisclosed</i> [‡]	25.05	46.85	50.09	22.40	36.25
	MediaPipe*	<i>undisclosed</i> [‡]	12.80	14.07	24.97	1.92	13.45
	FA [10]	LS3D-W [†]	12.95	27.43	21.71	19.07	20.39
	FA*	LS3D-W [†]	0.83	10.45	4.75	6.80	5.77
	TFW [52]	TFW	13.14	21.55	19.26	15.66	17.46
	SF-TL54 U-NET / Dlib [51]	SF-TL54	29.29	22.57	38.16	13.50	25.84
	Ours ($\bar{\sigma} < 6 \times 10^{-4}$)	FAKE [†] + T-FAKE	0.98	10.34	6.94	4.64	5.79
	Ours ($\bar{\sigma} < 2.2 \times 10^{-4}$)	FAKE [†] + T-FAKE	6.52	32.76	23.33	16.68	20.00
	P1 ($\bar{\sigma} < 15 \times 10^{-4}$)	FAKE [†] + T-FAKE-P1	0.18	17.04	10.41	7.28	8.84

Table 2.3: Results on CHARLOTTE dataset splits. Pre-processing with the pre-processing stack for RGB landmarks is indicated by *. Best NME value is in bold. We denote the confidence threshold with $\bar{\sigma}$. We include all predictions for $\bar{\sigma} < \infty$, Star and 3FabRec resulting in an optimal failure rate of 0 by design. Methods in gray used additional supervision from bounding boxes generated with the ground truth landmarks. RGB datasets are indicated by [†].

Metric	Method	Training Dataset	High	Low	Side	Front	Full
NME W/H ↓	FAKE [43, 44]	FAKE	0.1055	0.2675	0.1241	0.2534	0.1887
	FAKE*	FAKE	0.0933	0.2682	0.1312	0.2348	0.1824
	Thermal Only FT	T-FAKE	0.0832	0.1334	0.0677	0.1503	0.1090
	RGB + Thermal FT	FAKE + T-FAKE	0.0740	0.1346	0.0684	0.1420	0.1051

Table 2.4: Ablation results on CHARLOTTE dataset splits. Pre-processing with the pre-processing stack for RGB landmarks is indicated by *. The confidence threshold has been set to infinity. Note that RGB + Thermal and Thermal only models have been fine tuned (FT) from the FAKE models.

2.2.3.2 Benchmark datasets for landmark detection

To compare thermal performance, we use the CHARLOTTE dataset due to its high variability in resolution, pose, and background temperature. The dataset is annotated with either 73 2D landmarks for frontal images or 43 2D landmarks for lateral images. The 43-point landmarks represent the visible landmarks in profile pictures with our information on orientation. We use a total of 7684 annotated frames with different resolutions with a minimum resolution of around 70x60 from the dataset. For evaluation, we propose to split the CHARLOTTE dataset based on image size and perspective. The High condition contains all images larger than the median image width of 200 pixels and Low contains all smaller images. The condition Side includes all images with profile annotations (43 landmarks), and Front contains all frontal faces (73 landmarks). In addition, we evaluate on the established RGB landmark dataset 300W [47] to verify RGB landmark prediction performance. It uses the classical 68-point convention. Here, we crop images to the pre-defined bounding boxes around each face for our evaluation. We follow the protocol from [81] for 300W evaluation. The average image resolution is about 290 pixels.

2.2.3.3 Comparison to the state-of-the-art

We compare the landmarker performance with training on our full synthetic datasets against state-of-the-art pre-trained models on the CHARLOTTE dataset (see Table 2.2 for considered methods). Besides relevant thermal landmarkers we include RGB landmarkers. Performance of RGB landmarkers on thermal images can be boosted by applying a stack of unlearned linear filters before prediction [F1]. Consequently, application of this pre-processing stack allows us to include RGB landmarkers in our benchmark and to draw a fair comparison beyond task-specific thermal landmarkers. We use label adaptation for all methods to account for different landmark conventions for the image evaluation.

We train each label adaptation model with a 1000-image training split on CHARLOTTE similar to [43] and exclude the training split from the final evaluation. Note that this approach allows us to include our sparse model and our dense model in the evaluation. To benchmark the RGB performance of the multimodal model, we evaluate our sparse model predictions against state-of-the-art RGB sparse landmarkers on the 300W dataset.

CHARLOTTE performance. The NME performance as well as the failure rates of the different methods vary drastically for the different CHARLOTTE splits.

We list the dataset on which each of the methods has been trained in Table 2.3. We observe the lowest performance across the landmarks on the Front condition, followed by the low-resolution images. The FA and MediaPipe RGB methods show a performance that is comparable with the SF-TL54 and Aachen thermal landmarks. Using a pre-processing stack [F1], they consistently outperform them on the High condition in terms of NME and on Full, Front, Low and Side in terms of failure rate. Our model achieves state-of-the-art performance for sparse thermal landmarking across all different conditions. The only model with comparable performance is TFW which, however, only predicts a set of 5 landmarks. While the NME of TFW is slightly lower than our method for $\bar{\sigma} < \infty$ for Low, Front, and Full, our method outperforms TFW on all dataset splits in terms of NME and failure rate for $\bar{\sigma} < 6 \times 10^{-4}$. Interestingly, the performance of the model trained on the T-FAKE-P1 prototype still outperforms the model trained on the original FAKE dataset, in particular, on the low resolution condition. These results are presented in Table 2.3 and examples are visualized in Figure 2.7. The 300W-trained RGB methods Star and 3FabRec are close to our method and TFW, however, they have been supplied with the ground truth landmarks for tight bounding box computation. This gives an advantage for label adaptation of broken profile images as well as evaluation on broken and shifted ground truth in CHARLOTTE.

300W performance. Additionally, we evaluate the performance of our multimodal model on the 300W dataset in Table 2.6. Our model, trained on a combination of thermal and RGB facial images with label adaptation performs between MediaPipe and FA with label adaptation and all methods are outperformed by Star. It is important to note that the bounding box has only been optimized for Star and our model has been trained without centered faces and a large variability of head positions in the input frames.

Landmarking ablation study. To study the impact of our thermal data, we report the CHARLOTTE results of our landmarker trained with i) RGB images only, ii) fine tuned with thermal images and iii) fine tuned with both FAKE and T-FAKE. Again, we use label adaptation for all variations. Table 2.4 shows the results.

Training with the T-FAKE dataset significantly improves the accuracy of thermal landmarking across all conditions. In addition, multimodal training with the FAKE and T-FAKE datasets leads to better thermal landmarking performance than fine tuning only with the T-FAKE dataset.

Method	Inference Time (ms)	High	Low	Side	Front	Full
Sliding window ($\bar{\sigma} < \infty$)	88.73	0.0740	0.1346	0.0684	0.1420	0.1051
Refined bounding box ($\bar{\sigma} < \infty$)	10.41	0.0847	0.1329	0.0696	0.1494	0.1095
Whole image ($\bar{\sigma} < \infty$)	5.92	0.1091	0.1868	0.0842	0.2139	0.1490

Table 2.5: NME (W/H) ($\downarrow$) for CHARLOTTE splits with different strategies for landmark computation. The final results are estimated with sliding windows similar to [44], however, we achieved comparable results when computing the landmarks on input images rescaled to 224×224 . Here, we do not exclude high-uncertainty images and evaluate all images without failure rate, i.e., $\bar{\sigma} < \infty$. Average inference time per frame on the Full split has been benchmarked on a single NVIDIA H100 80GB GPU with a batch size of 1.



Figure 2.7: CHARLOTTE examples [37] with different RGB and gray scale landmark methods, including the multimodal model trained on T-FAKE-P1. While RGB methods can handle some challenging poses, they may produce broken landmarks on difficult images.

Inference Ablation. We analyze the impact of our inference strategy, see Table 2.5. We compare inference on i.) the complete image scaled to 224×224 (whole image), ii.) followed by refinement on a bounding box computed from the predictions obtained with i.) and finally iii.) with the sliding window approach described in section 2.2.3.1. Method iii.) produces the best results over all images except for CHARLOTTE Low, where ii.) performs slightly better while at the same time being also suitable for real-time estimation.

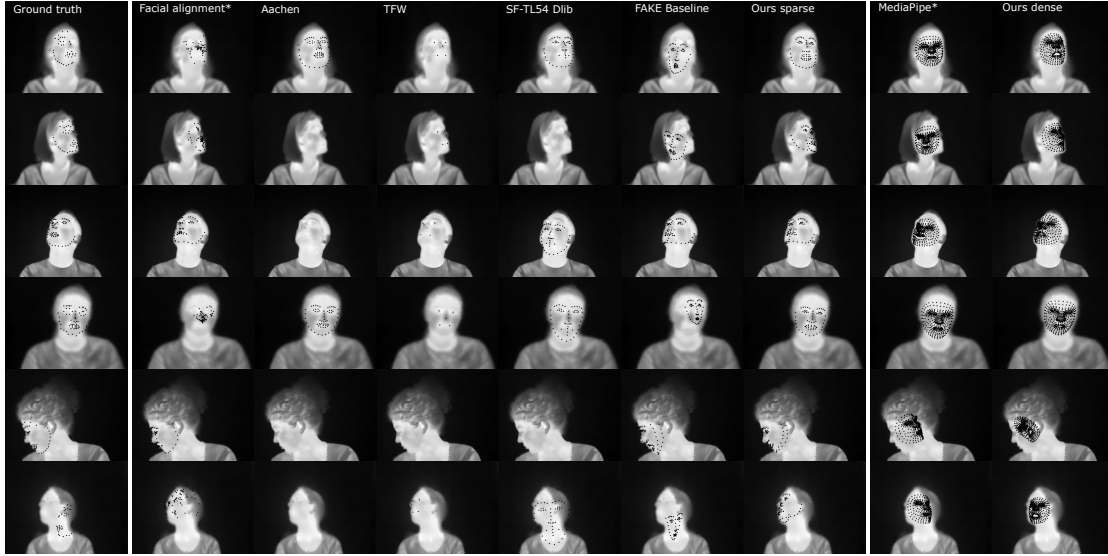


Figure 2.8: Results on examples from the CHARLOTTE dataset [37] with different RGB and thermal predictors and our models trained on T-FAKE. For images without landmarks, no faces were detected. The performance of RGB methods can be greatly improved when the images are inverted or sharpened, as indicated by *. The first column shows the limitations in the CHARLOTTE ground truth: profile annotation convention for frontal views (1st row), quantization artifacts for low resolution images (4th row), translated annotations (last row). © 2025 IEEE.

Method	300W (NME IO ↓)		
	Full	Comm.	Chal.
<i>MediaPipe</i> [45]	0.0562	0.0469	0.0947
<i>FA</i> [10]	0.0435	0.0380	0.0661
Star [81]	0.0287	0.0252	0.0430
<i>T-FAKE</i>	0.0522	0.0468	0.0743

Table 2.6: Comparison of RGB performance of our multimodal method. We trained the label adaptation on the 300W train split and evaluated each method on the 300W test split [82, 81]. For Star, we use the authors’ model pre-trained on 300W. NME has been normalized with respect to interocular distance [43, 47]. Methods with label adaptation are put in italic.

2.2.4 Conclusion

By training a probabilistic landmarker with the T-FAKE-P1 and T-FAKE datasets, we established the up-to-date first dense thermal landmarker and a state-of-the-art RGB+Thermal sparse landmarker. We tested these landmarkers with strict benchmarking by including established thermal and RGB methods and using label adaptation for each benchmarked method. Our work shows that synthetic data is sufficient for the training of thermal landmarkers and can be done on multiple imaging modalities.

2.2.4.1 Limitations

Dense Landmarks. For the presented approach, we relied on the ground truth 70-point landmarks of the FAKE dataset [43] and dense MediaPipe [45] annotations. Training with the original 320-point and 702-point landmarks could further boost accuracy and lead to an even denser landmarker. However, these landmarks are not publicly available. Furthermore, we only evaluate a MobileNet backbone. Performance for difficult poses and face variability could be achieved with denser models [44].

CHARLOTTE. The CHARLOTTE [37] dataset is among the largest datasets with thermal recordings of faces that contain different levels of image quality and have a high variability in poses such as side profile pictures and tilting which makes the dataset ideal as a benchmark. However, the annotation uses a non-standard convention where side profile images have a different number of landmarks than frontal faces. Furthermore, landmarks of low-resolution images are quantized and there are examples of shifted ground truth annotations (see Figure 2.8).

Label Adaptation. We retrain the label adaptation networks for each tested landmarker on its original predictions. Hence, for a given method, a high failure rate on CHARLOTTE means that fewer training images are available for that method. Also, very poor predictions on some of the images such as profile images without failure produce a low quality of adapted landmarks. It is important to note that the label adaptation does not use visual information from the benchmark dataset to translate landmarks. However, the same individuals were present in both the test and the training dataset splits which might allow the network to learn the facial statistics of individuals. This might give an advantage to landmarkers with only a few landmarks, e.g., TFW [52]. The high 68-point landmark performance of TFW on the CHARLOTTE thermal data does not necessarily mean that this generalizes and the advantage of increasing the number of predicted landmarks has been demonstrated for RGB images [44].

2.3 Lagrangian motion magnification revisited

Approaches for video motion magnification form a family of visualization techniques that aim to magnify imperceptibly small motions in videos. As discussed in Chapter 1, the most prominent techniques are based on Eulerian video processing. Eulerian approaches directly modify changes in intensity values on a fixed pixel grid rather than manipulating trajectories of moving parts in the video.

In this section, new approaches for Lagrangian motion magnification with dense *optical flow* (OF) and motion magnitude driven amplification are proposed and investigated. For different psychophysiological use-cases, motion magnitude driven amplification achieves very good performance with fewer artifacts than Eulerian approaches. Contrary to Eulerian motion magnification, the methods do not rely on temporal filtering and work in the presence of large motion.

The results from this section and the experiences with the suitability of dense OF for motion magnification informed the choice of the high-speed camera from section 2.1 and section 2.6.1. This section contains direct excerpts from [F9]³. The findings in this section, as published in [F9], serve as an exploratory foundation for more advanced and image driven motion decomposition approaches for Lagrangian motion magnification in later sections.

2.3.1 Introduction

Video *motion magnification* forms a group of visualization methods, that aim to magnify imperceptibly small motions in videos. We can group motion magnification techniques into Lagrangian and Eulerian methods [11]. Lagrangian approaches explicitly estimate motion and render videos with modified motion fields, while Eulerian ones modify the pixel time courses locally to achieve motion and change magnification.

The most prominent methods today follow the Eulerian motion magnification paradigm. They were first formally described by Wu et al. [11] who magnified changes in the video intensity on different spatial and temporal frequency bands. This enabled them to visualize the flow of blood with consumer cameras, which received broad media attention. Wadhwa et al. [15] extended this approach to the local phase domain, where they implemented *phase-based* motion magnification as local, spatial phase shifts. They have been shown to work in real-time when using an efficient filter bank [83].

³The code for the motion scaling operator described here is included in the implementation of [F12] and can be found at <https://github.com/phflot/momag>.

However, a challenge in the context of motion magnification is the amplification of small motion which is embedded in large ambient motion-fields. Phase-based motion magnification and other purely Eulerian approaches are unable to magnify relative changes in a scene with a lot of motion [13], [84]. Elgharib et al. [13] proposed a motion magnification pipeline which they call *dynamic video magnification* (DVMAG). They try to solve the problem via a sequence of explicit video stabilization and magnification steps. In particular, they apply phase-based motion magnification on a manually matted image inside a region of interest which they stabilize with respect to hand-selected points they expect to be stable. Holes in the image induced by the magnification are filled in with texture synthesis techniques. Even though this approach works in the presence of large motion, it is only applicable when all frames in the video are characterized by non-linear motion and are occlusion free.

The most recent work in motion magnification is called *motion acceleration magnification*, which was introduced by Zhang et al. [84] in 2017. They magnify the second order temporal gradients of image sequences in an Eulerian manner, which robustifies their approach with respect to linear motion. The technique has been shown to be robust under linear, large motion and as the time of the original paper [F9] (2018), performs well compared to existing methods.

In their pioneering work, Liu et al. [12] proposed a Lagrangian strategy for motion magnification, which formed one of the first works on video motion magnification. They interpolate a dense flow field from feature point trajectories in segmented regions of the input sequence. Ambient motion is removed by the parametric stabilization of supposedly steady points and automatically generated motion clusters are manually selected and magnified. Lagrangian approaches are usually criticized for their quality and speed due to their dependence on explicit motion estimation and compensation.

However, advances in dense OF estimation since the publication of [12] give us the opportunity to re-explore Lagrangian approaches for motion magnification in videos. Despite the ability of deep learning based approaches such as [85]⁴ to perform in real time, the flexibility that variational models provide enable us to use Lagrangian motion magnification for well controlled setups such as in laboratories for psychophysiological experiments.

In this section, we propose a robust, motion magnitude driven motion magnification strategy. The theoretical background is provided in section 2.3.2. In section 2.3.3, the proposed method is compared with Eulerian motion magnification approaches and applications of the proposed method are discussed.

⁴Note that significant advances in deep learning-based OF estimation such as PWC-Net [86] and RAFT [87] happened after or during submission of [F9] and will be considered in the next section.

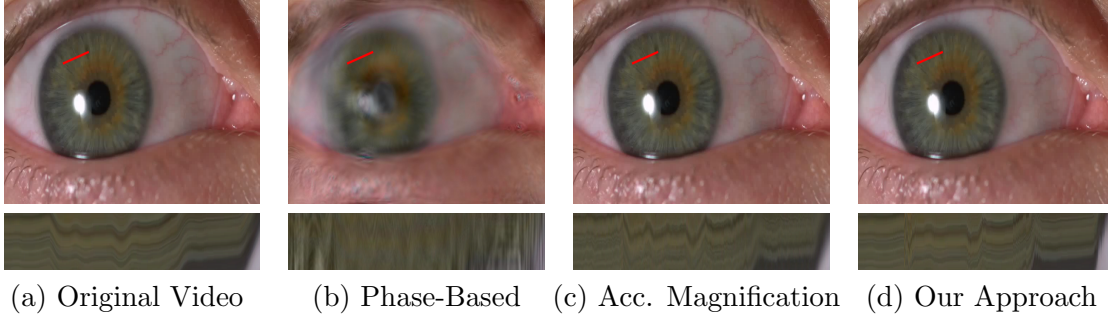


Figure 2.9: Visual comparison of phase-based motion magnification and acceleration magnification with the proposed Lagrangian magnification approach on the eye sequence from [13]. The temporal slice demonstrates how the Lagrangian magnification performs similarly well compared to the Eulerian acceleration magnification. © 2018 IEEE.

2.3.2 Materials and methods

We consider an image sequence $(f_t)_{t=1}^T$ with frames $f_t : \Omega \rightarrow [0, 1]$ defined on an image domain $\Omega \subset \mathbb{R}^2$. The discrete sampling on a rectangular pixel grid is denoted by $f_{i,j,t}$ with $(i, j) \in \Omega_h$ and $t \in \{1, \dots, T\}$, where

$$\Omega_h := \{1, \dots, m\} \times \{1, \dots, n\} \subset \mathbb{Z}^2, \quad \Omega_{h,T} := \Omega_h \times \{1, \dots, T\} \subset \mathbb{Z}^3.$$

Each frame consists of $m \times n$ samples and the sequence contains T frames. We denote spatial coordinates by $\mathbf{x} := (x, y)^\top$ and interpret $f_t(\mathbf{x})$ at off-grid locations via interpolation. Unless stated otherwise, operations on images and displacement fields are understood pointwise in $\mathbf{x}$.

Dense displacements are written as $\mathbf{w}_t(\mathbf{x}) := (u_t(\mathbf{x}), v_t(\mathbf{x}))^\top : \Omega \rightarrow \mathbb{R}^2$. To mitigate drift and amplify changes w.r.t. a baseline, we estimate reference-based flows $\mathbf{w}_t$ from a chosen reference frame f_{ref} to f_t . Inter-frame flows $\bar{\mathbf{w}}_t$ between successive frames f_{t-1} and f_t can also be used, which would always magnify with respect to a previous frame.

2.3.2.1 Motion estimation

A weakness of Lagrangian approaches is their dependence on the quality of underlying motion estimation. Other approaches, however, tend to fail in the presence of large motion. For example in [11], the limited linearization assumption leads to unsatisfactory results. To mitigate this, the approach used in [13] falls back to using motion estimation and image segmentation.

Groundbreaking advancements have been made with the estimation of OF over the last decades, e.g. [19, 87]. Also, given that most psychophysiological recordings are carried out in a laboratory setting, the benefits of having controlled conditions further motivate the use of better OF estimation methods for Lagrangian motion magnification.

At the time of writing the original paper [F9], variational methods for optical flow estimation were still among the best-performing techniques, particularly for small displacements [85]. Hence, the OF model of choice here is an ℓ_1 penalized linear combination of the brightness constancy assumption and gradient constancy assumption with separate robustification and a flow-driven regularizer [88] as well as the DIS OF model. The data term is normalized to reduce the influence of large gradients [89]. See Appendix F for more details.

Motion compensation. Subtle motions of interest are typically superimposed on larger, smooth ambient displacements such as head motion. While prior pipelines estimate and remove such ambient motion using parametric models from hand-selected points [12] or ROI-based stabilization [13], we directly work with dense displacement fields. Given a reference frame f_{ref} and a target frame f_t , we compute the forward displacement field $\mathbf{w}_t : \Omega \rightarrow \mathbb{R}^2$ with $\mathbf{x} := (x, y)^\top$ and $\mathbf{w}_t(\mathbf{x}) := (u_t(\mathbf{x}), v_t(\mathbf{x}))^\top$ such that $f_{\text{ref}}(\mathbf{x}) \approx f_t(\mathbf{x} + \mathbf{w}_t(\mathbf{x}))$. For readability, we fix t and omit the index in the following, i.e. we write $f := f_t$ and $\mathbf{w} := \mathbf{w}_t$. We remove the estimated motion by backward warping,

$$f_{\text{stable}}(\mathbf{x}) := f(\mathbf{x} + \mathbf{w}(\mathbf{x})).$$

To separate spatially smooth global motion from fine local motion, we estimate the displacement field twice with the same OF estimator but different smoothness weights: a fine field $\mathbf{w}_{\text{fine}}$ that uses lower regularization to capture detailed motion and a coarse field $\mathbf{w}_{\text{coarse}}$ that uses higher regularization to capture smooth global motion. We define the local residual as

$$\mathbf{w}_{\text{local}} := \mathbf{w}_{\text{fine}} - \mathbf{w}_{\text{coarse}}.$$

This decomposition is used in the subsequent magnification step by keeping $\mathbf{w}_{\text{coarse}}$ and selectively amplifying $\mathbf{w}_{\text{local}}$.

Magnitude scaling. We use a scaling operator similar to the gradient domain tone mapping operator of Fattal et al. [90]. It is a parametrized, exponential function, with a threshold parameter α that defines the upper magnitude threshold of motion we want to amplify and with a magnification parameter β that controls the steepness of the function (see Figure 2.10).

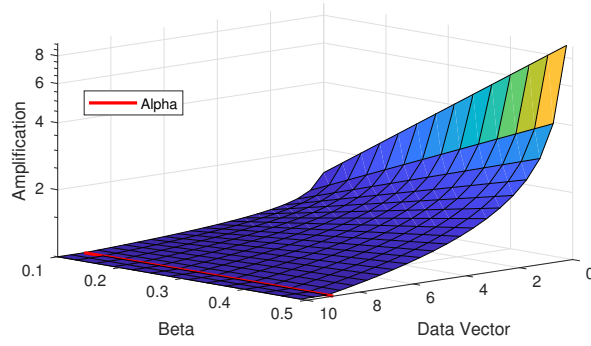


Figure 2.10: Our magnitude driven magnification chooses the amplification factor with a parametrized, exponential function. A vector with values $\in [0, \dots, 10]$ is scaled with $\alpha = 9$ and varying choices for β . Higher values of β increase the steepness of the amplification function. © 2018 IEEE.

Let $\alpha \in \mathbb{R}^+$, $0 < \beta < 1$ and $\varepsilon > 0$ be a small positive constant. Then the operator is defined as

$$\varphi_{\alpha,\beta}(w) = \alpha^\beta \left(\|w\| + \mathbb{1}_{\|w\| < \varepsilon} \varepsilon \right)^{-1} \left(\|w\| + \mathbb{1}_{\|w\| < \varepsilon} \varepsilon \right)^{(1-\beta)}, \quad (2.7)$$

with the indicator function

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}. \quad (2.8)$$

We limit the lower boundary of the function to 1, because we are not interested in attenuating any motion with higher magnitude than α . This finally results in

$$\Phi_{\alpha,\beta}(w) = \begin{cases} 1 & \text{if } \varphi_{\alpha,\beta}(w) < 1 \\ \varphi_{\alpha,\beta}(w) & \text{else} \end{cases}. \quad (2.9)$$

Motion magnification. Given an image sequence $(f_t)_{t=1}^T$ and a reference frame f_{ref} , let $\mathbf{w}_t(\mathbf{x}) := (u_t(\mathbf{x}), v_t(\mathbf{x}))^\top$ denote the forward displacement field from f_{ref} to f_t , with $\mathbf{x} := (x, y)^\top \in \Omega$. For readability, we fix t and omit it in the following, i.e. we write $f := f_t$.

To separate background motion that is assumed spatially more smooth from superimposed, subtle and highly localized motion, we compute two forward flow estimates between f_{ref} and f using the same OF model but different smoothness weights: A fine field $\mathbf{w}_{\text{fine}}$ with weaker regularization and potentially higher detail and a coarse field $\mathbf{w}_{\text{coarse}}$ with stronger regularization and only smoother, assumed global motion. Hence, we define the local component by

$$\mathbf{w}_{\text{local}} := \mathbf{w}_{\text{fine}} - \mathbf{w}_{\text{coarse}}.$$

We remove the motion by backward warping (evaluation at non-grid locations is done by interpolation),

$$f_{\text{stable}}(\mathbf{x}) := f(\mathbf{x} + \mathbf{w}_{\text{fine}}(\mathbf{x})).$$

Next, we construct the modified displacement field by selectively amplifying the detail component,

$$\tilde{\mathbf{w}} := \mathbf{w}_{\text{coarse}} + \Phi_{\alpha,\beta}(\mathbf{w}_{\text{local}}) \mathbf{w}_{\text{local}},$$

where $\Phi_{\alpha,\beta}$ is applied pointwise as defined in (2.9). Finally, we render the motion-magnified frame by forward warping with $\tilde{\mathbf{w}}$:

$$f_{\text{mag}}(\mathbf{x} + \tilde{\mathbf{w}}(\mathbf{x})) := f_{\text{stable}}(\mathbf{x}).$$

2.3.3 Results

We have presented a novel approach to motion scaling for Lagrangian motion magnification that can attenuate large motion while amplifying small motion. The visualization of micro-saccadic eye movements and pupil dilation (*hippus*) is commonly used as a benchmark for motion magnification techniques. Especially the relative motion of the iris due to saccadic eye movement (*iridodonesis*) is an interesting phenomenon which is hardly visible in healthy subjects. Physiologically, the quantification and visualization of microscopic eye-related movement is important, because these micromotions could be early markers for certain neuronal dysfunctions such as Parkinson's disease. The approach can handle scenes with large, local motions, which the methodology of DVMAG [13] does not. Our eye sequence is contaminated by an eye-blinking artifact and there is no background motion in that sense. Additional image examples can be found in [F9].

A representative comparison on the eye sequence from [13] is shown in Figure 2.9. The sequence is contaminated by a pronounced eye-blink artifact (large, local motion), which tends to induce visible blur/ghosting in purely Eulerian magnification results: both phase-based motion magnification [15] and Eulerian acceleration magnification [84] smear high-frequency details around the eyelids/eyelashes during the blink.

In contrast, our Lagrangian approach preserves fine facial texture while producing magnified dynamics in the temporal slice (Figure 2.9), and can therefore handle large, localized motion that breaks the assumptions of DVMAG [13]. Facial microexpressions are very fast and spatially small, subconscious movements of facial muscles which might be linked to different affective states in psychophysiological assessments. A common definition for microexpressions is usually hard to find, but recent work shows, that small movements, which last less than 500 ms can be considered as a microexpression [91].

For facial microexpression visualization, we use a smoothness-scale decomposition of the dense displacement field to separate global/background motion from localized motion (Figure 2.11). As illustrated in Figure 2.11 (A.4–A.6), the fine estimate $\mathbf{w}_{\text{fine}}$ captures both smooth ambient motion (e.g. head drift) and localized deformations (e.g. eyelids and mouth), while the strongly regularized coarse flow $\mathbf{w}_{\text{coarse}}$ retains mainly the smooth/global component. Their difference isolates the localized residual $\mathbf{w}_{\text{local}} = \mathbf{w}_{\text{fine}} - \mathbf{w}_{\text{coarse}}$ (A.5), which can be selectively amplified. If the full flow field is magnified, everything moves and the result is dominated by amplified background motion (Figure 2.11, B.1). After decomposition, magnifying only the localized component suppresses this global drift and restricts the apparent motion to expression-relevant regions (B.2). Applying the compressive scaling operator $\Phi_{\alpha,\beta}$ on the localized component further attenuates large residual displacements and reduces artifacts, while still amplifying subtle microexpression motion (B.3).

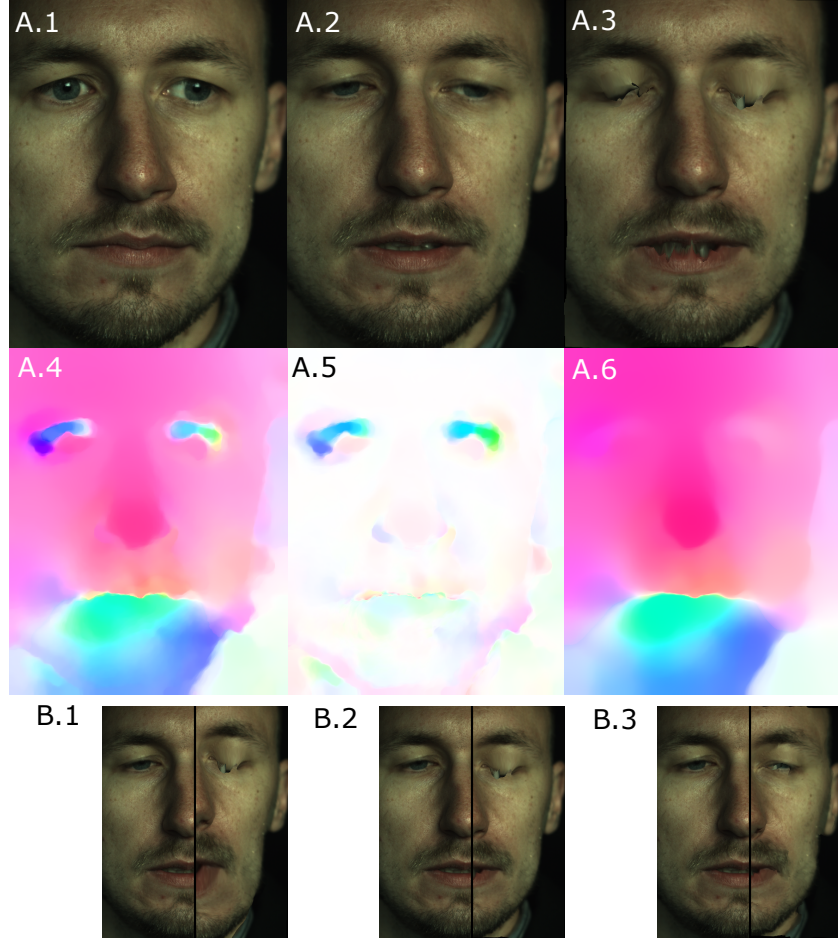


Figure 2.11: Smoothness-scale optical-flow decomposition and selective motion magnification between two frames (compare Figure 2.20) using DIS OF with 300 iterations variational refinement and two different smoothness parameters. (A.1) shows the reference frame and (A.2) the target frame. (A.4) Fine flow $\mathbf{w}_{\text{fine}}$ (lower regularization, $\alpha_{\text{fine}} = 6$) captures global and detailed motion and (A.6) coarse flow $\mathbf{w}_{\text{coarse}}$ (higher regularization, $\alpha_{\text{smooth}} = 15$) captures only global motion without local details such as eye blinking. (A.5) shows the local component $\mathbf{w}_{\text{local}} = \mathbf{w}_{\text{fine}} - \mathbf{w}_{\text{coarse}}$. (A.3) Linear local magnification with factor $m = 5$ using $\mathbf{w}_{\text{mag}} = \mathbf{w}_{\text{coarse}} + m \mathbf{w}_{\text{local}}$. (B.1–B.3) Split-screen results (left: original f_b , right: magnified) showing global magnification (full flow), linear local magnification, and local magnification with compressive scaling Φ (parameters $\alpha_{\Phi} = 15$, $\beta_{\Phi} = 0.8$, $\varepsilon_{\Phi} = 10^{-1}$) to suppress large/background motion artifacts. Flows are visualized with standard color coding.

2.4 Lagrangian motion magnification with landmark-prior and double sparse OF decomposition for microexpressions and facial micromovements

This section focuses on the visualization of motion in videos of faces. The motion in high framerate videos of faces can contain a wide variety of information that ranges from microexpressions over pulse or respiratory rate to cues on speech and the affective state. The approaches from the previous section process dense displacement fields for motion magnification. However, they are not fine tuned for the application on faces and do not include individual motion components. Furthermore, the presented magnification approach requires background motion compensation to magnify small motion that is embedded in large motion.

In this section, motion decomposition into global and local movements is extended to include awareness of facial features and the decomposition into individual motion components. To this end, this section and the related publications [F12, F5] make the following contributions:

- Integration of landmark information from the face to define global and local motion components.
- Fine tuning of *recurrent all-pairs field transforms* (RAFT) for OF with ground truth obtained from the variational *dense inverse search* (DIS) for OF algorithm applied to the *Chinese Academy of Sciences microexpression CASME II* video set of facial microexpressions on faces.
- Approximation of the OF field with sparse components both in space and time leading to a double sparse decomposition.
- Magnification of micromovements in specific areas of the face based on the double sparse and facial landmark-based decompositions.

This section contains direct excerpts from [F12, F5]⁵.

2.4.1 Introduction

Motion magnification describes a wide variety of algorithms to magnify and therefore visualize subtle, imperceptibly small motions in videos. In analogy to fluid dynamics, motion magnification techniques can be grouped into Eulerian and Lagrangian approaches.

⁵Our software and video results are available at <https://github.com/phflot/momag> and https://github.com/phflot/dsd_momag.

Eulerian methods were first introduced by Wu et al. [11]. These methods amplify motion or color changes by modifying the time-dependent color variation in the video using temporal or spatial filtering, e.g., by dealing with image pyramids of different resolutions [13, 15, 11, 84]. The term Eulerian emphasizes that information is processed on the fixed grid of pixels while Lagrangian methods manipulate point trajectories. Eulerian methods found applications for quantitative motion assessment as well. For example, Sarrafi et al. [92] and Eitner et al. [93] used point tracking on motion magnified recordings for the detection and modal parameter estimation of vibrations in different mechanical components. Fei et al. [94] applied motion magnification to the detection of AI-generated videos and Alinovi et al. [95] to respiratory rate monitoring. While Eulerian methods are often very easy and fast to evaluate and can therefore be readily employed in real-time applications, the type and degree of motions that can be magnified is limited. In particular, phase-based methods depend on local spatial frequencies which can lead to blurring around sharp image edges. Intensity-based Eulerian methods magnify motion by linearization, which is known to be limited to small movements.

Lagrangian approaches rely on the explicitly computed motion field of objects within the video which is in general done by means of OF estimation [96, 12, F9]. The OF fields enable not only selective and adaptive magnification of types of motion that vary spatially as, e.g. in [F9, 12], but also to attenuate them separately or completely remove all movement, see [96], [F3]. An advantage of Lagrangian methods is that the estimated motion information can not only be used for the motion magnification task, but also allow subsequent quantitative assessment [97, F7].

The advancements in OF estimation in the past decades resulted in high-accuracy methods with different model invariants as well as robustness and high computing speed. Variational methods optimize functions with explicitly modeled data terms and regularizers. They still produce state-of-the-art accuracy in many applications, especially in combination with advanced smoothness terms such as [97] and methods for the initialization of very large displacements: Chen et al. [98] lead (as of March 2022) the Middlebury OF benchmark [99] with respect to average endpoint error and average angular error. They use similarity transformations for a segmented flow field as initialization of large motion. In a second step, the variational method of Sun et al. [100] is used for subpixel refinement.

The widely available dense inverse search (DIS) OF method [71] initializes the OF field with patch correspondences and uses a variational method similar to Sun et al. and Brox et al. [19] for high resolution refinement. The availability of large-scale synthetic datasets such as MPI Sintel dataset [101] paved the way for deep learning-based methods such as RAFT [87] which computes a low resolution flow field from correlation volumes

which is subsequently upsampled via learned upsampling convolutions. Variants of the method are still among high performing methods in the competitive Sintel benchmark⁶ [102]. Its fast convergence [103] makes it in particular interesting for refinement tasks.

Non-contact monitoring of the affective and neurocognitive state is an emerging topic with many new applications in different areas of life sciences and engineering. Especially, estimation of the psychophysiological state can find applications in healthcare settings and even neuro-ergonomics in human machine interaction. In that context, facial microexpressions promise to allow an assessment of emotions from faces. In contrast to normal facial expressions, microexpressions can even occur when the person tries to conceal the emotion [91].

However, microexpressions are temporally and spatially very small movements and, therefore, difficult to observe. Recent work shows that small movements, which last less than 500 ms can be considered as microexpressions [91]. Besides appearance-based features that analyze the skin texture of faces, OF field-based features have long been used for the detection and analysis of microexpressions [104, 105, 106, 107]. Apart from the small, microexpression induced movements, appearance-based methods use face appearance features that describe skin texture of faces to capture changes in shading and texture for microexpression detection, see [104].

On a more general scope, micromovements of the face and head can contain other important physiological and psychophysiological information: the movement of the head can contain cues on heart rate [108], movement of the mouth can be applied for audio-visual speech recognition in noisy environments [109], and micromovements of the ear can contain cues on audio stimuli [F7]. While many of the analysis methods fall back on OF methods for analysis, there are not many annotated datasets for specific, local movements. The selective detection and magnification of micromovements of the face is a relatively untapped area of research with the potential to guide the annotation of such datasets and visualize the isolated movements that can be exploited for computer vision algorithms.

However, such underlying emotional implications are not the focus of this section. Instead, we will concentrate on the relatively untapped area of detecting and selectively magnifying micromovements. This exceeds the field of microexpressions and goes more in the direction of realistic magnifications in computer graphics.

⁶<http://sintel.is.tue.mpg.de/results>

Here, we rely on Lagrangian motion magnification, and make the following contributions. First, we fine tune the RAFT deep learning method [87] using variational approaches [19, 71] on a facial dataset to create new ground truth data to fine tune the network on recordings of faces with small displacements. Next, we employ ideas of sparse matrix factorization from [110], to decompose the OF field into meaningful sparse components. Variational approaches with sparsity promoting regularization terms are very common in inverse problems in imaging, see also [111, 112, 113, 114] for other sparsity based approaches than just sparse matrix factorization. Here we address the sparsity both in space and time leading to a double sparse representation. Although there exists a broad literature on sparse decomposition, also in connection with principal component analysis, see e.g. [111] and the references therein, to the best of our knowledge, such decomposition was not applied for motion magnification so far. Having the above decomposition at hand, we are finally able to magnify the motion in respective areas in an unsupervised way. For this, we use a sophisticated forward warping method. Image warping methods, usually simpler backward warping approaches, are used in the context of OF estimation, compression or image registration [19, 115]. We will explain forward and backward warping to illustrate why the first one is better suited for our purposes. Furthermore, we propose landmark guided decomposition of the facial OF field into global and local displacements. We use a diffusion-based image inpainting approach to remove the local displacements in areas of the face and gain an approximation of global motion. This way, small and local movements in the face can be analyzed in the presence of large movement.

2.4.2 Materials and methods

2.4.2.1 Motion estimation

A description of RAFT and variational OF estimation can be found in Appendix E.

We fine tune RAFT with additional training data obtained via variational OF methods. To this end, we replace the fine tuning on KITTI-2015 with a fine tuning on facial microexpression videos derived from the CASME II [116] and SMIC datasets [32]. First, we construct a *microexpression* (ME) dataset by annotating SMIC and CASME II with the *dense inverse search* (DIS) optimization approach from [71] implemented in the OpenCV library [117]. Note that DIS uses a variational OF technique close to [19, 100] (see Appendix E). Second, we enrich the ME dataset as follows: we use the already estimated OF $\mathbf{w}(\cdot, t)$ and apply it to the first frame $f(\cdot, 1)$ of each sequence in SMIC and CASME II to obtain new sequences $\hat{f}(\cdot + \mathbf{w}(\cdot, t), t) := f(\cdot, 1)$.

The reason for including both sequences f and $\hat{f}$ in ME is that

- i) for $\hat{f}$, an accurate ground truth displacement field exists, but it does not contain the changes in face appearance from the original video, because the color information is just propagated from the first frame of each sequence;
- ii) on the other hand, f contains changes in face appearance from the original video, but there is only a DIS estimation of the OF ground truth.

In summary, with this strategy, we ensure that the dataset contains accurate image and flow pairs with $\hat{f}$ as well as the original image sequence with dynamic face appearance information in f for which we have OF estimates. Including the dynamic face appearance information is particularly important in the context of microexpressions [104].

To synthesize the motion magnified image f_{mag} of f_t with respect to the reference f_0 , we warp the image into the grid of the reference frame and apply forward warping with a modified displacement field w_{mag} . Local movements are amplified by a factor $\mu > 1$ into $\mathbf{w}_{\text{mag}} = \mu \mathbf{w}_{\text{local}} + \mathbf{w}_{\text{global}}$ and the final magnified image satisfies $f_{\text{mag}}(\mathbf{x} + \mathbf{w}_{\text{mag}}) = f_t(\mathbf{x})$. The magnification factor μ can also be a non-linear function with respect to the motion magnitude, e.g. in [90], [F9] and section 2.3.

RAFT was improved with optimized training strategies, datasets and augmentations in various works. For a comparison of different methods see [103]. More precisely, the RAFT training schedule now combines different datasets in multiple training stages for pre-training and refinement as follows [87]: RAFT is pre-trained on FlyingChairs [118] for 100k iterations, which is followed by 100k iterations on FlyingThings3D [99]. Further, RAFT is fine tuned on 100k iterations on Sintel [101] (RAFT-Sintel), KITTI-2015 [119] and HD1k [120], which is finally followed by 50k fine tuning on KITTI-2015 (RAFT-KITTI). As mentioned above, here we replace the fine tuning on KITTI-2015 with a fine tuning on facial microexpression videos derived from the CASME II and SMIC datasets.

2.4.2.2 Landmark-based decomposition

To decompose the displacement field into global and local movements with relevance for microexpressions, we apply the face mesh landmark detector from MediaPipe [45] to automatically select *regions of interest* (ROIs) around mouth, eyes or forehead (see Figure 2.20) [F12].

For a displacement field $\mathbf{w}$, we define local displacements as $\mathbf{w}_{\text{local}} := \mathbf{w} - \mathbf{w}_{\text{global}}$ and employ the simplified motion assumption that missing parts of the global displacements in the image can be restored using diffusion-based inpainting (e.g. [121]).

The set of ROIs is given by $\mathcal{R} \subset \mathbb{R}^2$ and $\Omega \subset \mathbb{R}^2$ denotes the image domain. To separate local displacements from global displacements inside of the ROIs, we apply component-wise diffusion-based flow inpainting to $\mathbf{w}$ inside $\mathcal{R}$ to generate $\mathbf{w}_{\text{global}}$. We use Dirichlet boundary conditions $\mathbf{w}_{\text{global}}(\mathbf{x}) = \mathbf{w}(\mathbf{x}), \forall \mathbf{x} \in \partial\mathcal{R}$ and define $\mathbf{w}_{\text{global}}$ on Ω by setting $\mathbf{w}_{\text{global}}(\mathbf{x}) = \mathbf{w}(\mathbf{x}) \forall \mathbf{x} \in \Omega \setminus \mathcal{R}$ and then inpaint inside $\mathcal{R}$ by solving

$$\Delta \mathbf{w}_{\text{global}}(\mathbf{x}) = 0 \quad \text{for } \mathbf{x} \in \mathcal{R}, \quad \mathbf{w}_{\text{global}}(\mathbf{x}) = \mathbf{w}(\mathbf{x}) \quad \text{for } \mathbf{x} \in \partial\mathcal{R}. \quad (2.10)$$

This way, $\mathbf{w}_{\text{global}}$ contains smooth background motion such as head pose changes and $\mathbf{w}_{\text{local}}$ contains the residual local motion within the ROIs.

2.4.2.3 Double sparse optical flow decomposition

The complete derivation of the double sparse OF decomposition can be found in [F5] and is briefly summarized here.

The double sparse decomposition decomposes the sequence of OF fields into spatial components $G^k(x, y)$ and temporal components $d^k(t)$ [110]. The spatial components are enforced to be sparse using an L_1 -norm. This way, clear spatial regions (e.g., around the eyes or mouth) make up individual components (see Figure 2.15). The spatiotemporal components are grouped using the $L_{2,1}$ -norms which promotes group sparsity [114]. For each temporal profile $d^k(t)$, the norm calculates the sum of the L_2 -norms across time steps (which measures the total energy of the motion) and enforces sparsity by limiting the number of non-zero intervals [112]. The intuition behind the $L_{2,1}$ -norm is that it encourages components containing short time intervals instead of multiple movements as we would expect from microexpressions. It penalizes the widespread movement across time, encouraging temporal sparsity while allowing movement during select intervals.

The objective function is made up of both, the spatial and temporal components' sparsity constraints. It is solved with an alternating minimization, where the spatial and temporal components are updated iteratively [F5]. This decomposition is motivated by non-orthogonal PCA: The flow is approximated as a sum of rank-1 matrices of which each corresponds to a spatial and temporal component. For each step, either the spatial or temporal components are fixed while the other one is updated. Finally, the sparse representation of the OF field sequence is given by the *dictionary* with the spatial components $G^k(x, y)$.

