universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Variational Models of Visual Attention with a Special Focus on Dynamic Sequences“

verfasst von / submitted by

Aniello Raffaele Patrone

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Technischen Wissenschaften (Dr. techn.)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 786 880
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Informatik
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Ing. Dr. Otmar Scherzer

Contents

Abstract **vii**

Preface **ix**

I Preamble **1**

1 Challenges in visual attention estimation **3**

2 Variational optical flow **7**

3 Contributions of the thesis **11**
3.1 Visual attention analysis 11
3.2 On a spatial temporal decomposition of optical flow 14
3.3 Dynamical optical flow of saliency maps for predicting visual attention 16
3.4 Dejittering Models 18
3.5 Discussion and further research 20

II Publications **21**

4 Visual attention in edited dynamical images **23**
4.1 Introduction 24
4.2 Related work 25
4.3 Two-step model 27
4.4 Evidence 30
4.5 Open questions and potential Applications 34
4.6 Conclusion and comparative evaluation 34

5 The effect of cinematic cuts on human attention **37**
5.1 Introduction 38
5.2 Method 40
5.3 Results 42
5.4 Discussion 43
5.5 Conclusion 44

6 On a spatial-temporal decomposition of optical flow **45**
6.1 Introduction 46

6.2 Registration and optical flow 47
6.3 The optical flow equation in case of illumination disturbances . 48
6.4 Optical flow decomposition: basic setup and formalism 51
6.5 Optical flow decomposition in 1D 54
6.6 Numerics 56
6.7 Experiments 58
6.8 Conclusion 64

7 Dynamical optical flow of saliency maps for predicting visual attention 65
7.1 Introduction 66
7.2 Computational methods 68
7.3 Eye tracking experiment 75
7.4 Results 76
7.5 Conclusion 81
7.6 Acknowledgements 82

8 Infinite dimensional optimization models and pdes for dejittering 85
8.1 Introduction 86
8.2 Basic Notation and Problem Formulation 87
8.3 Line Dejittering 89
8.4 Line Pixel Dejittering 90
8.5 Pixel Dejittering 92
8.6 PDE Models as Formal Energy Flows 93
8.7 Numerical Results 93
8.8 Conclusion 96

III Appendix 97

Bibliography 99

Zusammenfassung 111

Curriculum Vitae 113

Abstract

Attention is the process of focusing our mental capacities on parts of the available information. This is because humans cannot process all available information at once. In this thesis, we focus on the visual attention and we try to simulate mathematically its behavior.
The diffusion of information through videos is more and more present in today's society though TV on demand, web-streaming, e-learning and online games, just to name a few. The present work focuses on the following research areas: the importance of cuts in movie sequences for visual attention, the attractiveness of a location in a video and the behavior of visual attention in the presence of distortions such as jitter.

In the following, we shall concentrate on the first research area and, more specifically, on cuts. They refer to an editing technique which leads to a strong change of the movie scene. In particular, object locations become uncorrelated through cuts. We initially analyze the behavior of viewers while watching a video containing a cut, from the point of view of cognitive science. We propose a two-step conceptual architecture and test it through eye-tracking experiments. The architecture is driven by the temporal coherence of the apparent movement, also known as *optical flow* and focuses on two cases: the viewer's reaction to a sequence without cuts and with cuts, respectively. We propose that the viewer's attention is attracted by novelty within a movie take not containing cuts. In this case, while the global flow is coherent, local incoherence indicates novelty. The viewer's behavior changes if a cut is encountered. In this case, the global flow is incoherent, signalizing the cut. The viewer's attention is attracted by repeated features such as repeated movement.
Mathematically, we formulate the two-step architecture as a variational optical flow problem. We start from the Horn-Schunck functional, conveniently modified in order to include the spatio-temporal extension by Weickert-Schnörr. We propose a decomposition of the flow into two optical fields: one characterizing the time-coherent flow and another referring to repeated movement, also known as *oscillating pattern*. In order to model the oscillating pattern, we propose a regularizer that is non-local in time, inspired by Meyer's book.

We delineate now the second research area, referring to the attractiveness of a certain location in a video. The target of a visual attention model is to estimate the attractiveness of a location for a viewer, translated numerically in a probability of interest. A map including the probability of interest for each point of a static image is called *saliency map*. In order to calculate the saliency of dynamic sequences, the standard approach is to calculate the saliency of each frame composing the video and the saliency of motion features, combining them through a weighting scheme. We propose an algorithm for calculating

the saliency of motion features in a dynamic sequence, called *dynamic saliency map*. Again, we formulate the motion features as a variational optical flow problem. In particular, we calculate the flow of a high-dimensional sequence composed by intensity or color channels complemented by the saliency map of each frame. This allows us to overcome the aperture problem. Moreover, we include a modified version of the spatio-temporal extension by Weickert-Schnörr in our functional. Thanks to the change we propose, our model is particularly effective in the case of occlusions. Indeed we simulate the human behavior continuously following motion of an object through occlusion in our dynamic saliency map.

We address the third and last research area, referring more specifically to the behavior of visual attention in the presence of distortions such as jitter. Humans are able to recognize shapes and objects up to a certain level of distortion. The human mind performs an automatic reconstruction of the original image. We simulate this reconstruction process in the case of static images and focus on a particular type of distortion, called *jitter*. Jitter arises when the time interval between sampling points of the signal is incorrect. We propose variational functionals to dejitter images affected by line, line pixel and pixel jitter.

The proposed algorithms allow cognitive scientists to test theories and perform quantitative evaluation. Eye-tracking experiments should be designed for testing the response of human visual attention compared to the result of our algorithms. A further step of mathematical interest could constitute the extension of our models towards a general one, able to simulate visual attention in all above-mentioned research areas at once. We claim that an appropriate formulation of the optical flow can deliver quantitative methods for the estimation of visual attention.

Preface

This manuscript is a cumulative dissertation, a collection of five articles, all of which have been published or submitted to scientific journals. The articles were written in the course of three years and are the result of a fruitful scientific collaborative effort. The purpose of this preface is to present the structure of this dissertation.

The present thesis is concerned with the modeling of visual attention of dynamic sequences. Moreover, we focus on a connection between visual attention and variational optical flow. The thesis is structured in three parts, as follows.

The first part, the preamble, constitutes the introduction and contains three chapters. In Chapter 1, we briefly present challenges in visual attention estimation. In Chapter 2, we introduce variational optical flow as a reliable and well-established method for motion estimation. Finally, we discuss the analyzed challenges, explain our ideas to solve them, and summarize the contributions of this thesis in Chapter 3.

The second part contains the five publications arranged in chapters. The first article [7] resulted from the collaboration with Ulrich Ansorge, Shelley Buchinger, Christian Valuch and Otmar Scherzer and is presented in Chapter 4. The second and the fourth articles [131, 102] resulted from the collaboration with Christian Valuch, Ulrich Ansorge and Otmar Scherzer and are presented in Chapters 5 and 7. The third article [101] was written together with Otmar Scherzer and is presented in Chapter 6. Finally, the fifth article [36] resulted from joint work with Guozhi Dong, Otmar Scherzer and Ozan Öktem and is presented in Chapter 8.

The third part is the appendix containing a single bibliography for the entire manuscript, the German translation of the abstract and a Curriculum Vitae with particular attention to my scientific career.

Let me express my gratitude and thanks to everyone who contributed in a direct or indirect way to this thesis. First, I would like to thank my supervisor, Otmar Scherzer. Thank you for your guidance, your critical spirit and your patience, in spite of things being sometimes "a disaster", like you said. I benefit from these years spent working together.
My gratitude goes to the people that I define "my Austrian family". This family is composed of Otmar - my supervisor, Min - his amazing assistant, Axel- our system administrator and all my colleagues of the CSC (or CNC for some). I learned a lot from everyone of them. They are more than simply colleagues, they became my friends. We shared not only interesting talks regarding our work, but also a part of our lives. We grew up together and we had a lot of fun. Thank you guys, this thesis would not have been possible without you.
I acknowledge the financial support of the project *"Modeling Visual Attention as a Key Factor in Visual Recognition and Quality of Experience"* by the Wiener Wissenschafts und Technologie Fonds - WWTF. This project gave me the possibility to meet amazing people such as Ulrich Ansorge, Christian Valuch and Shelley Buchinger.
Let me express my gratitude to my entire family, who supported me in this experience and was always there for me when needed.
Finally, I would like to thank Oana- the love of my life. You supported and pushed me when necessary. I could have never done this without you. Thank you.