2.4.2.4 Motion magnification

The detection of microexpressions and facial micromovements plays an increasingly important role for various applications from computer vision. Microexpressions and micromovements are often unobservable by the untrained and naked eye. This difficult task can be necessary in the context of the annotation of video datasets (e.g. compare [116]). The visualization of microexpressions and other micromovements in the face via motion magnification can potentially mitigate the difficulty of manual microexpression detection to enable manual annotation and microexpression analysis without professional or specific training. Furthermore, unsupervised selection and magnification of different motion components can help with exploratory investigation of facial micromovements for the development of new methods for remote assessment.

Having decomposed the OF field into meaningful spatial and temporal components, we show in this section how these components can be used to enhance the OF in the regions of interest and how this can be visualized in a new video sequence using a sophisticated warping method. Image warping is used in the context of OF estimation, compression or image registration (usually backward warping) or for applications from graphics such as texture mapping or novel view synthesis, among the rich literature, see, e.g. [19, 115].

2.4.2.5 Forward warping

For a given forward flow displacement $\mathbf{w}$ between two consecutive image frames f_1 and f_2 (so that we can skip the time variable), there are two prevalent methods for warping, forward warping and backward warping:

- i) Backward warping f_2 to yield f_{bw} :

$$f_{\text{bw}}(x, y) := f_2(x + u(x, y), y + v(x, y)).$$

Backward warping allows interpolation of pixel values from floating-point coordinates of the source image. In section 2.3 and here, we use backward warping to remove motion and move f_2 onto the same grid as f_1 .

- ii) Forward warping f_1 to yield f_{fw} :

$$f_{\text{fw}}(x + u(x, y), y + v(x, y)) := f_1(x, y).$$

A naive approach to go from f_1 to f_2 could be to compute $f_{\text{fw}}(x, y) := f_1(x - u(x, y), y - v(x, y))$, however, $\mathbf{w}$ is computed on the grid of f_1 and inverting implicitly evaluates the flow at the wrong coordinates. The effect can be seen in Figure 2.12.

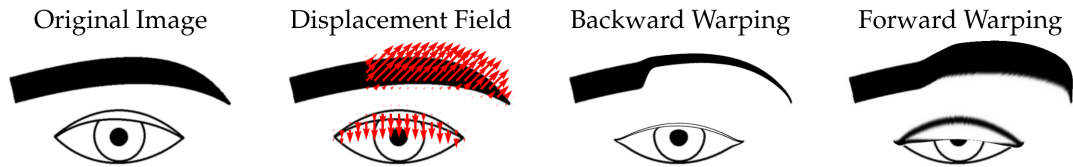


Figure 2.12: An illustration comparing forward warping of a forward displacement field with a naive sign-inverted backward warp. The latter evaluates the flow on the wrong grid and therefore produces geometric distortions. Figure originally published in [F5].

This figure shows the results of forward warping versus backward warping on a displacement field which corresponds to a blinking motion and raising of an eyebrow. In this case, the motion field is localized on the eyebrow and the eyelid. Therefore, when doing backward warping, the colors stay exactly the same everywhere the displacement field is zero. In this illustration, this yields a contraction of the eyebrow and a contraction of the eye itself.

However, in computer graphics, forward warping is significantly harder to do. This is because for the warped frame, we only have color information at the points $(x + u(x, y), y + v(x, y))$, which generally do not correspond to grid coordinates and produce scattered points. To overcome this problem, we use the following scheme: first, we consider a triangulation of the frame f_1 , which we aim to warp. The structure of this triangulation is shown in Figure 2.13 left. Each triangle connects three adjacent pixels and the vertices are equipped with the color values of the corresponding pixel. Then the triangles are displaced according to the vector field (u, v) yielding a displaced triangulation depicted in Figure 2.13 right. Finally, each of the displaced triangles is rasterized. More precisely, we take all pixels whose midpoints are contained in the displaced triangle and assign them a value interpolating the triangle vertex color values using barycentric coordinates.

Finally, it can be possible that multiple displaced triangles overlap. To resolve this ambiguity, the proposed algorithm allows for a depth map as input which indicates which triangles should be drawn above others. While for this application, an elaborate depth map based on facial features could be devised, we opted to set the depth map such that stronger motions always “overlap” weaker motions. For large divergence displacements, this approach has advantages over forward splatting approaches such as [122] because it does not produce holes as demonstrated in Appendix A.

To efficiently compute the triangulation, displacement and subsequent rasterization of the input image, we employ a shader pipeline written using the OpenGL programming language.

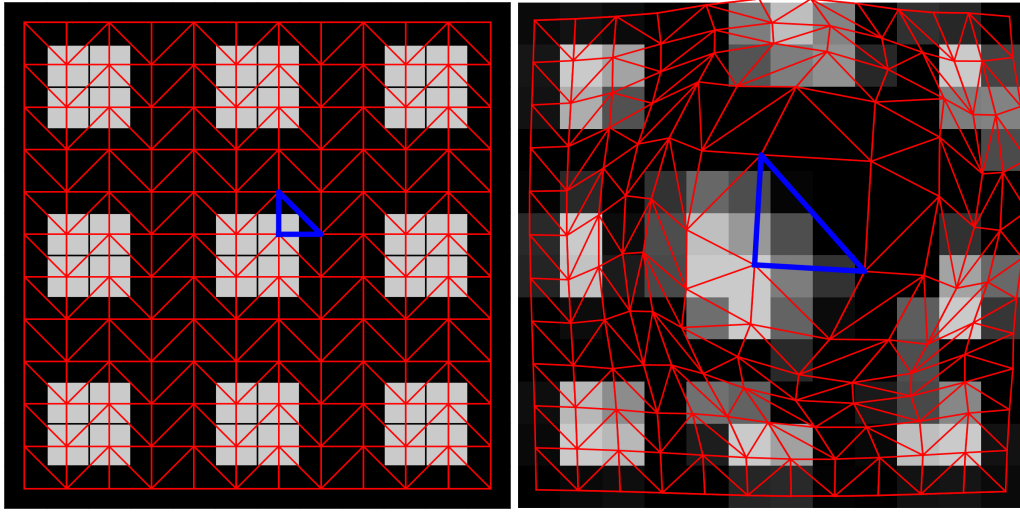


Figure 2.13: Illustration of the forward warping process. The image is first triangulated as shown on the left. The triangles are displaced and the warped image is then rasterized from these distorted triangles, see next figure. Figure originally published in [F5].

In short, this means that we make use of GPU native routines which are extremely efficient at handling such tasks and enable us to even employ this forward warping technique in real-time applications.

2.4.2.6 Motion magnification in selected regions

Sparse components. With all this at hand, we are now able to magnify selected movements in the input video. To this end, we still need to decide which components from the decomposition can be considered to be micromovements. In our case, this is done by simply evaluating for each $k \in \{1, \dots, K\}$, if the values of d^k lie in a specified range. However, as the d^k are all normalized due to our optimization procedure, we need to adjust them to better resemble the velocity magnitudes from the OF. In short, we do the following selection:

1. Compute the maximum of each domain G^k for all $k \in \{1, \dots, K\}$,

$$m^k := \max_{x,y} G^k(x, y).$$

2. Compute the maximal normalized motion magnitude for each component

$$c^k := m^k \max_{t \in \{1, \dots, T\}} \|d^k(t)\|_2.$$

3. For two pre-defined thresholds $0 < \lambda_1 < \lambda_2$, consider the k -th component $k \in \{1, \dots, K\}$ to belong to a micromovement if $\lambda_1 \leq c^k \leq \lambda_2$.

Let $\mathcal{I} \subset \{1, \dots, K\}$ be the set of components deemed to correspond to micromovements by the above selection. Then, we magnify their motion as follows.

For each, $t \in \{1, \dots, T-1\}$, we assume that $\mathbf{w}(\cdot, t)$ warps the frame f_1 into f_2 . Therefore, magnification of the micromovements by the magnification factor $\mu > 0$, we aim to warp f_1 by the flow field

$$\tilde{\mathbf{w}}(\cdot, t) := \mathbf{w}(\cdot, t) + \mu \sum_{k \in \mathcal{I}} d^k(t) G^k(\cdot).$$

Facial landmarks. For the motion magnification of facial landmarks, we use a slightly different approach. We do not have temporally isolated components but are interested in a decomposition into local and global displacements along the entire image sequence. Given an image sequence $(f_t)_{t=1}^T$ with T frames, we compute the forward displacements from a given reference frame or baseline f_{ref} to frame f_t as $\mathbf{w}_t := (u_t, v_t)^\top$. For readability, we fix an index t and omit it in the following, i.e. we write $f := f_t$ and $\mathbf{w} := \mathbf{w}_t$. In practice, the reference could be a neutral expression, or defined via fixed time offset, where an offset of 1 results in the flow used in the previous section. Subsequently, we proceed as follows:

1. Remove the motion $\mathbf{w}$ from f such that

$$f_{\text{stable}}(x, y) := f(x + u(x, y), y + v(x, y)).$$

2. Compute the global displacement $\mathbf{w}_{\text{global}}$ using diffusion-based inpainting, i.e. by initializing $\mathbf{w}_{\text{global}}$ with $\mathbf{w}$ outside the ROIs and then solving Equation (2.10).
3. Compute the local displacement $\mathbf{w}_{\text{local}} := \mathbf{w} - \mathbf{w}_{\text{global}}$.
4. Construct the modified displacement field $\tilde{\mathbf{w}} = \mathbf{w}_{\text{global}} + \Phi(\mathbf{w}_{\text{local}}) \mathbf{w}_{\text{local}}$, where Φ is an amplification function such as the compressive function in Equation (2.9) or constant magnification $\Phi_\mu : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\Phi_\mu(\mathbf{w}) = \mu$, with $\mu > 1$ (see Figure 2.20).
5. Render the magnified frame by applying forward warping using the modified displacement field $\tilde{\mathbf{w}}$ such that the motion magnified frame f_{mag} is defined by

$$f_{\text{mag}}(x + \tilde{u}(x, y), y + \tilde{v}(x, y)) := f_{\text{stable}}(x, y).$$

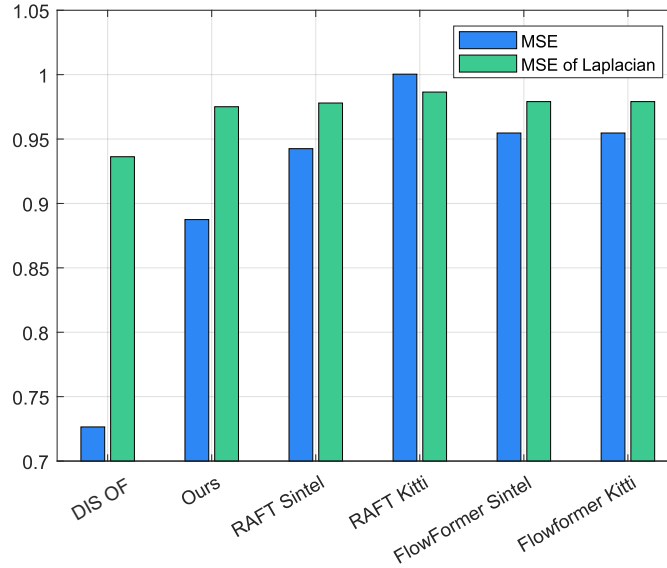


Figure 2.14: MSE for a microexpression sequence after motion compensation. The values are normalized with respect to the MSE of the raw recording. While DIS performs better than the refined RAFT, our RAFT refinement on the ME dataset improved the performance over other state-of-the-art models trained on Sintel and KITTI-2015 in our examples. Figure originally published in [F5].

2.4.3 Results

We illustrate our proposed method on an example video from CASME II [116] and a video from the medical dataset for the remote detection of respiratory infections (see Figure 2.19). All our experiments were performed on consumer-grade computers. RAFT training has been performed on a multi-GPU machine with Intel Xeon Gold processors and two Nvidia Tesla V100 GPUs.

2.4.3.1 Datasets and data acquisition

CASME II and SMIC are datasets of spontaneous microexpressions which include 35 (CASME II) and 16 (SMIC) participants respectively. Videos are recorded at 200 fps (CASME II) and 100 fps (SMIC). Each video is labelled with either observable or expected microexpressions, the latter based on self-reports of the participants. Additionally, we use a video from a dataset for the remote assessment of respiratory diseases (VI-SCREEN). For this dataset, participants were recruited at multiple recording sites, including Saarland University Clinic (UKS) Children’s Hospital and at the UKS central emergency admission, Germany. Recruitment happened from February 2022 to April 2023. This study includes 106 participants (median age 26 years, *interquartile range* (IQR) 24 years). Informed consent was obtained, and the study was approved by the Ethics committee of the Ärztekammer des Saarlandes (02/22).

Viral and bacterial respiratory pathogens were tested using multiplex polymerase chain reaction (Multiplex PCR) assays. The camera array consisted of stereo RGB, NIR and thermal cameras as well as microphones (compare [F1]). A duration of 5-10 minutes was recorded where participants answered questions in front of the camera. Core temperature was recorded with infrared ear thermometers. The videos in this dataset can contain microexpressions as well as other types of micromovements which are potentially related to symptoms of respiratory diseases. As a benchmark sequence, we selected a video which contains a large movement (eye blinking) as well as a small movement around the mouth.

2.4.3.2 Motion estimation

We compare our method against RAFT-Sintel and RAFT-KITTI as well as against FlowFormer⁷ after the same training schedules [123]. RAFT-Sintel performed better than RAFT-KITTI which is why we choose RAFT-Sintel as baseline model for refinement. Due to the lack of publicly available datasets of facial microexpressions with ground truth OF annotation, we compared the methods by estimating a motion compensation error on a CASME II sequence from participant number 14⁸. We used the *mean squared error* (MSE) of the gray values as well as on the Laplacian, see Figure 2.14. While the performance of DIS OF was better on the test sequence than all other methods, our RAFT refinement with the ME dataset could generally improve performance on facial microexpression videos.

2.4.3.3 Motion magnification

For the motion magnification experiments, we use a video of participant number 14 from CASME II as well as a recording from the VI-SCREEN dataset. Participant number 14 has been excluded from the OF training dataset. The videos contain microexpressions and micromovements. We perform the analysis outlined in section 2.4.2.3 using sparsity regularization parameters $\alpha_{\text{sp}} = 0.1$, $\beta_{\text{sp}} = 4$ and $K = 9$ components. A motion component $k \in \{1, \dots, K\}$ is selected for magnification if $\lambda_1 \leq c^k \leq \lambda_2$. Here, λ_1 suppresses noise-level components and λ_2 excludes large motions such as blinks (or gross head motion). The absolute threshold values depend on the scale of c^k and thus on image resolution and frame rate. For the experiments, we use $\lambda_1 = 0.1$ and $\lambda_2 = 0.3$.

⁷FlowFormer was among the top ten best performing methods on the Sintel benchmark [102] as of July 2023.

⁸The benchmark video is EP04.04f from participant number 14.

In general, the double sparse decomposition depends on the number of components K and on the sparsity parameters $(\alpha_{\text{sp}}, \beta_{\text{sp}})$. The parameter α_{sp} weights the spatial ℓ_1 penalty on the components G^k and therefore controls how localized each component becomes. A larger α_{sp} leads to more compact supports and smaller α_{sp} leads to more diffuse components. The parameter β_{sp} bounds the temporal coefficients d^k via the $\|\cdot\|_{2,1}$ constraint and promotes temporally concentrated activations, which matches the short duration of microexpressions at the given sampling rates. We chose $K = 9$, $\alpha_{\text{sp}} = 0.1$, and $\beta_{\text{sp}} = 4$ empirically on representative short clips around the events such that the decomposition yields stable and interpretable components, e.g., mouth versus eyes. For longer recordings with more expected events or different frame rates, K and β_{sp} need to be adapted accordingly.

The thresholds λ_1 and λ_2 in the selection criterion for c^k determine which motion scale is treated as a micromovement: λ_1 suppresses noise-level components, whereas λ_2 excludes large motions such as blinks or gross head motion. We set $\lambda_1 = 0.1$ and $\lambda_2 = 0.3$ based on the observed separation of c^k values and by visual inspection of the resulting magnified videos. Shifting the thresholds deliberately selects different motion types (compare Figure 2.17). Finally, the magnification factor μ trades off visibility against warping artifacts; we use $\mu = 4$ as a conservative default for the reported results, while larger values can be used for qualitative exploration depending on image resolution, frame rate, and the amount of divergence (compare Figure 2.19).

Our experiments show that our method is able to successfully decompose the OF field into reasonable components (see [F5]). Using this decomposition, we are able to perform facial motion magnification selectively in an unsupervised manner. For the case shown in Figure 2.15 and Figure 2.18, the threshold corresponds only to the components connected to the micromovements around the mouth. The blinking motion is unchanged. Changing the thresholds also lets us select different motions. In Figure 2.17, we illustrate the results of our algorithm when choosing instead the thresholds $\lambda_1 = 0.3$ and $\lambda_2 = \infty$. In this case, only the blinking motion is magnified.

For comparison, Figure 2.18 depicts the outcome of a similar experiment using just the $\|\cdot\|_2$ -constraint, (see Remark 2, [F5]). For this comparison, we use the same parameters $\alpha_{\text{sp}} = 0.1$, $K = 9$ and $\mu = 4$. Introduction of the sparsity-promoting constraint leads to a clearer separation of motion components in the image as well as between temporally disconnected motions. We can observe this in the visualization of the components in Figure 2.16, where the $\|\cdot\|_{2,1}$ -constraint produces smaller components with clearer edges.

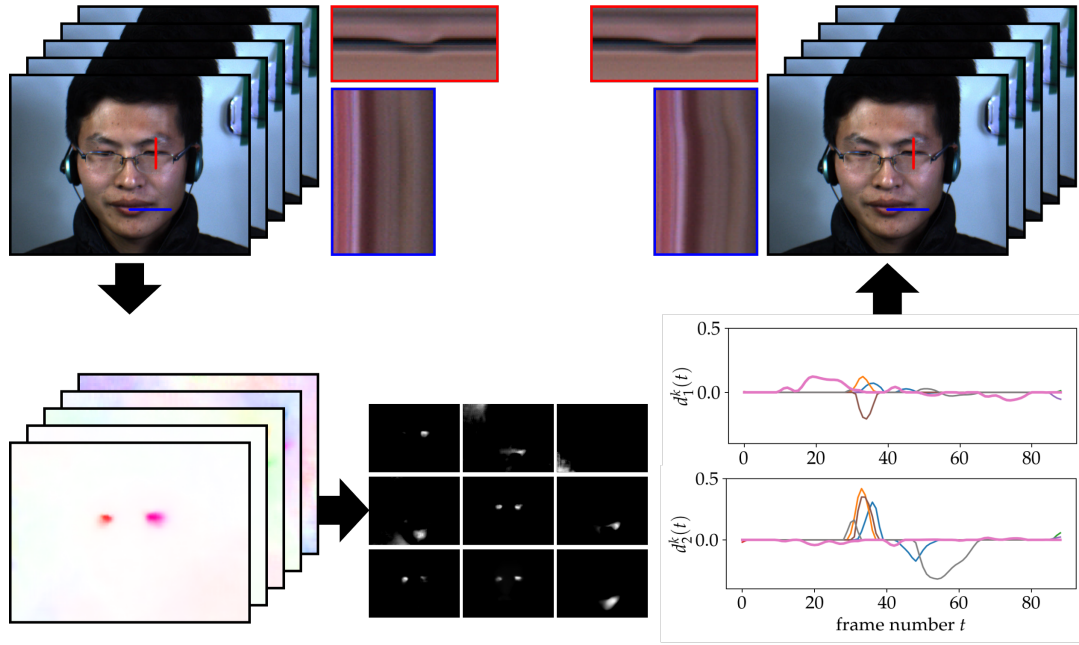


Figure 2.15: Illustration of our motion magnification procedure with the $\|\cdot\|_{2,1}$ -constraint. The algorithm starts with a sequence of 80 video frames of size 640×480 illustrated in the top left. The OF field computed for each time step is illustrated in the lower left. Then we use our method to decompose this OF field into $G^1, \dots, G^9$ shown in the bottom middle as well as the $(d_1^1, d_2^1), \dots, (d_1^9, d_2^9)$ in the bottom right. In this example, two components, shown by thicker lines in the plot, were selected by thresholding to be micromovements. The frames are warped accordingly to yield the video sequence in the top right. The pixels from the red and blue lines are plotted over time in the corresponding red and blue boxes. This shows that the motion around the mouth is magnified while leaving the blinking motion unchanged. Figure originally published in [F5].

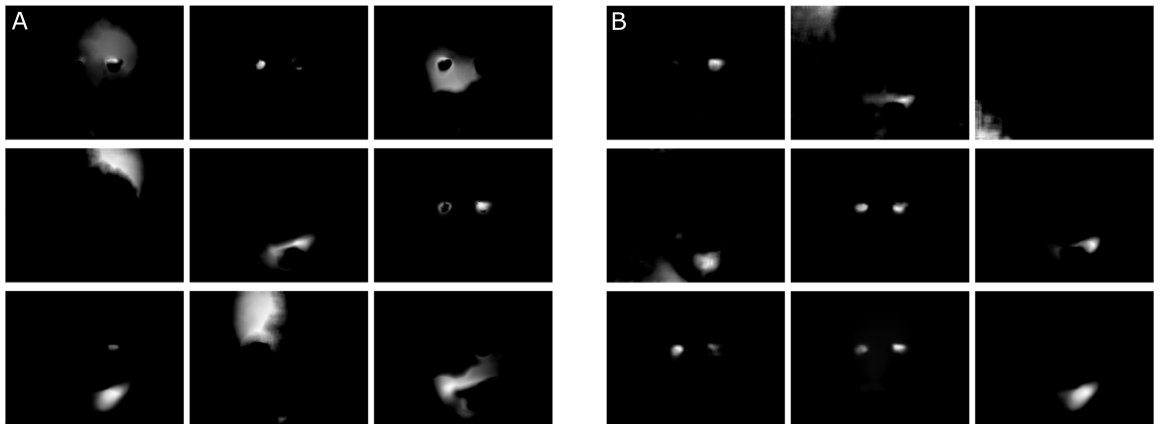


Figure 2.16: Illustration of the motion components $G^1, \dots, G^9$ with $\|\cdot\|_2$ -constraint (A) and with the $\|\cdot\|_{2,1}$ -constraint (B). The components in (B) show spatially smaller components with clearer edges. Figure originally published in [F5].

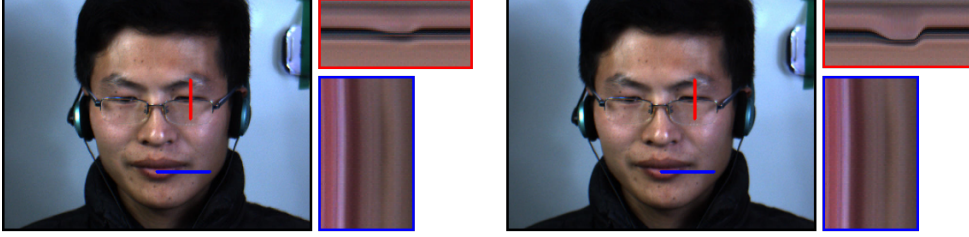


Figure 2.17: Illustration of the same video without (left) and with (right) motion magnification. The chosen thresholds ($\lambda_1 = 0.3, \lambda_2 = \infty$) only magnify the largest motions. This results in magnification of the blinking motion while leaving other facial movements unchanged. Figure originally published in [F5].

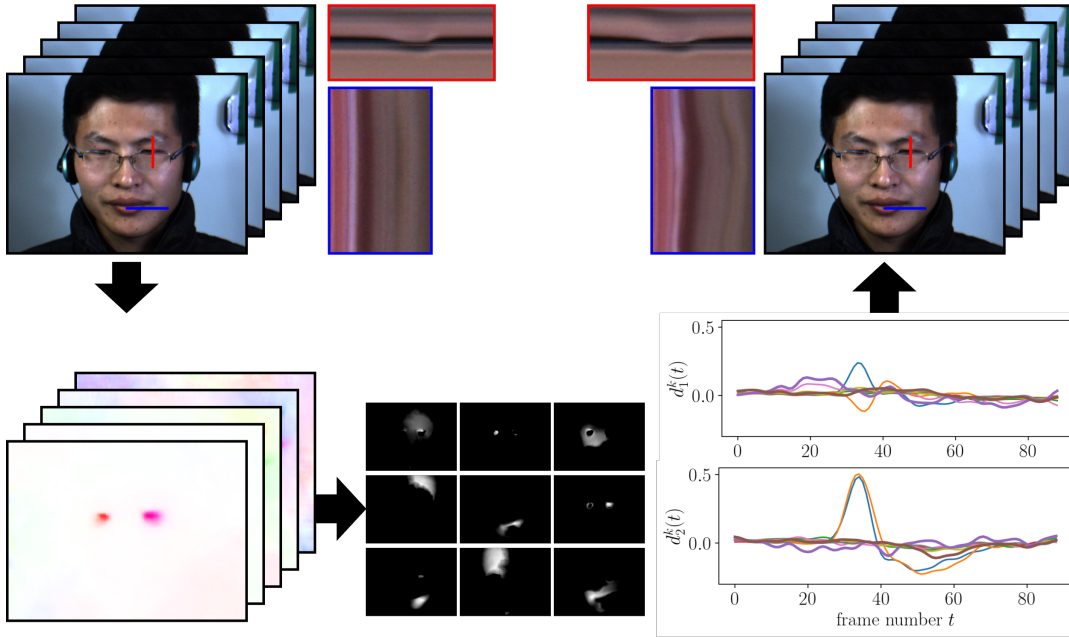


Figure 2.18: Illustration of our motion magnification procedure as in Figure 2.15, but with $\|\cdot\|_2$ -constraint. Figure originally published in [F5].

2.4.3.4 Landmark-based decomposition

As benchmark sequence, we use a HFR RGB recording from the DTSC dataset (see Table 2.1) that has been recorded at a SARS-CoV-2 drive-in testing station under real world conditions for micromovement analysis [F1] (see section 2.1).

The different motion decomposition strategies allowed us to selectively amplify local movements of facial features. Inpainting of the displacements in regions that have been removed with the help of facial landmarks allows the amplification of movements around the mouth and eyes which are embedded into global head movements (see Figure 2.20, (A)).

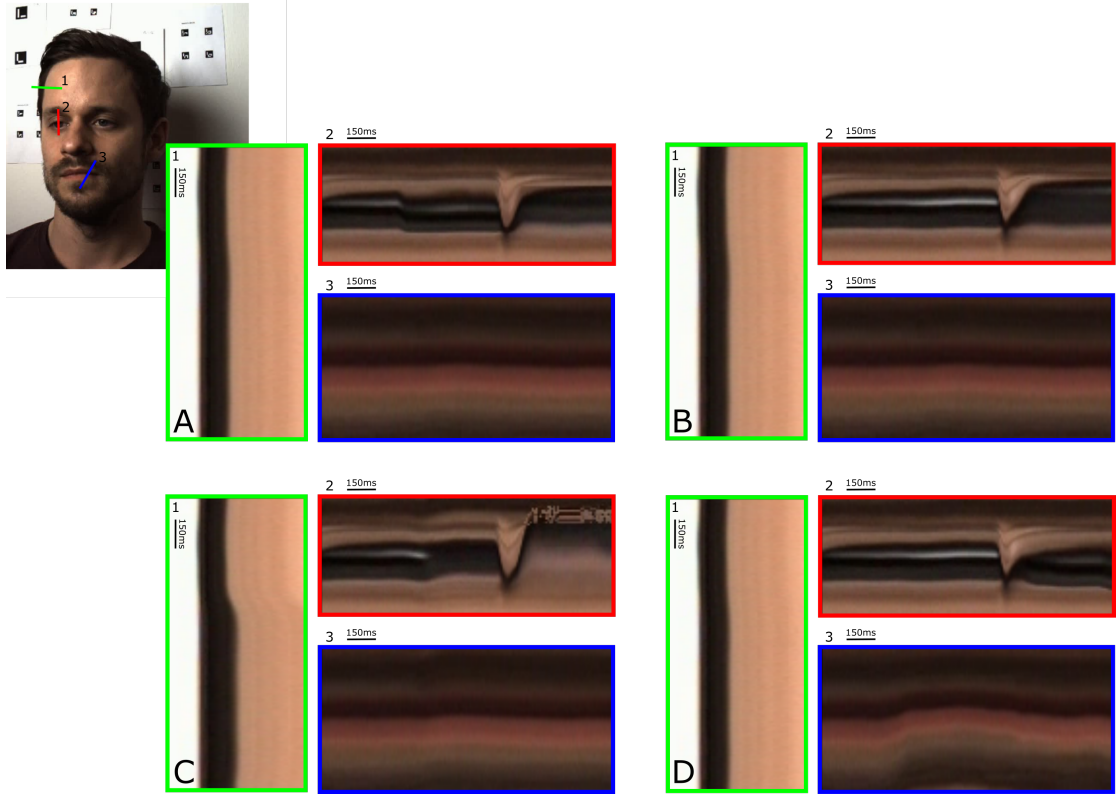


Figure 2.19: Experiments on the medical dataset with different parameters. The choice of parameters can be used to visualize different motion components when analyzing the video. (A,2): magnification of a subtle saccade before the blinking event with $(\beta_{sp}, \lambda_2, \mu) = (3, 0.2, 10)$; (B,2): magnification of a larger saccade after blinking with $(\beta_{sp}, \lambda_2, \mu) = (3, 0.4, 4)$; (C,1): magnification of the subtle head motion with $(\beta_{sp}, \lambda_2, \mu) = (4, 0.5, 8)$; (D,3): magnification of a micromovement of the mouth with $(\beta_{sp}, \lambda_2, \mu) = (8, 0.3, 10)$. All results are with $K = 9$ and $\lambda_1 = 0.1$. Figure originally published in [F5].

2.4.4 Conclusion

This section provides an unsupervised method for magnifying micro-motions in facial videos based on the Lagrangian approach with OF forward warping magnification. The actual regions for OF magnification were found by minimizing a double sparse decomposition model for OF and a thresholding procedure to detect the relevant regions. Our thresholding method for differentiating microexpression-related micromovements from other movements is open to further improvement. For example, the temporal extent of each motion could be taken into consideration. However, in case of the CASME II dataset, where eye movements are the only notable facial motions beside microexpressions, a simple thresholding was already sufficient. Furthermore, significant head-movements might disrupt the presented decomposition into time-independent regions of motion in the face. In these cases, a facial alignment pre-processing step might be necessary.

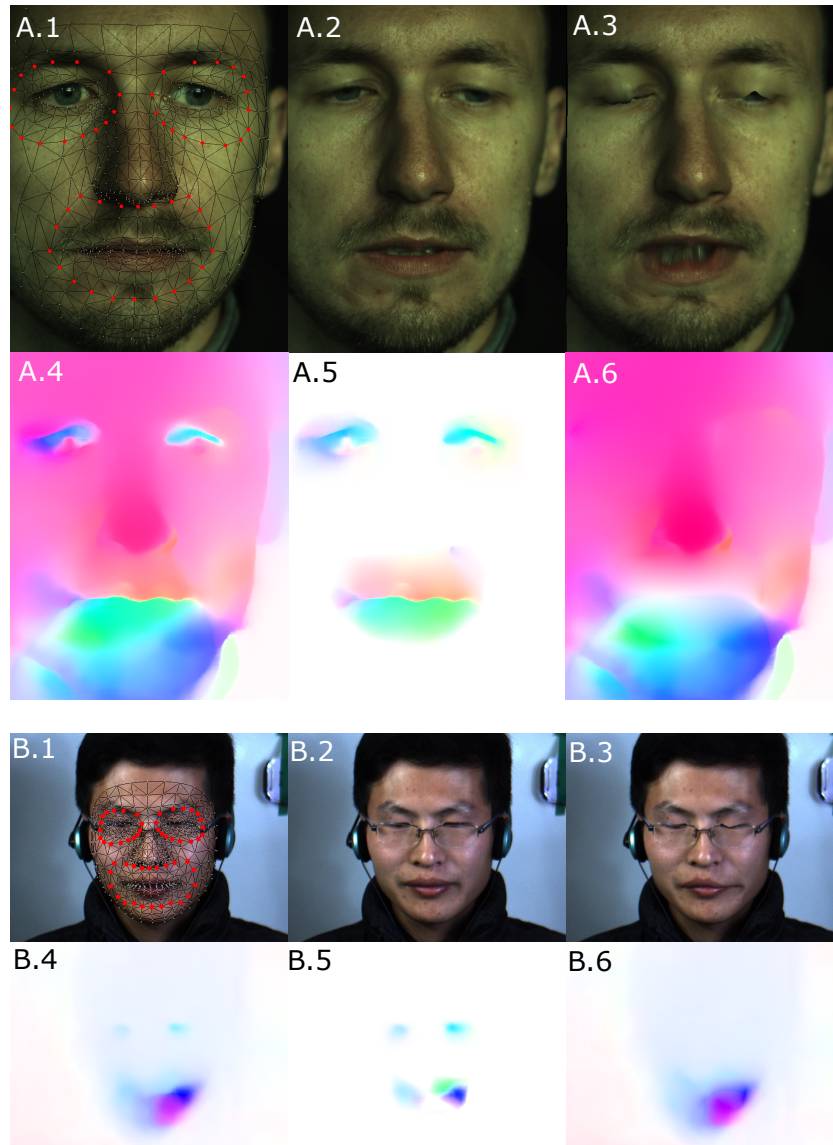


Figure 2.20: Landmark-based flow decomposition and magnification of the differences between two images f_1 and f_2 for two different videos A.1-A.6 and B.1-B.6. The area inside of the convex hull of the ROI face mesh vertices ((A.1, B.1), red dots) is removed from the calculated displacement field $\mathbf{w}$ (A.4, B.4) leading to the local (A.5, B.5) and global (A.6, B.6) movements. Local movements are amplified by a factor $\mu > 1$ into $\mathbf{w}_{\text{mag}} = \mu \mathbf{w}_{\text{local}} + \mathbf{w}_{\text{global}}$ and the final magnified image f_{mag} is computed by first aligning f_2 with f_1 and then forward warping the aligned image with $\mathbf{w}_{\text{mag}}$ (A.3, B.3). © 2022 IEEE.

For that purpose, we additionally proposed a facial landmark-based decomposition into local and global movements, where the local information is preserved within certain regions in the face [F12]. Our method has multiple parameters that need to be manually chosen. As a next step, statistics from real-world microexpression datasets could be included to learn suitable parameters for different motion types.

The presented methods have the potential to facilitate the annotation of datasets with such face recordings for microexpression detection. Subtle expression changes, where the detection would otherwise require expert training, could be magnified for untrained individuals, who could then annotate such data. Our preliminary results with a real world, medical recording demonstrated how different parameters would allow for the visualization of different motion components in videos.

To our knowledge, there exist no facial microexpression or micromovement datasets with ground truth OF. OF methods have been analyzed with respect to their predictive power of microexpressions [124] which does not allow conclusions on the overall quality of the OF predictions. Regarding the refinement of RAFT, future evaluations have to show the performance gain compared to OF methods tailored to faces with different training or transfer learning strategies such as [125] for which the models are currently not publicly available. Given the small movements in the SMIC and CASME II recordings as well as constant backgrounds, further data augmentations that introduce large displacements and different background textures could further improve the method. In this work, we only consider the forward flow for motion magnification. The combination of both flow directions could improve image quality in the context of occlusions and disocclusions in the future.

2.5 Markerless XROMM using DeepLabCut and marker masking

This section extends the approaches outlined in this chapter to a new modality, *X-ray reconstruction of moving morphology* (XROMM) in animal models. In particular, work with the multimodal camera setup as well as the landmark guided motion magnification laid the basis for the approaches in this chapter. The marker masking augmentation that is proposed here uses the same inpainting that is used in the previous section for the estimation of global movement components. The preparation of camera calibrations that were only implicitly available as well as the 3D reconstruction follow directly from the results in section 2.1. This section includes direct excerpts from [F13]⁹.

XROMM enables 3D tracking of implanted radiopaque markers in animals to study biomechanics. In recent years, *DeepLabCut* (DLC) has emerged as an important tool for markerless pose estimation in areas of the biological sciences such as neuroscience or behavioral research. While the XROMM markers can be detected and tracked using traditional image processing, DLC has been proven to improve XMALab-based workflows. However, these methods depend on precise marker positions and lack robustness when markers are missing.

The work in this section proposes to use XMALab results to reconstruct camera calibrations for the prediction of DLC-guided 3D XROMM markers in videofluoroscopic data. A novel marker masking augmentation enables tracking with incomplete marker sets. The results show that XROMM pose estimation is possible with subsets of visible beads and even without any visible markers without sacrificing accuracy.

2.5.1 Introduction

In recent years, markerless tracking via the deep-learning-based approach DLC [1] has become an essential tool for behavioral studies in biology and neuroscience. It allows tracking of user-defined pose models from a set of annotated examples using transfer learning. DLC natively supports 3D, markerless pose estimation [2] and there exists a rich model zoo with different pre-trained animal models. However, the predominant modality for DLC models is RGB and near infrared cameras. In contrast, studies of biomechanics often require accurate tracking of bone positions or soft-tissue deformations. *X-ray Reconstruction of Moving Morphology* (XROMM) enables 3D tracking of implanted radiopaque markers via high speed stereo x-ray cameras [126].

⁹[F13] was supported by NIH grants R03DE028395 and R03DE028395-02S1.

While the markers can be tracked using traditional low-level computer vision techniques, identification of marker positions remains challenging and tracking is prone to failure for intersecting marker trajectories [127].

Therefore, integration of XMA Lab with DLC was recently proposed [128]. While their approach integrates information from XMA Lab project files, the tool does not allow calibration of the X-ray cameras or 3D pose estimation without XMA Lab. Also, their approach is not robust concerning missing markers. XROMM experiments can last longer than a few months, during which the implanted markers need to stay in position. However, markers implanted in soft tissue can move and change their position, potentially leaving the implanted sites. Markers that are implanted in the bones can become exposed and even get lost during growth. In traditional deep learning approaches for RGB images, the state of the art heavily relies on image augmentations [43]. Image augmentations model real-world statistics that are not found in the training datasets and are important for generalization.

In this section, we propose a novel pipeline for using pre-annotated XMA Lab sessions for DLC dataset generation and training. Our approach does not require access to calibration information, instead, it exploits sets of pre-computed 3D points for the estimation of intrinsic and extrinsic camera parameters. We propose a diffusion in-painting-based masking augmentation for DLC training on XROMM videos. This improves XROMM tracking by allowing the use of previously discarded videos from animals with exposed beads and facilitates the training of markerless XROMM pose estimation models, eliminating the need for surgically implanted markers. In [F13], this work is furthermore extended by revealing new findings on 3D rodent tongue kinematics during natural feeding behavior.

2.5.2 Materials and methods

In this section, we will first describe the data acquisition process for XROMM recordings, train dataset generation, DLC interface, and the introduction of masking augmentation. Lastly, we outline analysis steps for extracting 3D tongue kinematics.

Data acquisition and dataset preparation. All surgical and experimental procedures were approved by the University of Chicago IACUC. Videos were recorded at 250 fps from two Sprague-Dawley rats during feeding. Radiopaque tantalum markers were surgically implanted into the cranium, hemimandibles, and tongue. Multiple recordings of each animal were annotated semi-automatically using XMA Lab. The marker positions were initially manually identified and semi-automatically tracked for each recording. Each annotated recording contains 2D marker positions per camera as well as

triangulated 3D positions and camera calibration parameters, all of which were saved in *.xma* project archives.

Fully manual annotation of large datasets can result in minor inconsistencies in naming conventions, in particular when marker names cannot be loaded from standardized definitions. We solve this by analyzing all combinations of substrings for each marker name and selecting the best-performing combination concerning a reference marker name via the Levenshtein distance.

2.5.2.1 3D XROMM DLC

Our pipeline exploits information in a dataset that has partially been labeled and calibrated with XMA Lab. The XMA Lab workflow includes camera calibration and estimation of 3D marker positions with calibration targets by matching markers on the epipolar lines. XMA Lab uses two different types of calibration patterns: A 3D calibration cube is used for estimating the extrinsic camera parameters and a 2D pattern that covers the region of interest for the estimation is used to estimate the intrinsic parameters of each camera. The results from the camera calibrations are stored and can be directly read from CSV files within the project archive. Instead of relying on calibration detections, we compute extrinsic parameters directly from the stored 3D marker positions via perspective-n-point pose computation through the OpenCV [117] stereo calibration framework and reuse the intrinsic parameters from the CSV files.

We undistort 2D points $\mathbf{p}_{\text{left}}$ and $\mathbf{p}_{\text{right}}$ using intrinsic parameters $\mathbf{K}_{\text{left}}$, $\mathbf{K}_{\text{right}}$ and distortion coefficients. The pose estimation process involves calculating the rotation $\mathbf{R}$ and translation $\mathbf{t}$ vectors that align 3D points to each camera's view, from which we compute the camera projection matrices $\mathbf{P}_{\text{left}}$ and $\mathbf{P}_{\text{right}}$. Integrating intrinsic parameters and projection matrices enables the triangulation of 3D points from corresponding 2D point pairs that are detected separately for each camera.

One of the strengths of DLC is its capability to generalize new data through transfer learning and, therefore, effectively reduce the required amount of training data. To train DLC, we constructed two datasets by subsampling the annotated recordings **DLC stride 100** (2704 frames) and **DLC stride 400** (728 frames). Additionally, we extend **DLC stride 100** and add images with a random subset of the markers being masked **DLC stride 100 augmented** (18564 frames) as described in the next section.

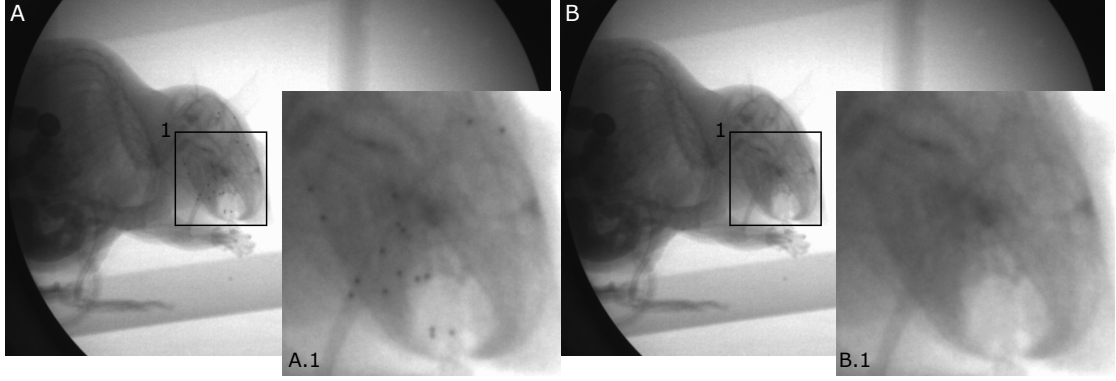


Figure 2.21: The implanted, radiopaque beads are visible as the small, grey ellipses (A.1). The diffusion-based masking operation completely removes the beads without visible manipulation of the X-ray morphology and texture (B.1). Implemented as a static augmentation, we extend the dataset with images where random subsets of markers have been removed. © 2024 IEEE.

2.5.2.2 Landmark-guided masking augmentation

The datasets generated from XMA Lab are limited to the marker positions of the implanted radiopaque beads. However, those beads can be exposed and get lost over time. To counter this, we propose a masking augmentation that removes and inpaints the area around the ground truth marker positions.

We apply diffusion-based image inpainting on ROIs around the radiopaque marker beads, defined by a 5-pixel radius neighborhood, for a random subset of marker positions (see Figure 2.21). Let $\Omega \subset \mathbb{R}^2$ denote the image domain and let $\mathcal{R} \subset \Omega$ denote the union of these ROIs. We define the inpainted image f_{inpaint} by keeping the background unchanged, i.e. $f_{\text{inpaint}}(\mathbf{x}) = f(\mathbf{x}) \forall \mathbf{x} \in \Omega \setminus \mathcal{R}$, and by diffusion-based inpainting inside $\mathcal{R}$. With Dirichlet boundary conditions specified as $f_{\text{inpaint}}(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x} \in \partial\mathcal{R}$, the inpainting process is formulated analogously to Equation (2.10) as solving the Laplace equation within the ROI:

$$\Delta f_{\text{inpaint}}(\mathbf{x}) = 0 \quad \text{for } \mathbf{x} \in \mathcal{R}, \quad f_{\text{inpaint}}(\mathbf{x}) = f(\mathbf{x}) \quad \text{for } \mathbf{x} \in \partial\mathcal{R}.$$

This process is applied for n randomly sampled markers as well as once for all marker positions, resulting in datasets that include frames with varying numbers of visible markers for each input video. This augmentation strategy aims to enhance the robustness of the models concerning experiments with partial sets of markers and, potentially, full markerless pose estimation.