Aniello Raffaele Patrone
University of Vienna
July, 2016

Part I

Preamble

Chapter 1

Challenges in visual attention estimation

Humans are able to focus on and process just few visual information at once. Visual attention can be defined as the process of focusing our mental capacities on selections of sensory inputs. This is done in order to allow the mind to successfully process the stimulus of interest.
In the past decades, an incredible effort was made in order to properly model visual attention. The first attempt was relative to static images [64], in which Itti et al. suggested to analyze low-level features of an image (intensity, colors and orientation) [64]. This approach is currently considered a standard in literature. Predicting the viewer's attention based on low-level features alone is a challenging task and numerous different solutions were proposed [3, 53, 63, 66, 105, 108]. We distinguish between *bottom-up* models, which are task-independent and driven mainly by the intrinsic features of the visual stimuli, and *top-down* models, which are task-dependent and driven by high-level processes (like face recognition, viewer preferences). Many attempts were made in order to properly combine bottom-up and top-down models [99, 104, 81].
In the last years there was an increasing interest in modeling visual attention for dynamic image sequences [42, 60, 62, 63, 77, 109]. The possibility to predict the visual attention regarding a dynamic sequence is of interest in many fields, like computer science, marketing, and cognitive science, just to name a few. This problem is much more challenging than for static images, due to the quantity of information involved in a video sequence. In spite of many attempts in literature [42, 60, 62, 63, 76, 77, 79, 81, 104, 109, 119, 148] little is known about how to appropriately model visual attention for dynamic image sequences.
Motion has proven to be a feature of strong attraction to the human eye in videos. There are however other factors to be considered in order to properly model visual attention. One of them is related to video recording techniques like for example visual *cuts*. These are a particular recording technique, or visual discontinuities, that require shifting attention from one location to another. This shift is motivated by uncorrelated object locations across the cut (see Figure 1). After a cut, current models of human attention emphasize the importance of new information, also known as *Bayesian surprise*. However, this is not necessarily true for cuts within edited videos. Evidence [7] suggests

Figure 1.1: Two examples of cut

that attention is attracted by repeated visual features, when correlation between two successive images is low [20, 80].
A repeated feature that attracts visual attention is movement. Indeed, there are movements that are smooth, like a person walking, and others that are repeated over time, like a pendulum. Humans are perfectly able to discern between a repeated and a smooth movement, thanks to their memory. We aim to simulate the ability to decompose the motion into a smooth part and a repeated one, an *oscillating pattern*. This motion decomposition is useful in the presence of cuts. In particular, we focus on a particular type of cut, well studied by cognitive scientists, called *flicker* [97]. When a dynamic sequence includes flicker, the observers experience the *change blindness* effect. In practice, as reported by J. K. O'Regan: "Change blindness is a phenomenon in which a very large change in a picture will not be seen by a viewer, if the change is accompanied by a visual disturbance that prevents attention from going to the change location". In the sequences used in our tests, each frame containing the visual information is alternated with a blank frame as visual disturbance, so that the viewer experiences the change blindness effect (see Figure 1.2). In this case, classical algorithms of motion estimation [57, 122] would fail. We aim to recognize the flicker as an oscillating pattern.

In order to consider repeated motion over time, it is necessary to consider the time evolution of motion. This is also relevant for modeling visual attention in the case of occlusions. Again, classical algorithms of motion estimation [27, 57, 122] would fail in presence of occlusions. Instead, humans are able

Figure 1.2: This is a flicker example. A white frame is interleaved between the two frames shown in a repetitive way. The area in the blue box is the area that changes.

to follow motion of an object though occlusions [4, 45, 118]. Again, this is thanks to human memory and human understanding of temporal coherence. Here, we aim to include the temporal coherence in our computational model and recognize occlusions as attractive, or *salient* for the viewer. The results of the model, a so-called *dynamic saliency mapping*, will be compared with eye-tracking data.

The basic framework used in this work is the motion estimation applied to problems of interest to cognitive scientists. In order to model the motion, we use the well-established variational optical flow estimation. Variational methods in image processing and computer vision are well known. The principal advantage of these methods is the mathematical formalism of their assumptions.