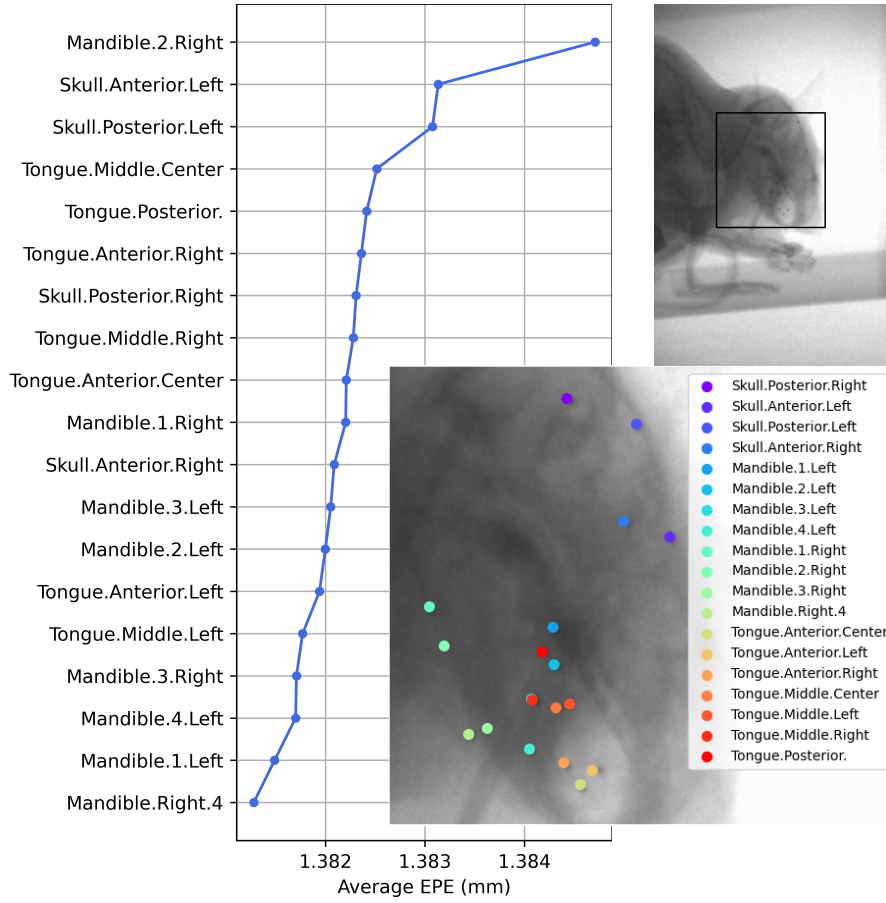


Figure 2.22: Average expected prediction error (EPE) with each respective marker masked out, sorted from highest to lowest error. While higher errors for soft tissue markers are plausible due to the lower contrast, the lower performance of the skull and mandible markers could either be explained by variation in marker placement compared to the train set or different skull texture in the X-ray image. © 2024 IEEE.

2.5.3 Results

Our pipeline reconstructs 3D marker positions from DLC predictions on unseen data, including data from an individual animal that has been excluded from the training. The error is 1.380 ± 0.014 mm for the smaller training dataset and 1.382 ± 0.013 mm for the larger dataset, see Table 2.7. We achieve a similar performance with the **DLC 100 augmented** model.

Markerless XROMM 3D. Our results indicate that DLC exploits the visible marker features in the XROMM recordings for pose estimation. The performance of the models without augmentation is significantly less without the markers while the performance with masking augmentation is similar on the original frames and less affected by the removal of the markers (see Table 2.7).

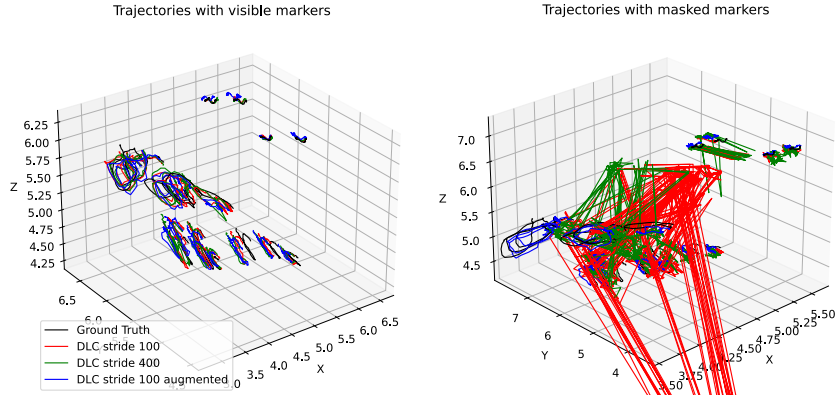


Figure 2.23: 3D trajectories with (left) and without (right) visible markers. Without the masking augmentation, DLC 100 and DLC 400 rely on the marker features and produce large outliers without visible markers that prevent correct triangulation. © 2024 IEEE.

In Figure 2.23, the trajectories form coherent clusters with visible markers, however, when inpainting the visible markers, only **DLC 100 augmented** aligns with the ground truth trajectories.

Model	Avg. EPE (markers)	Avg. EPE (no markers)
DLC 100	1.382 ± 0.013 mm	$1.830^* \pm 0.454$ mm
DLC 400	1.380 ± 0.014 mm	$1.490^* \pm 0.232$ mm
DLC 100 augmented	1.382 ± 0.017 mm	1.383 ± 0.018 mm

Table 2.7: Comparison of Tracking Performance with and without Markers over 50 frames. Significance has been computed with a paired t-test between the *marker* and *no marker* conditions and is indicated with * ($p < 0.01$).

Exposed and lost beads. Implanted radiopaque beads are often fixed in position over several months. However, markers can become exposed from the tissue over time, in particular, in soft tissue or the presence of underlying, physiological issues. Hence, the tracking of incomplete sets of markers carries high real-world significance and would enable an analysis of videos otherwise unattainable through manual annotation. We analyzed the performance of **DLC 100 augmented** with each of the 19 markers removed (see Figure 2.22).

The results indicate a low impact of missing markers on the overall tracking performance. The importance of different marker positions could be explained by variation in marker placement compared to the train set or different skull texture in the X-ray image for the bone markers and with lower contrast for markers in soft tissue.

2.5.4 Conclusion

3D marker XMA Lab workflow can be automated with DLC. We have proposed a marker masking augmentation and shown its ability to robustify models for missing markers or markerless videos.

The masking augmentations allow for XROMM experiments without or only with subsets of implanted radiopaque beads. This approach could eliminate hard-to-find, static positions, prioritizing soft tissues with minimal X-ray contrast.

While our results indicate the possibility of markerless XROMM 3D tracking with our pipeline, so far, we have only evaluated the model with artificially removed markers.

2.6 Broader impact

The methods developed in this chapter have resulted in multiple interdisciplinary collaborations and the described methods from this chapter found broader application in different projects, besides the ones described so far. This section contains preliminary work presented at conferences, as well as contributions to a study on auricular muscle activity, where the camera system demonstrated stimulus evoked, submillimeter motion of the ear [F7]. Conference contribution [F17] is concerned with an application of the proposed landmarking solution from section 2.2 within the detection of infections with respiratory diseases. Abstract [F16] discusses the impact of compression for the archival of large-scale datasets of faces. In particular, the accuracy of OF estimation on videos with and without different levels of compression is investigated.

2.6.1 Quantitative motion analysis in a study for auditory attention

Auditory attention research often focuses on accurately estimating the direction of a listener's attention. Traditional approaches analyze electroencephalography (EEG) measurements to assess auditory focus [129]. More recently, it has been suggested that vestigial ear muscles can contain cues on listening direction [130, F7]. These muscles, which actively orient the ears in many animals, typically do not produce visible ear motion in humans. In our study on auditory attention, we have applied a high-performance multi-camera setup to capture quantitative motion data. Stereo recordings of each ear were synchronized with electromyograms (EMGs) and Sounds were presented from speakers positioned at various angles around the participant (e.g., $\pm 30^\circ$, $\pm 120^\circ$) to elicit auriculomotor responses.

This section provides an overview over the camera-based aspects of publication [F7].

Camera setup. The camera setup consists of four Ximea MQ022CG-CM RGB cameras that are arranged in two stereo configurations, one on each side of the head. Calibration was performed with a checkerboard pattern mounted on a wooden board and utilizing the stereo camera calibrator app from the Mathworks Matlab Computer Vision System Toolbox. Calibration was performed after each modification of the camera setup. As optics, KOWA 35 mm macro lenses with an aperture of F0.4 were used. This allowed the cameras to record close-up views of each ear at enough distance to not through shadows and with acceptable depth of field.

The ears were uniformly illuminated with flicker free studio illumination and recordings were performed at 1936×1216 pixels resolution and 120 frames per second framerate and 2 ms exposure.

The cameras were synchronized through hardware triggering of all four cameras. The camera trigger signals were recorded to allow synchronization with the EMG data.

Results. The proposed setup was suitable for extracting sub-millimeter motion of the pinna after exogenic stimulation with transient sounds (responding to novel sounds while reading) [F7] and 3D motion magnification (see supplemental videos in [F7]).

The presented results in [F7] demonstrate the suitability of the proposed computer vision setup for video acquisition during studies from auditory attention to study ear motion.

2.6.2 Core temperature estimation

Remote respiratory disease screening has gained significant attention during the COVID-19 pandemic. Thermal cameras are a widespread modality for remote screening [39]. We propose a novel pipeline for core temperature estimation using RGB-thermal fusion by means of dense facial alignment. To analyze the potential for core temperature prediction, facial surface temperatures, estimated similarly as in section 2.1, are compared with *ground truth* (GT) in-ear temperatures of 106 participants in a study for remote respiratory disease screening (VI-Screen).

Data acquisition. The same VI-Screen data, which contains the samples in section 2.4, is used here [F5]. Participants were recruited at multiple recording sites, including Saarland University Clinic (UKS) Children’s Hospital and at the UKS central emergency admission, Germany. Recruitment happened from February 2022 to April 2023. This evaluation includes 106 participants (median age 26 years, *interquartile range* (IQR) 24 years). Informed consent was obtained, and the study was approved by the Ethics committee of the Ärztekammer des Saarlandes (02/22). Viral and bacterial respiratory pathogens were tested using multiplex polymerase chain reaction (Multiplex PCR) assays.

The camera array consisted of stereo RGB, NIR and thermal cameras as well as microphones, similar to section 2.1. A duration between 5-10 minutes was recorded where participants answered questions in front of the camera. Core temperature was recorded with infrared ear thermometers.

Core temperature prediction. We compute a spatial normalization of the faces in the RGB and thermal images by means of dense, facial landmark prediction (see Figure 2.24 (C)). We use MediaPipe detections for the RGB and the method for thermal landmark prediction described in section 2.2 for the thermal images (see Figure 2.24, A for MediaPipe and B for dense thermal landmarks — green dots correspond to high confidence and red dots to low confidence values).

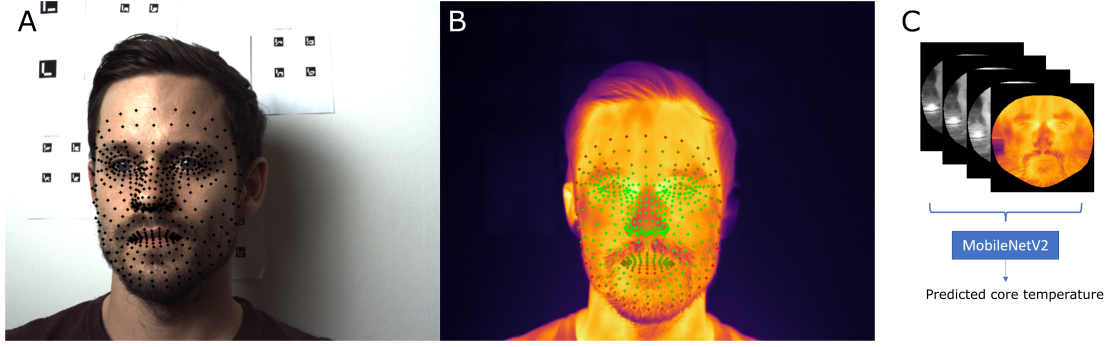


Figure 2.24: Proposed processing flow: RGB frame with predicted facemesh landmarks (A), thermal recording with predictions from our dense thermal landmark predictor and color coded uncertainty (B). The input frames are spatially normalized and fed into a MobileNetV2 with 4 input channels for temperature prediction [75] (C). © 2023 IEEE.

The spatially normalized and aligned images form a hyperspectral image stack that can be used to train a MobileNetV2 [75] head with 4 channel input to predict core temperature estimates.

Results. To assess the similarity of remote and GT temperatures, we correlate the ear GT temperature measurements with peri-orbital temperature estimates, a position in the face that is employed in the literature for thermal imaging-based body temperature assessment [F1]. Contactless measurements (naive max temperature) show an elevated temperature of 37.5°C in 15% of the negative group, 24% for the positive group (including SARS-CoV-2) and 10% for SARS-CoV-2 positive patients. GT ear measurements show 6% for the negative group, 21% for positive and 0% for SARS-CoV-2. Preliminary evaluations indicated that the proposed architecture can improve temperature prediction over naive temperature, however, an in depth ablation study is part of future work. Future work will evaluate the method on the final dataset with more than 200 participants with an ablation study to demonstrate advantages of RGB-thermal fusion for remote core temperature estimation.

2.6.3 Influence of video compression on OF estimation

This chapter shows how video recordings of faces are an important modality for biomedical research. However, one of the challenges with videos is the requirement for storage space. The size of the datasets collected with the multi-camera setup (see Table 2.1) ranges from tens to hundreds of terabytes. While lossy compression of video datasets can achieve significant compression ratios, information that is relevant in a psychophysiological context may be lost. Hanfland et al. [131] analyzed the effect of different compression algorithms on rPPG signal extraction. For that purpose, they analyzed Motion JPEG, MPEG-4 and Motion JPEG 2000 (lossless) on different ROIs of short recordings at 20 fps. Here, we analyze the effect of compression on the endpoint error (EPE) of dense optical flow estimated with RAFT [87].

We use 60s of a high-frame-rate (HFR) video recording of the face that has been recorded at the SARS-CoV-2 testing station (see section 2.1 [F1]). Compression is performed with lossless HDF5 compression (gzip) as well as with motion JPEG (MJPEG) and MPEG-4 codecs with different parameters. We use the Lvc59.18.100 libx264 encoder for H.264 with ffmpeg which uses global motion compensation and a variable bitrate for compression as well as the Lvc59.18.100 mjpeg encoder.

For the uncompressed sequence $f(x, y, t)$, compressed sequences $f_c(x, y, t)$ and frames $0 \leq t < T$ we compute the optical flow field $w(x, y, t)$ between $f(x, y, 0)$ and $f(x, y, t)$ as well as w_c between $f(x, y, 0)$ and $f_c(x, y, t)$. Mean EPE is estimated as the mean distances between the vectors in w and w_c .

Results. Evaluations were performed on a Windows machine with Python 3.8.11. Qualitative results show that high compression ratios for MPEG-4 result in large block artifacts at the beginning of the recording that get better towards the end of the recording. This indicates that higher maximum bitrate could improve the overall performance of recordings with many redundant frames. For MJPEG, higher compression resulted in JPEG compression artifacts on all frames equally.

Using the lossless data compression filters in hdf5, we could achieve compression ratios of 84% (lzf) and 60% (gzip) where LZF compression took around 52.1s, gzip around 1403.8s against 26.3s (AMD Ryzen 9 3900X and 860 EVO 1TB) for writing the file without compression. For lossy compression, the images were demosaiced first and the single raw channel is interpolated to the three RGB channels with identical resolution. For lossy compression, AVI with 983040kbps achieved a compression ratio of 5.7%.

Average EPE calculated on the first 5 s shows that less compression correlates with lower EPE values which indicates that the lowest possible compression is required here as well: A bitrate of 200k MPEG-4 resulted in an average EPE of only 4.8 pixels, while 5M is already only at 0.28 and 100M is at 0.14. The highest MJPEG compression setting resulted in 0.36 pixels EPE and the lowest in 0.17.

Discussion. Not only are rPPG signals impacted by lossy compression. This section discusses how motion estimation requires low compression ratios to preserve small movements. Average EPE is a common performance metric for OF estimation and the EPE values correlate with compression settings, while higher MJPEG compression had less impact on the results than high MPEG-4 compression which could be due to the fact that MJPEG does not encode motion information.

2.7 Summary

This chapter describes a multimodal camera system suitable for the acquisition of large RGB-NIR-thermal datasets in psychophysiological studies. The datasets acquired with this system are among the largest datasets available in the context of thermal imaging. The pairwise calibrated cameras allow the computation of dense correspondences between frames of different modalities which in turn allows full annotation with facial landmarks from the RGB images. Evaluations on the dataset recorded in the context of SARS-CoV-2 drive through screening centers, which were installed in the first year of the COVID-19 pandemic throughout Germany [F1], demonstrate the suitability of the setup for the acquisition of several relevant vital parameters. At the same time, these evaluations revealed limitations of existing approaches. The performance of existing facial landmarks in thermal images is a limiting factor for out-of-lab facial temperature assessment. Building on those insights, a multimodal model for dense facial landmark detection with state-of-the-art performance in RGB and thermal images was developed.

To analyze facial micromovements, this chapter proposes new approaches for Lagrangian motion magnification. Due to the spatiotemporal properties of those non-periodic and isolated, small movements in the face, we focused on the development of methods for the manipulation of displacement fields and motion decomposition. The methodological advancements in this chapter solve the problem of motion magnification of small, local movements in the face during head movement. Furthermore, the methods allow the unsupervised identification of spatiotemporal motion components.

The insights and strategies towards multimodal imaging and cross-domain challenges span various applications beyond the core contributions. Principles of camera geometry have been applied for learning markerless pose estimation on stereo XROMM recordings. A multi-camera setup was used for motion quantification in a study exploring the relationship between vestigial ear muscles and auditory attention direction. Finally, we have outlined a potential use case for dense thermal landmarks in core temperature assessment and discussed the impact of image compression on preserving small motions for archiving.

This chapter takes a consistent approach towards cross-domain adaptation in each of the presented methods. Limitations of existing methods for neuroimaging applications are identified, and the gap between current state-of-the-art methods, datasets, algorithms, and task requirements is addressed with methodological advancements. Additionally, the methods in this chapter build on robust baseline approaches rather than reinventing the wheel. This chapter makes an important contribution to large-scale image acquisition with multi-camera systems in out-of-lab settings and to dynamic image analysis in RGB and thermal imaging within medical and neuroscience applications.

Chapter 3

Motion correction

The removal of unwanted motion in videos is called motion correction or motion compensation. It is an important pre-processing step that is part of most neuroimaging processing pipelines. In contrast to motion magnification, where the goal is to visualize tiny movements, the goal of motion correction is the removal of all motion in videos. There are many different sources for unwanted motion or displacements in videos and related data with temporal resolution such as linescans or single trial matrices (see section 3.1). While there exist many methods to correct simple movements such as translations, existing methods struggle with a combination of low *signal-to-noise ratio* (SNR) and difficult movements such as the displacements that happen during injection into tissue.

This chapter describes new methods for motion correction by means of dense displacement fields. Similar to the previous chapter, the methodological advancements are developed by analyzing the gap between existing methods and the requirements of the neuroimaging videos. The methods are based on *optical flow* (OF) estimation that is optimized for the statistics of different neuroimaging data.

This chapter contributes to motion correction of *2-photon* (2P) microscopy videos with the development of two variational OF-based methods along with efficient implementations in an easy-to-use toolbox and imaging frameworks. The first method handles registration of optical imaging 1D linescans as well as the non-constant alignment of single trial *evoked response potentials* (ERP) matrices. The second method corrects 2D 2P microscopy videos with state-of-the-art accuracy. The corresponding *Flow-Registration* toolbox handles various types of videos and works with different channel configurations. Building on the experiences with *recurrent all-pairs field transforms* (RAFT) for OF and OF dataset generation from the previous section, an approach for synthetic label generation tailored to RAFT-based 2P motion correction is proposed.

The chapter is structured as follows: The first section presents results from the publication [F4] and the second section from [F3]. Building on the insights from 2P motion videos, a deep learning-based approach with synthetic label generation for motion correction of 2P imaging data is proposed. Finally, the broader impact of the methods is discussed and the motion correction approaches and optical imaging methodology are applied to psychophysiological RGB videos from [F8].

The chapter contains direct excerpts from the first author publications [F4, F3] (see 1.3.1 for the unaltered, signed, and published author contributions for each of the journal publications). In particular, section 3.1 is taken from [F4] and section 3.2 is taken from [F3]. Section 3.3 contains preliminary results for deep learning-based motion compensation of 2P videos. Section 3.4 contains excerpts from [F8].

3.1 Variational 1D alignment

In the context of signal analysis and pattern matching, alignment of 1D signals for the comparison of signal morphologies is an important problem. For image processing and computer vision, 2D OF methods find wide application for motion analysis and image registration and variational OF methods have been continuously improved over the past decades. Existing methods for 1D alignment target mostly constant displacements, do not allow native subsample precision or precise control over regularization or are slower than the proposed method.

This section proposes a variational method for the alignment and displacement estimation of 1D signals. The estimation of non-flat displacements is posed as an optimization problem with a similarity and smoothness term similar to variational OF estimation and solved with efficient optimization strategies that allow real-time applications on consumer-grade hardware.

The method is applied to two downstream tasks from functional neuroimaging: The alignment of 2-photon imaging line scan recordings and the denoising of evoked and event-related potentials in single trial matrices. For both, it achieves state-of-the-art results in terms of alignment quality and computing speed.

This section contains direct excerpts from [F4]¹.

3.1.1 Introduction

OF estimation methods form an important family of image processing algorithms with applications in many areas where motion analysis, tracking or registration is needed. In the context of OF estimation of 2D natural images and image registration, *variational methods* form a powerful group of well-studied methods that perform with very high accuracy in the presence of small displacements [100]. Up until today, they find applications in state-of-the-art OF algorithms, where they can be used for the refinement of small displacements [98] or for the generation of ground truth displacement fields [132], [133]. They are able to estimate non-parametric, dense and smooth displacement fields with subpixel accuracy [100].

Variational methods in image processing are a group of image processing methods, that aim to find solutions for ill-posed image processing problems by minimization of an energy functional.

¹The approach is implemented as a MATLAB toolbox and available online from https://github.com/phflot/variational_aligner. It is suitable for 1D alignment problems, where high accuracy and high speed is needed and non-constant displacements occur.

The optimization variable is an image with the desired properties such as smoothness or similarity to a given input. They find application in many image processing fields, for example image denoising [134]. In variational OF estimation, the optimization variables are the components $(u, v)^\top$ of a dense displacement field (see Equation 3.1). A data term D defines the similarity of a moving image to a reference (constancy assumption) and a regularization term S assures smooth solutions in the presence of ambiguities (aperture problem) [19]. The data term and regularizer are modeled with respect to the data characteristics and the expected type of motion.

$$E(u, v) = \int_{\Omega} \underbrace{D(u, v)}_{\text{Data term}} + \underbrace{S(u, v)}_{\text{Regularizer}} dx dy \quad (3.1)$$

Most commonly, variational OF methods are formulated as the minimization of this energy functional and analyzed via the Euler-Lagrange framework. The Euler-Lagrange equations (ELEs) provide first-order necessary optimality conditions for a minimizer of the energy functional E . The constancy assumption in the data term is linearized. However, the linearized constancy assumption is usually only valid for small displacements. Hence, the resulting incremental ELEs are solved with a coarse-to-fine strategy on a Gaussian pyramid whose warping scheme can be interpreted as minimizing the non-linearised constancy assumption [19]. For strictly convex penalizers in the data and smoothness terms, this yields a sequence of convex programming problems [100, 19, 135].

Imaging areas, that require very high frame rates or spectral rather than spatial resolutions are suitable for the application of line scan sensors. Line scans are sequences of 1D recordings typically of a scene under static conditions or of a scene with camera or object movement orthogonal to the spatial dimension. Among others, they find application in industrial quality inspection [136] and biomedical imaging [137]. There are many sources for movement artifacts in line scan recordings: In the context of moving objects, poor synchronization, or jitter in the object movement results in non-uniform sampling. Brosch et al. [136] have proposed a variational formulation to estimate the warping function in the context of movement jitter for light field line scan data to reconstruct a recording that is sampled at even intervals.

Movements parallel to the spatial dimension result in shifts of the signal, that can often be approximated by constant displacements. Depending on the camera and lens / illumination setup, movements towards the camera can have additional effects such as linear displacements due to zooming or changes in signal morphology.

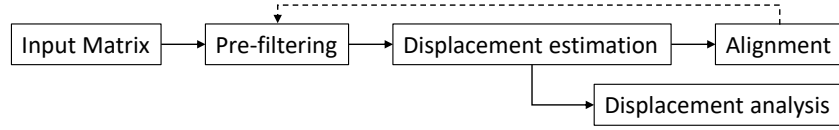


Figure 3.1: Exemplary flow chart for the alignment of 1D signals. A 2D input matrix is filtered and the non-flat displacements are estimated. The displacement map can be further analyzed and used for the alignment of the raw signal.

Within the scope of single trial analysis of *evoked and event-related potentials* (EP and ERP) in *electroencephalography* (EEG) recordings, event locked responses are analyzed [138]. The average response may have a high *signal-to-noise ratio* (SNR) but also a lower resolution of the signal. From the average signal it is also not possible to analyze temporal variations across trials. Popular approaches to increase SNR in ERP recordings for single trial analysis are often motivated by image processing methods [139]. This is because the single trial matrix representation of ERP recordings resembles the 2D representation of line scan data: Each line is a measurement with high temporal resolution of a stimulus locked event.

To remove constant displacements in 1D data, cross-correlation maximization with respect to a reference profile finds wide application. Methods to compare two arbitrary 1D signals are often based on *dynamic time warping* (DTW) [140]. The idea of DTW is to repeat individual samples of both signals to minimize a distance metric. This way, distinct features of the two signals can be directly compared. Nielsen et al. [141] introduced a method where they extend DTW with piecewise cross-correlation. Their correlation optimized warping approach is applied for the alignment of chromatographic data [142], [143].

In their preliminary work, Deriso et al. [144] have presented a variational formulation of the DTW problem. In their method they optimize the complete non-parametric warping function. They solve the non-convex optimization problem with dynamic programming in a multigrid framework.

In this section, we investigate non-parametric displacement estimation for 1D data and describe a variational method for non-parametric, dense displacement estimation of 1D signals, which is motivated by variational OF estimation (see Figure 3.1).

By exploiting the linear motion assumption within a coarse-to-fine strategy, our approach solves a sequence of convex problems. We demonstrate the performance of our algorithm on two problems from neuroimaging with alignment of optical imaging line scan recordings and the compensation of latency jitter in ERP recordings. Our implementation is fast enough to allow real-time alignments on consumer-grade hardware.

3.1.2 Materials and methods

In this section, we describe our method for the alignment and displacement analysis of 1D signals. To this end, we propose a variational model and pre-processing strategy and describe our implementation.

We work with 2D matrices consisting of multiple 1D signals which can be temporal (EP and ERP alignment) or spatial (line scan alignment). Let M be the length (spatial or temporal) of the 1D signal, N be the number of measurements and $c \in \{1, \dots, C\}$ is the channel index for data with $C \in \mathbb{N}^+$ channels. Then an individual measurement for channel c is given by $f^c(x) : \Omega_1 \rightarrow \mathbb{R}$ with $x \in \Omega_1 = (0, M) \subset \mathbb{R}$. The 2D matrix representation of multiple measurements is given by the function $f^c(x, j) : \Omega_2 \rightarrow \mathbb{R}$ with $\Omega_2 = \Omega_1 \times \{1, \dots, N\} \subset \mathbb{R} \times \mathbb{N}^+$. For $j \in \{1, \dots, N\}$ we denote the j -th measurement by $f_j^c(x) := f^c(x, j)$. We depict the reference profile with $g^c(x) : \Omega_1 \rightarrow \mathbb{R}$. The discrete version of g is denoted by g_i^c and of f by $f_{i,j}^c$. For simplicity, we depict partial derivatives with lower case letter x and use the lower case letters i, j exclusively for discretized access. Continuous access to the discrete data is realized via bicubic interpolation as recommended by [100].

3.1.2.1 Variational model for 1D alignment

We formulate the problem of aligning 1D signals as a variational problem, where we want to find displacements $u(x, j) : \Omega_2 \rightarrow \mathbb{R}$ of 1D signals f with respect to a reference g . We assume gradient constancy between f and g as well as small gradients in u . We penalize both assumptions with the generalized Charbonnier penalty function Ψ_a . Let $a > 0$ and $s \geq 0$ be a squared residual, then, for a small $\varepsilon > 0$, the generalized Charbonnier penalty function [100] is given by

$$\Psi_a(s) = (s + \varepsilon)^a. \quad (3.2)$$

Following the authors in [100], we choose the penalization of the data and smoothness terms with $a_s, a_d \in [0.45, 1] \subset \mathbb{R}$, where $\Psi_{0.5}$ gives the Charbonnier penalty function and Ψ_1 is quadratic penalization. $\Psi_{0.45}$ is slightly non-convex but has been shown to perform well in the context of OF estimation [100].

The energy functional for 1D displacements of single channel c and measurement j is then given by

$$E(u(\cdot, j)) = \int_{\Omega_1} \Psi_{a_d} \left((g_x^c(x) - f_x^c(x + u(x, j), j))^2 \right) + \alpha \Psi_{a_s} \left((u_x(x, j))^2 \right) dx \quad (3.3)$$

where the relation of the data and smoothness term is weighted by the smoothness weight parameter $\alpha \in \mathbb{R}_{>0}$. The displacement field u is then calculated as $\text{argmin } E(u)$. To achieve that, we linearize the constancy assumption into $u \cdot f_{xx}^c + f_x^c - g_x^c$ and calculate incremental updates on each level of a Gaussian pyramid.

The linearized assumption is a good approximation for small displacements and it has been shown that the final solution of such a scheme approximates the unlinearized constancy assumption [19]. Gradient constancy with subquadratic penalization is invariant under global additive changes of the signal and robust with respect to outliers. We normalize the input data, which can additionally compensate for multiplicative changes. In the context of variational multimodal image registration, there exist intensity similarity measures [145], that allow affine (cross correlation), functional (correlation ratio), and arbitrary (mutual information) dependence of the two signals. Cross correlation is a commonly used metric for 1D signal analysis. Additionally to the gradient constancy assumption, we implement a cross correlation data term as presented in [145] with isotropic smoothness term, however the gradient descent scheme from [145] has much slower convergence than our gradient constancy model.

The variational formulation allows a straightforward extension to data with multiple channels. Given multichannel data $(f^c)_{c=1}^C$ and a measurement j , we can extend our data term with separate penalization on each channel as in

$$\sum_{c=1}^C p^c(x) \Psi_{a_{\text{data}}} \left((g_x^c(x) - f_x^c(x + u(x, j), j))^2 \right) \quad (3.4)$$

The weight term $p^c : \Omega_1 \rightarrow \mathbb{R}$ is shared across measurements and multiplied pointwise with each channel and can be used to integrate knowledge about the information content / SNR per channel and to change the weight of each channel or sample.

To our knowledge, there exist no methods for the alignment of 1D signals that natively support the integration of multiple channels to calculate displacements.

Pre-processing. In the following, we apply 1D Gaussian convolution with a fixed standard deviation σ_1 along the first dimension before calculating the displacements. This reduces high frequency noise and makes the signals differentiable. We normalize the input data to compensate for different scaling.

On top of that, due to potentially low SNR of the input data, we need an additional pre-processing strategy. We propose a novel, anisotropic kernel for the processing of 2D single trial matrices (see Figure 3.2).

We choose a filter that smooths orthogonally to the discrete time sequences and line scans, while allowing displacements of the individual measurements. Our kernel is a 2D Gaussian function G_σ with zero mean and standard deviation $\sigma = (\sigma_1, \sigma_2)^\top$

$$K_\sigma(x, y) = \frac{\omega(x, y)G_\sigma(x, y)}{\|\omega G_\sigma\|_1} \quad (3.5)$$

that is weighted by the angular distance to the 1D signals

$$\omega(x, y) = \begin{cases} 1 & \text{if } x = y = 0 \\ \left| \frac{y}{\sqrt{x^2 + y^2}} \right| & \text{else.} \end{cases} \quad (3.6)$$

For the rest of this section, we use equal standard deviations $\sigma_1 = \sigma_2$.

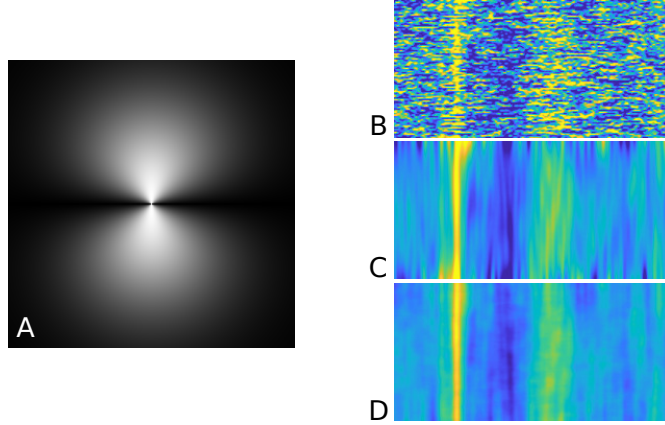


Figure 3.2: The anisotropic kernel (A) defined by equation 3.5 which we use for pre-smoothing: A normalized Gaussian kernel K_σ with zero mean and standard deviation σ that is weighted by the spatial angle. Applied to an ERP map (B), we see that traces across trials are accentuated (D) when comparing with Gaussian convolution (C). For (C) and (D) we used a standard deviation of $\sigma = 10$.

In the context of real-time alignment of data with low SNR, at each point in time we have only access to the past measurements. In this case, we can use half the kernel and weight G_σ with

$$\tilde{\omega}(x, y) = \begin{cases} 0 & \text{if } y > 0 \\ 1 & \text{if } x = y = 0 \\ \left| \frac{y}{\sqrt{x^2 + y^2}} \right| & \text{else.} \end{cases} \quad (3.7)$$

3.1.2.2 Implementation

We discretize the Euler-Lagrange equations of the linearized functional following Brox et al. [19] and solve it with an iterative solver with lagged diffusivity on a Gaussian scale space. It has been empirically shown that median filtering of 2D flow increments improves the estimation accuracy [100]. Accordingly, at each pyramid level, we apply a 1D median filter to the estimated displacement increments. For the results presented in this section, we choose a filter size of 5. We use cubic interpolation for the compensation of displacement increments as well as for the final alignment. For multimodal data terms, we follow Hermosillo et al. [145] and implement a gradient descent algorithm on a Gaussian scale space with cross-correlation in the data term.

Our model requires a common reference and computes displacements and alignment for all measurements with respect to this reference. Analysis of the displacements of subsequent measurements can be realized from the displacement field without explicitly comparing two signals. However, for certain data, registration to a fixed reference can result in large displacements. Large displacements can violate the linearized constancy assumption on the highest level of the pyramid and hinder the estimation of the joint pdf for the multimodal data terms.

For this reason, we assume that we can find a row order with gradually increasing displacement magnitudes, meaning for all $i \in \{1, \dots, M\}$ we want to find the bijective function $\pi : \mathbb{N} \rightarrow \mathbb{N}$ such that the displacement magnitudes are gradually increasing $|u_{i,\pi(1)}| \leq |u_{i,\pi(j)}| \leq |u_{i,\pi(n)}|$ with respect to the reference profile. This mapping π integrates experiment specific a priori knowledge into the alignment calculation or can be estimated heuristically. For example, we expect line scan data with high temporal resolution to show small inter line movements making the identity a reasonable choice for π . In terms of implementation, for all $i \in \{1, \dots, M\}$ we initialize $u_{i,\pi(1)}$ with a constant displacement field and then calculate the displacements row by row, where we initialize the highest pyramid level of row $\pi(j)$ with the previous result $u_{i,\pi(j-1)}$.

This strategy results in the smallest possible displacement increments, as long as we can provide some good π . The parameters of our method are on the one hand the solver specific parameters such as warping depth, number of iterations, update frequency of the non-linear term and the pyramid rescaling factor $\eta \in (0, 1)$ which can be increased for better accuracy. On the other hand, we have model specific parameters which modify the objective function: The smoothness parameter α defines the weight of the smoothness term with respect to the data term. The penalizer exponents of the generalized Charbonnier function $0.45 \leq a_d, a_s \leq 1$ control how deviations from the model assumptions are penalized: Large values of a_s enforce smooth displacements and low values allow discontinuities. Smaller values of a_d make the data term more robust with respect to outliers and noise.

We have implemented our method in MATLAB with the MATLAB Image Processing Toolbox and C++. The warping loop is written in MATLAB and on grounds of computing speed, the iterative solver is written in C++ and compiled into a `mex` function.

3.1.3 Results

In this section, we report the performance of our alignment approach on EEG single trial matrices and functional 2-photon imaging line scan recordings. We use real physiological as well as synthetic data. For the sake of visualization, we report the results for line scans (section 3.1.3.2) always transposed, meaning the displacements have been estimated along the y-axis. For quantitative analysis, we compare variational 1D alignment with the removal of constant shifts by maximizing cross correlation (no subsample refinement) and correlation optimized warping as implemented in [143].

3.1.3.1 Synthetic experiments

We use two different synthetic datasets: The first one simulates a line scan recording with a non-constant motion artifact, that is difficult to remove. We use it to demonstrate the compensation by our method of motion artifacts that cannot be parametrized by constant shifts. To investigate overfitting and clustering of signal features as well as to benchmark computing speed, we use random 2D matrices.

Line scan with diverging structure. For our first synthetic data experiment, we simulate a line scan recording that has been contaminated by non-constant motion noise. The data consists of 800 line measurements with 150 samples each.

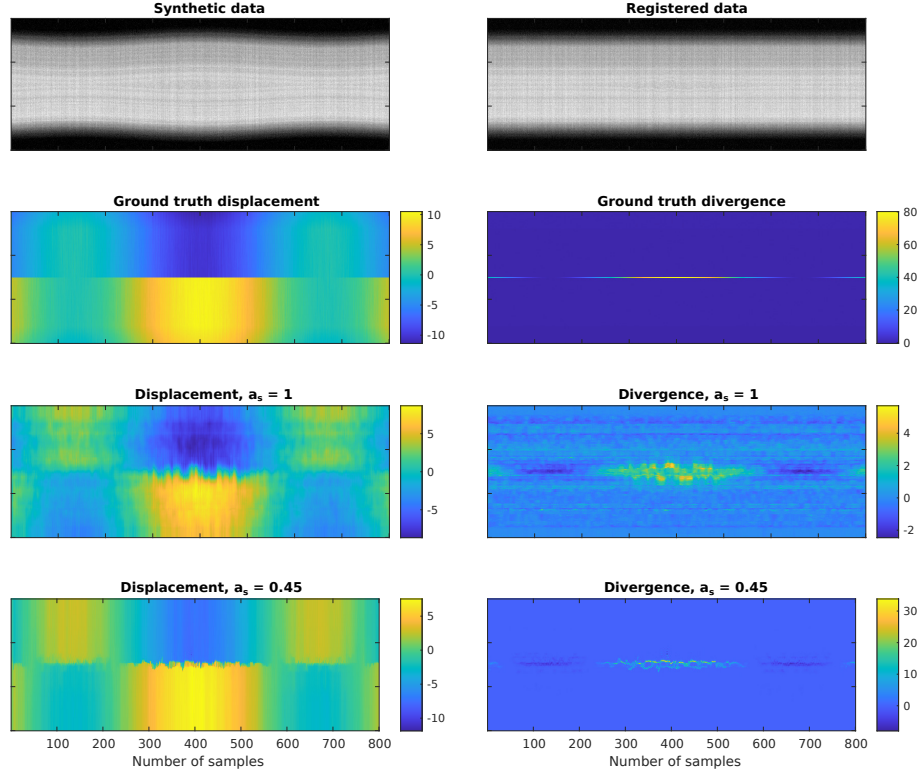


Figure 3.3: Synthetic line scan recording with converging and diverging features (top left) which can be almost completely compensated by our method (top right). The bottom left plots depict the ground truth displacement map and the estimated displacement map with $a_s = 1$ and $a_s = 0.45$. On the right is the first derivative along the spatial dimension of the displacement map. The ground truth displacement map contains discontinuities which is why $a_s = 0.45$ yields a better estimation than $a_s = 1$.

The average of the ground truth displacement field u_{gt} is zero, which means the motion cannot be parametrized by constant shifts. The signal features converge and diverge in the center as seen by changes in u_x (see Figure 3.3).

For the evaluation, we use different values of a_s and choose parameters with less iterations for a fast estimation (see Table 3.1). We choose the average profile as reference, which in the case of the variational method is refined one time. As expected, the dataset cannot be denoised with constant displacements (see Figure 3.4 and Table 3.1). Constant alignment with maximization of the cross correlation does not improve the signal. Quantitatively, correlation optimized warping performs even worse than constant alignment, which might be due to the high amount of noise in our dataset. Because of the discontinuities in the displacement fields, a low choice of a_s performs better than quadratic penalization of the smoothness.

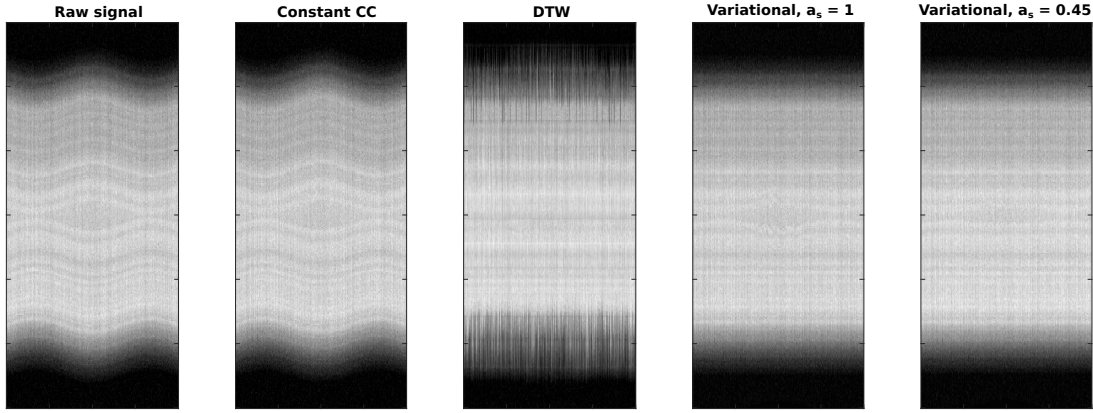


Figure 3.4: Qualitative registration performance on the first synthetic line scan recording. Left to right is the *raw signal*, removal of constant shifts with *cross correlation* (Constant CC), *correlation optimized warping* (DTW) and *variational alignment* with different parameter choices a_s . Due to absence of constant shifts, constant CC cannot improve the input data. DTW contains many outliers and misalignments while variational alignment with $a_s = 0.45$ can almost perfectly align the sequence.

Alignment of random noise with structured reference. We observe that our method tends to overfit noisy data if we do not apply pre-smoothing. This results in clustering of vertical traces (see figure 3.5 and 3.6), that is, if the SNR is low in a single line, features in the noise are aligned with respect to features in the reference. When we provide a structured pattern such as a sine as reference, overfitting can reproduce the pattern in the signal average after alignment (see figure 3.5). Regarding the average of an aligned matrix with the average profile as reference, overfitting seems to accentuate features in the average signal after alignment.

In our method, overfitting is regularized with the size of our filter from equation 3.5 across trials and the size of the 1D filter along trials. Additionally, the subquadratic data term reduces the influence of outliers.

Method	STD	PSNR	Time (ms)
Correlation optimized warping	0.048	26.09	58.5
Raw signal	0.046	26.51	–
Constant displacement cross correlation	0.046	26.52	0.1
Variational fast $a_s = 0.45$	0.034	29.25	1.2
Variational $a_s = 1$	0.033	29.43	7.8
Variational $a_s = 0.45$	0.033	29.55	9.3

Table 3.1: Performance of our method in comparison with cross correlation based constant alignment (no subsample refinement) and correlation optimized warping. Smaller values for the average *standard deviation* (STD) and larger values for the average *peak signal-to-noise ratio* (PSNR) are better. *Variational fast* uses less iterations, smaller η and no refinement of the reference.

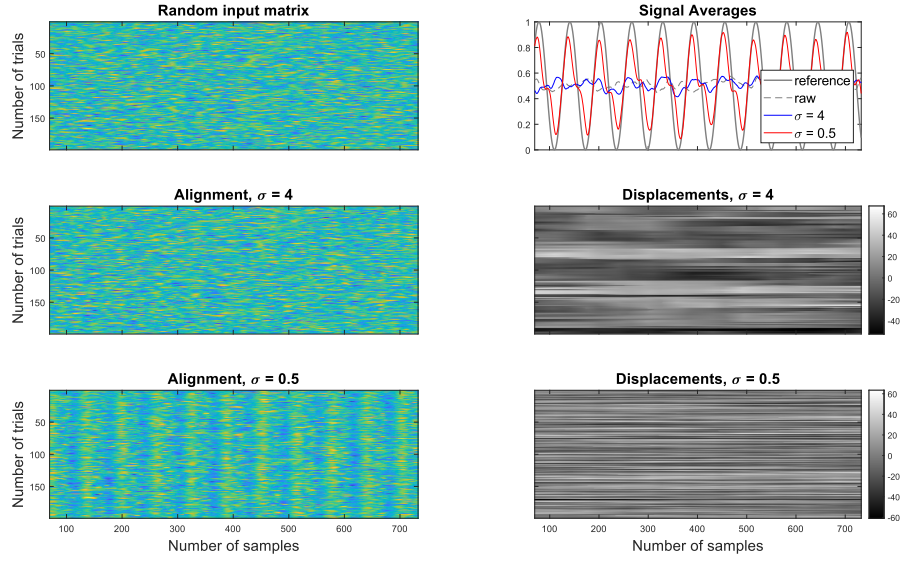


Figure 3.5: Alignment of random signals: We have aligned a matrix containing random, temporally low-pass filtered values with a sine wave as reference. It demonstrates the importance of integrating temporal coherence. When pre-filtering with $\sigma = 4$, the displacement field is relatively smooth across trials. With $\sigma = 0.5$ our method overfits and can reconstruct a reference that is not present in the data.

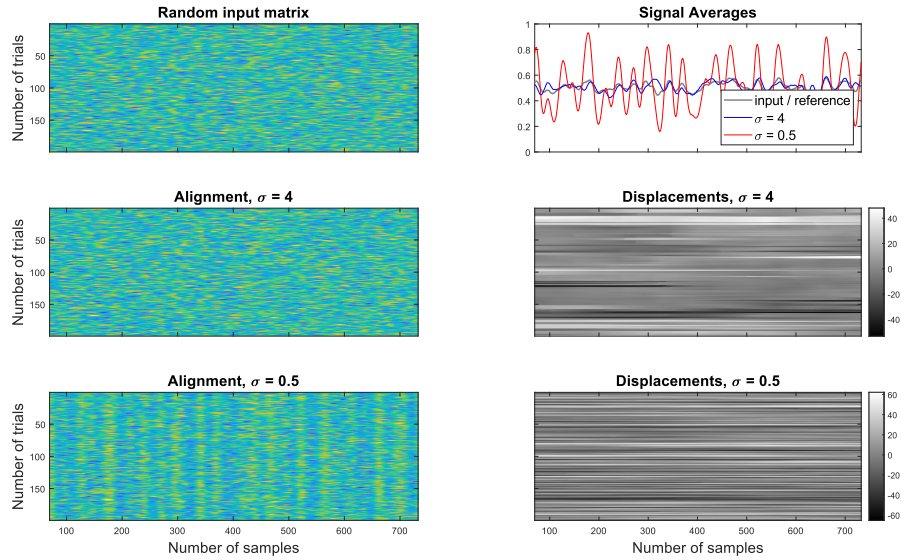


Figure 3.6: Alignment of random signals: When using the average line profile as reference, we observe a clustering effect when using a small filter size.