If we want to model the attention relative to visual information, it is important to consider that images are often affected by noise and distortions. They change our scanning pattern because humans process the image in order to recognize (or reconstruct) the original image [40, 139]. Therefore, the denois-

Figure 1.3: This is a line jitter example. In the first image (left) we show the original image and in the second image (right) we show the line jittered version of the original image.

ing problem is of interest for cognitive scientists and especially for the visual attention modeling community. It is possible to use variational methods also for the denoising problem, in order to reconstruct the original image. In this thesis, we focus on a special type of denoising called *dejittering*, that is usual for static images. Dejittering is defined as the process of assigning pixel positions to image data recorded with pixel displacements. *Jitter* is a type of distortion which arises frequently in signal processing, when the distance (over time) between sampling points vary, rendering signal errors. There are three different forms of jitter: line jitter, line pixel jitter and pixel jitter. *Line jitter* consists of horizontal shifts of each row (line) of an image. The shift is the same for the entire row. This may typically happen when digitizing analog video frames. In case of line registration problems due to bad synchronization pulses, the line jitter is present. The effect is that the image lines are (randomly) shifted with respect to their original location (see Figure 1.3 for an example of jitter). If pixels in a row are shifted differently, we speak about *line pixel jitter*. Finally, there is *pixel jitter*, where one also experiences vertical shifts. Viewers are able to reconstruct distorted images, up to a certain level of distortion [40, 139]. Here, we aim at developing a model to simulate this behavior.

Chapter 2

Variational optical flow

Motion estimation is a key problem in the analysis of dynamic sequences. Humans perceive motion when consecutive frames of a recorded scene are shown. In particular due to the nature of the human eye, viewers observe motion by variations in intensity. These allow us to perceive also a sense of the scene depth, e.g. when objects are occluded. However, when we observe the motion through a camera, only the *apparent motion* can be observed through intensity variations. This apparent motion is known as the *optical flow* [57].

In the optical flow we assume that points moving through a scene preserve their intensity. This assumption is called *brightness-constancy assumption*. Let us consider a dynamic sequence to be a time continuous recording of images. Each image is described by a function $f(\vec{x},t)$, where $\vec{x}=(x_1,x_2)^t$ in the planar plane $\Omega\subset\mathbb{R}^2$ and $t\in[0,1]$ is the time interval in which the dynamic sequence is taking place. The value of $f(\vec{x},t)$ is the recorded image intensity value at a point $\vec{x}$ and at a time t. In the brightness-constancy assumption this value is considered constant along a smooth trajectory $\vec{\gamma}(\vec{x},t)$. That is:

$$f(\vec{\gamma}(\vec{x},t),t)=f(\vec{x}_0,t_0) \tag{2.1}$$

for some initial point $\vec{x}_0$ and time t_0. If the equation (2.1) holds, it follows by differentiation with respect to time (see for instance [57])

$$\frac{df(\vec{\gamma}(\vec{x},t),t)}{dt}=\nabla f(\vec{\gamma}(\vec{x},t),t)\cdot\partial_t\vec{\gamma}(\vec{x},t)+\partial_t f(\vec{\gamma}(\vec{x},t),t)=0 \tag{2.2}$$

must hold true for all $\vec{x}\in\Omega$ and for all times $t\in[0,1]$. In (2.2) we denote by $\nabla f=(\frac{\partial}{\partial x_1}f,\frac{\partial}{\partial x_2}f)$ and $\partial_t f$ the partial derivatives in space and time of the function f. For simplicity, let us denote by

$$\vec{u}(\vec{\gamma}(x,t),t)\coloneqq\partial_t\vec{\gamma}(x,t)$$

the velocity of a point moving along γ. In order to estimate the optical flow we need to find a time-dependent vector field $\vec{u}:\Omega\times[0,1]\to\mathbb{R}^2$ satisfying

$$\nabla f\cdot\vec{u}+\partial_t f=0 \tag{2.3}$$

for all $(\vec{x},t)\in\Omega\times[0,1]$. For the sake of brevity we have omitted the arguments of the functions.

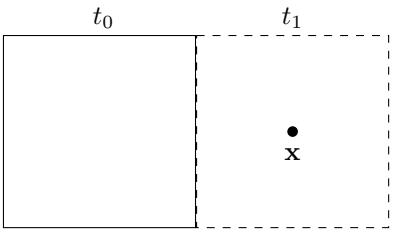

Figure 2.1: We show an example for which it does not exist a solution for the optical flow equation (2.3). A cube of uniform color, or in other words with $\nabla f = 0$ for the internal part, is moving from an initial position at time t_0 to a target position at time t_1 (dashed in figure). We notice that the function f for the point x and at time t_1 has time derivative $\partial_t f \neq 0$ and $\nabla f = 0$ therefore it is not possible to calculate a solution of (2.3).