The underlying assumption for the filtering across trials is a self-similarity in terms of a row coherence with respect to the provided permutation mapping $\pi(j)$. Our anisotropic filter enhances alignment relevant regularities, while suppressing fluctuations that can be attributed to noise. Therefore, depending on the application, other denoising methods can be applied as well (e.g. compare [146], [147]).

It is beyond the scope of this work to present a universal solution to avoid overfitting with this method. However, a general good practice is to adjust parameters, based on SNR. Problems where we generally deal with higher SNRs, or where temporal filtering can increase the SNR produce good results without much inter-trial filtering (see Figure 3.9). Problems with inherently low SNR are prone to overfitting and require pre-filtering, e.g., compare [146] and Figure 3.11. We recommend filtering spatiotemporally such that the average SNR of signal and reference / average is positive. This way, the alignment only compensates visible displacements in the data. Also traces that appear after alignment without any visible correlates in the input data should be interpreted with caution and according to the respective application.

Real-time applications. Our method is fast enough to run in real-time on consumer grade hardware for different use cases. Given a varying sample size n , we benchmark the processing time of our MATLAB implementation on $512 \times n$ matrices as well as on single lines with n samples and with a fixed number of 1000 samples for varying pyramid scale factors η with 20 iterations (see Figure 3.7). The alignment of the matrices took less than a few seconds with the average time per line in the order of milliseconds and no more than 120 ms for an isolated registration of a single line. The machine used for the estimation was a desktop computer with an AMD Ryzen 3900X processor (3.8 GHz, 12 cores, 24 logical processors) and 64 GB of memory. The alignment of an isolated line can be applied in the context of EPs and ERPs (see section 3.1.3.3), where we experiment with inherent delays of more than 100 ms in-between events. The registration of a full matrix might be applied to the online alignment of line scan data with block-wise acquisition and processing / display of line scan data.

For the alignment of short segments with $\eta = 0.9$ or for 1000 samples with small changes in resolution between pyramid levels, our method performs with up to 1000 Hz. When aligning ERPs (see section 3.1.3.3), $\eta = 0.9$ with 1000 samples allows the compensation of a 2 s segment in 4.5 ms given enough room for other processing steps. In their paper, Deriso et al. [144] report an average speed of 0.25 s for a 1000 sample segment. Even though we cannot directly compare the results due to different hardware architectures, the sequential version of our method can perform around two magnitudes faster on average for multiple lines and depending on the choice of η can be one magnitude faster for the alignment of an isolated line.

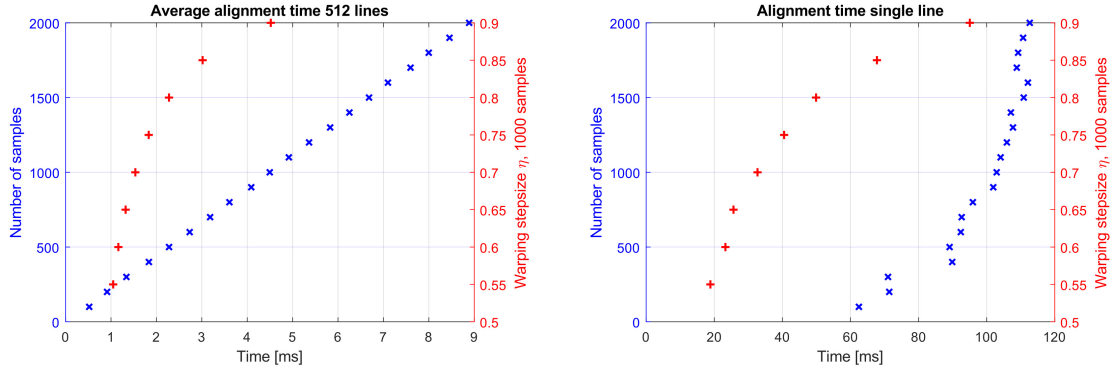


Figure 3.7: Alignment speed of our method per line with respect to varying number of samples and pyramid scale factors. We used 20 iterations and $\eta = 0.9$ for the varying sample size and 1000 samples for varying η . Left is the average time per line when using a multitrial matrix as input and right is the performance for the alignment of an isolated line.

When comparing our method with popular DTW approaches, on average our system needs 4.6ms for 1000 samples and default parameters in MATLAB. On our system, the python package `fastdtw` takes around 229.6ms with default parameters and 6s with radius of 50. We attribute the slower alignment of isolated lines mainly to the characteristics of our MATLAB implementation: On the one hand, for multitrial input, filtering and resizing methods benefit from parallelized implementations in MATLAB. On the other hand, MATLAB overhead for calling external functions amortizes for larger matrices and is much more dominant for the alignment of individual lines. Also given the fact that the type of solver we use can be implemented on graphics cards and can easily be parallelized [135], we expect the performance of an optimized implementation of our method for alignment of a single line to be close to the reported average performance.

3.1.3.2 Line scan data

Method	STD	PSNR	Time (ms)
Raw signal	821.48	-57.58	—
Cross correlation	536.90	-54.81	0.1
Correlation optimized warping	494.46	-54.04	24.8
Variational alignment 20 iterations	470.15	-53.77	0.6
Variational alignment 100 iterations	466.97	-53.75	2.2

Table 3.2: Performance of our method on the 2-photon imaging line scan data (unnormalized) in comparison with cross correlation based constant alignment (no subsample refinement) and correlation optimized warping. Lower values for the average *standard deviation* (STD) and higher values for the average *peak signal-to-noise ratio* (PSNR) are better.

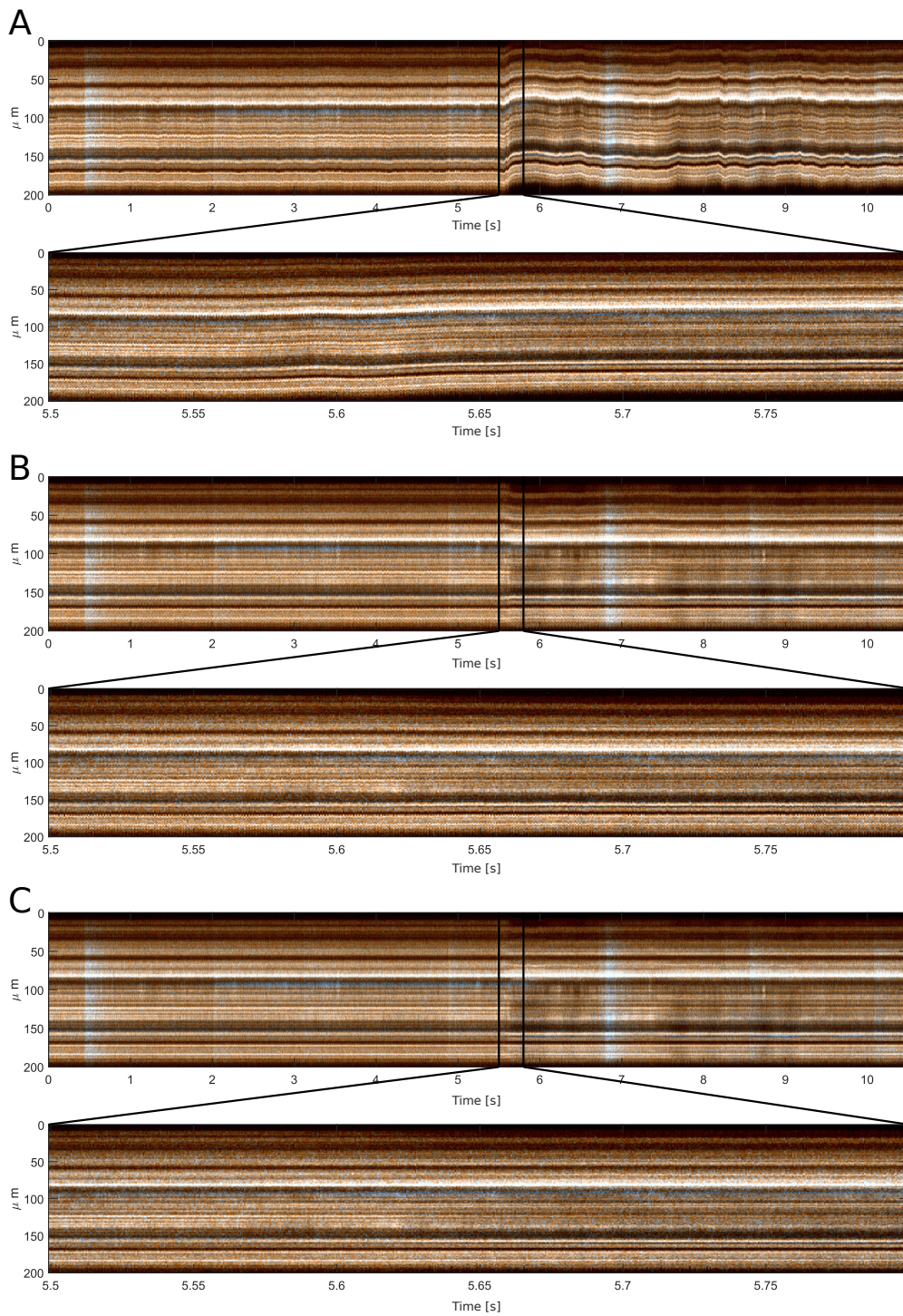


Figure 3.8: 2-photon imaging line scan recording *without alignment* (A), with *cross correlation alignment* (B) and with *variational alignment* (C). The cross correlation approach fails to remove high frequency motion. During the strong motion artifact (5.5 s), removal of constant displacements cannot account for the converging structures. This results in an average temporal standard deviation of 241.2 for (A), 190.6 for (B) and 179.7 for (C). The recording is taken from Roome et al. [137] and used with the authors' permission.

We apply our method to a 2-photon line scan recording from Roome et al. [137] of Purkinje neuron dendrites in the cerebellum of awake mice.

The recording contains two simultaneously recorded line scan channels, the first reporting calcium changes, the second voltage changes, and was recorded at 2 kHz. Imaging artifacts occurred because of animal movement and are composed of different components, including a slight convergence of the structures after movement onset, see Figure 3.9 and 3.8 around 5.5 s. In the same way as for section 3.1.3.1, the performance of our method is compared with cross correlation based alignment of constant displacements and correlation optimized warping.

Optical imaging recordings can contain multiple channels [137] and sometimes contain a structural as well as a functional channel. We register on both channels simultaneously and weigh each channel by structure content.

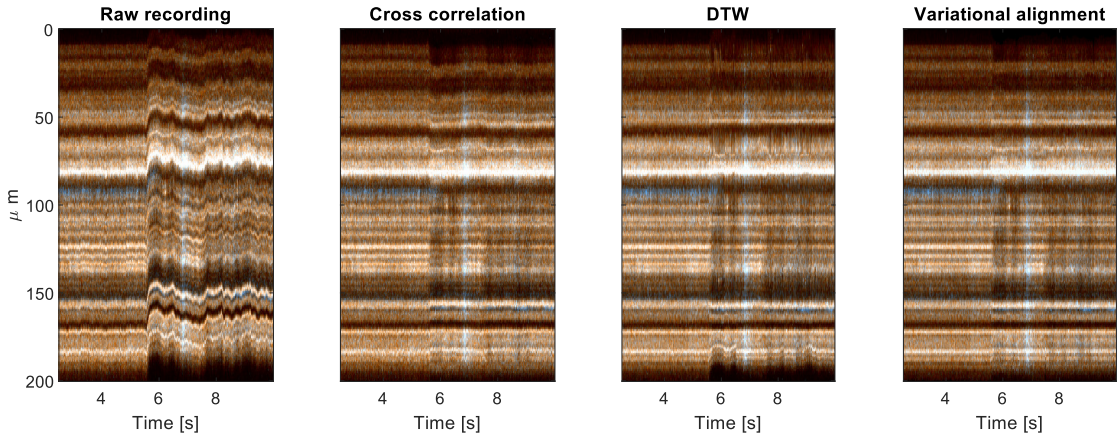


Figure 3.9: Section from the 2-photon imaging line scan recording from left to right *without alignment*, with *cross correlation alignment*, with *correlation optimized warping alignment* and with *variational alignment*.

For the alignment of the data in Figure 3.8 and 3.9 our method took on average 2.2 ms per line with 100 iterations and with 20 iterations only 0.6 ms. From table 3.2 we see that there is not a large increase in performance for 100 compared to 20 iterations. Thus, state-of-the-art results for a line scan of width 270 can be achieved at a little more than 1.6 kHz, which conforms with our results from the previous section.

However, difficult motion artifacts in two-photon imaging data as seen from 5.5 s onwards can coincide with shifts orthogonally to the imaging plane. This way, structures can completely disappear and concentration gradients of dyes can be misinterpreted as functional activity. Therefore, such points in time are usually detected and discarded systematically in evaluations. With an improvement in registration accuracy, it might be possible to use the previously discarded data for analysis.

Importantly, the improved registration accuracy does not come at the cost of much higher computation time. Therefore, we see our method as a superior alternative to alignment approaches that estimate constant displacements.

3.1.3.3 ERP single and group trial analysis

Analysis of 2D matrix representations of ERP in EEG recordings has been suggested for the analysis of trial to trial variances in critical wave morphologies, (e.g. see [139]). Inter-trial and inter-subject variances not only distort signal and grand averages, but also cannot be analyzed after averaging – even though they can contain different cues on dynamics during experiments, in particular, when analyzing endogenous processes and exogenous parameter variations. Because of the low SNR in individual trials, single trial analysis requires denoising strategies, which, due to the possibility of 2D representation of the data, are often motivated by image denoising techniques [139, 146].

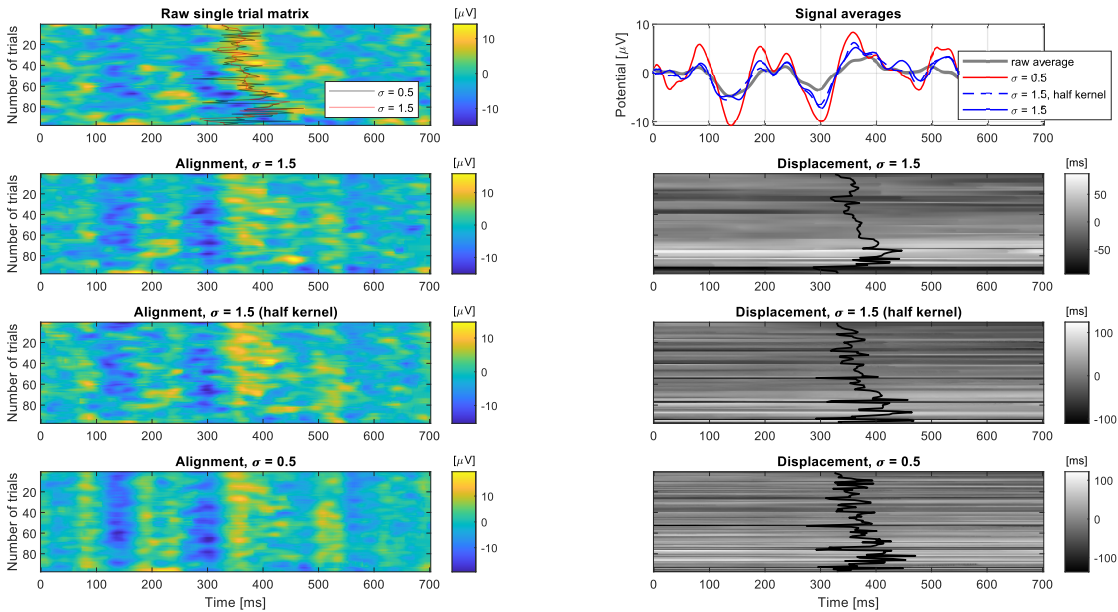


Figure 3.10: Alignment of a single trial ERP recording during an auditory odd-ball paradigm sorted by the subject’s response time, see [148] and section 3.1.3.4 for more details. Each line in the single trial matrix corresponds to an individual measurement. Using only past trials results in a similar performance as the full kernel. However, a low level of temporal integration results in strong clustering of the signal. All ERP maps were filtered with $\sigma = 1.5$ for better visualization. On the top left, the trace of the P300 wave derived as the trajectory through the displacement fields is indicated by the black ($\sigma = 0.5$) and red ($\sigma = 1.5$) lines. On the right, the trajectories for each respective displacement map are indicated by the black line. The trace is also indicated in the respective Displacement maps on the right.

Knowledge of the displacement or latency jitter in ERP matrices on a per trial basis has the potential to greatly improve ERP single trial analysis: In the context of denoising, aligning signals with respect to their morphology allows smoothing vertically to the time dimension with linear filters. Compared to anisotropic filtering approaches, the direction of inter-trial changes does not need to be estimated. Filtering methods based on non-local means estimation such as [139] produce state-of-the-art results for two dimensional ERP denoising. However, they do not aim to recover the implicitly estimated spatial relation of the self-similarity. Also, the displacement maps in our method allow the analysis of the estimated traces (see section 3.1.3.1 and Figure 3.10), while boosting important features in the signal average after alignment².

We use recordings of auditory late responses (ALRs) to an auditory oddball paradigm from Bernarding et al. [148] with the author's permission (for more details on data acquisition see section 3.1.3.4). Our method could help with studying such data and the underlying cognitive neurodynamics on a single trial basis [149], [150], [151].

The same holds true for EP alignment: We apply our method to auditory brainstem responses (ABR) collected in Corona-Strauss et al. [152] (see Figure 3.11). The data was recorded from 10 subjects. The ABRs were evoked by chirps with a bandwidth between 0.1-10 kHz and intensity levels of 40, 30 and 20 dB peak equivalent sound pressure level.

Before processing, the data was averaged over 10 samples and filtered with the BM3D filter [153]. We observe, that the diagnostically most relevant wave V [154] is easier to recognize and its amplitude is enhanced after alignment as compared to classical averaging without alignment, see Figure 3.11. This has the potential to reduce the required number of trials for analysis.

3.1.3.4 ERP dataset and acquisition details

We use data from Bernarding et al. [148] that consists of recordings of 13 student volunteers (27 ± 4.25 years, 9 males and 4 females) with no history of hearing problems, but due to a lack of an identifiable P300 wave, 5 subjects were excluded.

The recordings are auditory late responses (ALRs) to an easy and difficult auditory oddball paradigm. Data were recorded at 512 Hz and bandpass filtered 1 – 30 Hz and thresholded below 50 μ V for artifact rejection. Ag/AgCl-electrodes were attached ipsilateral to the stimulus at the right mastoid, the common reference at the vertex and the ground at the upper forehead.

²For an in depth analysis of the application of our method to ERP data, we refer the reader to Thinnes et al. [146].

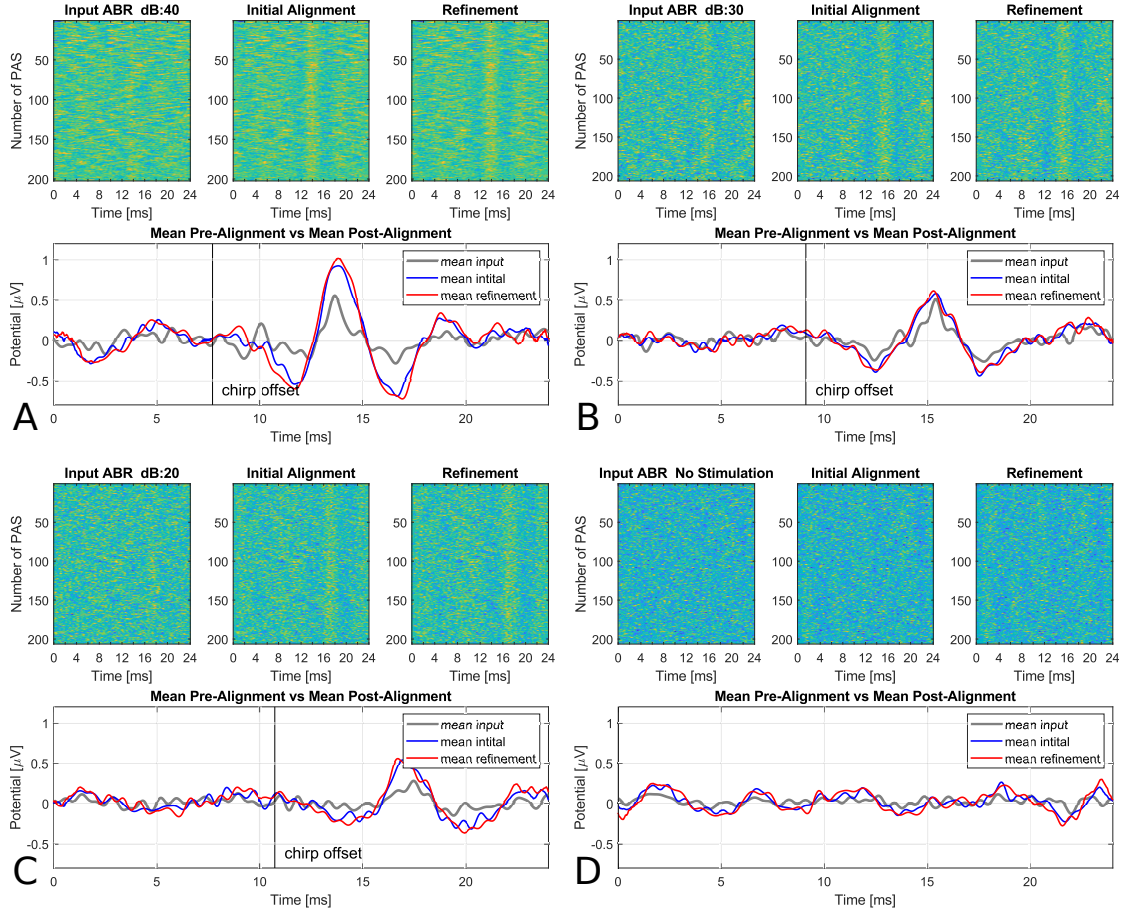


Figure 3.11: Application of our method to *auditory brainstem responses* (ABRs) data. Each row represents partial averages of 10 consecutive sweeps (PAS). In the average response, the diagnostically most relevant wave V is easier to recognize and its amplitude is enhanced after alignment as compared to classical averaging without alignment. However, we can also observe the clustering effects described in section 3.1.3.1 for lower stimulation amplitudes. Stimulation is done with 40 dB (A), 30 dB (B), 20 dB (C) peak equivalent sound pressure level and (D) is spontaneous activity. The refinement step is registration with the average of the matrix after initial alignment as reference.

As reported by Bernarding et al., the paradigm consisted of 100 deviant syllables, which was in both paradigms always the same syllable with a probability of 20%. For the data reported here, the syllables had the same vowel, and the plosives had the same place of articulation (difficult paradigm). A complementary grand-average example for the difficult paradigm, including the before/after alignment comparison, is shown in Figure 3.12.

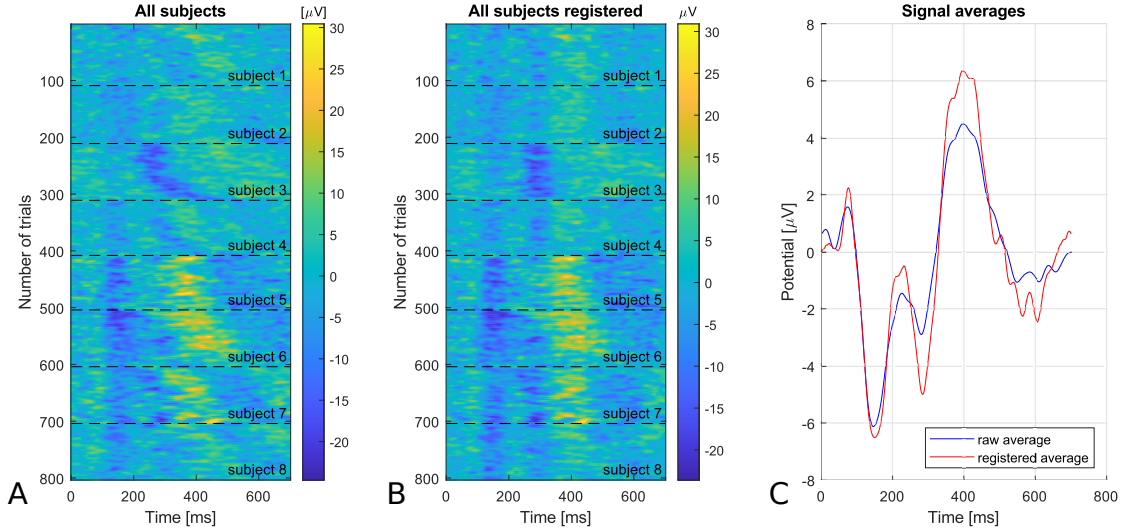


Figure 3.12: Difficult paradigm of an auditory oddball experiment, filtered with $\sigma = 1$ (A), after alignment with the grand average as reference (B) and grand average of before and after alignment (C).

3.1.4 Discussion

In this work, we have presented a novel variational method for the alignment and estimation of non-flat and non-parametric displacements of 1D data with applications in neuroimaging. Our method produces state-of-the-art results for the alignment of 1D signals and is fast enough to enable real-time applications on consumer-grade hardware. The method is implemented as an accessible MATLAB toolbox and publicly available. Our continuous formulation is motivated by 2D OF estimation and thereby builds on a rich framework of different regularizers and constancy assumptions as well as efficient optimization strategies. Our model uses robust data and smoothness terms penalized by generalized Charbonnier penalties, with subquadratic penalization in the data term and either quadratic or subquadratic penalization in the smoothness term. It natively supports weighted multichannel data.

We have demonstrated the performance of our approach on 2-photon line scan recordings as well as 2D single trial matrix representations of ERPs and EPs. The alignment of ERP matrices opens the way for improved analysis with reduced number of trials. Due to its high accuracy and fast implementation it forms an alternative to established methods such as DTW and constant template matching. In future work it can be additionally evaluated against methods currently under development such as Deriso et al. [144].

Averaging signals to improve SNR requires a linear relationship and independence of signal and noise [155] that can be violated by jitter. For short stimulus intervals, adjacent response techniques such as Woldorff et al. [156] try to remove jitter by shifting the waveforms.

The explicit estimation of the displacements for each trial could allow for a quantitative characterization of the jitter in EEG single and group trial recordings and our method solves this task for non-uniform jitter with subsample precision.

This could, for example, find application in the research of auditory processing and perception by means of EP and ERP alignment [154], [157]. Jitter due to non-linear responses affects averaged ERPs and application of our method could break the dependence of averaged ERPs from the level of stimulus-locking. Additionally, the temporal integration parameter σ could allow different types of investigations: High σ values with the implicit assumption of high temporal coherency could enable the analysis of slow non-linearities, while smaller values of sigma could enable the analysis of faster mechanisms in EP data [157]. However, our method estimates global displacements without spectral resolution. An extension to the compensation of subband decompositions could allow analysis and denoising of waves with different neural generators in different frequency bands in the future.

For line scan recordings, our method is an alternative to constant displacement estimation which offers higher accuracy at only slightly higher computation time and the ability to compensate non-flat displacements. In addition, we believe that this method could find many further applications for the alignment of line scan data beyond the field of neuroimaging.

3.2 Flow-Registration: Non-parametric image registration of 2-photon imaging data

Functional 2P microscopy is a key technology for imaging neuronal activity and 2P microscopy plays a role in the previous section’s approach to aligning linescan recordings. While linescans enable very high temporal resolutions, the lack of spatial resolution limits processing of such data. For example, movement orthogonal to the line direction cannot be compensated with traditional approaches.

2D image sequences, however, can contain non-rigid movement artifacts that require high-accuracy movement correction. Variational OF estimation is a group of methods for motion analysis with established performance in many computer vision areas. However, it has yet to be adapted to the statistics of 2-photon neuroimaging data. This section presents the motion compensation method *Flow-Registration* that outperforms previous alignment tools and allows to align and reconstruct even low *signal-to-noise ratio* (SNR) 2-photon imaging data and is able to compensate high-divergence displacements during local drug injections. The method is based on statistics of such data and integrates previous advances in variational OF estimation. Flow-Registration is available as an easy-to-use ImageJ / FIJI plugin as well as a MATLAB toolbox with modular, object oriented file IO, native multi-channel support and compatibility with existing 2-photon imaging suites.

This section contains direct excerpts from [F3]³.

3.2.1 Introduction

2-photon microscopy in combination with synthetic or genetically encoded indicators allows imaging a wide range of different aspects of neuronal activity with cellular or even sub-cellular resolution in anesthetized as well as behaving animals [158, 159]. Importantly, small signal changes might carry crucial information. However, the imaging data can be afflicted with different types of noise and artifacts. On one hand, due to the low number of generated photons with 2-photon excitation, the shot noise is typically significant. On the other hand, movement noise can be introduced during acquisition. Motion artifacts can be caused by heart beat, breathing, as well as motor behavior in awake animals.

³Documentation, the MATLAB code for Flow-Registration and the ImageJ plugin can be found on the GitHub repository https://github.com/phf1ot/flow_registration. The version used for the evaluations in this section can additionally be found as supplemental code of the respective publication [F3]. It contains MATLAB scripts to reproduce the results reported here as well as the precompiled ImageJ / FIJI plugin.

Also, some experimental paradigms inherently result in large, non-rigid and non-elastic deformations, for example, local drug injections. While many established tools exist for the correction of small vibrations and rigid drifts, the compensation of large and/or non-uniform motion is still a challenge.

Furthermore, for high accuracy alignment, subpixel registration is necessary: Small displacements can be approximated by a linearized motion assumption [14]. This means even small residual motion can induce large artifacts around image edges due to a proportional relation of spatial image gradient and motion magnitude with respect to induced brightness changes.

While *optical flow* (OF) based image registration methods were used for 2-photon imaging data before [160], they do not incorporate the advances in OF estimation developed in recent years in the context of computer vision areas. As a consequence, they can perform poorly when compared with state-of-the-art motion correction tools for 2-photon imaging and dense OF has been reported to be too sensitive to low *signal-to-noise ratios* (SNR) typical of calcium imaging [17]. In many calcium-imaging pipelines, motion correction is instead performed by rigid or piecewise-rigid template matching. NoRM-Corre registers overlapping patches against an updated template and interpolates the inferred translations to a per-frame motion field [17]. In the previous section, we developed a variational motion compensation strategy for 1D linescans (e.g. [161]) and could already demonstrate that metrics from variational OF estimation can be robust enough for low SNR applications [F4].

Non-rigid image registration is an extensively studied topic in medical imaging, with a large body of work on variational/PDE-based formulations, transformation models, and multiresolution optimization [162, 163, 18]. Toolkits such as elastix [164] and *advanced normalization tools* ANTs [165, 166, 167] provide flexible implementations of these concepts with a variety of similarity measures and transformation models. ANTs has proven performance for brain image registration and labeling of brain scans [166, 167]. Inter-subject or cross-modality image registration often uses regularization that enforces smooth, physically plausible deformations, commonly via diffusion- or linear elastic regularization of the displacement field, and in some settings via diffeomorphic constraints to preserve topology [18, 163]. However, in contrast to typical medical image registration, motion correction in 2-photon microscopy is dominated by Poisson noise, bleaching, temporally varying functional signals and the motion can contain non-elastic deformations, e.g. during local drug injections. Also, frames within single 2P videos are often much more similar than images in inter-subject or multimodal registration which makes the setting closer to OF estimation between frames. Therefore, in this chapter we formulate the problem as non-parametric, OF registration of each frame to a fixed reference.

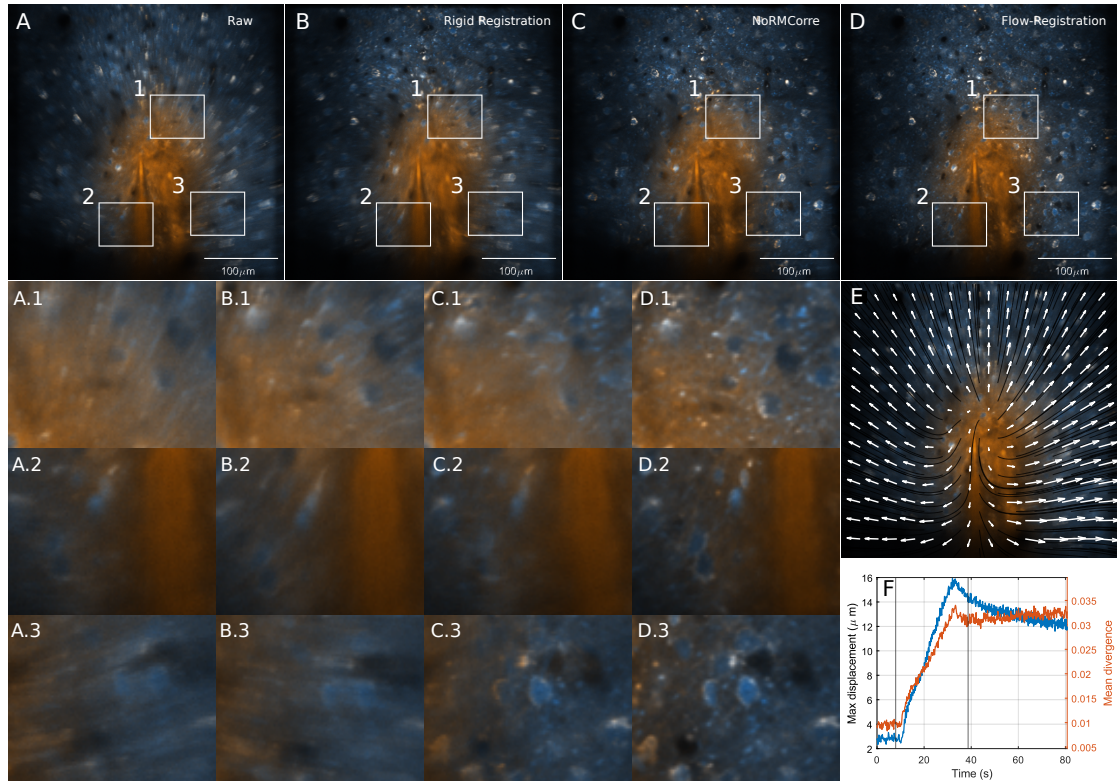


Figure 3.13: Comparison of registration performance on a very challenging two-channel 2-photon recording during drug injection in vivo. The challenges of the sequence are a very low signal-to-noise ratio together with brightness changes (orange) and non-elastic deformations due to the injection. Average of (A) raw images, (B) after rigid registration, (C) after registration with NoRMCorre, and (D) after Flow-Registration. The tissue expands from the injection point (E) resulting in large displacements as well as in a high divergence (F) in the displacement field. Flow-Registration can recover fine structures with much more detail, allowing region-of-interest selection along fine structures, while the blur and double images in (A-C) indicate residual motion. This is well visible for the fine structures and grains in D.1 which are smeared by motion blur in A.1-C.1 and become more difficult to observe. The channels in the false color representation have all been normalized with respect to the min and max intensity values of the averaged raw recording. Neurons express the PKA sensor GAKdY-mut (green channel, mapped to blue) and the red chromophore TurboRFP (red channel, mapped to orange); the injection contained the fluorescent dye Alexa 594 (red channel, orange).

There has been a great deal of work on OF techniques in the past decades with the goal of improving accuracy, model invariants as well as robustness and computing speed. The recent advancements in OF estimation tackle the problem of large displacements with discontinuous motion fields and motion layers. Chen et al. [98] lead (as of July 2021) the Middlebury optical flow benchmark [99] with respect to average *endpoint error* (EPE) and average angular error. They use similarity transformations for a segmented flow field as initialization of large motion. In a second step, the variational method of Sun et al. [100] is used for subpixel refinement.

In this work, we propose a novel, OF based image registration approach for the motion compensation of 2-photon neuroimaging data. We build on the well studied framework for variational OF estimation [14, 19, 100] and adapt this framework to 2-photon imaging data by techniques which have recently been developed in visual computing. We demonstrate the performance on challenging 2-photon imaging data and can report state-of-the-art results in terms of registration quality, competitive computation speed and easy accessibility. Our method is available as an easy-to-use MATLAB toolbox as well as an ImageJ / FIJI plugin.

3.2.2 Materials and methods

In this section, we will first briefly describe the acquisition methods for the data that is used in the benchmark dataset and then present the theory as well as implementation details of the proposed motion compensation method Flow-Registration⁴.

3.2.2.1 Motion correction method

Statistics in 2-photon microscopy. The motion statistics for 2-photon imaging differ from the motion we encounter in natural images in several ways: Due to the illumination strategy, there is only a single imaging plane and thus no different motion layers. Also, the imaged object is usually soft, biological tissue and, as a consequence, we expect smooth motion fields without discontinuities. We expect the image to contain a fixed *region of interest* (ROI) with small displacements between frames and, potentially, large drift over time which can result in large displacements with respect to the reference. Due to the nature of the scanning method, horizontal displacements may occur. Also, the images are usually not represented in perceptual color spaces, restricting the use of classical photometric invariances, such as in [89], but, often, there exist multiple signal and or structural channels.

⁴We refer the reader to the original publication [F3] for the details on image acquisition, dyes and animals which have been written by the contributing authors.

Depending on the indicators used, the different channels can contain disjoint structures and different SNR characteristics. Due to technical limitations of current generation imaging, high speed imaging can often only be realized on narrow fields of view (FOV).

Our image registration method builds on those observations: Due to the absence of different motion layers and temporally coherent recordings, we do neither need an initialization strategy with a more robust method such as in [98] nor a regularizer that preserves discontinuities (e.g. [100, 89]).

Also, the assumptions of elastic regularizers which penalize the divergence of the OF field, e.g. compare [145], are violated by recordings during drug injection, see Figure 3.13 (F), where the displacement field has a high divergence after injection. Additionally, we generalize our previous linescan alignment [F4] to recordings with narrow FOV and propose a non-uniform warping strategy where we perform more warping steps along the larger image dimension during optimization.

Optical flow method. Let $f : \Omega \rightarrow \mathbb{R}^C$ be a moving and $f_{\text{ref}} : \Omega \rightarrow \mathbb{R}^C$ be a fixed image where $\Omega = (0, M) \times (0, N) \subset \mathbb{R}^2$ with width M , height N and $C \in \mathbb{N}^+$ the number of channels. We then depict the displacement field that points from the reference to the moving image as $\mathbf{w} : \Omega \rightarrow \mathbb{R}^2$. In the following, we use a compact notation and omit the function arguments if not needed. We use coordinates $\mathbf{x} = (x, y)^\top$ and depict the components of the displacement field with $\mathbf{w}(\mathbf{x}) = (u(\mathbf{x}), v(\mathbf{x}))^\top$ and one channel of f for readability with $f^c : \Omega \rightarrow \mathbb{R}, c \in \{1, \dots, C\}$. The goal of variational OF is to estimate $\mathbf{w}$ such that $f_{\text{ref}}(\mathbf{x})$ depicts the same scene point as $f(\mathbf{x} + \mathbf{w}(\mathbf{x}))$ by minimizing an energy functional that consists of a data term $D(\mathbf{w})$, a smoothness term $S(\mathbf{w})$ and usually a smoothness weight α . The data term penalizes deviations from constancy assumptions and the smoothness term imposes smoothness on the results in image areas that do not contain enough information (aperture problem [89]). The energy functional for variational OF usually has the form

$$E(\mathbf{w}) = \int_{\Omega} D(\mathbf{w}) + \alpha S(\mathbf{w}) \, dx dy \quad (3.8)$$

We model our observations as a variational OF method with gradient constancy, robust, separate channel penalization of the data term (e.g., [88]) and first-order regularizer that we set to homogeneous diffusion. We normalize the data term according to Zimmer et al. [89]. The regularizer and data term are penalized with a generalized Charbonnier penalty function which is given by $\Psi_a(s) = (s + \varepsilon)^a$, for $a > 0$, a squared residual $s \geq 0$ and a small $\varepsilon > 0$ [100]. We use a_{smooth} to depict the parameter for the smoothness and a_{data} to depict the parameter for the data term. Gradient constancy assumes that the gradients of the imaged structures do not change with respect to the reference.

For a moving image f and a fixed image f_{ref} , a data term that quadratically penalizes constancy of the image gradients for a single channel $c \in \{1, \dots, C\}$ is given by

$$\begin{aligned} d^c(\mathbf{x}, \mathbf{w}(\mathbf{x})) = & \left(f_x^c(\mathbf{x} + \mathbf{w}(\mathbf{x})) - (f_{\text{ref}}^c)_x(\mathbf{x}) \right)^2 \\ & + \left(f_y^c(\mathbf{x} + \mathbf{w}(\mathbf{x})) - (f_{\text{ref}}^c)_y(\mathbf{x}) \right)^2 \end{aligned} \quad (3.9)$$

In the variational framework, we can minimize the joint energy of the structural and the functional channels.

This has the advantage that, in theory, we get a better SNR, if the same structures are visible in both channels and, additionally, considering the aperture problem, we have potentially more information on the motion whenever disjunct structures are visible in the different image channels, such as in the data in Figure 3.14 (A)-(D). Additionally, the energy functional can naturally incorporate a manual weighting term per channel, such as ROIs, to enforce a low value of the data term around important image structures. Because of potentially different SNR properties of the two channels and the likelihood of disjoint structures, we evaluate the subquadratic penalizer on each channel separately which is known as separate robustification [89]. With a channel specific weight function $(p^c)_{c=1}^C$, a moving image $(f^c)_{c=1}^C$ and a fixed image $(f_{\text{ref}}^c)_{c=1}^C$ with C channels, the final data term with separate robustification is given by

$$\begin{aligned} D_{a_{\text{data}}}(\mathbf{x}, \mathbf{w}) = & \sum_{c=1}^C p^c(\mathbf{x}) \Psi_{a_{\text{data}}} \left(\left(f_x^c(\mathbf{x} + \mathbf{w}(\mathbf{x})) - (f_{\text{ref}}^c)_x(\mathbf{x}) \right)^2 \right. \\ & \left. + \left(f_y^c(\mathbf{x} + \mathbf{w}(\mathbf{x})) - (f_{\text{ref}}^c)_y(\mathbf{x}) \right)^2 \right) \end{aligned} \quad (3.10)$$

and the smoothness term as

$$S_{a_{\text{smooth}}}(\mathbf{w}) = \Psi_{a_{\text{smooth}}} \left(|\nabla u|^2 + |\nabla v|^2 \right) \quad (3.11)$$

For the numerical approximation, we follow the framework of Bruhn et al. [168], Brox et al. [19] and Papenberg et al. [169] and discretize the Euler-Lagrange equations in the compact *motion tensor* notation to solve them with an iterative multiscale solver (rescaling factor $\eta \in (0, 1)$) with lagged non-linearities with an update in every 5 iterations.

We apply the recommended good practices proposed by Sun et al. [100], namely, we use bicubic interpolation for the warping steps and 5×5 median-filtering (mirror boundary) of the flow increments for each level to increase accuracy (see Appendix F for algorithm details).

Pre-processing and registration. Low-pass filtering is an important pre-processing step for local and global methods [170]. It makes the images derivable and integrates over temporal changes, while removing image noise. We found convolution with a 3D Gaussian kernel together with subquadratic penalization, to be robust enough to deal with the noise in the benchmark data, e.g., compare the layer 5 data (see Figure 3.14).

In the context of optical microscopy recordings, the recorded image region is usually fixed in the paradigm which means that only little relative movement of the object and camera is expected and the scene content does not change significantly. Let the video sequence be denoted by $(f_t)_{t=1}^T$, where each $f_t : \Omega \rightarrow \mathbb{R}^C$ is a moving image frame at discrete time t . Let $f_{\text{ref}} : \Omega \rightarrow \mathbb{R}^C$ be the fixed reference image and $(\mathbf{w}_t)_{t=1}^T$ with $\mathbf{w}_t = (u_t, v_t)$ be the displacement field pointing from f_{ref} to f_t . Then $f_t(\mathbf{x} + \mathbf{w})$ depicts the same scene points as $f_{\text{ref}}(\mathbf{x})$. For the final motion compensation step, we apply backward warping with bilinear or bicubic interpolation to calculate the motion compensated frame $\hat{f}$ at time t such that $\hat{f}_t(x, y) = f_t(x + u_t(x, y), y + v_t(x, y))$.

Parameter selection. The main parameters of our method are the regularization parameter α , the penalizer powers a_{data} and a_{smooth} , the warping depth, the rescaling factor η and for each channel the Gaussian kernel size $\sigma = (\sigma_1, \sigma_2, \sigma_3)$ for pre-processing. For the regularizer, larger values of a_{smooth} reduce discontinuities in the flow field (staircasing artifacts) and for the data term, smaller values of a_{data} reduce the influence of outliers and render the method more robust under noise. For 2-photon recordings, we set $a_{\text{smooth}} = 1$, which results in a homogeneous diffusion regularizer. Note that for $a = 0.5$ we get a regularized ℓ^1 norm, while the function becomes non-convex for values of $a < 0.5$. However, Sun et al. encourage a choice of $a_{\text{data}} = 0.45$ on the Middlebury benchmark, which we apply here. The choice of α presents the compromise of correctly registering smaller structures that deviate from the global motion direction and a globally smooth solution. Using a different value of α in each of the image dimensions renders the smoothness term anisotropic. This might improve the estimation of motion artifacts induced by horizontal scanning.

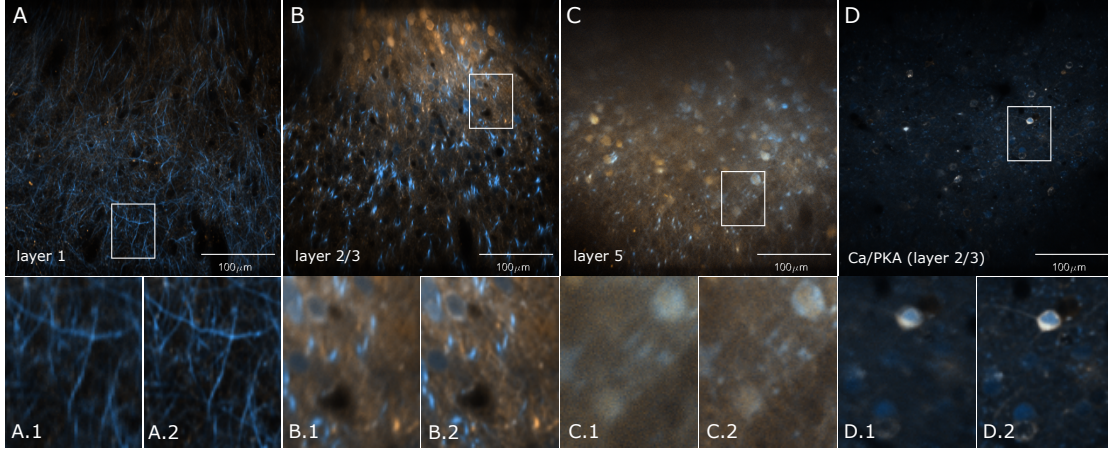


Figure 3.14: The benchmark datasets from layer 1 (A), layer 2/3 (B), layer 5 (C) and the Ca/PKA (layer 2/3) (D) dataset. Temporal average of raw recordings (A.1-D.1) and after application of Flow-Registration (A.2-D.2), both during the motion contaminated frames. Neurons express GAKdYmut (orange) and TurboRFP (blue) (A), GAKdYmut (blue) and TurboRFP (orange) (B-C) and GAKdYmut (orange) and jRGECO (blue) (D).

Toolbox design and implementation. The ImageJ / FIJI plugin makes use of the ImageJ file formats and therefore can resort to all supported file types. Our software design for the MATLAB toolbox consists of modular file readers and writer classes that can be automatically instantiated or supplied as parameters to an options object that defines registration jobs. By default, MDF, tiff image stacks, MATLAB mat files or hierarchical dataformat (HDF5) files are supported. Additionally, the output formats support HDF5 files that are readable by the 2-photon imaging suites Begonia [171] and CaImAn [172]. The file IO is designed for multi-channel processing and supplies 4D matrices in the format $\text{height} \times \text{width} \times \text{channel} \times \text{time}$ independently of the actual data representation on disk. To compensate a recording, the file reader supplies n subsequent frames (batches of size n) with on-the-fly binning to the Flow-Registration engine until the end of the video is reached. The frames in each batch are concurrently compensated and the average displacement of the last frames is used to initialize the lowest pyramid level of the displacement estimation in the following batch.

Reference frames can either be supplied directly or are computed from a specified set of frames, where the set of frames is pre-aligned with respect to the temporal average and then temporally averaged again. For this pre-alignment, the parameters α and σ are increased in size to make the result more robust under noise and reduce overfitting. Given that each batch can have varying average values, we normalize the data with respect to the minimum and maximum values of the reference frame after applying the Gaussian filter. Joint normalization is performed by default but channel-wise normalization is supported as well.

The displacements of each frame with respect to the reference on the low-pass filtered sequences are computed and the raw frames are then registered via backward warping with bicubic interpolation, where out-of-bounds values are replaced with the values from the reference. To reduce quantization due to the interpolation, the results are stored with double precision as default but can also be stored with the precision of the input file.

To increase computational speed, the finest pyramid level for the OF calculation can be reduced in the options and the resulting low-resolution displacement field will be upsampled to the input image resolution in the final computation step. The Flow-Registration plugin implements this in the *Registration quality* setting, where only the highest quality setting calculates the solution on all levels. With settings $\eta = 0.8$ and minimum level of 6, we get almost tenfold faster computation time on the injection sequence with almost the same accuracy (PSNR 42.523 vs. 42.519). With those parameters, the highest resolution at which the displacements are estimated is given by $512 \cdot 0.8^6 = 134$.

Batch processing of multiple input files is possible via a batch processor class where a list of filenames is supplied that will be compensated either with individual references or the same reference.

To run the ImageJ / FIJI plugin, it needs to be installed via Plugins→Install Plugin which adds a Flow-Registration entry under Plugins. In its current version, the ImageJ / FIJI plugin can only be used for the compensation of shorter sequences depending on host system memory, the extension for virtual stacks is planned for the future. The MATLAB toolbox does not have such restrictions and can be used for the batch compensation of big data. For the MATLAB toolbox, a C++ compiler is required and the code has been tested for MATLAB R2018a onwards. The folder *demos* in the MATLAB toolbox contains scripts to reproduce the video results and quantitative results presented here as well as examples on how to use the code for different application scenarios. The *jupiter* demo compensates an amateur Jupiter recording and demonstrates different aspects of multi-channel tiff and ROI processing as well as a potential application beyond the scope of neuroimaging.

3.2.3 Results

In this section, we first describe the benchmark dataset as well as the metrics we use for quantitative registration performance assessment. We then present the performance of our method in terms of registration quality and computational speed.

3.2.3.1 Benchmark datasets and metrics for evaluation

We used multiple datasets that have been recorded with the described setup previously. The dataset names indicate the imaging depth (see Figure 3.14 and Figure 3.15), where deeper layers (e.g. 5) correspond to lower SNR. For the layer 1-5 recordings, timepoints during movement onset which were contaminated with movement artifacts were selected from continuous recordings. For the datasets during drug and saline injections, we selected the time points around the injection events (e.g. compare Figure 3.13 (E)). From each dataset, we selected 2500 frames (80.9 s). A common approach for the evaluation of 30 Hz 2-photon microscopy data is temporal binning to increase the SNR. Therefore, our final benchmark dataset contains the data with binning over five frames (6.2 Hz, 500 frames total), as well as on a subset of the raw recordings, indicated by the suffix 30 Hz. For the dataset *saline injection*, a total of 5 frames from the beginning and end of the experiment were excluded due to artifacts.

Those data are real-world, low SNR datasets without ground truth displacements, therefore, metrics such as AEE or AAE are not applicable. The same holds true for perception based metrics due to the overall small movements and high image noise in the data. For the evaluation, we calculate reference based PSNR (see Table 3.4) and performance factors based on MSE as well as temporal STD (see Figure 3.15). To calculate the metrics, we first applied 2D Gaussian convolution with $(\sigma = (3, 3)^T)$. As reference, we used the temporal average over the reference frames from the motion compensation. The MSE and STD performance factors are then calculated as the fold change of the average MSE and STD values $MSE(\text{raw}) \cdot MSE(\text{compensated})^{-1}$ and $STD(\text{raw}) \cdot STD(\text{compensated})^{-1}$. They indicate the factor by which the MSE or STD value is higher on the raw sequence than the MSE or STD of each method (see Figure 3.15 (A) and (B)) or the factor by which the MSE value is higher for NoRMCorre when compared to Flow-Registration on a frame-wise basis (see Figure 3.15 (C)). We excluded all frames that contributed to the reference, as well as a boundary of 25 pixels. This boundary ensures that only valid regions are considered in the evaluation even with the large displacements of the injection sequences. Low-pass filtering reduced the influence of image noise on the results.

For the compensation of all sequences, the first 100 frames (6.2 Hz) and 500 frames (30.9 Hz) were used as reference and supplied to the respective method. We construct a pair of synthetic frames from the real-world recording that match their statistics with synthetic ground truth displacements and noise. On this synthetic data, we compare the average *endpoint error* (EPE) for Flow-Registration, NoRMCorre as well as elastix and ANTs with different configurations.

For this purpose, we use the injection sequence and add simulated injection, xyz-shifts as well as noise to the clean image we obtain after temporal averaging of the motion compensated recording. We compose the synthetic motion field of a component for line jitter from the scanning as well as high divergence displacements to simulate injection. The injection displacements are modeled as (in pixel coordinates with 512×512 resolution, noting that image coordinates are used where the y-axis points downwards):

$$\mathbf{w}_{\text{inj}}(x, y) = 0.05 \cdot \begin{cases} (x - 256, 0.2 \cdot (y - 280))^\top, & \text{if } y \geq 280 \\ (x - 256, y - 280)^\top, & \text{otherwise} \end{cases}.$$

For the jitter, we apply $\mathbf{w}_{\text{jit}}(x, y) = (\sin(0.001x\pi), 0)^\top$ which results in the final displacement $\mathbf{w} = \mathbf{w}_{\text{inj}} + 2\mathbf{w}_{\text{jit}}$ (see Figure 3.20 (B), left in section 3.3). We adapt the noise model and implementation of Poisson image noise for 2-photon microscopy from Zhang et al. [173] and apply noise with target PSNR of 30 dB and 35 dB.

To simulate changes in the intensity of the visible structures for example due to the injected indicators or due to z-shifts, we apply positive and negative multiplicative intensity changes with a scaled, 2D Gaussian kernel. See Figure 3.20 (A) in section 3.3 for a comparison of the benchmark frames and the original video.

3.2.3.2 Experiments with synthetic data

We compare our method with the 2-photon motion compensation suite NoRMCorre and with the traditional image registration toolboxes ANTs and elastix. For ANTs and elastix, we use the respective Python (version 3.9.7) implementations, namely, *antspy* (version 0.3.1) and *itk-elastix* (itk version 5.2.1, itk-elastix version 0.13.0) and for NoRMCorre we use the official MATLAB implementation. Evaluations are performed with a python interpreter on the *Windows subsystem for Linux* (WSL) 2 (hardware acceleration). ANTs and elastix allow for similar choices of data terms and transformations. While elastix supports extensive registration configurations in config files that are backed by a model zoo, ANTs has multiple presets and allows easy adaptation of the registration method.

For elastix, we report the results with different metrics and simulated gradient constancy. For ANTs, we report the performance of the default *SyN* presets with *cross-correlation* (CC) and *mutual information* (MI) metric. We report the results on the individual channels for all methods and on multichannel configurations for our method as well as for elastix. For elastix, we simulated gradient constancy as a multichannel input on the up to four gradient channels (directional gradients in x and y direction on both channels) with *AdvancedMeanSquares* metric. The results indicate that gradient constancy is superior over other data terms. With lower PSNR, subquadratic penalization becomes vital, such that Flow-Registration outperforms all other methods — for some methods more than an order of magnitude. Our results also indicate that for users of elastix, it could make sense to use MI and CC on the gradients of the images in certain situations.

The evaluations confirm our assumption that the combination of two channels can improve the performance on noisy data: Our method performs consistently better with both channels as compared to evaluations on the individual channels. The same is not true for elastix and ANTs, indicating that there are parameters such as metric weights that could be further optimized for an overall better performance. While some of the traditional image registration methods implemented in elastix outperform NoRMCorre in terms of EPE, NoRMCorre’s performance seems to be relatively robust under noise and does not benefit from Gaussian pre-smoothing which confirms the results from their paper [17]. However, their performance cannot match Flow-Registration, which is up to a factor of 38 better on the clean recording and still 2.6 on the noisy 30 dB recording.

In terms of running time, we cannot directly compare our method with elastix and ANTs due to the fact that the evaluations have been done on an emulated environment. Qualitatively, all evaluations of the other methods were significantly slower than Flow-Reg. fast which took around 0.5 s for the complete estimation without parallelization.

3.2.3.3 Experiments on real recordings

For the real-world evaluation, we include our method, NoRMCorre as well as elastix with gradient constancy and MSE metric. For NoRMCorre, we calculate the displacements for each channel individually and for elastix, we use the joint estimation because the EPE on both channels was lower than the average EPE of the separate channels.

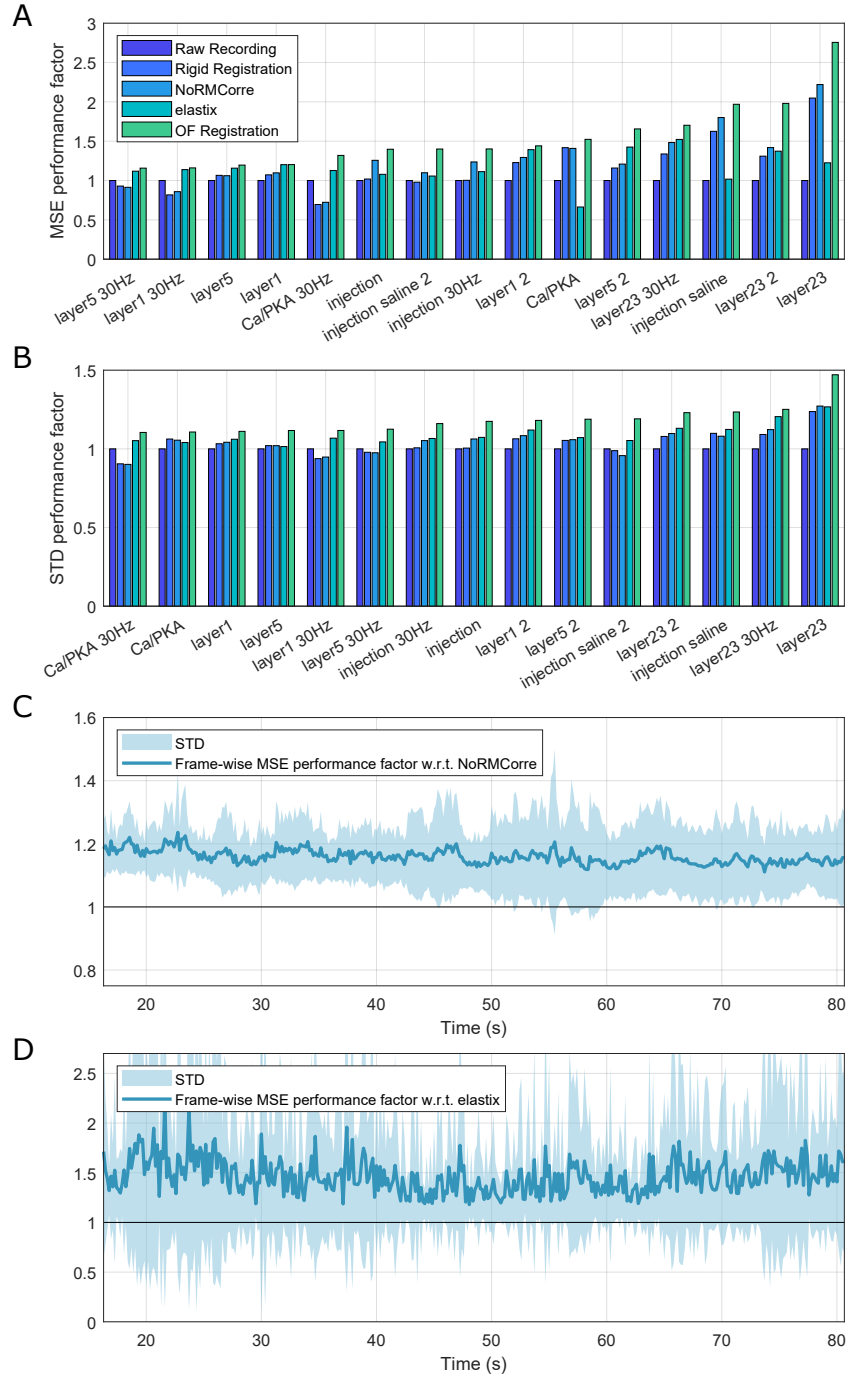


Figure 3.15: Application of Flow-Registration to 15 different datasets. We use 2-photon recordings from layer 5, layer 2/3, layer 1 and three sequences during in vivo drug or saline injection at 6.2 Hz and 30.9 Hz. In all applied metrics, Flow-Registration performs consistently better than rigid registration and NoRMCorre. The metrics are averaged MSE (A) and STD (B) performance factors (see section 3.2.3.1) with respect to the raw recording. (C) displays the frame-wise MSE performance factor of Flow-Registration with respect to NoRMCorre over all 6.2 Hz datasets with 500 frames (layer1 - layer 5, injection, injection saline 2; values > 1 indicate better performance of Flow-Registration). The frame-wise performance is significantly better for Flow-Registration when compared with NoRMCorre (C) and elastix (D) ($p < 0.00001$, paired, two-sided Wilcoxon signed rank test).

Flow-Registration consistently outperforms elastix [164], NoRMCorre [17] and rigid image registration on the benchmark datasets in terms of the reduced temporal STD and MSE ratios (see section 3.2.3.1) as well as *peak signal-to-noise ratio* (PSNR), see Figure 3.15 and Table 3.4. Frame-wise comparisons of frames registered with Flow-Registration and NoRMCorre as well as with elastix show a significantly higher ($p < 0.00001$, paired, two-sided Wilcoxon signed rank test) performance of Flow-Registration, see Figure 3.15 (C) and (D). The graph indicates the factor by which the NoRMCorre MSE is higher than the Flow-Registration MSE (i.e. > 1 indicates better performance of our method). While NoRMCorre has a better frame-wise MSE performance on the 6 Hz recordings than elastix on average, the difference was not found to be significant ($p > 0.05$, paired, two-sided Wilcoxon signed rank test). Qualitatively, our method produces much crisper average frames compared to other methods during challenging motion events such as local drug injections with large intensity changes in the different channels, see Figure 3.13, Figure 3.16 and Figure 3.17.

For a fast approximation of the solution, the minimum warping depth can be used to define the maximum resolution at which the solution is computed. We found around 10x speedup to be possible without reducing compensation quality (avg PSNR 42.523 with the precise and 42.519 with the approximated solution, min level = 6).

In terms of computation time, the MATLAB toolbox is similar to existing methods and even faster with the approximated solution. On a single channel version of the 500 frame injection sequence (consumer workstation, 12 cores @ 3.8 GHz, 64GB of memory), our method (50 iterations, no pre-processing / IO, min level 0) takes 115s and only 15s with min level 6, while NoRMCorre (grid size 32, one iteration) takes 160s and with grid size 16 even 763s (setting used for the benchmark solution). Flow-Registration was run for all datasets with default parameters and $\alpha = 1.5$, for the 6.2 Hz datasets with $\sigma = (1, 1, 0.1)^\top$ and on the 30.9 Hz datasets with $\sigma = (1.5, 1.5, 0.5)^\top$. For the injection sequence, additionally the channel weight was set to $(1.15, 0.85)^\top$. The batch size was set to the size of the datasets.

3.2.4 Discussion

We have presented the new motion compensation method Flow-Registration and can report state-of-the-art registration results. The MATLAB implementation outperforms NoRMCorre ANTs and elastix in terms of registration quality as well as computation speed with comparable parameters.

NoRMCorre is an established 2-photon neuroimaging motion compensation method and the default method in recent 2-photon and calcium imaging suites such as CaImAn (2019) [172], EZcalcium (2020) [174] or Begonia (2021) [171] and can be considered the state of the art for the alignment of 2-photon videos. Our toolbox can produce output files that are compatible with CaImAn as well as with Begonia and can be easily extended for other frameworks and workflows via object oriented IO file readers and writers. ANTs and elastix are intensity based image registration methods which implement different data terms and metrics for mono- and multimodal image registration [164, 166, 165, 167]. Our method outperforms ANTs and elastix by at least a factor of 2 in terms of EPE on the synthetic data. Flow-Registration also shows a consistently higher performance on the real-world recordings than elastix. The best performing elastix configurations were simulated gradient constancy with Gaussian smoothing which is close to the Optimization target in Flow-Registration but without a subquadratic penalizer. Elastix seems to be less robust under changes in high-frequency jitter or motion blur resulting in conflicting performance on the real-world datasets before and after binning. On some of the real-world recordings, elastix shows a better performance than NoRMCorre, however, the frame-wise performance difference for NoRMCorre and elastix was not found to be significant on the 6 Hz recordings. Pnevmatikakis et al. found that registration on structural channels does not necessarily result in better registration performance when compared to registration on the functional channel [17]. While we also could not identify a clear trend on the synthetic data, our evaluations show that the joint optimization of a structural and functional channel consistently outperformed the single channel configurations for Flow-Registration. A potential explanation could be an effective reduction in the SNR for joint structures as well as a higher amount of structural information in the presence of disjoint structures in the different channels. Those effects could be even more relevant for multispectral images with a higher number of channels.

The evaluations on synthetic data have shown that the current state of the art can be outperformed with metrics and regularizers common in variational OF estimation. Particularly, the experiments presented here could demonstrate the benefit of a subquadratic penalizer of the image gradients for low SNR 2-photon imaging data which can perform better than metrics with more invariances under gray value transformations such as mutual information or cross-correlation based metrics. This also suggests that the implementation of subquadratic penalizers and gradient or higher order based constancies could be a useful extension to other image registration toolboxes such as elastix and ANTs.

Flow-Registration reliably corrects movement artifacts even when mixed with intensity changes, as local calcium signals, spreading calcium waves, or calcium responses triggered by local drug injections. The method will record all movements also those of physiological origin, for example, the constriction or dilation of blood vessels. While this implies that those changes cannot be analyzed in the compensated recording, the complete motion information is contained in the estimated displacement fields. Decomposition of the displacements into movements with different transforms or fixed with respect to certain points could enable the selective visualization of movement components of interest for example with Lagrangian motion magnification [F9] as well as quantification of those changes.

We did not systematically quantify the sensitivity of extracted functional signals to the choice of final bilinear or bicubic resampling kernel. All reported results use bicubic, while bilinear is available as a conservative, range-preserving alternative. Motion compensation methods often incorporate many regularization parameters that require difficult fine tuning for different motion statistics. For example, NoRMCorre has a total of 7 regularization and 2 subpixel refinement parameters [17]. A limitation of Flow-Registration as well as of the compared methods is the influence of the parameters on the results making the parameter choice a crucial step in the pipeline. However, our method needs only a single regularization parameter and one parameter for pre-smoothing, while the results have native subpixel precision because of the continuous modeling. We have extended the concept for line scan alignment from our previous work [F4] to the compensation of recordings with narrow FOV by implementing non-uniform warping that allows more pyramid steps along the larger dimension of the recording.

Generally, a drawback of 2D motion compensation approaches for 2-photon imaging is the lack of z-shift correction. While there exist methods for high-speed, online 3D compensation [175], they require complicated setups and the current generation is limited to rigid 3D motion compensation — which might therefore benefit from refinement with a method such as Flow-Registration as well. While our method does not aim for online processing, deep learning-based approaches could unlock real-time applications of our method. Even state-of-the-art self-supervised methods often require annotated data as initialization [176]. Flow-Registration has already been applied to many state-of-the-art 2-photon imaging recordings. The explicit, high-accuracy estimation of displacements with our method can be used for the generation of datasets for the training of efficient, deep learning-based motion compensation methods in the future.

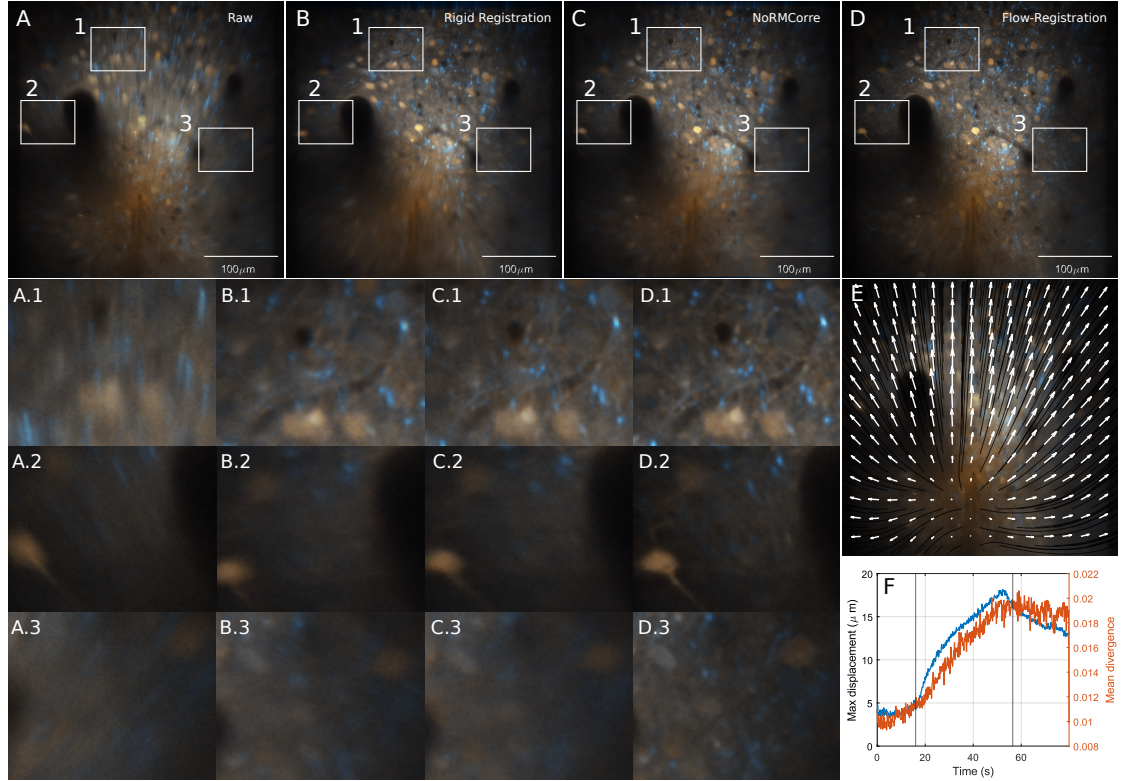


Figure 3.16: Qualitative results of the first in vivo saline injection sequence (dataset *injection saline*). Average of (A) raw images, (B) after rigid registration, (C) after registration with NoRMCorre, and (D) after Flow-Registration. The tissue expands from the injection point (E) resulting in large displacements as well as in a high divergence (F) in the displacement field. Flow-Registration can recover fine structures with much more detail, allowing region-of-interest selection along dendritic structures (blue), while the blur in (A-C) indicates residual motion.

3.2.5 Conclusion

The solutions presented in this section solve the pre-processing problem of motion contamination in 2-photon microscopy and multichannel video recordings with difficult, non-linear motion.

Our evaluations show that subquadratic metrics with gradient constancy and homogeneous smoothness can outperform methods with more advanced data invariances such as given by CC or MI constancies and deformations on this data, while, at the same time, it is easier to optimize as demonstrated in the OF literature. We implemented this method into an accessible MATLAB toolbox and ImageJ plugin.

The core software design paradigm of our toolbox was the easy yet versatile integration into different workflows and toolboxes for 2P imaging. We have developed a MATLAB toolbox that supports common file formats such as MDF, tiff image stacks, MATLAB

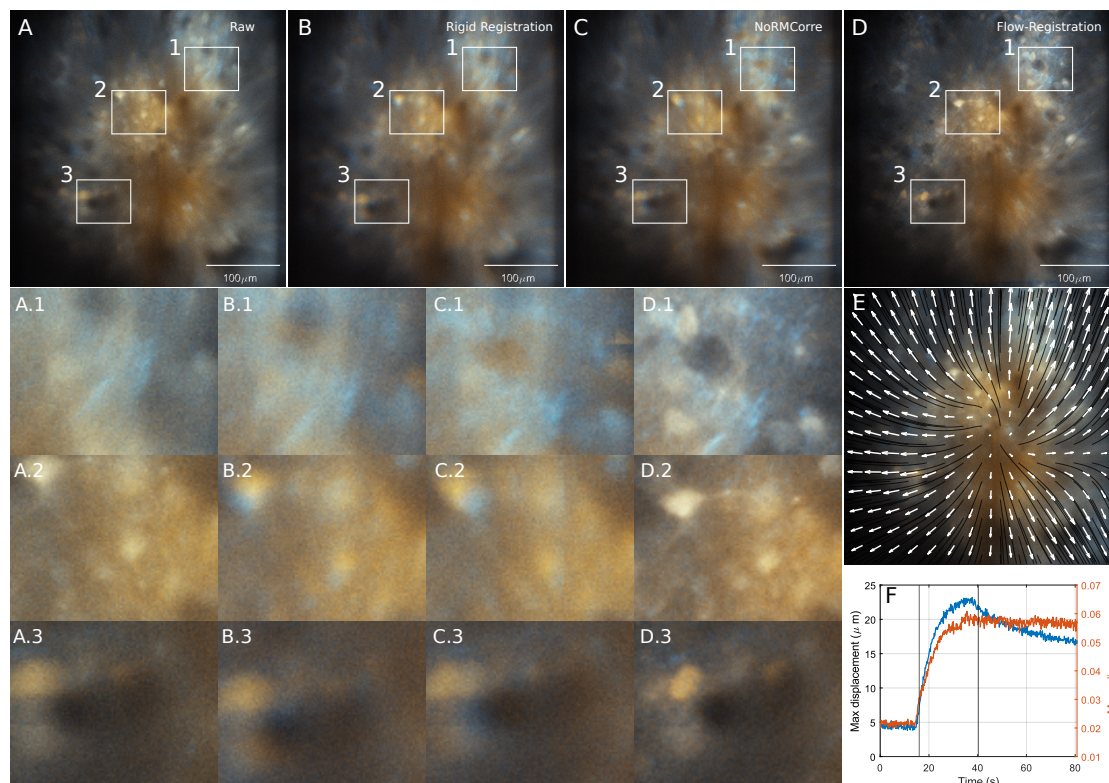


Figure 3.17: Qualitative results of the second in vivo saline injection sequence (dataset *injection saline 2*). Average of (A) raw images, (B) after rigid registration, (C) after registration with NoRMCorre, and (D) after Flow-Registration. The tissue expands from the injection point (E) resulting in large displacements as well as in a high divergence (F) in the displacement field. Flow-Registration can recover fine structures with much more detail, allowing region-of-interest selection along dendritic structures (blue), while the blur in (A-C) indicates residual motion.

mat files or HDF5 files in single file or multichannel configurations. The image IO is designed in a modular, object oriented way, such that the toolbox can easily be extended with new data formats and embedded pre-processing. The code is memory efficient and allows compensating big-data recordings with specified pre-processing methods and on-the-fly binning. The computationally heavy code is written in C++ which allows the implementation of Python wrappers in the future.

The ImageJ / FIJI plugin builds on the ImgLib2 [177] library which supports most common image formats through the bio-formats plugin. The implementations support weighted, multichannel input with weight masks to enforce higher weight on the data term inside of ROIs. The ImageJ / FIJI plugin is integrated with the MATLAB toolbox, such that the plugin can export parameters and reference frame configurations that can be loaded as MATLAB bulk registration jobs.

3.3 RAFT-2P meets Flow-Registration: Synthetic label generation for 2P motion compensation

The previous section discusses the challenges of non-rigid movement artifacts and presents a solution to the problem: The adaptation of variational OF to the statistics of 2P videos that outperforms traditional methods. The Flow-Registration toolbox shows state-of-the-art performance for 2P motion correction. However, the mathematical model behind Flow-Registration has theoretical limitations: It requires spatiotemporal smoothing and relies on a gradient constancy assumption which, due to its dependence on directional information, cannot be used to compensate rotational movement. Furthermore, while variational methods produce very high-accuracy OF fields for small displacements, in recent years, deep learning-based methods for OF estimation such as RAFT achieve state-of-the-art performance on most benchmarks with difficult motion.

Consequently, targeting the limitations of Flow-Registration requires investigating the potentialities and limits of deep learning-based OF estimation for motion compensation in 2P videos. In Chapter 2, RAFT plays a crucial role when fine tuning deep learning-based OF for faces. In the context of 2P imaging, RAFT is a good candidate for deep learning-based motion correction as well: It generalizes well and outperforms related methods on fine tuning tasks [103]. In this section, a modified RAFT OF model is trained with a deep learning strategy and synthetic label generation for 2P motion correction. The results from the previous section lay the foundation for the related dataset as well as image augmentations. The 2P data from the previous section is characterized by a low SNR and high divergence and other displacements. Here, those are modeled into displacement generators, such that this synthetic label generation approximates the motion statistics in 2P microscopy videos and can be applied online during training or to generate large-scale, synthetic training datasets. The performance of the fine tuned model is evaluated on the Flow-Registration benchmarks.

The method presented in this section outperforms Flow-Registration on synthetic benchmarks, in particular, for relevant, low *signal-to-noise-ratio* (SNR) videos and achieves slightly lower but comparable performance on the real-world benchmarks. The method is implemented into an easy-to-use Python toolbox that will be made available upon journal submission of the results described in this section.

3.3.1 Introduction

In recent years, the emergence of deep learning-based methods for OF estimation has led to advancements in OF performance as well as a high availability of high-accuracy methods for real-world data tasks. However, one of the core challenges for deep learning-based OF estimation is the need for datasets with dense ground truth displacements. Different real-world and synthetic datasets have been proposed for that purpose. Common approaches can be to use synthetic sequences rendered from 3D scenes or real-world images. While synthetic videos and images allow the generation of highly accurate ground truth displacements, generalization of algorithms trained on synthetic data to real-world scenes often depends on the type of image augmentations that are applied to the dataset during training [43, 118]. At the same time, estimation of ground truth displacements from real-world images is a non-trivial task and often contains manual annotation steps. Early in-lab approaches to compute ground truth displacements for real-world image pairs made use of special fluorescent dyes which are invisible under visible light. With those dyes, *ultraviolet* (UV) light textures were painted onto the scenes. Each scene was then captured twice, once under visible and once under UV light. The fluorescent textures satisfy the brightness constancy and do not have the aperture problem which allowed easy estimation of the ground truth displacements with existing OF methods. Under visible light, difficult illumination conditions and brightness changes were then modeled. This approach has been used for the Middlebury OF benchmark [99] which has been the state-of-the-art OF benchmark for many years⁵. Another real-world dataset with accurate ground truth is the KITTI dataset and KITTI benchmark suite [178, 179, 119]. KITTI contains city driving scenes that have been annotated with ground truth in a semi-automated fashion. The authors used recordings from 3D laser scans, semi-manual annotation and in the case of KITTI-2015 [119] 3D vehicle models were fitted into scenes with multiple cars. KITTI-2015 is an important dataset to train OF models and is often included into the training schedules of deep-learning-based methods for refinement (e.g. [180, 123, 87]).

Another way to generate OF datasets is via generation of synthetic data. Synthetic sequences from rendered, 3D scenes have the advantage that a ground truth displacement with subpixel accuracy can be derived directly from the scene geometry. The MPI Sintel benchmark makes use of the open source movie *Sintel*. The dataset contains frames from different scenes with various degrees of shading and visual effects and ground truth displacements generated from the scene and camera geometry [102]. The Sintel benchmark⁶ is one of the most important benchmarks for OF estimation and the Sintel dataset is often employed for refinement in state-of-the-art methods (e.g. [180, 123, 87]).

⁵As of November 2024, the latest entries OF EPE in the Middlebury benchmark are from 2021.

⁶<http://sintel.is.tue.mpg.de/results>

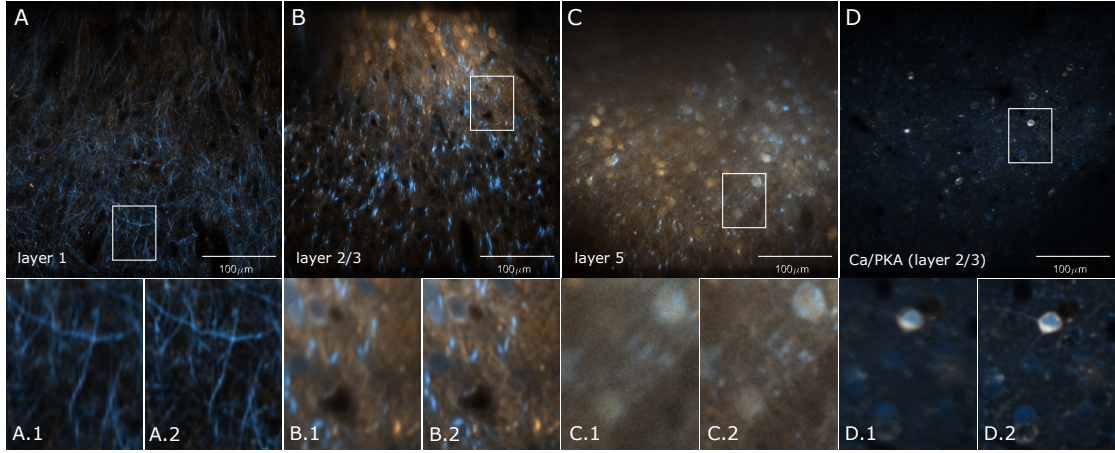


Figure 3.18: The benchmark datasets from layer 1 (A), layer 2/3 (B), layer 5 (C) and the Ca/PKA (layer 2/3) (D) dataset. Temporal average of raw recordings (A.1-D.1) and after application of Flow-Registration (A.2-D.2), both during the motion contaminated frames. Neurons express GAKdYmut (orange) and TurboRFP (blue) (A), GAKdYmut (blue) and TurboRFP (orange) (B-C) and GAKdYmut (orange) and jRGECO (blue) (D). This figure was repeated from the previous section for convenience.

A simple way to create synthetic OF datasets is to combine different 2D images and apply 2D transformations to those. The FlyingChairs dataset had been introduced for the training of FlowNet [118].

FlyingThings3D [101] was one of the first synthetic 3D datasets that was developed exclusively for the training of CNNs for OF estimation, stereo and scene flow. FlyingThings3D contains random objects with random motion transformations with different lighting and in front of different backgrounds. FlyingChairs, FlyingThings3D, Sintel and KITTI-2015 have emerged to be among the most important datasets for supervised OF training and are employed in state-of-the-art methods (e.g. [180, 123, 87]).

The current trend on dataset innovations goes in the direction of self-supervised OF estimation such as self-supervised RAFT [181]. A general strategy is to generate image pairs with ground truth displacements on the fly by means of warping or to use advanced modification of displacements as image augmentations during training.

3.3.2 Materials and methods

This section first reiterates the observational model on which Flow-Registration was built. Based on this, we propose different transformations and motion generators that model the motion statistics of 2P videos. Finally we discuss training of a modified RAFT model for 2P motion compensation.

3.3.2.1 Motion statistics in 2P microscopy videos

The statistics of motion in 2P videos of the brain and other biological tissues differs from the motion in natural videos to the point, where methods that are tailored to 2P applications are required. In the previous work in this chapter [F4, F3], we have identified the following observations in motion statistics of 2P microscopy videos:

- 2P illumination results in a single imaging plane without different motion layers.
- The imaged object is usually soft, biological tissue resulting in smooth motion fields without discontinuities.
- Fixed *region of interest* (ROI) with small displacements between frames and, potentially, large drift over time which can result in large displacements with respect to the reference.
- Due to the nature of the scanning method, horizontal displacements may occur.
- Movement orthogonal to the scanning direction can result in expansion and compression along the y-axis of the image.
- Photometric invariances, such as in [89], cannot be applied due to variable channel numbers. There can exist multiple signal and or structural channels.
- Each of the imaging channels can contain disjoint structures and different SNR characteristics.
- High speed imaging can often only be realized with narrow FOVs.
- There exist paradigms and imaging setups in awake, behaving animals, where large, rotational displacements occur [182].

3.3.2.2 Training data

While the constancy assumptions and motion statistics in variational OF methods are explicitly modeled in the data and smoothness terms, deep learning-based approaches rely on an appropriate representation of the motion statistics in the training data. This dataset is built from static reference frames to which ground truth movement and noise are added. Reference frames are chosen to be averages over video segments after motion compensation with Flow-Registration (see Figure 3.18 for examples).

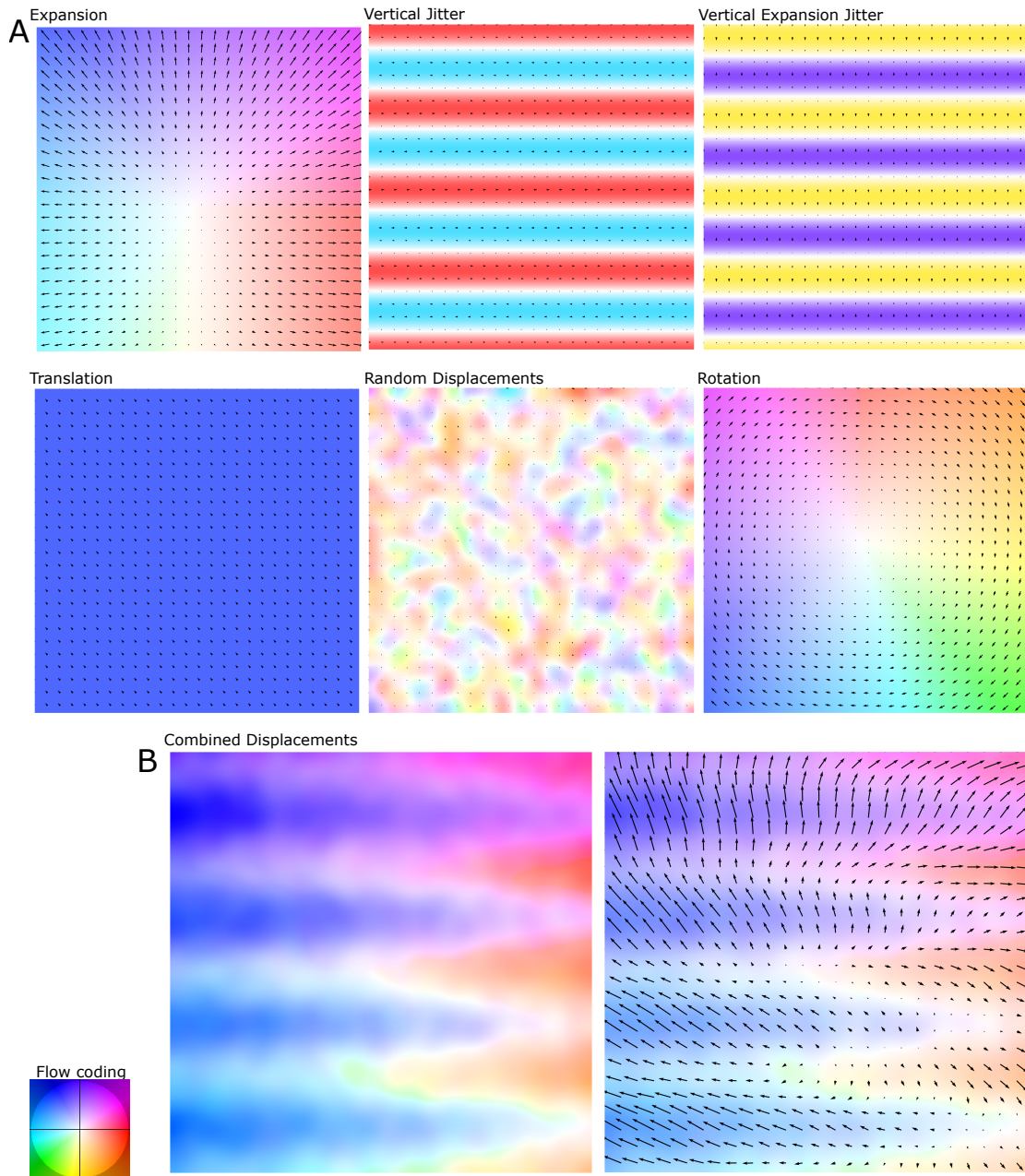


Figure 3.19: The different generated motion components for the RAFT-2P dataset in a combined visualization of color coding and vector plot (A). The combined displacement field is the sum over each of the individual motion components (B) with the color coded OF visualization on the left and the combined visualization on the right.

As it is not possible to reconstruct ground truth displacements from the recordings discussed in the previous section, the approach here is the generation of synthetic ground truth optical flow labels. Similarly to section 2.4, the ground truth labels are applied to the references via forward warping.

There are multiple motion components that were revealed from analyses in the previous section. Based on those components, we propose the following motion generators for synthetic label generation:

- i) **Injection / recoil.** High positive or negative divergence displacement similar to the benchmark dataset in [F3]

$$\mathbf{w}_{\text{inj}}(x, y) = \delta \cdot \begin{cases} (x - x_c, 0.2(y - y_c))^\top, & \text{if } y \geq y_c \\ (x - x_c, y - y_c)^\top, & \text{otherwise} \end{cases},$$

where $(x_c, y_c)^\top$ is the center point of the expansion or contraction and $\delta \sim \text{Uniform}(-\max_\delta, \max_\delta)$. To match the observed, non-uniform displacements after drug injection, the expansion is reduced by a factor of 0.2 below the injection center in the direction of the sink or source.

- ii) **Rotation.** Rotational transformation around a jittered center point. Let $\theta \sim \text{Uniform}(0, 2\pi)$ be a random rotation angle and $\mathbf{x} := (x, y)^\top$, then the displacement for rotation is given by

$$\mathbf{w}_{\text{rot}}(x, y) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x - x_c \\ y - y_c \end{pmatrix} + \begin{pmatrix} x_c \\ y_c \end{pmatrix} - \mathbf{x}.$$

In the final dataset, the center point (x_c, y_c) is randomly jittered.

- iii) **Translation.** Shift in the position of pixels in both the x and y directions. Let $\Delta x, \Delta y \sim \text{Uniform}(-\max_\Delta, \max_\Delta)$ be the displacements in the x and y directions. Then the translation augmentation is given by

$$\mathbf{w}_{\text{trans}}(x, y) = (\Delta x, \Delta y)^\top.$$

- iv) **Horizontal scanning jitter.** Horizontal jitter based on a sinusoidal function. Let $\zeta(y) = \zeta_{\text{magnitude}} \cdot \sin\left(\frac{2\pi y}{T} + \varphi\right)$, where $\zeta_{\text{magnitude}} \sim \text{Uniform}(0, \max_{\zeta})$, $T \sim \text{Uniform}(\min_T, \max_T)$, and $\varphi \sim \text{Uniform}(0, 2\pi)$ denote the magnitude, period, and phase of the jitter. Then the scanning jitter is given by

$$\mathbf{w}_{\text{jitter}}(x, y) = (\zeta(y), 0)^\top.$$

- v) **Vertical expansion jitter.** Vertical jitter based on a sinusoidal function.

Let $\xi(y) = \xi_{\text{magnitude}} \cdot \sin\left(\frac{2\pi y}{T} + \varphi\right)$, where $\xi_{\text{magnitude}} \sim \text{Uniform}(\min_{\text{magnitude}}, \max_{\text{magnitude}})$, $T \sim \text{Uniform}(\min_{\text{periods}}, \max_{\text{periods}})$, and $\varphi \sim \text{Uniform}(0, \pi)$ denote the magnitude, period, and phase of the jitter, respectively. Then the expansion jitter is given by

$$\mathbf{w}_{\text{v-jitter}}(x, y) = (0, \xi(y))^\top.$$

- vi) **Random jitter.** 2D Gaussian filtered uniform random noise $\mathbf{w}_{\text{rand}}$.

Let $E = \{\mathbf{w}_{\text{inj}}, \mathbf{w}_{\text{rot}}, \mathbf{w}_{\text{trans}}, \mathbf{w}_{\text{jitter}}, \mathbf{w}_{\text{v-jitter}}, \mathbf{w}_{\text{rand}}\}$ be the set of all random displacement field operators. A synthetic displacement field for a new sample is then the sum over a random subset $S \subseteq E$:

$$\mathbf{w}(x, y) = \sum_{e \in S} e(x, y).$$

Examples for the different motion components and a generated synthetic ground truth displacement field are shown in Figure 3.19. In this section, a dataset with 20480 image pairs is pre-rendered with random displacements containing all of the above generators. Each motion is added with a probability ranging from 0.2-0.9, where horizontal and vertical jitter are applied with the highest probability.

3.3.2.3 RAFT-2P

Pre-trained RAFT models exist for three channel RGB images. Similarly to section 2.2, where RGB images improved model performance of the thermal landmark predictors, here, the interest is on fine tuning RAFT rather than retraining it. To benefit from pre-training on large datasets, we add a pre-processing layer to RAFT that maps multichannel images to 3-channels. This pre-processing layer is a linear combination with learnable weights of all input channels.

For single channel images, we initialize the input layer weights uniformly. For images with 2 channels, we initialize the weights similar to the visualization 2P representation from the previous section [F3]. Given the minimum intensity values $l = (l_1, l_2)^\top$ and maximum intensity values $h = (h_1, h_2)^\top$ from the temporal average of the raw recordings, the color $(R, G, B)^\top$ at a given pixel $\mathbf{x}$ with intensity values $s := (s_1, s_2)^\top = f(\mathbf{x})$ is calculated as

$$\begin{aligned} R &= (s_2 - l_2)/(h_2 - l_2) \\ G &= 0.5 \cdot (s - l)^\top \cdot \left((h_1 - l_1)^{-1}, (h_2 - l_2)^{-1} \right)^\top \\ B &= (s_1 - l_1)/(h_1 - l_1) \end{aligned}$$

In the context of OF, models are fine tuned to adapt a model for the motion and domain statistics in new data. RAFT was improved with optimized training strategies, datasets and augmentations in various works. For a comparison of different methods see [103]. The RAFT training schedule combines now different datasets in multiple training stages for pre-training and refinement as follows [87]: RAFT is pre-trained on FlyingChairs [118] for 100k iterations, which is followed by 100k iterations on FlyingThings3D [99]. Further, RAFT is fine tuned on 100k iterations on Sintel [101] (RAFT-Sintel), KITTI-2015 [119] and HD1k [120], which is finally followed by 50k fine tuning on KITTI-2015 (RAFT-KITTI). In our experiments, we follow this schedule up to the RAFT-Sintel model and replace the final fine tuning on KITTI-2015 (RAFT-KITTI) with fine tuning on our microexpression (ME) dataset. In the same way as in section 2.4, we use the RAFT-Sintel weights as initialization for fine tuning [F5]. During training, samples from the Sintel final pass [101] are included as well with a probability of 0.2. We apply a set of common image augmentations commonly used for optical flow [87]. Additionally, we model potential activity as a 2D Gaussian function multiplied with the image and add Poisson noise to approximate the shot noise encountered in HFR 2P recordings.

3.3.3 Results

RAFT-2P P1 is benchmarked on the original Flow-Registration real-world and synthetic benchmark from the previous section [F3]. While the new approach does not outperform Flow-Registration, the performance lies above traditional methods (see Table 3.4). Furthermore, RAFT-2P P1 achieves comparable performance without the need for spatiotemporal filtering. Importantly, RAFT-2P P1 is an improvement over vanilla RAFT but vanilla RAFT performs comparable to or better than traditional methods, except for the very low SNR condition layer5 30Hz.

On the synthetic benchmark, RAFT-2P P1 outperforms all existing methods for low SNR (see Table 3.3). In particular, RAFT-2P P1 improves over the vanilla RAFT baseline.

Looking at qualitative results, we notice that both RAFT-2P P1 and vanilla RAFT produce very smooth displacement fields (see Figure 3.20 (B)). However, vanilla RAFT only captures the expansion displacements and fails to capture the horizontal scanning jitter displacements.

3.3.4 Conclusion

This section presents a proof-of-concept for deep learning-based motion correction and motion estimation in 2P neuroimaging videos. The method is based on synthetic label generation and uses RAFT with a pre-processing layer. The approach builds directly on the insights from section 3.2: Various motion generators are proposed based on the motion statistics of 2P neuroimaging videos. While the approach does not outperform Flow-Registration, RAFT-2P offers multiple advantages. Firstly, it does not rely on spatiotemporal filtering and, secondly, it is invariant to rotational movements.

Limitations. The current RAFT-2P P1 results do not disentangle which parts of the synthetic training pipeline drive the observed improvements. In particular, we did not quantify the contribution of the Poisson/noise augmentation relative to the geometric motion generators. The gains in low-SNR conditions may be dominated by noise modeling rather than improved coverage of 2P motion statistics. A controlled ablation is required to attribute these effects. This also limits conclusions about generalization across different microscopy structures: without isolating the effect of noise modeling versus texture/motion augmentation, we cannot yet infer how much structural and texture variability the training data must cover to generalize robustly to other tissue types and imaging conditions.

Future work. The presented results form a foundation to develop the method further. When retraining RAFT for faces in section 2.4, variational OF was used for ground truth prediction and the final dataset consists of a mix of real and generated input images. In the same way, displacements estimated with Flow-Registration could be used to enrich the training data. Furthermore, augmentations and the parameters of the synthetic motion generators need to be fine tuned to the data statistics and RAFT’s capacity to learn from difficult conditions. To this end, a full ablation study is required to study the effects of motion generator parameters, RAFT training and real-world samples on motion correction quality. Furthermore, the performance of the RAFT-2P model architecture for narrow ROI recordings has not been evaluated yet.

Method	Synthetic clean			Synthetic noisy 35dB			Synthetic noisy 30dB		
	ch1+ch2	ch1	ch2	ch1+ch2	ch1	ch2	ch1+ch2	ch1	ch2
RAFT-2P P1	0.26	0.25	0.26	0.48	0.45	0.93	0.74	0.70	1.44
Flow-Reg.	0.05	0.10	0.07	<u>0.52</u>	0.56	1.04	<u>0.90</u>	0.99	1.64
Flow-Reg. fast	<u>0.14</u>	0.22	0.16	<u>0.59</u>	0.66	1.17	<u>1.06</u>	1.10	1.73
vanilla RAFT	0.13	0.12	<u>0.11</u>	1.20	<u>1.07</u>	1.76	1.66	<u>1.59</u>	6.35
NoRMCorre	—	<u>1.93</u>	2.62	—	<u>1.98</u>	3.09	—	<u>2.34</u>	4.27
NoRMCorre*	—	<u>2.26</u>	3.06	—	<u>2.28</u>	3.23	—	<u>2.43</u>	3.89
elastix GC*	0.23	0.59	<u>0.17</u>	1.71	<u>1.18</u>	2.59	2.93	<u>2.00</u>	3.86
elastix MSE	12.32	45.27	<u>6.55</u>	13.64	11.75	<u>7.04</u>	16.64	<u>14.13</u>	35.79
elastix MSE*	<u>1.41</u>	42.84	3.27	<u>2.87</u>	18.86	3.42	4.89	27.48	<u>4.18</u>
elastix CC	7.25	8.65	6.46	<u>6.53</u>	8.17	6.29	6.41	8.21	<u>6.16</u>
elastix CC*	<u>1.20</u>	6.23	2.57	<u>2.65</u>	6.12	3.57	<u>3.70</u>	6.68	4.09
elastix CC + GC*	0.28	0.53	<u>0.21</u>	1.43	<u>1.25</u>	2.67	2.34	<u>1.99</u>	3.76
elastix MI	6.28	<u>5.59</u>	6.91	5.98	<u>5.62</u>	7.16	<u>5.92</u>	6.29	7.28
elastix MI*	<u>1.42</u>	4.06	3.02	<u>2.75</u>	4.66	4.83	<u>4.07</u>	4.90	5.55
elastix MI + GC*	0.36	0.67	<u>0.34</u>	1.42	<u>1.16</u>	2.66	2.03	<u>1.95</u>	3.66
ANTs ElasticSyN	—	<u>1.38</u>	1.54	—	<u>1.68</u>	2.72	—	<u>1.93</u>	3.15
ANTs ElasticSyN*	—	<u>1.44</u>	1.57	—	<u>1.74</u>	2.63	—	<u>1.96</u>	3.30
ANTs SyN	—	1.51	<u>1.42</u>	—	<u>1.60</u>	2.78	—	<u>2.21</u>	3.27
ANTs SyN*	—	<u>1.50</u>	1.51	—	<u>1.61</u>	2.60	—	<u>2.06</u>	3.34
ANTs SyNCC	—	<u>0.13</u>	0.14	—	<u>2.12</u>	3.22	—	<u>3.26</u>	4.41
ANTs SyNCC*	—	0.19	<u>0.17</u>	—	<u>2.23</u>	3.41	—	<u>3.39</u>	4.52

Table 3.3: Average endpoint errors of the estimated displacements (the lower the better). The value indicates the average distance in pixels of the estimated displacements with respect to the ground truth. The best performance per column is put in bold and the best performance per dataset is underlined. Channel ch1 simulates the functional channel with positive multiplicative changes (drug injection) and ch2 the structural with negative multiplicative changes (z-shift). 35 dB is comparable to the 6 Hz and 30 dB to the raw 30 Hz videos. Methods where we applied Gaussian convolution with $\sigma = 1.5$ are indicated by *. NoRM-Corre required Gaussian filtering with at least $\sigma = 0.5$ or would throw an error. Flow-Registration outperforms all traditional methods in terms of EPE with an improvement between one to two orders of magnitude. Those results are reported in [F3] and are extended with the RAFT performances.

While the RAFT-2P P1 prototype outperforms established methods for 2P motion correction, future evaluations will include comparisons with more recent approaches such as REALS, CIAtah, FIFER and others [183, 184, 185, 186].

Finally, a controlled noise ablation is required to disentangle the contribution of noise augmentation from geometric motion synthesis, and to quantify how much structural and texture variability in the training data is required for robust generalization across microscopy structures and imaging conditions. This will inform the final data composition and training schedule. Furthermore, future work will extend the approach to estimate spatially resolved axial drift (z-shift) in addition to in-plane motion, exploiting RAFT features to guide the registration of frames to a reference volume.

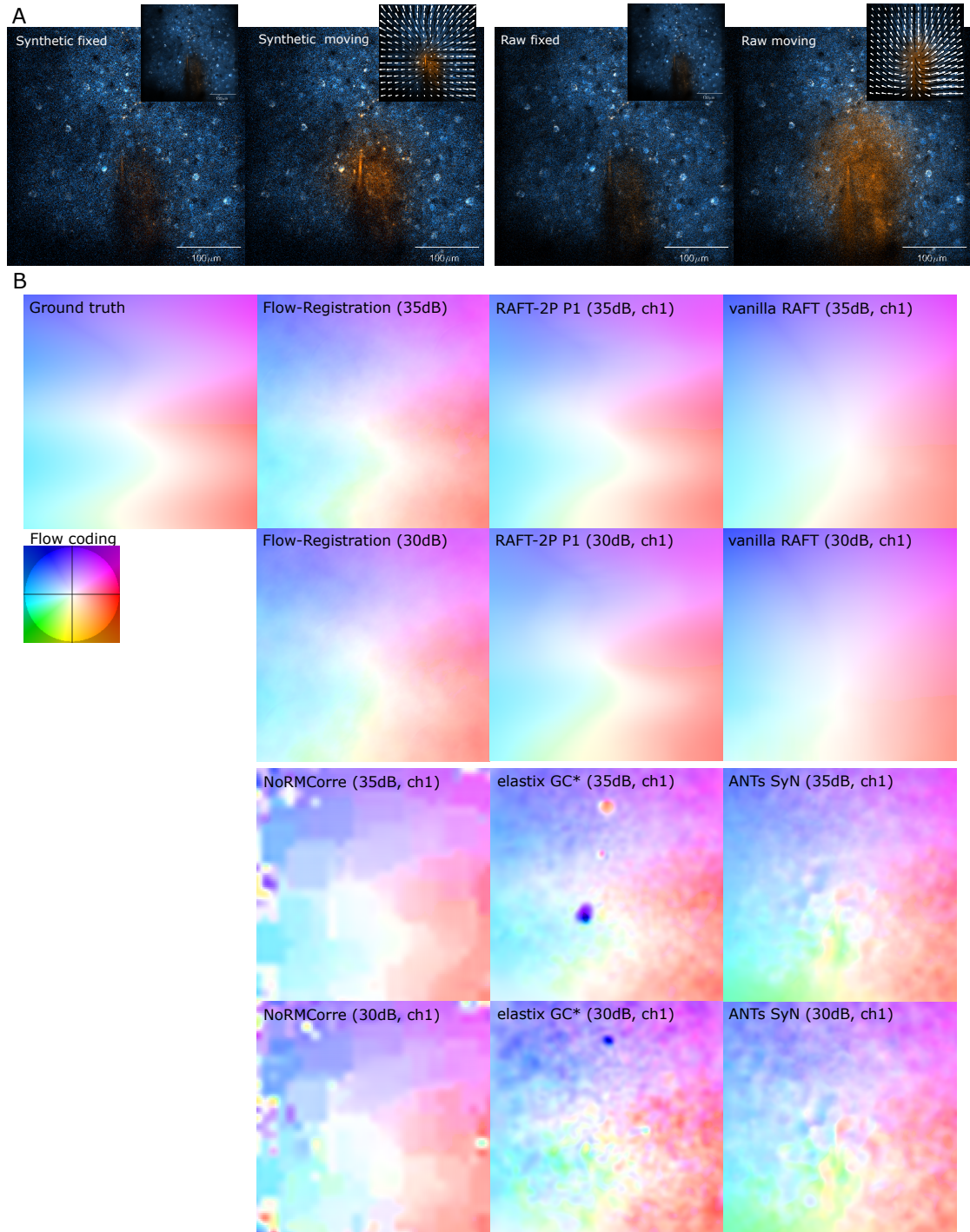


Figure 3.20: The two frames from the synthetic benchmark (A, left) with 35 dB PSNR in comparison with raw frames from the 6 Hz injection data (A, right). The small images show the clean frames without motion and with the synthetic displacement field (A, left) and the estimated displacement field (A, right). The clean frames of the real recording have been calculated by averaging frames from the motion compensated injection sequence. (B) shows the color-coded displacement fields as estimated by each method alongside the ground truth displacement. We report the channel configuration that performed best (see Table 3.3). RAFT-2P P1 is generally able to produce very smooth displacement fields but only 2P refinement recovers the horizontal displacements that can result from the scanning method. This figure has been adapted from [F3] and we have extended it with the RAFT-2P P1 performance.

<i>Dataset</i>	<i>Raw</i>	<i>Rigid</i>	<i>NoRMCorre</i>	<i>elastix GC*</i>	<i>Flow-Registration</i>	<i>vanilla RAFT</i>	<i>RAFT-2P P1</i>
Ca/PKA	86.16	87.57	87.56	84.90	87.75	87.33	87.40
Ca/PKA 30Hz	82.09	80.64	80.79	82.60	83.01	82.35	82.62
injection	38.98	39.60	41.46	39.14	42.52	42.01	42.33
injection 30Hz	38.40	38.84	40.25	38.66	41.09	40.38	40.73
injection saline	78.21	80.30	80.85	77.99	81.34	80.88	81.21
injection saline 2	78.63	78.61	79.18	78.77	80.16	79.03	80.00
layer1	91.66	91.87	91.96	92.36	92.37	91.44	91.64
layer1 2	87.53	88.13	88.31	88.56	88.69	88.24	88.25
layer1 30Hz	85.13	84.26	84.44	85.69	85.76	84.89	85.26
layer23	77.14	79.47	79.81	77.82	80.79	76.94	77.29
layer23 2	89.24	90.24	90.60	90.30	91.87	91.15	91.10
layer23 30Hz	73.25	74.30	74.79	74.89	75.45	73.89	74.08
layer5	90.11	90.34	90.32	90.70	90.83	90.38	90.33
layer5 2	88.73	89.16	89.29	90.01	90.6	89.89	90.31
layer5 30Hz	83.76	83.45	83.38	84.24	84.39	75.76	83.97

Table 3.4: Average PSNR values (the higher the better) on each dataset for the different methods. To reduce the influence of shot noise and isolate changes due to residual movement, the recordings have been filtered with a 2D Gaussian with $\sigma = 3$ before estimating the PSNR which results in large PSNR values and images that only differ in (sub-)pixel displacements. PSNR has been calculated with respect to the maximum value 2^{16} . The highest PSNR value in each row is put in bold. Those results are reported in [F3] and are extended with the RAFT performances. While RAFT could not outperform Flow-Registration on any of the real-world datasets, RAFT-2P P1 almost always outperformed vanilla RAFT and often had a higher performance than the traditional image registration techniques.

3.4 Broader impact

This section discusses the broader impact of the methods developed in this chapter. Since its publication, Flow-Registration has been continuously improved with feedback from the research community. The next section summarizes the most important improvements. Furthermore, OF-based motion correction has been applied as a proof-of-concept to the quantification of EDA correlates on the human palm [F19, F8] with optical imaging methodology. With the right illumination, changes in EDA can produce tiny sweat dots on top of the sweat glands. Those dots have a diameter of only a few pixels, and high-accuracy motion correction enables the computation of relative metrics such as the fractional change.

3.4.1 Flow-Registration

Flow-Registration is used for 2P motion correction within neuroscience research [20, 21, 22] and is optimized for specific use-cases. The following are advancements of Flow-Registration since publication.

- **Narrow ROI compensation.** 2P recordings in linescan mode are an important tool to record high speed activity such as voltage changes. However, linescans offer only limited capacity for motion compensation. While we have shown that the problem can be solved with variational methods in one dimension in this chapter [F4], movement orthogonal to the scanning direction cannot be corrected. An alternative is to use ROIs that cover only a narrow spatial band around the object of interest such as dendritic structures. Alignment of these narrow ROIs poses a challenge to multiscale approaches, where downsampling usually happens uniformly in both dimensions. Therefore, the smaller dimension of the image guides traditional multiscale approaches. The Flow-Registration pyramid solver is optimized to handle narrow ROI recordings. We apply non-uniform scaling and continue the pyramid in the larger dimension until a minimum size is reached. To further mitigate the large displacements parallel to the large axis of narrow ROI scans, CC-based initialization of displacements proved to be an important addition for difficult data with large motion changes over time. Here, x and y displacements are first corrected for translational movement via cross-correlation and subsequently refined with the OF-based motion compensation.
- **Energy-based metric and clustering.** Flow-Registration supports a multi-reference compensation mode that has been adapted and extended from our previous work [187].

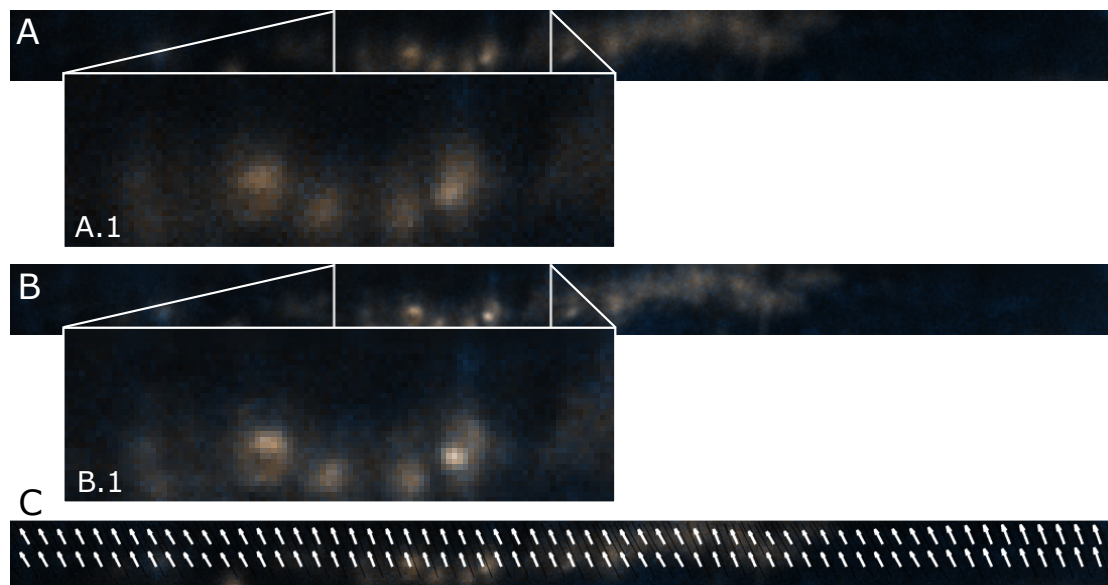


Figure 3.21: Average frame of videos with narrow FOV around dendritic structures before compensation (A) and after compensation (B). The estimated displacements are visualized in (C). There is only a small movement orthogonal to the larger image dimension. CC-based pre-alignment and registration with Flow-Registration results in crisp structures (B.1) when compared with the video before alignment (A.1) allowing stable extraction of the signal of interest.

In 2P videos, movements orthogonal to the imaging plane can result in vanishing image structures. Based on the observation that shifts in z-direction correlate with the objective function of OF, i.e. a lower similarity indicates z-shift, each frame can be clustered with respect to the difference of the gradients. Each of the clusters is subsequently compensated with respect to the cluster mean as reference, resulting in stable recordings for each imaging plane.

- **Compatibility.** Flow-Registration exists as an ImageJ/FIJI plugin as well as a self-contained MATLAB toolbox with rich IO capabilities. Supported input file types cover the most common imaging types. The toolbox is compatible with the MATLAB-based 2P imaging suites such as Begonia [171], CaImAn [172] and NANSEN [182].
- **Wide-field microscopy.** In the context of wide-field microscopy, the image contains contributions from different depths. The narrow focus plane induces a relationship between depth and the spectral content of an image. The observation that there is a relation between the movement components and spectral image content inspired the registration on a Laplacian pyramid [188] which we extended into a semi-synthetic dataset to benchmark motion compensation methods for wide-field microscopy recordings [F14].

3.4.2 Motion compensation for remote EDA from RGB videos of the human palm

The *electrodermal activity* (EDA), which measures skin conductance changes, is an important modality in psychophysiological research. Changes in EDA reflect sympathetic nervous system activity primarily through eccrine sweat gland activation and large changes can result in surface sweat [189]. In our work, we have demonstrated that RGB videos of the human palm can visually capture correlates of EDA [F8, F10, F19] under optimized lighting conditions. Observable reflectance depends on the angle of the camera and illumination sources with respect to the imaged object's surface normals. Therefore, when imaging the curved surface of a palm, different illumination angles increase the reflectance contrast in different regions, depending on the skin surface.

One way to extract EDA correlates from such videos is the frame-by-frame analysis and detection of individual sweat dots, using, for instance, photometric and gradient-based analysis techniques [F8]. While this is robust with respect to movement noise, it does not allow the analysis of individual sweat dots over time nor the qualitative visualization of changes. Temporal analysis of the videos is not straightforward. In contrast to microscopy recordings of the brain, the palm surface contains a distinct, spatially high-frequency texture that changes with activity. As a consequence, residual motion introduces large, localized variation across frames. Here, we build on the insights from this chapter for OF-based motion correction to remove all motion from palm recordings and extract EDA correlates. To this end, we use the same metrics for relative activity analysis that are used for the analysis of optical recordings in neuroscience research [137]. Computing the fractional change not only results in relative signal changes but also allows spatiotemporal analysis of the signals.

3.4.2.1 Materials and methods

For a detailed description of the paradigm and study cohort, we refer to [F8]. Here, we use the recordings of participant 18 during the auditory paradigm. The following contains direct excerpts from [F8].

Data acquisition. Ten short audio clips were selected from the International Affective Digitized Sounds (IADS-2) database [190]. These sounds are standardized and designed to elicit emotional and physiological responses, including changes in electrodermal activity (EDA). Each audio clip was played for 7 seconds. The inter-stimulus interval (ISI) was randomly varied between 40 and 60 seconds to prevent habituation and maintain participant engagement. During the ISI, participants viewed a fixation cross displayed on a monitor to maintain visual focus and minimize distractions. Participants wore calibrated headphones (Sennheiser HDA 300) to ensure consistent and accurate sound delivery. The sound pressure level (SPL) was calibrated to 60 dB to standardize the auditory stimulus intensity across participants. The experiment was conducted in a darkened room to reduce ambient light interference with the specialized illumination setup used for the palm recordings. The controlled environment helped enhance the visibility of sweat-induced luminance changes on the palm surface. A multimodal trigger system was used to synchronize the onset of auditory stimuli with both the sensor-based EDA recordings and the camera-based video recordings. This synchronization ensured accurate alignment of stimuli presentation with physiological responses in subsequent analyses.

Fractional change computation. Functional optical imaging in related areas quantifies intensity changes of 2D signals in terms of the fractional change [137]. For an image f and baseline f_b , the fractional change can be computed as $\Delta f = \frac{f - f_b}{f_b}$. However, this requires that f has been registered onto f_b by means of high-accuracy motion correction. To this end, we use an OF-based motion compensation approach similar to the methods described in this chapter. We use the RAFT optical flow backend [87] that has been fine tuned on faces in Chapter 2 [F5].

The recording uses three light directions that are temporally multiplexed to enable the recording of different specular reflection areas with the best possible SNR. Hence, the largest area of activity is obtained by combining all light directions into a single recording and the high-accuracy motion correction allows the analysis of spatial activity patterns in the same way as for optical imaging videos.

The challenges with applying OF to high-resolution images include memory limitations for deep-learning-based methods such as RAFT and scaling problems with $\mathcal{O}(n^2)$ algorithms in variational OF solvers as well as the accumulation of image changes over time. Therefore, we compute one initial OF field on the downsampled image, apply motion compensation and refine the results on a predefined ROI. This is done for all three lights, where separate references are kept for each light and each sequence separately.

After frame-wise motion correction, we combine signals from the three light sources. Let the set of light directions be denoted by $\mathcal{L} = \{0, 1, 2\}$. The raw video sequence consists of frames $(f_k)_{k=1}^{3T}$, where T is the number of final, fused images. The light direction cycles sequentially such that for a given fused time step $t \in \{1, \dots, T\}$, the corresponding raw frames are indices $k = 3(t - 1) + l + 1$ for each $l \in \mathcal{L}$. We define the fused frame F_t at time step t by taking the maximum intensity value across the corresponding registered pixels for the three light directions:

$$F_t(\mathbf{x}) = \max_{l \in \mathcal{L}} \hat{f}_{3(t-1)+l+1}(\mathbf{x})$$

where $\hat{f}_k(\mathbf{x})$ represents the pixel intensity at location $\mathbf{x} = (x, y)^\top$ in the motion-corrected frame corresponding to the raw frame f_k .

The expected signals in the fused recording F are tiny reflectance changes at the locations of the eccrine sweat glands on the skin surface. At the same time, a *remote photoplethysmography* (rPPG) signal as a measure of blood perfusion is expected below the skin surface. We apply a simple observation model to separate the recording into these two components, where the rPPG signal is approximated by the spatially median filtered frame $\tilde{F}$. This allows us to define the reflectance signal as $F_{\text{refl}} := F - \tilde{F}$.

To compute EDA correlates from the fractional change ΔF , we first calculate the median filtered image $\tilde{F}$. For a baseline F_{base} , defined as the temporal average of the pre-stimulus period, the fractional change is calculated as $\Delta F_t = (F_t - F_{\text{base}})/F_{\text{base}}$. We compute ΔF and $\Delta \tilde{F}$ accordingly. Finally, we report the average of the high-pass filtered changes $\Delta F - \Delta \tilde{F}$ (see Figure 3.22). To reduce sensor noise, the signal F is pre-smoothed with a Gaussian filter ($\sigma = 1$ px), and the final activity maps are obtained by averaging the fractional changes across all color channels.

3.4.2.2 Results

We have applied the image processing pipeline for the quantitative analysis of optical imaging microscopy recordings to videos of the human palm in a paradigm to evoke EDA activity. We could demonstrate the feasibility of remote EDA estimation from RGB videos in the past [F10] and motion compensation in combination with ROI analysis and the estimation of relative intensity changes has been applied to those data. While quantitative measures of the activity can be extracted from single frames, motion compensation enables spatiotemporal analysis in the same way as for optical neuroimaging videos. Figure 3.22 C and D show clear activity patterns on the palm in a high activity condition with respect to a baseline. The spatial high-frequency component corresponds to sweat activity, while the low frequency component reflects changes in blood perfusion.

For the evaluation of the palm recordings, RAFT-based motion correction has been applied to 20 recordings of the hand during stimulation with 2 s pre-stimulus baseline and 15 s post-stimulus recording for analysis. The average over the 2 s baseline has been used as a reference for fractional change estimation. While the recordings also contain a resting period at the beginning of the experiments, using an average over the resting period as baseline performed poorly due to z-drift and slight rotations of the hand over time which induced changes in shading with respect to the resting period. Figure 3.22, E shows similar signals for EDA activity and the video-based correlates in the spatial averages, while the fractional change analysis without motion correction results mostly in noise (light gray plot). Pearson's correlation analysis using a 0.5 Hz low-pass filter on both signals revealed a weak, non-significant correlation between EDA and ΔF **without** motion correction ($r = 0.287 \pm 0.348$, $p = 0.155$), but a significant, positive correlation between ΔF **with** motion compensation and EDA ($r = 0.855 \pm 0.112$, $p < 0.001$).

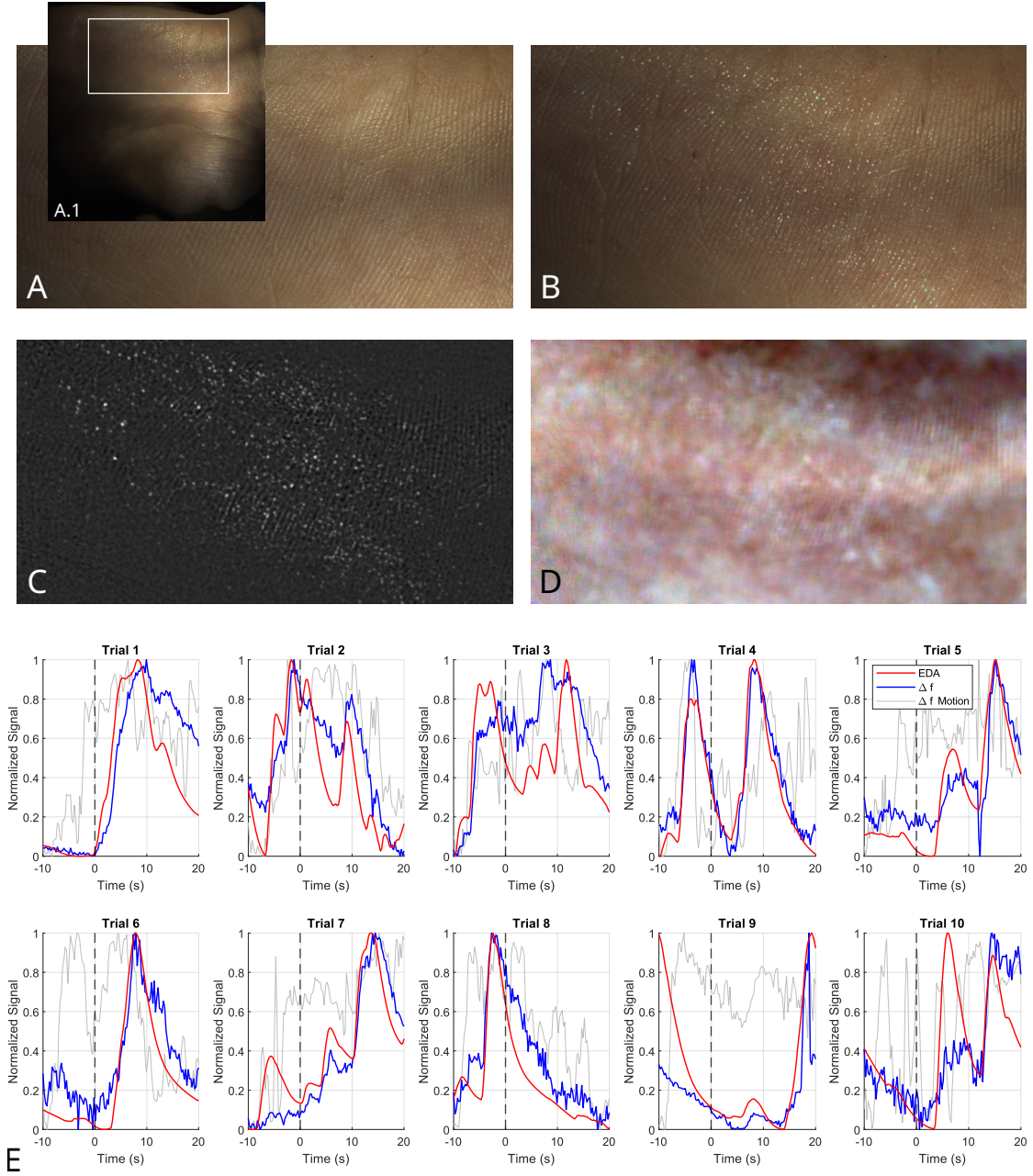


Figure 3.22: Recordings of the human palm with low (A) and high (B) EDA activity leading to surface sweat. OF-based motion correction of a ROI with uniform illumination (A.1) allows the analysis of the fractional change ΔF . Here, the spatially high-pass filtered ΔF contains sweat activity (C), while the spatially low-pass filtered ΔF reveals changes in subcutaneous blood perfusion. The spatial average of the high-pass filtered ΔF contains video correlates of the EDA signal (E): Pearson's correlation analysis using a 0.5 Hz low-pass filter on both signals revealed a weak, non-significant correlation between EDA and ΔF **without** motion correction ($r = 0.287 \pm 0.348$, $p = 0.155$), but a significant, positive correlation between ΔF **with** motion compensation and EDA ($r = 0.855 \pm 0.112$, $p < 0.001$). For details on image acquisition and quantitative evaluation see [F8]. The presented samples are from the auditory paradigm of participant 18.

3.5 Summary

This chapter presents new motion correction methods for functional 2P microscopy imaging and related neuroimaging modalities. The proposed variational approach for 1D alignment allows fast alignment of 2P line scan recordings with difficult displacements as well as the correction of drift in multi trial matrices in EEG measurements.

The proposed variational OF model for 2P motion correction is tailored to the statistics of 2P videos. For the task of OF-based motion correction of 2-photon microscopy neuroimaging videos with difficult movement, the presented method outperforms the current state of the art. The implementation in the Flow-Registration framework interfaces with popular 2-photon imaging suites and is continuously improved with the feedback from the research community.

Furthermore, preliminary results in this chapter demonstrate a proof-of-concept for deep learning-based 2P motion correction with synthetic label generation. The motion types from the observational model used in the variational OF method have been used to define different motion generators. Using these generators, we presented a proof-of-concept for deep learning in optical flow with synthetic label generation. This approach fine tunes RAFT for 2P motion estimation. While the model outperforms many traditional methods for 2P motion correction, it could not significantly improve upon Flow-Registration.

In Chapter 2, limitations of existing methods for neuroimaging applications were revealed and the presented solutions target the gap between the performance of state-of-the-art methods, datasets, algorithms, and the requirements of the tasks. At the same time, a focus is on finding suitable baselines instead of reinventing the wheel. This chapter follows the same cross-domain adaptation approach. While methods such as NoRMCorre [17] propose a completely new approach for non-linear motion correction developed only for 2P videos, Flow-Registration builds on a rich framework of OF estimation that has been studied in the context of many different types of image data. By targeting the differences in image statistics and optimizing the model, Flow-Registration shows state-of-the-art performance, serving as a valuable resource for neuroscience research and being used in high-impact studies such as [20, 21, 22].

Chapter 4

Outlook, discussion and limitations

This thesis makes contributions to visual computing and biomedical imaging in the context of systems neuroscience and neuroimaging. It solves cross-domain challenges for problems related to **motion analysis**, **motion visualization**, **motion correction**, and **multimodal image alignment** for sensor fusion. It identifies and closes gaps between existing methods and the unique requirements of neuroimaging data and videos recorded during psychophysiological experiments. The proposed methods use state-of-the-art baselines and difficult benchmarks, e.g., for *2-photon* (2P) motion correction in the same way as for the estimation of thermal facial landmarks. The evaluations in this thesis show that vanilla computer vision methods developed for RGB can outperform domain-specific approaches on challenging benchmarks. However, the methods presented here go further by achieving state-of-the-art performance.

Chapter 2 introduced a fully calibrated multimodal multi-camera setup that has been used for large-scale data acquisition in out-of-lab scenarios. It laid the groundwork for developing methods to remotely assess human affective and neurocognitive states. First evaluations demonstrated its suitability for the extraction of vital parameters with relevance for the detection of respiratory diseases, which laid the groundwork for interdisciplinary and multi-institute research projects and collaborations. The multi-camera setup has been used at the Systems Neuroscience & Neurotechnology Unit to acquire large-scale datasets, including some of the largest paired RGB-NIR-thermal facial datasets available today [F1, F8, F6, F2].

Motion magnification research has been dominated by Eulerian methods for years [11, 191, 106]. By identifying their limitations for the motion magnification of faces, this thesis revisited the Lagrangian approach to motion magnification. Modern methods for OF estimation, in particular, when fine tuned on faces with the approach proposed in this thesis, are well suited for OF-driven, Lagrangian motion magnification. Here, the local motion magnitude and a priori image information such as facial landmarks guide the amplification and produce magnified videos with excellent visual quality. Furthermore, the developed methods allow motion magnification of facial micromovements during head motion. In particular, the method for facial landmark-based motion decomposition is a step towards the development of methods for identification and analysis of micromovements in the face for novel medical applications. In the context of respiratory diseases, rhinosinusitis, clearing of the throat, coughing, suppressed sneezing, and other symptoms could lead to distinct micromotions in the face.

An important contribution of this thesis is the analysis of the statistics in different types of 2P microscopy data for 2D videos as well as line-scan and narrow ROI recordings to develop new motion correction approaches. The methods described in Chapter 3 also work for non-visual neuroimaging data such as EEG recordings. Flow-Registration outperforms the previous state-of-the-art and is a valuable resource for neuroscience research [22, 21, 20].

4.1 Outlook and limitations

4.1.1 2P motion compensation and motion analysis

While Flow-Registration demonstrates state-of-the-art performance w.r.t. methods from traditional image processing, the model choices and implementation details imply theoretical limitations: The data term uses gradient constancy, which, on the one hand, ensures invariance under functional intensity changes of the indicators but on the other hand, contains directional information of image structures. Therefore, gradient constancy is violated by rotational movement. Additionally, the gradients are sensitive to noise. Therefore, as recordings become noisier, they will require increasingly larger integration of the spatiotemporal neighborhood. We have implemented this with 3D Gaussian convolution. A larger σ value, however, assumes some level of spatial and temporal coherence in the videos. This means that more filtering will limit the detection of fast movements such as scanning jitter as well as motion of small structures.

We were unable to match the performance of Flow-Registration with the preliminary RAFT-2P P1 prototype on the original Flow-Registration benchmark. However, a RAFT-based motion compensation promises higher performance with large displacements and higher performance under noise and for fast changing displacements such as scanner jitter. RAFT-2P P1 works without the help of any spatiotemporal filtering as a pre-processing step and does not require any temporal coherence in the videos. In its current state, the method could already be a suitable alternative for Flow-Registration on very noisy data with fast motion noise. To more closely match the performance of Flow-Registration, the RAFT-2P training strategy as well as the parameters for synthetic data, including better background textures, are investigated next. Furthermore, the synthetic data might benefit from the same approach presented in section 2.4, where variational OF estimates are included in the training data.

A limiting factor in motion compensation for 2P videos is the thin focus plane that is only defined by the point spread function of the scanning laser. Movement along the z-axis can completely change image content resulting in movements that cannot be compensated with 2D motion compensation approaches. However, z-movement can correlate with x-y image movement. Hence, prediction of a z-shift error could be included in the motion prediction process of RAFT-2P. Additionally, pre-recorded 3D volumes around the imaging plane can be exploited for intelligent baseline prediction [192]. Extending this to registration of the 2D scans into 3D volumes could further improve 2P baseline correction.

4.1.2 Lagrangian motion magnification and motion decomposition

With this work, we have made an important contribution to the motion decomposition of facial movements which helps with the analysis of local movements in the face. However, there are limitations to the landmark-based decomposition presented here. While the landmark-based motion compensation works well for small head motion, difficult motion such as large head rotations can lead to fake, parallax induced motion components that cannot be inpainted with the current strategy. Furthermore, the sparse-PCA-based approach only estimates non-overlapping spatiotemporal components. Hence, both methods are limited in the context of large non-translational head movements. Future improvements could involve developing methods that do not rely on explicitly chosen facial landmarks for interpolation-based decomposition. Instead, these methods could focus on separating parameterized 3D head movements, such as rotations and translations, from local facial motions.

In this work, we have only discussed the visualization aspect in the form of motion magnification. However, the strength of motion decomposition for motion magnification lies in the explicit estimation of 2D motion components in the face. The potential for motion decomposition in the face in highly dynamic scenes for motion magnification and microexpression analysis still seems to be largely untapped. Consequently, those estimated features could also form a valuable resource for microexpression and micro-movement quantification and estimation. Extension of the methods to 3D via scene-flow estimation and 3D face reconstruction (e.g. [44]) potentially further improves the meaningfulness of estimated motion components.

The sparse PCA-based approaches discussed here enable decomposition into spatially and temporally sparse motion components. This approach works best in the absence of large, global motion. This, however, poses a core challenge for micromotion analysis in-the-wild. Future developments for sparse dictionary learning on displacement fields should aim at learning sparse representations that allow the reconstruction of global head movements from displacement fields that contain both local and global displacements analogous to other image and video reconstruction tasks [110].

None of the developed methods can properly compensate for parallax induced motion. This means that under head motions such as rotations, the global flow of areas with higher or lower profile than the rest of the face cannot properly be interpolated. As a result, facial parts such as the nose could contain residual movement in the local flow after decomposition. In the future, this could be solved by parameterized head motion estimation that maps the movement onto 3D face models.

Backprojection into the image coordinate system of the magnified frame would allow parallax compensated decomposition into local and global displacements.

Eulerian motion magnification is part of many pipelines for microexpression detection. After exploration of Lagrangian motion magnification and the high performance of modern optical flow methods, the self-evident next step is to replace Eulerian motion magnification with a decomposed movement field that contains only displacements of interest. For such a method, motion magnification would be a visualization in the sense of model explainability that is not inherently required by the pipeline and classification or detection approach itself.

4.1.3 Facial landmarks on thermal images

With this work, we could demonstrate how dense landmarks can be trained jointly for RGB and thermal images with state-of-the-art performance. However, we had to resort to annotation of the FAKE dataset with RGB landmarks due to limited availability of the dense ground truth from the dataset [43]. Future work on thermal landmarking will benefit from synthetic datasets with highly available dense ground truth landmarks.

However, post-thermalizing a synthetic RGB dataset, while effective, as proposed here, diverges slightly from the typical paradigm of large-scale synthetic data, which is traditionally created directly within the target imaging domain. Instead of working with simulated RGB images, an emission-based thermal renderer would allow directly outputting thermal images for a synthetic dataset from simulated temperatures. This would require automated analysis of large databases with thermal faces for robust thermal face representation learning. Our dense landmarker, therefore, could serve as a key technology that would allow analysis of large, thermal datasets of faces to develop generative models for simulating thermal identities similar to [43].

Abbreviations

2P	2-Photon
CASME	C hinese A cademy of S ciences M icroexpression
DAN	D eep A lignment N etwork
DIS	D ense I nverse S earch
DLC	D eep L ab C ut
DTSC	D rive T hrough S creening C enter
DTSCs	D rive T hrough S creening C enters
DVMAG	D ynamic V ideo M AGnification
ECG	E lectro c ardiography
EDA	E lectrodermal A ctivity
EEG	E lectro e ncephalography
ELE	E uler- L agrange E quations
EPE	E nd P oint E rror
EMG	E lectro m yogram
EyeAR	E ye- A spect R atio
FAN	F acial A lignment N etwork
GAN	G enerative A dversarial N etwork
GNLL	G aussian N egative L og L ikelihood
GT	G round T ruth
HFR	H igh F ramerate
IEEE	I nstitute of E lectrical and E lectronics E ngineers
ISI	I nter- S timulus I nterval
KITTI	K arlsruhe I nstitute of T echnology and T oyota T echnological I nstitute
MAE	M ean A bsolute E rror
ME	M icroexpression

MHR	M ean H eart R ate
MPI	M ax- P lanck- I nstitute
MSE	M ean S quared E rror
NANSEN	N euro A nalysis S oftware E Nsemble
NIR	N ear I nfrared
NME	N ormalized M ean E rror
OF	O ptical F low
OIST	O kinawa I nstitute of S cience and T echnology
OI	O ptical I maging
PCA	P rincipal C omponent A nalysis
PCR	P olymerase C hain R eaction
PPE	P ersonal P rotective E quipment
PSNR	P eak S ignal-to- N oise R atio
RAFT	R ecurrent A ll- P airs F ield T ransforms (for optical flow)
RGB	R ed G reen B lue
RI	R edness I ndex
ROI	R egion of I nterest
rPPG	R emote P hotoplethysmography
RMSE	R oot M ean S quare E rror
SARS-CoV-2	S evere A cute R espiratory S yndrome C oronavirus T ype 2
SMURF	S elf- T eaching M ulti- F rame U nsupervised RAFT with F ull- I mage W arping
SNNU	S ystems N euroscience and N eurotechnology U nit
SNR	S ignal-to- N oise R atio
SPL	S ound P ressure L evel
SVD	S ingular V alue D ecomposition
TFW	T hermal F aces in the W ild
w.r.t.	w ith r espect t o

Appendix A

Comparison of forward warping approaches

In this appendix, we compare splatting-based forward-warping strategies against the triangle-based approach proposed in section 2.4. In particular, we compare our triangle-based forward warping [F5] against forward-splatting baselines such as average splatting and softmax splatting [122] in combination with Navier-Stokes (NS) image inpainting [193]. We use a synthetic benchmark with controlled expansion and translation of a circle on a 512×512 grid. Specifically, the circle is expanded by factors $(1.3, 1.5)^\top$ and $(3, 1.5)^\top$ and optionally translated to the right by 60 pixels (see Figure A.1).

Figure A.2 and Figure A.3 illustrate characteristic artifacts of forward splatting under large divergence motion (expansion/disocclusions). Forward splatting scatters discrete source samples with finite kernel support. Under large divergence motion from expansion, the target region becomes under-sampled, so some target pixels receive no contributions. This leads to visible holes such as the regions between the vertical lines in Figure A.3, bottom right. Increasing kernel support can reduce holes, but tends to blur boundaries and introduce haloing. Inpainting can fill remaining holes (black regions in Figure A.2) but introduces structured artifacts as well as errors in the target shape (e.g. Figure A.3 middle row, right).

Our triangle-based warping mainly shows artifacts when no depth buffering is used (see Figure A.2, third row): overlapping warped triangles at motion discontinuities are resolved by the fixed rasterization order rather than geometry. This leads to direction-dependent layering. With depth buffering, the warp is hole-free, but can still exhibit mild boundary blur where triangles interpolate across foreground/background discontinuities (Figure A.3, middle left).

Overall, the comparison highlights that splatting-based forward warping is prone to under-sampling artifacts under large divergence motion, whereas triangle rasterization yields hole-free warps. Occlusion ordering is handled robustly when a depth buffer (or depth map) is available, making it ideal for large divergence motion that could occur in magnified facial videos during events such as opening of the mouth or blinking of the eye.

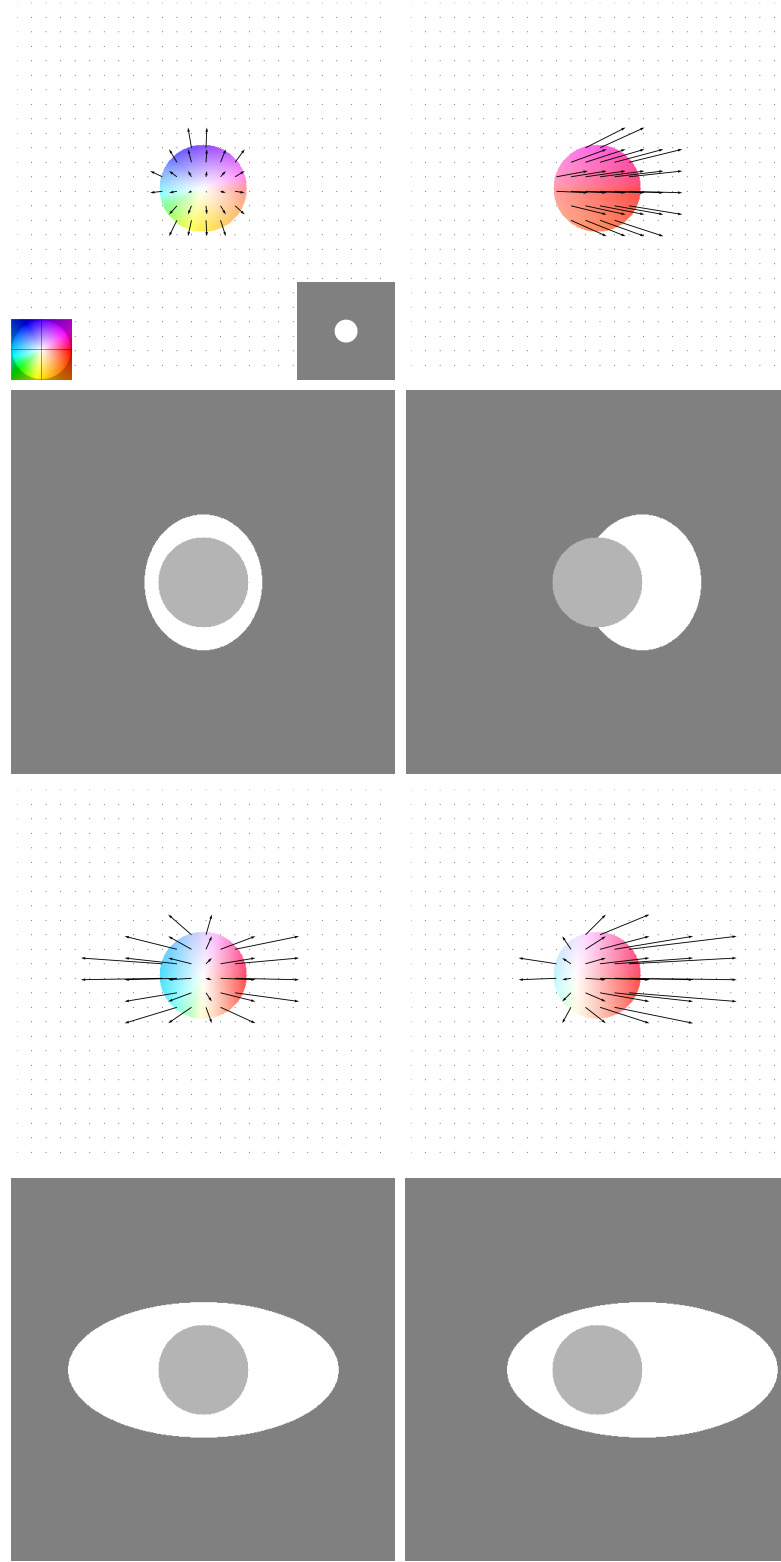


Figure A.1: Forward warping task with difficult displacements where the points on the still image project to sparse positions in the moving image. The four displacements enlarge and translate a white circle on a gray background. Displacements contain different levels of expansion, namely by factors $(1.3, 1.5)^\top$ (top) and $(3, 1.5)^\top$ (bottom), on the right with and on the left without translation by 60 pixels. The image resolution is 512×512 . The grayscale images show the ground truth circle after warping in white as well as the original circle in light gray. The OF fields are visualized with color-coding and arrow visualization.

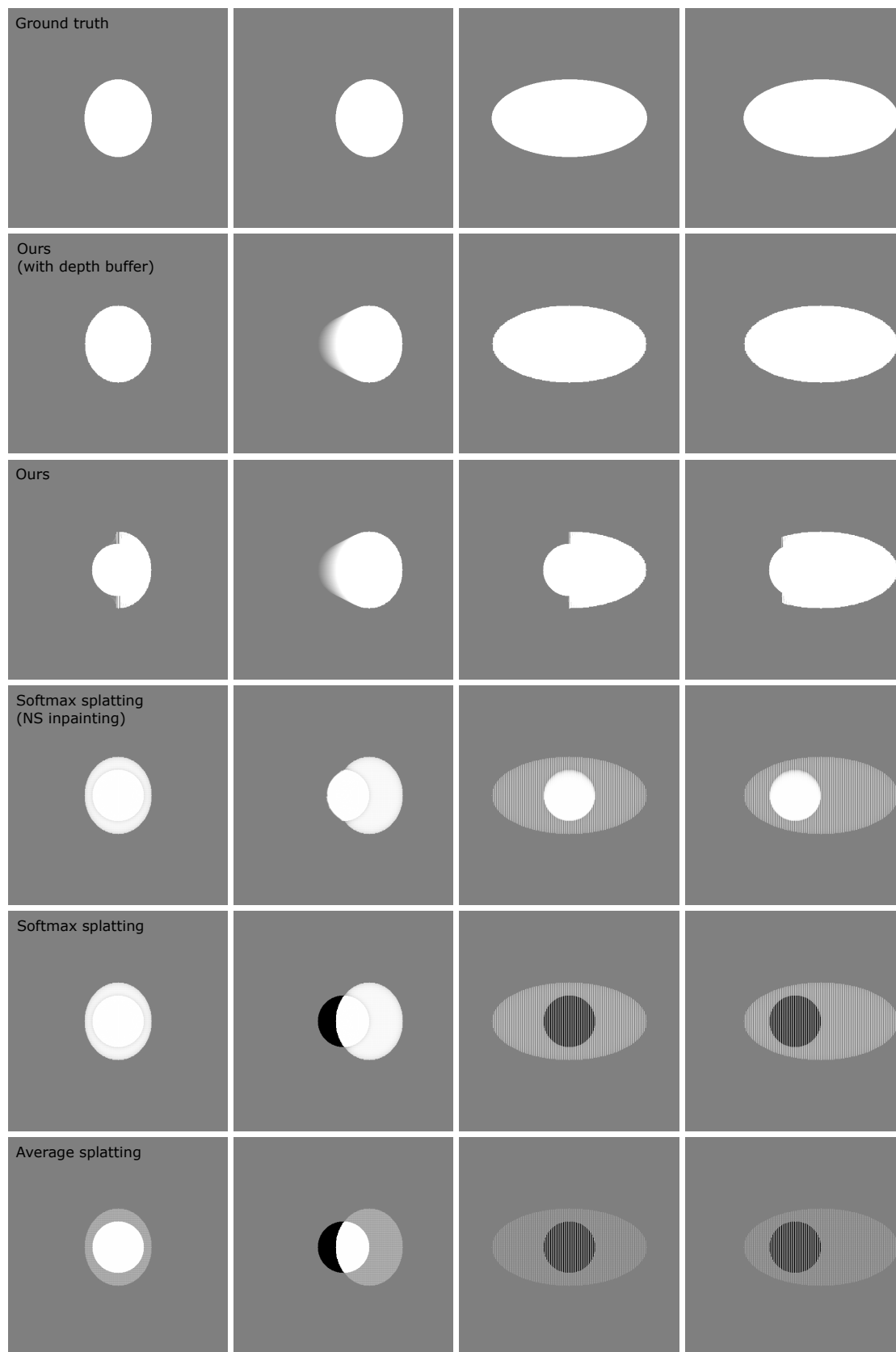


Figure A.2: Comparison of different forward splatting approaches with our proposed triangle based approach. Columns show four synthetic deformations combining expansion and translation for a different method on each row. Average splatting tends to show artifacts in each image while softmax splatting produces visible holes in particular for the motion with large divergence. Black regions indicate unmapped pixels from disocclusions. NS inpainting removes these holes but introduces structured haloing/stripping and boundary bias.

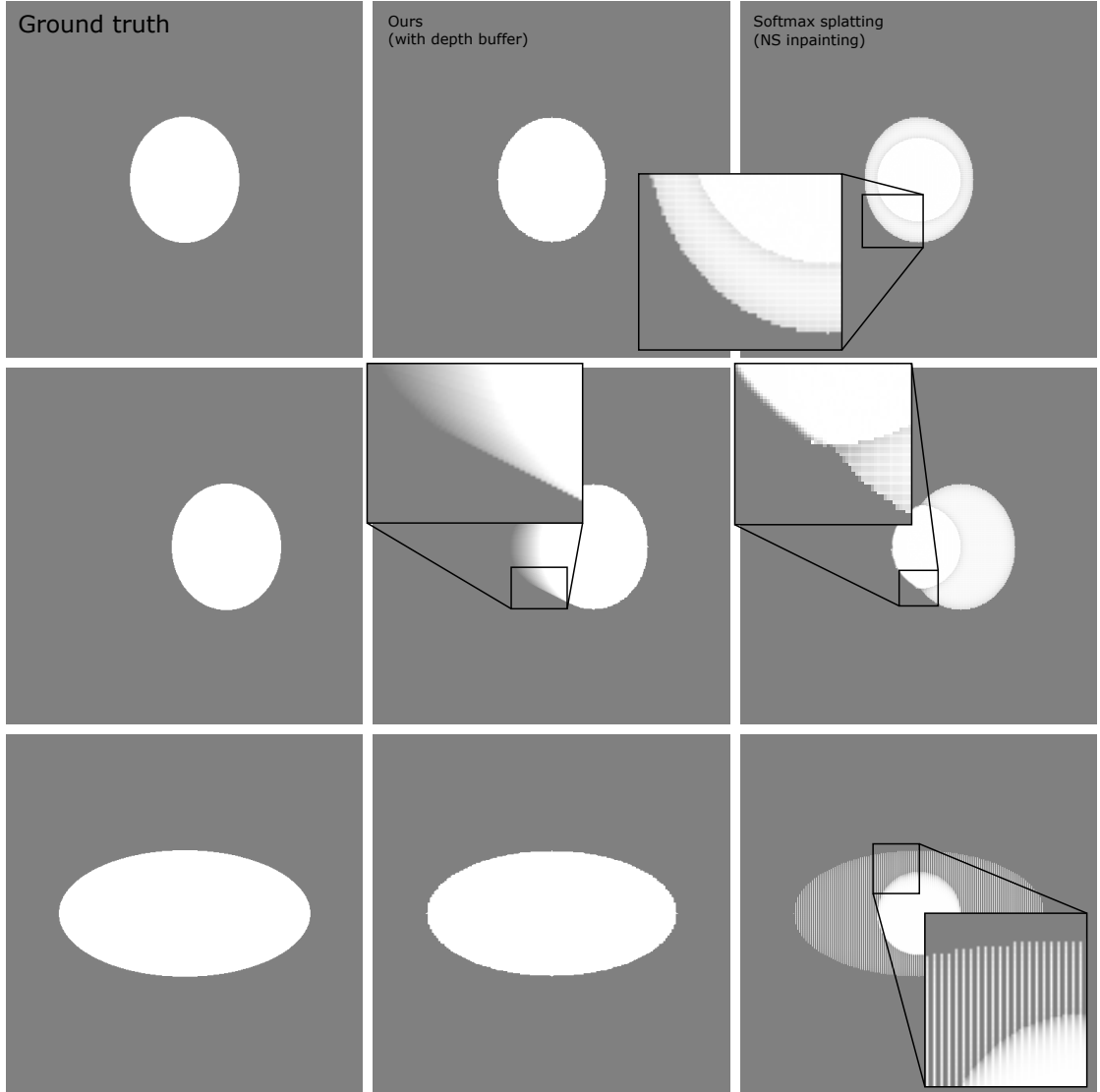


Figure A.3: Comparison of triangle-based forward warping with depth buffer (left) and softmax splatting with NS inpainting (right). The rows show the results on the first three synthetic examples. Softmax splatting exhibits characteristic splatting artifacts even after hole filling via NS inpainting: Haloing around expanding regions and grid-aligned striping in the large divergence case. While our approach is hole-free, it exhibits mild boundary blur where triangles interpolate across foreground/background discontinuities (middle left).

Appendix B

Motion-induced dependence of signal and noise

We show that in pixel time courses, residual motion artifacts are signal-dependent and therefore generally violate additive noise models that assume independence between signal and noise.

Lemma 1. *Let $(f_t)_{t=1}^T$ be an image sequence with $T \in \mathbb{N}^+$, where each frame $f_t : \Omega \rightarrow \mathbb{R}$ is defined on $\Omega \subset \mathbb{R}^2$. Let $c : \Omega \rightarrow \mathbb{R}$ be a spatially non-uniform (time-invariant) background texture and let $G : \Omega \rightarrow \mathbb{R}_{>0}$ be a spatial gain function. Let $(s_t)_{t=1}^T$ be a spatially non-uniform, time-varying signal of interest with $s_t : \Omega \rightarrow \mathbb{R}$.*

Suppose the sequence is affected by motion described by displacement fields $(\mathbf{w}_t)_{t=1}^T$ with

$$\mathbf{w}_t(\mathbf{x}) := (u_t(\mathbf{x}), v_t(\mathbf{x}))^\top : \Omega \rightarrow \mathbb{R}^2, \quad (\text{B.1})$$

and assume $\mathbf{x} + \mathbf{w}_t(\mathbf{x}) \in \Omega$ for all $\mathbf{x} \in \Omega$ and $t \in \{1, \dots, T\}$. The observed frame at time t is

$$f_t(\mathbf{x}) = G(\mathbf{x}) \left(c(\mathbf{x} + \mathbf{w}_t(\mathbf{x})) + s_t(\mathbf{x} + \mathbf{w}_t(\mathbf{x})) \right). \quad (\text{B.2})$$

Then, for any fixed pixel $\mathbf{x} \in \Omega$, the motion-induced term in an additive decomposition of the pixel time course $(f_t(\mathbf{x}))_{t=1}^T$ depends on the underlying scene content (signal and background) and is, in general, not statistically independent of the signal of interest.

Proof. Fix $\mathbf{x} \in \Omega$ and consider the gain-normalized measurement

$$g_t(\mathbf{x}) := \frac{f_t(\mathbf{x})}{G(\mathbf{x})} = (s_t + c)(\mathbf{x} + \mathbf{w}_t(\mathbf{x})). \quad (\text{B.3})$$

Assume that $s_t + c$ is differentiable near $\mathbf{x}$ and is approximately linear over the region traversed from $\mathbf{x}$ to $\mathbf{x} + \mathbf{w}_t(\mathbf{x})$, so that $(s_t + c)(\mathbf{x} + \mathbf{w}_t(\mathbf{x}))$ can be approximated by its first-order Taylor expansion about $\mathbf{x}$. Then

$$(s_t + c)(\mathbf{x} + \mathbf{w}_t(\mathbf{x})) \approx (s_t + c)(\mathbf{x}) + \mathbf{w}_t(\mathbf{x})^\top \nabla(s_t + c)(\mathbf{x}). \quad (\text{B.4})$$

Hence, in the gain-normalized additive model

$$g_t(\mathbf{x}) = (s_t + c)(\mathbf{x}) + n_t(\mathbf{x}), \quad (\text{B.5})$$

the motion-induced term satisfies the first-order approximation

$$n_t(\mathbf{x}) \approx \mathbf{w}_t(\mathbf{x})^\top \nabla(s_t + c)(\mathbf{x}). \quad (\text{B.6})$$

Therefore, unless $\nabla(s_t + c)(\mathbf{x}) = \mathbf{0}$ (locally constant scene) or $\mathbf{w}_t(\mathbf{x}) = \mathbf{0}$ (no motion), the motion-induced term is a non-trivial function of the underlying (signal+background) field through its spatial derivatives. Consequently, it cannot, in general, be statistically independent of the signal of interest. $\square$

In particular, the linearized expression shows that motion artifacts scale with the local spatial gradient of the scene: even subpixel residual motion can induce large temporal fluctuations near edges or other high-frequency structures. Hence, residual motion can violate additive, signal-independent noise assumptions used in downstream processing.

Appendix C

Motion magnification error

We show that spatially varying gain combined with residual motion can introduce a multiplicative artifact in gain-normalized measurements that can both scale the recovered intensity and magnify the apparent displacement of activity patterns. Through the linearization that also underlies Eulerian video magnification [11], where small spatial shifts appear as temporal intensity modulations, we explain why such normalization can exaggerate motion signatures of the signal of interest such as propagating waves of activity.

To quantify the motion error, we regard a simplified OI model, where the recording is contaminated by motion and the signal of interest contains additional motion, such as a propagating wave. Let $\Omega \subset \mathbb{R}^2$ and let $(f_t)_{t=1}^T$ be an OI sequence with $f_t : \Omega \rightarrow \mathbb{R}$. We model the static structure of the scene as a spatially varying baseline intensity field $B : \Omega \rightarrow \mathbb{R}_{>0}$, which acts as a multiplicative gain for the active signal. We assume very small noise motion and a smooth moving signal of interest such that motion noise in the signal is negligible.

Suppose the gain function is affected by motion described by displacement fields $(\mathbf{w}_t)_{t=1}^T$ with

$$\mathbf{w}_t(\mathbf{x}) := (u_t(\mathbf{x}), v_t(\mathbf{x}))^\top : \Omega \rightarrow \mathbb{R}^2, \quad (\text{C.1})$$

and assume $\mathbf{x} + \mathbf{w}_t(\mathbf{x}) \in \Omega$ for all $\mathbf{x} \in \Omega$ and $t \in \{1, \dots, T\}$. The motion magnitude of $\mathbf{w}_t$ is assumed to be in the range of single- to subpixel displacements, such that linearization assumptions hold true.

The signal of interest may contain additional motion (e.g. a propagating wave) described by displacements $(\hat{\mathbf{w}}_t)_{t=1}^T$ with

$$\hat{\mathbf{w}}_t(\mathbf{x}) := (\hat{u}_t(\mathbf{x}), \hat{v}_t(\mathbf{x}))^\top : \Omega \rightarrow \mathbb{R}^2, \quad (\text{C.2})$$

and we define the (motion-containing) signal term

$$F_t(\mathbf{x}) := s_t(\mathbf{x} + \hat{\mathbf{w}}_t(\mathbf{x})). \quad (\text{C.3})$$

The observed frame at time t is modelled as

$$f_t(\mathbf{x}) = B(\mathbf{x} + \mathbf{w}_t(\mathbf{x})) F_t(\mathbf{x}). \quad (\text{C.4})$$

Except for the motion, this model fulfills the assumptions of fractional change and normalization with a temporally constant baseline-frame $f_{\text{base}} : \Omega \rightarrow \mathbb{R}_{>0}$, which we define as the motion-free gain field $f_{\text{base}}(\mathbf{x}) := B(\mathbf{x})$. Linearization of B and normalization then yields

$$\frac{f_t(\mathbf{x})}{f_{\text{base}}(\mathbf{x})} = \frac{B(\mathbf{x} + \mathbf{w}_t(\mathbf{x}))}{B(\mathbf{x})} F_t(\mathbf{x}) \approx F_t(\mathbf{x}) + F_t(\mathbf{x}) \cdot \underbrace{(\mathbf{w}_t(\mathbf{x})^\top \nabla B(\mathbf{x})) B(\mathbf{x})^{-1}}_{c_t(\mathbf{x})}.$$

That means the error induced only by motion after normalization is given by $F_t(\mathbf{x}) \cdot c_t(\mathbf{x})$.

If we now again linearize F_t in $\hat{\mathbf{w}}_t(\mathbf{x})$, we get

$$\frac{f_t(\mathbf{x})}{f_{\text{base}}(\mathbf{x})} \approx c_t(\mathbf{x}) \cdot s_t(\mathbf{x}) + s_t(\mathbf{x}) + (1 + c_t(\mathbf{x})) \hat{\mathbf{w}}_t(\mathbf{x})^\top \nabla s_t(\mathbf{x}).$$

This approximates the function

$$\left[\frac{f_t}{f_{\text{base}}} \right] (\mathbf{x}) \approx c_t(\mathbf{x}) \cdot s_t(\mathbf{x}) + \underbrace{s_t \left(\mathbf{x} + (1 + c_t(\mathbf{x})) \hat{\mathbf{w}}_t(\mathbf{x}) \right)}_{\text{motion-magnified signal}}.$$

So the error term $(\mathbf{w}_t(\mathbf{x})^\top \nabla B(\mathbf{x})) B(\mathbf{x})^{-1}$ not only modulates s_t resulting in intensity changes, but apparently it also modulates the displacements of the signal of interest. This way, the extracted signal is a superposition of a scaled signal $c_t(\mathbf{x}) \cdot s_t(\mathbf{x})$ and signal movements which are not $\hat{\mathbf{w}}_t(\mathbf{x})$, but contain displacements

$$\left(1 + (\mathbf{w}_t(\mathbf{x})^\top \nabla B(\mathbf{x})) B(\mathbf{x})^{-1} \right) \hat{\mathbf{w}}_t(\mathbf{x}).$$

This effect is illustrated in Figure C.1, where synthetic data with only horizontal motion of the background texture and gain function (max. 0.3 pixels) already produces systematic deviations in the fractional-change time courses measured at fixed locations. This means we get a potential quantitative error when estimating motion patterns in the signal such as propagating waves, which depends on the noise motion magnitude and the spatial gradient of the gain function. This holds true even if the spatial gain function is fully known. A temporally constant gain function is also part of more sophisticated methods, such as Raguet et al. [194], and therefore those methods might benefit from more robust motion compensation approaches. Therefore, especially in the context of the estimation of propagating waves, where the propagation speeds are subject to academic debate, we can advise the application of strong motion compensation approaches based on our theoretical investigations.

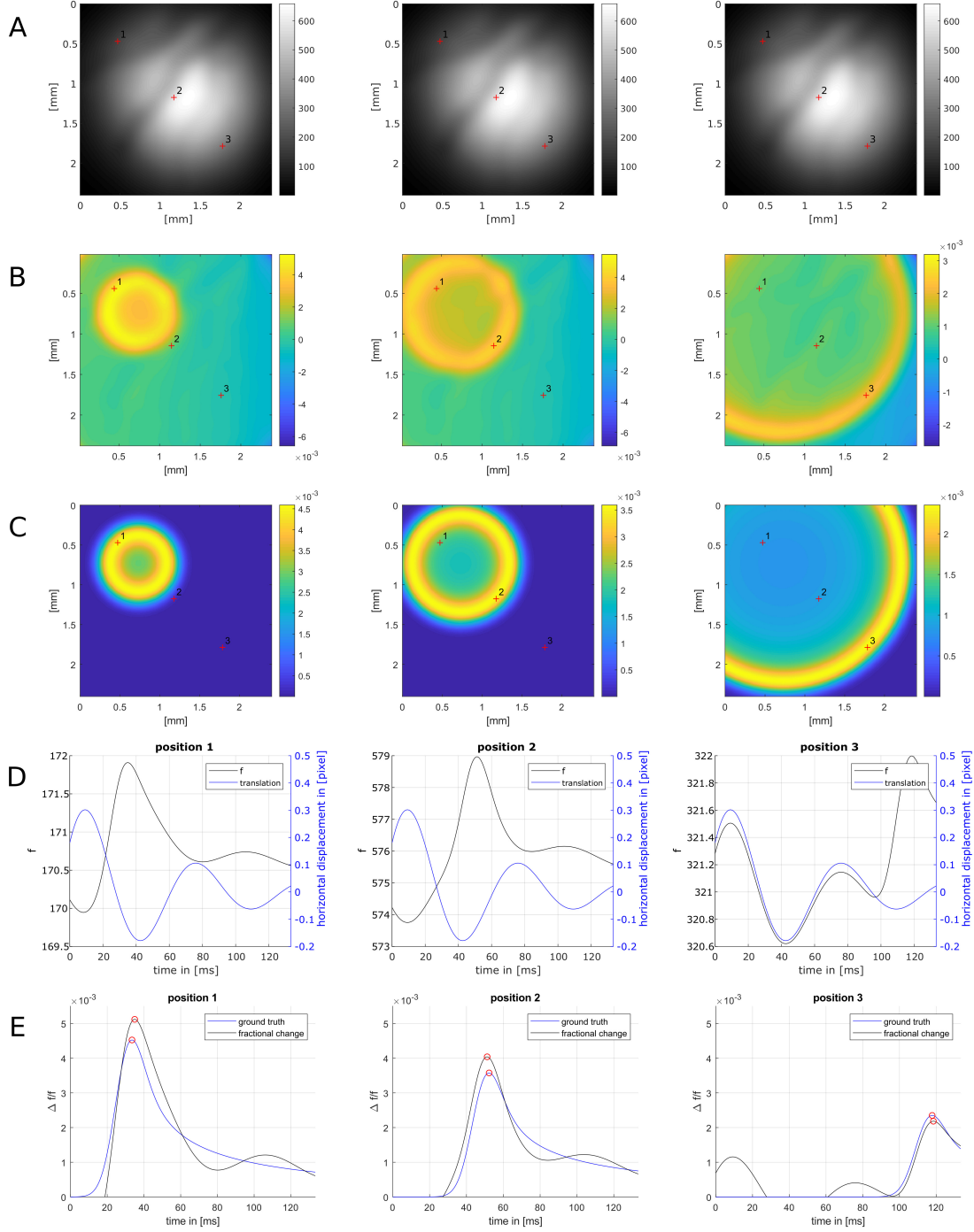


Figure C.1: Experiments with synthetic data: (A) are the uniformly scaled input frames at the peak times of the ground truth. (B) is the simulated activity, reconstructed with differential imaging in (C). The only noise is horizontal movement of the background texture and the gain function with a maximum magnitude of 0.3 pixels (D). (E) are the time courses of the ground truth and fractional change calculated as $(f_t - f_{\text{base}})/f_{\text{base}}$ measured at the positions 1–3.

Appendix D

Parameter selection and data presentation for Flow-Registration

D.1 Parameter selection

For Flow-Registration, parameters were chosen based on visual inspection. For NoRMCorre, parameters from the NoRMCorre repository demo scripts were modified based on the statistics such as maximum displacements that have been calculated from the OF results. NoRMCorre incorporates different regularization approaches that make it more robust under noise and implicitly define the properties of the compensated displacements. There are a total of 7 parameters for regularization and 2 for subpixel refinement against the single regularization parameter α and additional solver specific parameters for Flow-Registration, while subpixel accuracy is natively supported. We found it necessary to modify 4 of the NoRMCorre parameters for our datasets (see Table D.1). Dynamic update of the reference was turned off due to the reference based metrics that were used for evaluation. In practice, on the injection sequences, this reduced drift, but slightly increased jitter at the same time.

dataset	grid_size	max_shift	max_dev	update_template
layer 1	$(32, 32)^\top$	—	—	false
layer 1 30Hz	$(64, 64)^\top$	—	—	false
layer23	$(32, 32)^\top$	—	—	false
layer23 30Hz	$(64, 64)^\top$	—	—	false
layer5	$(64, 64)^\top$	—	—	false
layer5 30Hz	$(128, 128)^\top$	—	—	false
CaPKA	$(32, 32)^\top$	—	—	false
injection	$(16, 16)^\top$	25	$(25, 25)^\top$	false
injection	$(16, 16)^\top$	25	$(25, 25)^\top$	false
injection 30Hz	$(64, 64)^\top$	25	$(25, 25)^\top$	false
injection saline	$(128, 128)^\top$	10	$(26, 26)^\top$	false
injection saline 2	$(32, 32)^\top$	15	$(25, 25)^\top$	false

Table D.1: NoRMCorre parameters used for the compensation of the benchmark data. All unmentioned parameters have been kept at the default setting of the MATLAB code (October 2021). If the second version of the dataset is not mentioned, parameters match the first version.

D.2 False color representation

The microscopy images presented in Chapter 3 aim to visualize artifacts in the average frames caused by residual motion in two imaging channels. Therefore, we make use of an inclusive (linear) false color representation that allows distinction of the channels both with deuteranopia and protanopia. Given the minimum intensity values $l = (l_1, l_2)^\top$ and maximum intensity values $h = (h_1, h_2)^\top$ from the temporal average of the raw recordings, the color $(R, G, B)^\top$ at a given image position $f(x, y) = (p_1, p_2)^\top$ is calculated as

$$\begin{aligned}
 R &= (p_2 - l_2)/(h_2 - l_2) \\
 G &= 0.5 \cdot (p - l)^\top \cdot \left((h_1 - l_1)^{-1}, (h_2 - l_2)^{-1} \right)^\top \\
 B &= (p_1 - l_1)/(h_1 - l_1)
 \end{aligned}$$

This means that the reference normalized red imaging channel is mapped onto B , the green channel onto R and G is the average of the other two. Therefore, turquoise areas correspond to channel 2 of the input, orange areas correspond to channel 1 and white areas indicate that both channels are active.

Appendix E

OF estimation

This appendix is a reworked excerpt from [F5] and describes the variational OF model as well as RAFT OF estimation used throughout this thesis.

OF computation by a variational model. Our Lagrangian motion magnification approach uses the rather accurate variational detection of the OF between image frames in a video. This is indeed highly nontrivial, since facial micromotions are in general highly local both in space and time and are therefore often imperceptible to a human observer. OF models for videos are usually based on a gray-value constancy assumption, meaning that in a discrete-time model each pixel keeps its value when moving over time, i.e.,

$$f(\mathbf{x} + \mathbf{w}(\mathbf{x}, t), t + 1) = f(\mathbf{x}, t)$$

where $\mathbf{x} := (x, y)^\top$ and $\mathbf{w}(\mathbf{x}, t) := (u(\mathbf{x}, t), v(\mathbf{x}, t))^\top$. A first-order Taylor expansion of the left-hand side around $(\mathbf{x}, t)$ yields

$$f(\mathbf{x} + \mathbf{w}(\mathbf{x}, t), t + 1) \approx f(\mathbf{x}, t) + \nabla f(\mathbf{x}, t)^\top \mathbf{w}(\mathbf{x}, t) + \frac{\partial f}{\partial t}(\mathbf{x}, t),$$

and hence the linearized OF constraint

$$\frac{\partial f}{\partial t}(\mathbf{x}, t) + \frac{\partial f}{\partial x}(\mathbf{x}, t) u(\mathbf{x}, t) + \frac{\partial f}{\partial y}(\mathbf{x}, t) v(\mathbf{x}, t) \approx 0. \quad (\text{E.1})$$

Clearly, digital video sequences are discrete both in space and time. Given

$$f_{i,j,t}, \quad (i,j) \in \Omega_h := \{1, \dots, m\} \times \{1, \dots, n\} \subset \mathbb{Z}^2, \quad t = 1, \dots, T$$

for a discrete rectangular pixel grid Ω_h , we are interested in the OF fields $\mathbf{w}_{i,j,t}$ between consecutive video frames $t = 1, \dots, T - 1$. For readability, we define $f_t := (f_{i,j,t})_{(i,j) \in \Omega_h}$ and $\mathbf{w}_t := (\mathbf{w}_{i,j,t})_{(i,j) \in \Omega_h}$.

Then for a given $t = 1, \dots, T - 1$ the derivatives in (E.1) are replaced by finite differences and the approximation (E.1) is used to form the data term $D(f_t, f_{t+1}, \mathbf{w}_t)$ within a variational model

$$\arg \min_{\mathbf{w}_t} \{D(f_t, f_{t+1}, \mathbf{w}_t) + \lambda R(\mathbf{w}_t)\}, \quad \lambda > 0.$$

Here $R(\mathbf{w}_t)$ is a regularizing term or prior of the OF field which is necessary to make the underdetermined problem well-posed.

Different data and regularization terms were used in the literature. Starting with the Horn-Schunck model [14], there now exist many sophisticated variational OF models for different purposes. For an overview, we refer to [195].

OF detection by RAFT. While recently introduced deep-learning based OF techniques can outperform variational OF methods in terms of accuracy on many modern benchmarks, their performance depends not only on architectural innovations but also on the quality of the training data and training strategies. In Chapter 2 and Chapter 3, we rely on the RAFT approach from [87] which has demonstrated state-of-the-art performance. Training and inference of the original RAFT is briefly described in the next remark.

Remark 2 (Original RAFT). Consider two consecutive image frames $f_1, f_2 \in \mathbb{R}^{m \times n}$ (corresponding to $f(\cdot, t)$ and $f(\cdot, t+1)$). First, these are encoded into feature maps at one eighth of their resolution, resulting in feature maps $\tilde{f}_1, \tilde{f}_2 \in \mathbb{R}^{C \times \tilde{m} \times \tilde{n}}$, where $\tilde{m} = m/8$ and $\tilde{n} = n/8$ and C is the number of feature channels. Then RAFT is a deep-learning based model which computes an update $\Delta \tilde{\mathbf{w}}$ for an initial OF field $\tilde{\mathbf{w}}$ based on the features $\tilde{f}_1, \tilde{f}_2$:

$$\begin{aligned} \Phi: \mathbb{R}^{C \times \tilde{m} \times \tilde{n}} \times \mathbb{R}^{C \times \tilde{m} \times \tilde{n}} \times \mathbb{R}^{2 \times \tilde{m} \times \tilde{n}} &\rightarrow \mathbb{R}^{2 \times \tilde{m} \times \tilde{n}}, \\ (\tilde{f}_1, \tilde{f}_2, \tilde{\mathbf{w}}) &\mapsto \Delta \tilde{\mathbf{w}}. \end{aligned}$$

For an initial flow equal to zero, the method realizes the following iteration for $r = 0, \dots, R-1$:

$$\begin{aligned} \tilde{\mathbf{w}}_0 &:= \mathbf{0}, \\ \Delta \tilde{\mathbf{w}}_r &:= \Phi(\tilde{f}_1, \tilde{f}_2, \tilde{\mathbf{w}}_r), \\ \tilde{\mathbf{w}}_{r+1} &:= \tilde{\mathbf{w}}_r + \Delta \tilde{\mathbf{w}}_r. \end{aligned}$$

For inference, the authors use the output flow field after a fixed number R of such iterations. For the r -th iterate, learned upsampling is applied to bring the computed flow back to the original image resolution, i.e., $\tilde{\mathbf{w}}_r \mapsto \mathbf{w}_r$. Then, for training this model, the authors use the loss function

$$\mathcal{E} = \sum_{r=1}^R \alpha^{R-r} \|\mathbf{w}^{(gt)} - \mathbf{w}_r\|_1,$$

where $\mathbf{w}^{(gt)}$ denotes the ground truth and $\alpha = 0.8$ is chosen.

Appendix F

Variational OF optimization

A minimizer of an energy functional solves the respective Euler-Lagrange equations. In variational optical flow, brightness or gradient based constancy assumptions can be linearized and written in motion tensor notation with a motion tensor J . Compare Bruhn [168] for motion tensors that describe different constancy assumptions as well as a compact notation of the Euler-Lagrange equations and discretization scheme. Our implementation can easily be extended with other constancies by implementing the respective motion tensor.

Zimmer et al. scale alpha with $\alpha_i = \alpha \cdot \eta^{-i}$ to counter the effects of the data term normalization [89]. We use a slightly slower rate and scale alpha on each level of the solver scale space such that at a level i we use $\alpha_i = \alpha \cdot \eta^{-i/2}$. In practice, this made the result more robust under lower choices of α while enabling the compensation of high-frequency jitter.

Algorithm 1 Multiscale OF solver

Compute the maximum possible warping depth into max_level_x and max_level_y
for $\max(\text{max_level_x}, \text{max_level_y}) \geq i \geq \text{min level}$ **do**
 level_width = $\eta^{\min(i, \text{max_level_x})} \cdot \text{width}$
 level_height = $\eta^{\min(i, \text{max_level_y})} \cdot \text{height}$
 if $i = \max(\text{max_level_x}, \text{max_level_y})$ **then**
 Initialize u and v with supplied initialization or zeros
 else
 Resize previous u and v to $(\text{level_width}, \text{level_height})^\top$
 end if
 Compensate f with rescaled u and v
 Compute the motion tensor J
 Compute the flow increments $(du, dv)^\top$ with smoothness weight $\alpha \cdot \eta^{-i/2}$
 Median filter du and dv with a 5×5 kernel
 $(u, v)^\top \leftarrow (u, v)^\top + (du, dv)^\top$
end for
 Resize $(u, v)^\top$ to the input size

Appendix G

Additional DTSC results

This appendix collects supplementary proof-of-concept analyses from our multimodal recordings at a SARS-CoV-2 *drive-through screening center* (DTSC) as described in [F1]. The analyses document modality-specific data quality and feasibility for remote extraction of physiological signals under DTSC constraints.

Figure G.1 summarizes symptom self-reports from the standardized yes/no questionnaire. Since participants were admitted to the DTSC according to the applicable RKI recommendations at the time which included symptoms and/or contact history, the cohort contains many flu-like or common-cold symptoms. Figure G.2 provides an overview of the modality-wise evaluations and summarizes the results for each of these modalities, in particular how the contactless physiological parameters compare to gender and age specific norm values and groups them by the presence/absence of symptoms.

Remote plethysmography. We applied the method of Wang et al. [36] to extract *remote photoplethysmography* (rPPG) signals from skin-segmented super-pixels of the *high-frame-rate* (HFR) and 4K recordings. Peak detection and summary statistics were computed using custom scripts. The analysis indicates that mean heart rate decreases with age (Figure G.2 (A)). Female participants show slightly higher mean heart rate and lower heart-rate variability measures than male participants, consistent with contact-based sensor studies on large cohorts [196, 197].

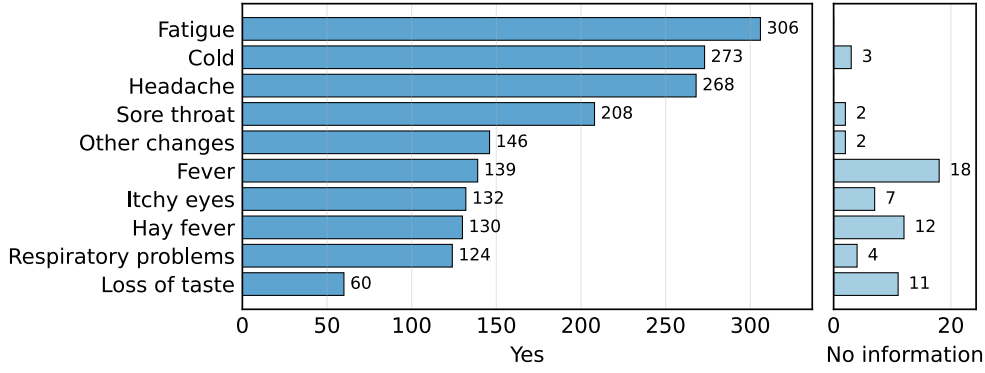


Figure G.1: Counts of symptoms self-reported during the yes/no section of the DTSC questionnaire. Since participants were admitted based on the applicable RKI recommendations which included symptoms and/or contact history, most participants reported flu-like or common-cold symptoms. The most common were fatigue (306), cold (273), headache (268), and sore throat (208). Results are reported in [F1].

Remote eye analysis. We precomputed facial landmarks for the HFR and 4K recordings using dlib [58]. We analyzed blink rate as a marker for drowsiness [198] and sclera redness as an indicator for conjunctivitis [199] (Figure G.2, B). We computed the *eye-aspect ratio* (EyeAR) [200] from the landmarks and detected blinks via peak counting. In our cohort, blink rate differed between symptom groups (e.g. itchy eyes vs. fever). For sclera redness, we extracted an eye region of interest, applied grayscale-based segmentation, and computed the redness index (RI) [199]. There was no significant RI difference found between the compared symptom groups.

Computer audition. Audio recordings obtained under real-world conditions can carry diagnostic value despite signal degradations [24]. With our setup, we obtained reproducible audio quality from uniform hardware (Figure G.2 (D)). Using established speech feature extraction schemes [201], we observed that the voice quality was sufficient for a gender classification sanity check, reaching above 90% cross-validated accuracy with a support vector machine [F1].

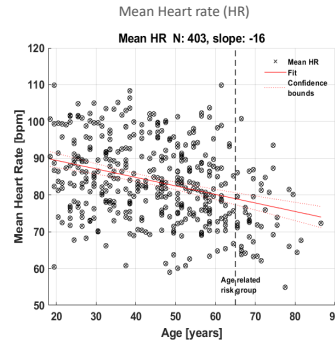
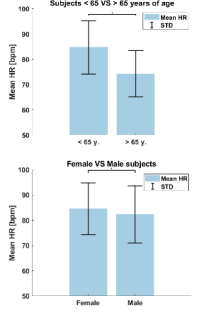
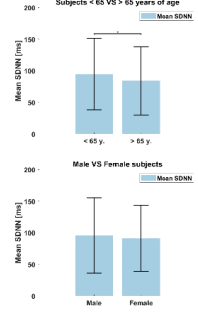
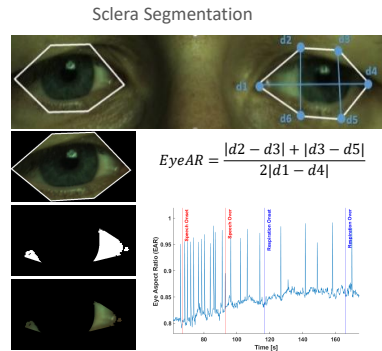
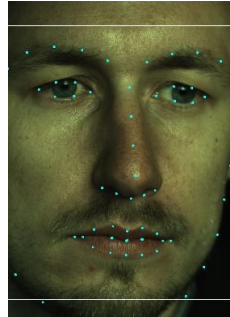
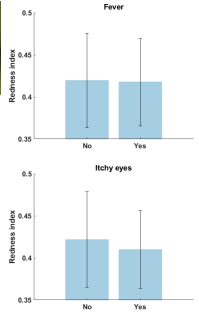
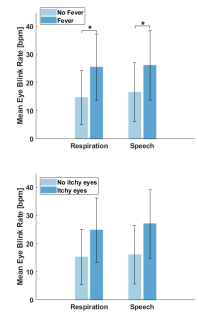
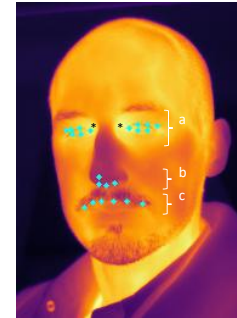
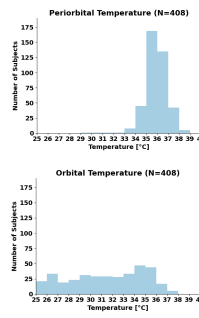
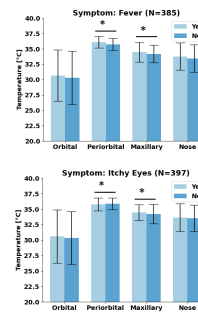
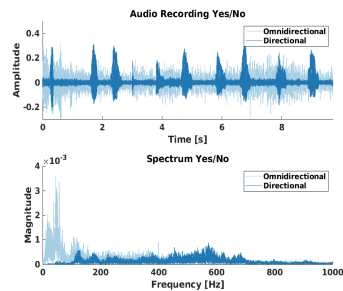
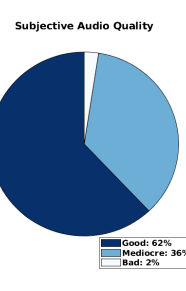
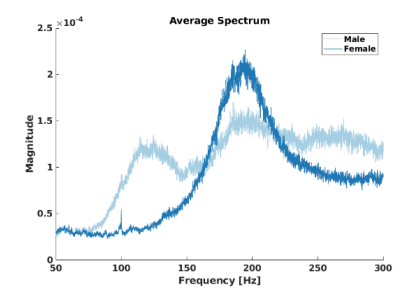
A Remote Plethysmography**Sex and age differences in HR****Standard Deviation of NN Intervals (SDNN)****B Remote Eye Analysis****Redness-Index****Blinking-Rate****C Functional Thermography****Temperature Extraction****Temp. Distribution****Mean Temperatures****D Computer Audition****Audio Quality****Average Spectrum**

Figure G.2: Results obtained in the reported evaluation period from the different modalities [F1]. rPPG (A) was used for the extraction of heart rate (HR) and heart-rate variability (HRV). For remote optometry (B) we analyzed EyeAR and sclera redness. For static temperatures from the face (C), we looked at four points of interest. Manual assessment of our audio recordings (D) shows that 97.6% of our recordings had very good to acceptable sound quality, while for the rest the voice was rated unintelligible. Individual panels as originally published in [F1].

Personal Contributions

- [F1] P. Flotho, M. J. Bhamborae, T. Grün, C. Trenado, D. Thinnies, D. Limbach, and D. J. Strauss. Multimodal data acquisition at SARS-CoV-2 drive through screening centers: Setup description and experiences in Saarland, Germany. J Biophotonics, 14(8):e202000512, 2021.
- [F2] E. N. Schneider, B. Buchheit, P. Flotho, M. J. Bhamborae, F. I. Corona-Strauss, F. Dauth, M. Alayan, and D. J. Strauss. Electrodermal responses to driving maneuvers in a motion sickness inducing real-world driving scenario. IEEE Trans Hum Mach Syst, 52(5):994–1003, 2022.
- [F3] P. Flotho, S. Nomura, B. Kuhn, and D. J. Strauss. Software for non-parametric image registration of 2-photon imaging data. J Biophotonics, page e202100330, 2022. e202100330 jbio.202100330.R2.
- [F4] P. Flotho, D. Thinnies, B. Kuhn, C. J. Roome, J. F. Vibell, and D. J. Strauss. Fast variational alignment of non-flat 1D displacements for applications in neuroimaging. J Neurosci Methods, 353:109076, 2021.
- [F5] P. Flotho, C. Heiss, G. Steidl, and D. J. Strauss. Lagrangian motion magnification with double sparse optical flow decomposition. Front Appl Math Stat, 9:1164491, 2023.
- [F6] P. Flotho, M. Piening, A. Kukleva, and G. Steidl. T-FAKE: Synthesizing thermal images for facial landmarking. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 26356–26366, 2025.
- [F7] D. J. Strauss, F. I. Corona-Strauss, A. Schroeer, P. Flotho, R. Hannemann, and S. A. Hackley. Vestigial auriculomotor activity indicates the direction of auditory attention in humans. Elife, 9:e54536, 2020.

- [F8] M. Bhamborae, E. N. Schneider, P. Flotho, A. L. Francis, and D. J. Strauss. LumEDA: image luminance based contactless correlates of electrodermal responses. Physiol Meas, 13(2):025010, 2025.
- [F9] P. Flotho, M. J. Bhamborae, L. Haab, and D. J. Strauss. Lagrangian motion magnification revisited: Continuous, magnitude driven motion scaling for psychophysiological experiments. In Proc Annu Int Conf IEEE Eng Med Biol Soc, pages 3586–3589, 2018.
- [F10] M. J. Bhamborae, P. Flotho, A. Mai, E. N. Schneider, A. L. Francis, and D. J. Strauss. Towards contactless estimation of electrodermal activity correlates. In Proc Annu Int Conf IEEE Eng Med Biol Soc, pages 1799–1802, 2020.
- [F11] L. Haab, P. Flotho, D. Eckert, K. Schwerdtfeger, M. Hülser, D.J. Strauss, and M. Möller. Time-multiplexed illumination for simultaneous structural and functional voltage sensitive dye recordings with a single photo sensor. In Proc Annu Int Conf IEEE Eng Med Biol Soc, pages 1899–1902, 2020.
- [F12] P. Flotho, C. Heiss, G. Steidl, and D.J. Strauss. Lagrangian motion magnification with landmark-prior and sparse PCA for facial microexpressions and micromovements. In Proc Annu Int Conf IEEE Eng Med Biol Soc, pages 2215–2218, 2022.
- [F13] P. Flotho, M. El-Tabbal, T. E. Lever, and K. Takahashi. Rodent tongue kinematics with markerless XROMM using DeepLabCut and marker masking. In Proc Annu Int Conf IEEE Eng Med Biol Soc, pages 1–4, 2024.
- [F14] P. Flotho, L. Haab, D. Eckert, K. Takahashi, K. Schwerdtfeger, and D. J. Strauss. Semi-synthetic dataset for the evaluation of motion compensation approaches for voltage sensitive dye imaging. In Proc Int IEEE EMBS Conf Neural Eng, pages 1134–1137, 2019.
- [F15] S. K. Pandey, D. J. Strauss, and P. Flotho. Impact of medical face masks on facial alignment performance. Annu Int Conf IEEE Eng Med Biol Soc, 2022.
- [F16] P. Flotho, M. J. Bhamborae, C. Karls, M. Lehser, and D. J. Strauss. Comparing video compression for archiving camera-based physiological measurements for rPPG and motion analysis. Annu Int Conf IEEE Eng Med Biol Soc, 2022.

- [F17] P. Flotho, D. Limbach, A. Sternjakob, R. Bals, R. Beinert, M. Hannig, J. Nguyen, S. Smola, G. Steidl, M. Zemlin, and D. J. Strauss. Towards RGB-thermal fusion for improved core temperature estimation in remote respiratory disease screening (VI-Screen). *Annu Int Conf IEEE Eng Med Biol Soc*, 2023.
- [F18] P. Flotho, C. Haxton, T. E. Lever, D. A. Reed, B. Sagl, L. Avivi-Arber, D. J. Strauss, and K. Takahashi. XROMM 3D pose estimation with DeepLabCut. *Annu Int Conf IEEE Eng Med Biol Soc*, 2023.
- [F19] M.J. Bhamborae, P. Flotho, A. Mai, E.N. Schneider, A.L. Francis, and D.J. Strauss. Functional skin imaging: Towards contactless estimation of electrodermal activity correlates. In *Psychophysiology*, volume 57, pages S35–S35. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2020.

Bibliography

- [1] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci, 21(9):1281–1289, 2018.
- [2] T. Nath, A. Mathis, A. C. Chen, A. Patel, M. Bethge, and M. W. Mathis. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat Protoc, 14(7):2152–2176, 2019.
- [3] J. Lauer, M. Zhou, S. Ye, W. Menegas, S. Schneider, T. Nath, M. M. Rahman, V. Di Santo, D. Soberanes, G. Feng, V. N. Murthy, G. Lauder, C. Dulac, M. W. Mathis, and A. Mathis. Multi-animal pose estimation, identification and tracking with DeepLabCut. Nat Methods, 19(4):496–504, 2022.
- [4] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In Proc Comput Vis ECCV, pages 34–50. Springer, 2016.
- [5] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In Med Image Comput Comput Assist Interv, pages 234–241. Springer, 2015.
- [6] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst, 33:6840–6851, 2020.
- [7] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge University Press, 2003.
- [8] Reinhard Klette. Concise computer vision, volume 233. Springer, 2014.

- [9] L. Valgaerts, A. Bruhn, H. Zimmer, J. Weickert, C. Stoll, and C. Theobalt. Joint estimation of motion, structure and geometry from stereo sequences. In Comput Vis ECCV, pages 568–581, 2010.
- [10] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks). In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 1021–1030, 2017.
- [11] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman. Eulerian video magnification for revealing subtle changes in the world. ACM Trans Graph, 31(4):1–8, 2012.
- [12] C. Liu, A. Torralba, W. T. Freeman, F. Durand, and E. H. Adelson. Motion magnification. ACM Trans Graph, 24(3):519–526, 2005.
- [13] M. Elgharib, M. Hefeeda, F. Durand, and W. T. Freeman. Video magnification in presence of large motions. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 4119–4127, 2015.
- [14] B. K. P. Horn and B. G. Schunk. Determining optical flow. Artif Intell, 17(1-3):185–203, 1981.
- [15] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman. Phase-based video motion processing. ACM Trans Graph, 32(4), 2013.
- [16] T. Ebrahimi and C. Horne. MPEG-4 natural video coding—an overview. Signal Process Image Commun, 15(4-5):365–385, 2000.
- [17] E. A. Pnevmatikakis and A. Giovannucci. NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. J Neurosci Methods, 291:83–94, 2017.
- [18] J. Modersitzki. Numerical methods for image registration. Oxford University Press, 2003.
- [19] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In Comput Vis ECCV, pages 25–36. Springer, 2004.

- [20] F. Fiore, K. Alhalaseh, R. R. Dereddi, F. Bodaleo Torres, I. Çoban, A. Harb, and A. Agarwal. Norepinephrine regulates calcium signals and fate of oligodendrocyte precursor cells in the mouse cerebral cortex. Nat Commun, 14(1):8122, 2023.
- [21] M. Li, M. Eltabbal, H.-D. Tran, and B. Kuhn. Scn2a insufficiency alters spontaneous neuronal Ca²⁺ activity in somatosensory cortex during wakefulness. iScience, 26(11), 2023.
- [22] C. Cecchetto, S. Vassanelli, and B. Kuhn. Simultaneous two-photon voltage or calcium imaging and multi-channel local field potential recordings in barrel cortex of awake and anesthetized mice. Front Neurosci, 15:741279, 2021.
- [23] A. S. Manolis, A. A. Manolis, T. A. Manolis, E. J. Apostolopoulos, D. Papatheou, and H. Melita. COVID-19 infection and cardiac arrhythmias. Trends Cardiovasc Med, 30(8):451–460, 2020.
- [24] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li. COVID-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 corona crisis. Front Digit Health, 3:564906, 2021.
- [25] B. J. Quilty, S. Clifford, CMMID nCoV working group, S. Flasche, and R. M. Eggo. Effectiveness of airport screening at detecting travellers infected with novel coronavirus (2019-nCoV). Euro Surveill, 25(5):2000080, 2020.
- [26] Y.-F. Tu, C.-S. Chien, A. A. Yarmishyn, Y.-Y. Lin, Y.-H. Luo, Y.-T. Lin, W.-Y. Lai, D.-M. Yang, S.-J. Chou, Y.-P. Yang, M.-L. Wang, and S.-H. Chiou. A review of SARS-CoV-2 and the ongoing clinical trials. Int J Mol Sci, 21(7):2657, 2020.
- [27] R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirgmaier, C. Drosten, and C. Wendtner. Virological assessment of hospitalized patients with COVID-2019. Nature, 581(7809):465–469, 2020.
- [28] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. C. Hui, B. Du, L.-J. Li, G. Zeng, K.-Y. Yuen, R.-C. Chen, C.-L. Tang, T. Wang, P.-Y. Chen, J. Xiang, S.-Y. Li, J.-L. Wang, Z.-J. Liang, Y.-X. Peng, L. Wei, Y. Liu, Y.-H. Hu, P. Peng, J.-M. Wang, J.-Y. Liu, Z. Chen, G. Li,

- Z.-J. Zheng, S.-Q. Qiu, J. Luo, C.-J. Ye, S.-Y. Zhu, and N.-S. Zhong. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med, 382(18):1708–1720, 2020.
- [29] J.-P. O. Li, D. S. C. Lam, Y. Chen, and D. S. W. Ting. Novel coronavirus disease 2019 (COVID-19): The importance of recognising possible early ocular manifestation and using protective eyewear. Br J Ophthalmol, 104(3):297–298, 2020.
- [30] K. T. Kwon, J.-H. Ko, H. Shin, M. Sung, and J. Y. Kim. Drive-through screening center for COVID-19: a safe and efficient screening system against massive community outbreak. J Korean Med Sci, 35(11), 2020.
- [31] Robert Koch Institute. COVID-19-Verdacht: Maßnahmen und Testkriterien - Orientierungshilfe für Ärzte. Accessed in 2020 at www.rki.de.
- [32] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In IEEE Int Conf Autom Face Gesture Recognit Workshops, pages 1–6, 2013.
- [33] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap. SAMM: A spontaneous micro-facial movement dataset. IEEE Trans Affect Comput, 9(1):116–129, 2018.
- [34] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof. A thermal infrared face database with facial landmarks and emotion labels. IEEE Trans Instrum Meas, 68(5):1389–1401, 2019.
- [35] C. Palmero, A. Clapés, C. Bahnsen, A. Møgelmoose, T. B. Moeslund, and S. Escalera. Multi-modal RGB–depth–thermal human body segmentation. Int J Comput Vis, 118(2):217–239, 2016.
- [36] W. Wang, A. C. Den Brinker, S. Stuijk, and G. De Haan. Algorithmic principles of remote PPG. IEEE Trans Biomed Eng, 64(7):1479–1491, 2017.
- [37] R. Ashrafi, M. Azarbayjani, and H. Tabkhi. Charlotte-ThermalFace: A fully annotated thermal infrared face dataset with various environmental conditions and distances. Infrared Phys Technol, 124:104209, 2022.
- [38] E. Y.-K. Ng. Is thermal scanner losing its bite in mass screening of fever due to SARS? Med Phys, 32(1):93–97, 2005.

- [39] D. Bitar, A. Goubar, and J.-C. Desenclos. International travels and fever screening during epidemics: A literature review on the effectiveness and potential use of non-contact infrared thermometers. Euro Surveill, 14(6):19115, 2009.
- [40] S. Ioannou, V. Gallese, and A. Merla. Thermal infrared imaging in psychophysiology: Potentialities and limits. Psychophysiology, 51(10):951–963, 2014.
- [41] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman, X. Di, C. N. Fondje, V. M. Patel, N. J. Short, B. S. Riggan, N. M. Nasrabadi, and S. Hu. A large-scale, time-synchronized visible and thermal face dataset. In Proc IEEE Winter Conf Appl Comput Vis, pages 1559–1568, 2021.
- [42] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 4873–4882, 2016.
- [43] E. Wood, T. Baltrusaitis, C. Hewitt, M. Johnson, J. Shen, N. Milosavljevic, D. Wilde, S. Garbin, T. Sharp, I. Stojiljkovic, T. Cashman, and J. Valentin. Fake it till you make it: face analysis in the wild using synthetic data alone. In Proc IEEE Int Conf Comput Vis, pages 3681–3691, 2021.
- [44] E. Wood, T. Baltrusaitis, C. Hewitt, M. Johnson, J. Shen, N. Milosavljevic, D. Wilde, S. Garbin, T. Sharp, I. Stojiljkovic, T. Cashman, and J. Valentin. 3D face reconstruction with dense landmarks. In Comput Vis ECCV, 2022.
- [45] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172, 2019.
- [46] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung. Deepfake detection: A systematic literature review. IEEE Access, 10:25494–25513, 2022.
- [47] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. Image Vis Comput, 47:3–18, 2016.
- [48] D. Poster, S. Hu, N. Nasrabadi, and B. Riggan. An examination of deep-learning based landmark detection methods on thermal face imagery. In IEEE Comput Soc Conf Comput Vis Pattern Recognit Workshops, pages 980–987, 2019.

- [49] K. Mallat and J.-L. Dugelay. Facial landmark detection on thermal data via fully annotated visible-to-thermal data synthesis. In Proc IEEE Int Joint Conf Biometrics, pages 1–10, 2020.
- [50] T.-H. Ho, C.-Y. Yu, T.-Y. Ko, and W.-T. Chu. The NCKU-VTF dataset and a multi-scale thermal-to-visible face synthesis system. In Multimed Model, pages 463–475. Springer, 2023.
- [51] A. Kuzdeuov, D. Koishigarina, D. Aubakirova, S. Abushakimova, and H. A. Varol. SF-TL54: A thermal facial landmark dataset with visual pairs. In Proc IEEE/SICE Int Symp Syst Integr, pages 748–753, 2022.
- [52] A. Kuzdeuov, D. Aubakirova, D. Koishigarina, and H. A. Varol. TFW: Annotated thermal faces in the wild dataset. IEEE Trans Inf Forensics Secur, 2022.
- [53] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The Menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking. Int J Comput Vis, 127:599–624, 2019.
- [54] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. Image Vis Comput, 28(5):807–813, 2010.
- [55] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In IEEE Int Conf Comput Vis Workshops, pages 2144–2151, 2011.
- [56] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. Face alignment in full pose range: A 3D total solution. IEEE Trans Pattern Anal Mach Intell, 41(1):78–92, 2019.
- [57] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 1867–1874, 2014.
- [58] D. E. King. Dlib-ml: A machine learning toolkit. J Mach Learn Res, 10:1755–1758, 2009.
- [59] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In Conf Comput Vis Pattern Recognit Workshops, pages 88–97, 2017.

- [60] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng. LUVLi face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 8236–8246, 2020.
- [61] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, W. Zhao, X. Zhu, Z. Yan, X.-Y. Zhang, J. Wu, Z. Lei, S. Tripathi, M. Kothari, M. H. Zama, D. Deb, B. Biesseck, P. Vidal, R. Granada, G. Fickel, G. Führ, D. Menotti, A. Unnervik, A. George, C. Ecabert, H. Otroshi Shahreza, P. Rahimi, S. Marcel, I. Sarridis, C. Koutlis, G. Baltso, S. Papadopoulos, C. Diou, N. Di Domenico, G. Borghi, L. Pellegrini, E. Mas-Candela, Á. Sánchez-Pérez, A. Atzori, F. Boutros, N. Damer, G. Fenu, and M. Marras. FRCSyn-onGoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems. Inf Fusion, 107:102322, 2024.
- [62] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. Int J Comput Vis, 127(6):845–862, 2019.
- [63] D. D. Poster, S. Hu, N. J. Short, B. S. Riggan, and N. M. Nasrabadi. Visible-to-thermal transfer learning for facial landmark detection. IEEE Access, 9:52759–52772, 2021.
- [64] B. S. Riggan, N. J. Short, and S. Hu. Thermal to visible synthesis of face images using multiple regions. In Proc IEEE Winter Conf Appl Comput Vis, pages 30–38, 2018.
- [65] T. Zhang, A. Wiliem, S. Yang, and B. Lovell. TV-GAN: Generative adversarial network based thermal to visible face recognition. In Proc IEEE Int Conf Biometrics, pages 174–181, 2018.
- [66] N. G. Nair and V. M. Patel. T2V-DDPM: Thermal to visible face translation using denoising diffusion probabilistic models. In Proc IEEE Int Conf Autom Face Gesture Recognit, pages 1–7, 2023.

- [67] K. Mallat and J.-L. Dugelay. A benchmark database of visible and thermal paired face images across multiple variations. In Int Conf Biometrics Spec Interest Group, pages 1–5. IEEE, 2018.
- [68] U. Cheema and S. Moon. Sejong face database: A multi-modal disguise face database. Comput Vis Image Underst, 208:103218, 2021.
- [69] V. V. Kniaz, V. A. Knyaz, J. Hladuvka, W. G. Kropatsch, and V. Mizginov. ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In Comput Vis ECCV Workshops, pages 606–624. Springer International Publishing, 2018.
- [70] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 1125–1134, 2017.
- [71] T. Kroeger, R. Timofte, D. Dai, and L. Van Gool. Fast optical flow using dense inverse search. In Comput Vis ECCV, pages 471–488. Springer, 2016.
- [72] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Adv Neural Inf Process Syst, 26, 2013.
- [73] J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trounev, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In Proc Int Conf Artif Intell Stat, pages 2681–2690. PMLR, 2019.
- [74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In Proc IEEE Conf Comput Vis Pattern Recognit, pages 248–255, 2009.
- [75] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 4510–4520, 2018.
- [76] M. Godi, C. Joppi, A. Giachetti, F. Pellacini, and M. Cristani. Texel-Att: Representing and classifying element-based textures by attributes. arXiv preprint arXiv:1908.11127, 2019.

- [77] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 2879–2886, 2012.
- [78] M. Kopaczka, L. Breuer, J. Schock, and D. Merhof. A modular system for detection, tracking and analysis of human faces in thermal infrared recordings. Sensors, 19(19):4135, 2019.
- [79] B. Browatzki and C. Wallraven. 3FabRec: Fast few-shot face alignment by reconstruction. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, 2020.
- [80] I. Grishchenko, A. Ablavatski, Y. Karynnik, K. Raveendran, and M. Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time. Conf Comput Vis Pattern Recognit Workshops, 2020.
- [81] Z. Zhou, H. Li, H. Liu, N. Wang, G. Yu, and R. Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 15475–15484, 2023.
- [82] Y. Huang, H. Yang, C. Li, J. Kim, and F. Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In Proc IEEE Int Conf Comput Vis, pages 3080–3090, 2021.
- [83] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman. Riesz pyramids for fast phase-based video magnification. In Proc IEEE Int Conf Comput Photography, pages 1–10, 2014.
- [84] Y. Zhang, S. L. Pinteá, and J. C. van Gemert. Video acceleration magnification. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 529–537, 2017.
- [85] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 2462–2470, 2017.
- [86] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 8934–8943, 2018.
- [87] Z. Teed and J. Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In Comput Vis ECCV, pages 402–419. Springer, 2020.

- [88] A. Bruhn and J. Weickert. Towards ultimate motion estimation: combining highest accuracy with real-time performance. In IEEE Int Conf Comput Vis, volume 1, pages 749–755 Vol. 1, 2005.
- [89] H. Zimmer, A. Bruhn, J. Weickert, L. Valgaerts, A. Salgado, B. Rosenhahn, and H.-P. Seidel. Complementary optic flow. In Energy Minimization Methods Comput Vis Pattern Recognit, pages 207–220. Springer, 2009.
- [90] R. Fattal, D. Lischinski, and M. Werman. Gradient domain high dynamic range compression. In ACM Trans Graph, volume 21, pages 249–256. ACM, 2002.
- [91] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu. How fast are the leaked facial expressions: The duration of micro-expressions. J Nonverbal Behav, 37(4):217–230, 2013.
- [92] A. Sarrafi, Z. Mao, C. Niezrecki, and P. Poozesh. Vibration-based damage detection in wind turbine blades using phase-based motion estimation and motion magnification. J Sound Vib, 421:300–318, 2018.
- [93] M. Eitner, M. Musta, L. Vanstone, J. Sirohi, and N. Clemens. Modal parameter estimation of a compliant panel using phase-based motion magnification and stereoscopic digital image correlation. Exp Tech, 45(3):287–296, 2021.
- [94] J. Fei, Z. Xia, P. Yu, and F. Xiao. Exposing AI-generated videos with motion magnification. Multimed Tools Appl, 80(20):30789–30802, 2021.
- [95] D. Alinovi, G. Ferrari, F. Pisani, and R. Raheli. Respiratory rate monitoring by video processing using local motion magnification. In Proc Eur Signal Process Conf EUSIPCO, pages 1780–1784. IEEE, 2018.
- [96] J. Bai, A. Agarwala, M. Agrawala, and R. Ramamoorthi. Selectively de-animating video. ACM Trans Graph, 31(4), 2012.
- [97] C. Hartmann, H. A. Weiss, P. Lechner, W. Volk, S. Neumayer, J. H. Fitschen, and G. Steidl. Measurement of strain, strain rate and crack evolution in shear cutting. J Mater Process Technol, 288:116872, 2021.
- [98] Z. Chen, H. Jin, Z. Lin, S. Cohen, and Y. Wu. Large displacement optical flow from nearest neighbor fields. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 2443–2450, 2013.

- [99] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. Int J Comput Vis, 92(1):1–31, 2011.
- [100] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 2432–2439, 2010.
- [101] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 4040–4048, 2016.
- [102] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, Comput Vis ECCV, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.
- [103] D. Sun, C. Herrmann, F. Reda, M. Rubinstein, D. J. Fleet, and W. T. Freeman. Disentangling architecture and training for optical flow. In Comput Vis ECCV, pages 165–182. Springer, 2022.
- [104] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Transactions on Affective Computing, 10(1):32–47, 2019.
- [105] F. Xu, J. Zhang, and J. Z. Wang. Microexpression identification and categorization using a facial dynamics map. IEEE Trans Affect Comput, 8(2):254–267, 2017.
- [106] N. A. C. Le, A. Johnston, R. C.-W. Phan, and J. See. Micro-expression motion magnification: Global Lagrangian vs. local Eulerian approaches. In IEEE Int Conf Autom Face Gesture Recognit Workshops, pages 650–656, 2018.
- [107] L. Zhang and O. Arandjelović. Review of automatic microexpression recognition in the past decade. Mach Learn Knowl Extr, 3(2):414–434, 2021.
- [108] L. Shan and M. Yu. Video-based heart rate measurement using head motion tracking and ICA. In Proc Int Congr Image Signal Process, volume 1, pages 160–164. IEEE, 2013.

- [109] Y. Mroueh, E. Marcheret, and V. Goel. Deep multimodal learning for audio-visual speech recognition. In Proc IEEE Int Conf Acoust Speech Signal Process, pages 2130–2134, 2015.
- [110] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In Annu Int Conf on Mach Learn, ICML '09, page 689–696, New York, NY, USA, 2009. Association for Computing Machinery.
- [111] R. Beinert and G. Steidl. Robust PCA via regularized reaper with a matrix-free proximal algorithm. J Math Imaging Vis, pages 626–649, 2021.
- [112] R. Cabral, F. De la Torre, J. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In Proc IEEE Int Conf Comput Vis, 2013.
- [113] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Journal of the ACM, 58(3):1–37, 2011.
- [114] M. Holler and K. Kunisch. On infimal convolution of TV-type functionals and applications to video and image reconstruction. SIAM J. Imag. Sci., 7(4):2258–2300, 2014.
- [115] G. Wolberg. Digital Image Warping. IEEE Computer Society Press, Los Alamitos, CA, 1990.
- [116] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PLoS One, 9(1):e86041, 2014.
- [117] G. Bradski. The OpenCV Library. Dr. Dobb's Journal of Software Tools, 2000.
- [118] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In Proc IEEE Int Conf Comput Vis, 2015.
- [119] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 3061–3070, 2015.
- [120] D. Kondermann, R. Nair, K. Honauer, K. Krispin, J. Andrulis, A. Brock, B. Gusefeld, M. Rahimimoghaddam, S. Hofmann, C. Brenner, and B. Jähne. The HCI

- benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In Conf Comput Vis Pattern Recognit Workshops, pages 19–28, 2016.
- [121] C. Ballester, M. Bertalmio, V. Caselles, and G. Sapiro. Image inpainting. In Proc ACM SIGGRAPH, pages 417–424, 2000.
- [122] S. Niklaus and F. Liu. Softmax splatting for video frame interpolation. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 5437–5446, 2020.
- [123] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li. FlowFormer: A transformer architecture for optical flow. In Proc Comput Vis ECCV, pages 668–685. Springer, 2022.
- [124] B. Allaert, I. R. Ward, I. M. Bilasco, C. Djeraba, and M. Bennamoun. A comparative study on optical flow for facial expression analysis. Neurocomputing, 500:434–448, 2022.
- [125] M. Alkaddour, U. Tariq, and A. Dhall. Self-supervised approach for facial movement based optical flow. IEEE Trans Affect Comput, 13(4):2071–2085, 2022.
- [126] B. J. Knörlein, D. B. Baier, S. M. Gatesy, J.D. Laurence-Chasen, and E. L. Brainerd. Validation of XMALab software for marker-based XROMM. J Exp Biol, 219(23):3701–3711, 2016.
- [127] F. Bunyak, N. Shiraishi, K. Palaniappan, T. E. Lever, L. Avivi-Arber, and K. Takahashi. Development of semi-automatic procedure for detection and tracking of fiducial markers for orofacial kinematics during natural feeding. In Proc Annu Int Conf IEEE Eng Med Biol Soc, pages 580–583, 2017.
- [128] J. D. Laurence-Chasen, A. R. Manafzadeh, N. G. Hatsopoulos, C. F. Ross, and F. I. Arce-McShane. Integrating XMALab and DeepLabCut for high-throughput XROMM. J Exp Biol, 223(17):jeb226720, 2020.
- [129] P. J. Schäfer, F. I. Corona-Strauss, R. Hannemann, S. A. Hillyard, and D. J. Strauss. Testing the limits of the stimulus reconstruction approach: auditory attention decoding in a four-speaker free field environment. Trends Hear, 22:2331216518816600, 2018.

- [130] S. A. Hackley. Evidence for a vestigial pinna-orienting system in humans. Psychophysiology, 52(10):1263–1270, 2015.
- [131] S. Hanfland and M. Paul. Video format dependency of ppgi signals. In Proc of the Int Conf on Ele Eng, volume 1, page 2, 2016.
- [132] D. Maurer, M. Stoll, and A. Bruhn. Order-adaptive and illumination-aware variational optical flow refinement. In BMVC, 2017.
- [133] D. Maurer and A. Bruhn. ProFlow: Learning to predict optical flow. BMVC, 2018.
- [134] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Physica D, 60(1-4):259–268, 1992.
- [135] P. Gwosdek, H. Zimmer, S. Grewenig, A. Bruhn, and J. Weickert. A highly efficient GPU implementation for variational optic flow based on the Euler-Lagrange framework. In Comput Vis ECCV Workshops, pages 372–383. Springer, 2010.
- [136] N. Brosch, S. Štolc, and D. Antensteiner. Warping-based motion artefact compensation for multi-line scan light field imaging. Electron Imaging, 2018(15):273–1, 2018.
- [137] C. J. Roome and B. Kuhn. Simultaneous dendritic voltage and calcium imaging and somatic recording from Purkinje neurons in awake mice. Nat Commun, 9(1):3388, 2018.
- [138] T. C. Handy. Event-related potentials: A methods handbook. MIT Press, 2005.
- [139] D. J. Strauss, T. Teuber, G. Steidl, and F. I. Corona-Strauss. Exploiting the self-similarity in ERP images by nonlocal means for single-trial denoising. IEEE Trans Neural Syst Rehabil Eng, 21:576–583, 2013.
- [140] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process, 26(1):43–49, 1978.
- [141] N.-P. Vest Nielsen, J. M. Carstensen, and J. Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J Chromatogr A, 805(1-2):17–35, 1998.

- [142] G. Tomasi, F. Van Den Berg, and C. Andersson. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. J Chemometr, 18(5):231–241, 2004.
- [143] T. Skov, F. van den Berg, G. Tomasi, and R. Bro. Automated alignment of chromatographic data. J Chemometr, 20(11-12):484–497, 2006.
- [144] D. Deriso and S. Boyd. A general optimization framework for dynamic time warping. Optim Eng, 24(2):1411–1432, 2023.
- [145] G. Hermosillo, C. Chef d’Hotel, and O. Faugeras. Variational methods for multi-modal image matching. Int J Comput Vis, 50(3):329–343, 2002.
- [146] D. Thinnies, P. Flotho, F. I. Corona-Strauss, D. J. Strauss, and J. F. Vibell. Compensation of ERP latency jitter using fast variational 1D displacement estimation. Authorea Preprints, 2025.
- [147] M. C. Kohl, E. Schebsdat, E. N. Schneider, and D. J. Strauss. Denoising of single-trial event-related potentials by shrinkage and phase regularization of analytic wavelet filterbank coefficients. In Proc Int IEEE EMBS Conf Neural Eng, 2019.
- [148] C. Bernarding, F. I. Corona-Strauss, M. Latzel, and D. J. Strauss. Auditory streaming and listening effort: An event-related potential study. In Conf Proc IEEE Eng Med Biol Soc, volume 2010:1, pages 6817–6820, 2010.
- [149] Z. Morteza pouraghdam, L. Haab, F. I. Corona-Strauss, G. Steidl, and D. J. Strauss. Assessment of long-term habituation correlates in event-related potentials using a von Mises model. IEEE Trans Neural Syst Rehabil Eng, 23:363–373, 2015.
- [150] Z. Morteza pouraghdam, R. C. Wilson, L. Schwabe, and D. J. Strauss. Bayesian modeling of the dynamics of phase modulations and their application to auditory evoked responses. Front Comput Neurosci, 10(2):1–10, 2016.
- [151] T. W. Picton. The P300 wave of the human event-related potential. J Clin Neurophysiol, 9(4):456–479, 1992.
- [152] F. I. Corona-Strauss, W. Delb, B. Schick, and D. J. Strauss. Phase stability analysis of chirp evoked auditory brainstem responses by Gabor frame operators. IEEE Trans Neural Syst Rehabil Eng, 17(6):530–536, 2009.

- [153] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Transactions on image processing, 16(8):2080–2095, 2007.
- [154] T. W. Picton. Human auditory evoked potentials. Plural Publishing, 2010.
- [155] M. Ihrke, H. Schrobsdorff, and J. M. Herrmann. Denoising and averaging techniques for electrophysiological data. In Coordinated activity in the brain, pages 165–189. Springer, 2009.
- [156] M. G. Woldorff. Distortion of ERP averages due to overlap from temporally adjacent ERPs: Analysis and correction. Psychophysiology, 30(1):98–119, 1993.
- [157] J. T. Valderrama, A. de la Torre, I. Alvarez, J. C. Segura, A. R. D. Thornton, M. Sainz, and J. L. Vargas. A study of adaptation mechanisms based on ABR recorded at high stimulation rate. Clin Neurophysiol, 125(4):805–813, 2014.
- [158] F. Helmchen and W. Denk. Deep tissue two-photon microscopy. Nat Methods, 2(12):932–940, 2005.
- [159] P. Theer, B. Kuhn, D. Keusters, and W. Denk. Two-photon microscopy and imaging. Encycl Mol Cell Biol Mol Med, 15, 2006.
- [160] D. S. Greenberg and J. N. D. Kerr. Automated correction of fast motion artifacts for two-photon imaging of awake animals. J Neurosci Methods, 176(1):1–15, 2009.
- [161] C. J. Roome and B. Kuhn. Dendritic coincidence detection in Purkinje neurons of awake mice. Elife, 9:e59619, 2020.
- [162] J. B. A. Maintz and M. A. Viergever. A survey of medical image registration. Med Image Anal, 2(1):1–36, 1998.
- [163] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. IEEE Trans Med Imaging, 32(7):1153–1190, 2013.
- [164] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P.W. Pluim. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging, 29(1):196–205, 2009.
- [165] B. B. Avants, N. Tustison, and G. Song. Advanced normalization tools (ANTs). Insight J, 2(365):1–35, 2009.

- [166] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal, 12(1):26–41, 2008.
- [167] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage, 54(3):2033–2044, 2011.
- [168] A. Bruhn. Variational optic flow computation: accurate modelling and efficient numerics. PhD thesis, Saarland University, 2006.
- [169] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert. Highly accurate optic flow computation with theoretically justified warping. Int J Comput Vis, 67(2):141–158, 2006.
- [170] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. Int J Comput Vis, 61(3):211–231, 2005.
- [171] D. M. Bjørnstad, K. S. Åbjørnsbråten, E. Hennestad, C. Cunen, G. H. Hermansen, L. Bojarskaite, K. H. Pettersen, K. Vervaeke, and R. Enger. Begonia—a two-photon imaging analysis pipeline for astrocytic Ca²⁺ signals. Front Cell Neurosci, 15:176, 2021.
- [172] A. Giovannucci, J. Friedrich, P. Gunn, J. Kalfon, B. L. Brown, Sue A. Koay, J. Taxidis, F. Najafi, J. L. Gauthier, P. Zhou, B. S. Khakh, D. W. Tank, D. B. Chklovskii, and E. A. Pnevmatikakis. CaImAn an open source tool for scalable calcium imaging data analysis. Elife, 8:e38173, 2019.
- [173] Y. Zhang, Y. Zhu, E. Nichols, Q. Wang, S. Zhang, C. Smith, and S. Howard. A Poisson-Gaussian denoising dataset with real fluorescence microscopy images. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 11710–11718, 2019.
- [174] D. A. Cantu, B. Wang, M. W. Gongwer, C. X. He, A. Goel, A. Suresh, N. Kourdougli, E. D. Arroyo, W. Zeiger, and C. Portera-Cailliau. EZcalcium: open-source toolbox for analysis of calcium imaging data. Front Neural Circuits, 14:25, 2020.

- [175] V. A. Griffiths, A. M. Valera, J. Y.N. Lau, H. Roš, T. J. Younts, B. Marin, C. Baragli, D. Coyle, G. J. Evans, G. Konstantinou, T. Koimtzis, K. M. N. S. Nadella, S. A. Punde, P. A. Kirkby, I. H. Bianco, and R. A. Silver. Real-time 3D movement correction for two-photon imaging in behaving animals. Nat Methods, 17(7):741–748, 2020.
- [176] P. Liu, M. Lyu, I. King, and J. Xu. SelFlow: Self-supervised learning of optical flow. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 4571–4580, 2019.
- [177] T. Pietzsch, S. Preibisch, P. Tomančák, and S. Saalfeld. ImgLib2 – generic image processing in Java. Bioinformatics, 28(22):3009–3011, 2012.
- [178] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 3354–3361, 2012.
- [179] M. Menze, C. Heipke, and A. Geiger. Joint 3D estimation of vehicles and scene flow. In ISPRS Ann Photogramm Remote Sens Spatial Inf Sci, volume II-3/W5, 2015.
- [180] X. Shi, Z. Huang, D. Li, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li. FlowFormer++: Masked cost volume autoencoding for pretraining optical flow estimation. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 1599–1610, 2023.
- [181] A. Stone, D. Maurer, A. Ayvaci, A. Angelova, and R. Jonschkowski. SMURF: Self-teaching multi-frame unsupervised RAFT with full-image warping. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, pages 3887–3896, 2021.
- [182] E. Hennestad, A. Witoelar, A. R. Chambers, and K. Vervaeke. Mapping vestibular and visual contributions to angular head velocity tuning in the cortex. Cell Rep, 37(12), 2021.
- [183] J. Cho, S. Han, E.-S. Cho, K. Shin, and Y.-G. Yoon. Robust and efficient alignment of calcium imaging data through simultaneous low rank and sparse decomposition. In Proc IEEE Winter Conf Appl Comput Vis, pages 1939–1948, 2023.

- [184] B. Ahanonu, A. Crowther, A. Kania, M. Rosa-Casillas, and A. I. Basbaum. Long-term optical imaging of the spinal cord in awake behaving mice. Nat Methods, 21(12):2363–2375, 2024.
- [185] A. Flores-Valle and J. D. Seelig. Axial motion estimation and correction for simultaneous multi-plane two-photon calcium imaging. Biomedical Opt Express, 13(4):2035–2049, 2022.
- [186] W. Liu, J. Pan, Y. Xu, M. Wang, H. Jia, K. Zhang, X. Chen, X. Li, and X. Liao. Fast and accurate motion correction for two-photon Ca^{2+} imaging in behaving mice. Front Neuroinform, 16:851188, 2022.
- [187] P. Flotho, A. Romero, K. Schwerdtfeger, M. Hülser, and D. J. Strauss. Energy based selective averaging approach for multi-trial optical imaging recordings. In Proc IEEE Int Symp Biomed Imaging, pages 244–247, 2017.
- [188] P. Flotho, A. Romero-Santiago, K. Schwerdtfeger, M. Hülser, L. Haab, and D. J. Strauss. Pyramid approach for the reduction of parallax-related artefacts in optical recordings of moving translucent volumes. In Annu Int Conf IEEE Eng Med Biol Soc, pages 4411–4414, 2017.
- [189] Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson, and D. L. Filion. Publication recommendations for electrodermal measurements. Psychophysiology, 49(8):1017–1034, 2012.
- [190] M. M. Bradley and P. J. Lang. The international affective digitized sounds (IADS-2): Affective ratings of sounds and instruction manual. Technical report, Technical Report B-3. University of Florida, Gainesville, FL, 2007.
- [191] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, S.-W. Tan, and X. Li. Effective recognition of facial micro-expressions with video motion magnification. Multimed Tools Appl, 76(20):21665–21690, 2017.
- [192] T. M. Ryan, A. J. Hinojosa, R. Vroman, C. Papasavvas, and L. Lagnado. Correction of z-motion artefacts to allow population imaging of synaptic activity in behaving mice. J Physiol, 598(10):1809–1827, 2020.

- [193] M. Bertalmio, A. L. Bertozzi, and G. Sapiro. Navier-Stokes, fluid dynamics, and image and video inpainting. In Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit, volume 1, pages I–I, 2001.
- [194] H. Raguet, C. Monier, L. Foubert, I. Ferezou, Y. Fregnac, and G. Peyré. Spatially structured sparse morphological component separation for voltage-sensitive dye optical imaging. J Neurosci Methods, 257:76–96, 2016.
- [195] F. Becker, S. Petra, and C. Schnörr. Optical flow. In Handbook of Mathematical Methods in Imaging, pages 1945–2004. Springer, New York, 2015.
- [196] R. Zhao, D. Li, P. Zuo, R. Bai, Q. Zhou, J. Fan, C. Li, L. Wang, and X. Yang. Influences of age, gender, and circadian rhythm on deceleration capacity in subjects without evident heart diseases. Ann Noninvasive Electrocardiol, 20(2):158–166, mar 2015.
- [197] J. Koenig and J. F. Thayer. Sex differences in healthy human heart rate variability: A meta-analysis. Neurosci Biobehav Rev, 64:288–310, may 2016.
- [198] R. Martins and J. M. Carvalho. Eye blinking as an indicator of fatigue and mental load—a systematic review. In Occupational Safety and Hygiene III. 2015.
- [199] I. Sárándi, D. P. Claßen, A. Astvatsatourov, O. Pfaar, L. Klimek, R. Mösges, and T. M. Deserno. Quantitative conjunctival provocation test for controlled clinical trials. Methods Inf Med, 53(4):238–244, 2014.
- [200] T. Soukupova and J. Cech. Real-time eye blink detection using facial landmarks. In Luka Cehovin, Rok Mandeljc, and Vitomir Struc, editors, Comput Vis Winter Workshop, pages 1–8, Rimske Toplice, Slovenia, February 2016.
- [201] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile - the munich versatile and fast open-source audio feature extractor. Proc ACM Int Conf Multimed, pages 1459–1462, 2010.

Acknowledgements

There are many people that I want to thank for their support during the time of this thesis and with the moving end-of-year deadlines of this “Christmas project”.

First of all, I thank my supervisor **Prof. Dr. Dr. Daniel J. Strauss** for his guidance throughout this thesis. He presented and we discussed many interesting problems and use cases from systems neuroscience and neurotechnology that shaped this thesis and my research interests. When I joined his group, he provided me with an environment with academic freedom, where I could find my research direction and could always stick with the latest developments in image processing and computer vision. Together, we shaped Lagrangian motion magnification for faces, and he gave me the opportunity to build a visual computing lab from scratch and gather experience with grant writing which ultimately also resulted in large financial freedom in designing our systems and approaches. The amount of opportunities to work on a broad range of topics, the exposure to state-of-the-art setups, cameras as well as the trust and responsibility to contribute to the group’s development at the chair was a valuable opportunity. It left me with a unique learning experience where, at the end of the PhD, I feel I could demonstrate way more than only technical expertise. The interdisciplinary and international collaborations I developed during this time, and some of which Daniel initiated, have been invaluable.

I thank **Prof. Dr. Bernd Kuhn** who leads the Optical Neuroimaging Unit at the Okinawa Institute for Science and Technology. Although our first project happened only during a short visit of 3 months, our work together laid the foundation for multiple joint publications. The visit was an important period during my PhD, where I have learned about the compatibility of my work with real-world research needs. Seeing our toolbox for motion correction being used across the globe and the value it adds was a rewarding experience and shaped my interest for interdisciplinary and cross-domain work. I am grateful to Bernd for his support during the revision of this thesis: While I was staying in the course of a JSPS fellowship, he supported me both with a working place and strong mentoring through a clear timeline which allowed me to focus on the new additions.

Equally, I want to thank **Prof. Dr. Gabriele Steidl** who leads the applied math group at the Technical University Berlin. She was the driving force behind our joint publications. I have learned a lot from her and her clear approach to research is one to look up to. I am thankful for the visits over the years that lead to great scientific discussions and helped me shape my trajectory and vision for interdisciplinary and cross-domain research!

My thanks extend to the reviewers of this thesis. They provided detailed, and actionable feedback that strengthened the thesis and helped transforming it from a collection of papers into a coherent document and helped me improve the overall thesis with additional examples and refinements.

A great thanks goes to **Prof. Dr. Karsten Schwerdtfeger** and **Dr. Lars Haab** who both sparked my interest in optical imaging. Lars has been a great mentor and I have spent a considerable time of this thesis in his lab. I want to thank **Dr. Kazutaka Takahashi** for our collaboration and interesting data and challenges. Many thanks also to **Dr. Anna Kukleva**, **Moritz Piening** and **Cosmas Heiß** who made our joint publications great papers.

At OIST, I want to thank **Dr. Shinobu Nomura**, **Dr. Miles Desforges**, **Dr. Mohammed el Tabbal** and **Dr. Soumen Jana** for the great and close collaborations that pushed our tools forward.

At SNNU, I want to thank my past and present colleagues **Dr. Corinna Bernarding**, **Dr. Manuel Kohl**, **Dr. Farah I. Corona-Strauss**, **David Thinnies**, **Elena Schneider**, **Benedikt Buchheit**, **Andreas Schröer**, **Tobias Grün** and most of all **Mayur J. Bhamborae** who helped shaping the visual computing lab from day 1 and who will push computer vision within SNNU to new frontiers!

A deep thank you to my friends and family (if not mentioned above).

Dorothee, thank you for the years of support, motivation and love — and thank you for enduring the long working nights, workcations, and the challenging phases of my thesis with me.

Adrian, **Stefan** and **Kathi** thank you for your friendship. **Dushyant**, you really carried me over the finish line. Thank you **Nik**, **Junaid**, **Joachim**, **Sascha**, **Torsten**, **Andy**, **Julian** and **Christian** for all the friendship and time together over the years.

Last but not least, I thank my parents **Ulrike** and **Eckehard** and, in particular, my two brothers **Georg** and **Matthias**. They have been a huge support throughout the years and laid the foundation for my interest in science and research - **Matthias** I don't know why but I am happy that you are doing the same and I know you will rock it.