The equation (2.3) is called *optical flow equation*. It is equivalent to (2.1) under suitable assumptions and with appropriate initial and boundary conditions. Moreover, let us notice that equation (2.3) is linear in the unknown $\vec{u}$, therefore it is reasonable for sufficiently small displacements.

Looking more carefully at equation (2.3), we realize that it is underdetermined and existence and uniqueness of a solution are not guaranteed. A solution to (2.3) does not exist if $\partial_t f \neq 0$ and $\nabla f = 0$ (see Figure 2.1). Regarding the uniqueness we notice that a trivial solution, also called *normal flow*, is:

$$\vec{u}^\dagger = \begin{cases} -\dfrac{\partial_t f}{|\nabla f|^2} \nabla f, & \text{if } \nabla f \neq 0, \\ 0, & \text{else,} \end{cases} \tag{2.4}$$

This solution is not unique. Indeed, if we add to (2.4) a flow $c\,\nabla f(x,t)^\perp$ which is orthogonal to the image gradient, this new flow solves the equation (2.3) for every $c \in \Omega$ and even for every $c : \Omega \times [0,1] \to \mathbb{R}$. In other words, from (2.3) we can infer only the movement along the direction of the image gradient. This issue is called *aperture problem* (see also the illustrations in [9, Sec. 5.3]).

Due to the above reason, the estimation of the optical flow can be interpreted as an *ill-posed* inverse problem. In [41, 116] can be found a general treatment of inverse problems and regularisation theory, where [116] has a particular focus on imaging.

In order to solve (2.3), one has to overcome the above mentioned problems. We refer to [9, 17, 44] for an introduction and a comparison of various techniques. Moreover, it is worth to cite the benchmark framework created by Baker et al. [14]. The corresponding website of the Middlebury College is an important reference point of the optical flow community for comparison of methods.[1]

In this thesis, we choose a *variational approach* for estimation of the optical flow. For the above-mentioned problems, the equation (2.3) cannot be solved directly. A common approach [57] in order to obtain *well-posedness* of the

[1] `http://vision.middlebury.edu/flow/`

optical flow problem, and with it uniqueness of a solution, is *Tikhonov regularization* [126]. Using an approach as in [126] for the optical flow estimation results in finding the unique minimizer of

$$\|\nabla f \cdot \vec{u} + \partial_t f\|^2_{L^2(\Omega)} + \alpha \mathcal{R}(\vec{u}), \tag{2.5}$$

where $\mathcal{R}$ is used to enforce smoothness of the minimizer and $\alpha > 0$ is a weight parameter. Usually, the first term of (2.5) is called *data term*, whereas the second is called *regularizer*. Using an approach as in (2.5) we can restrict the space of solutions to a desirable one.

In their seminal work [57], Horn and Schunck proposed to compute the minimizer of

$$\|\nabla f \cdot \vec{u} + \partial_t f\|^2_{L^2(\Omega)} + \alpha \left|\vec{u}\right|^2_{H^1(\Omega)}, \tag{2.6}$$

for one pair of frames only. The $H^1(\Omega)$ Sobolev seminorm is a good choice because penalizes first derivatives with respect to space and favors spatial regularity of the solution. In other words, the assumption is that the velocity field should vary smoothly in space. This is perfectly reasonable if we think for example of a rigid object moving. In this case we expect that the velocity field values are similar for all the points of the object.
The functional in (2.6) is easy to solve numerically, but entails isotropic regularity and does not allow discontinuities in the flow field. Schnörr [117] was the first to show the well-posedness of (2.6), making additional assumptions on ∇f.

In order to calculate a solution to (2.6) we solve the corresponding set of Euler-Lagrange equations following the calculus of variations [33, Chap. IV]. These equations form a system of second-order elliptic partial differential equations:

$$\begin{aligned} (\nabla f \cdot \vec{u} + \partial_t f)\nabla f - \alpha \Delta \vec{u} = 0 \quad &\text{in } \Omega, \\ \partial_{\vec{n}} \vec{u} = 0 \quad &\text{on } \partial\Omega. \end{aligned}$$

Here, $\partial_{\vec{n}}$ is the normal derivative along the outward unit normal $\vec{n}$ on $\partial\Omega$.

Weickert and Schnörr suggested in [142] to extend the smoothness of the flow field also in time. They proposed the functional:

$$\|\nabla f \cdot \vec{u} + \partial_t f\|^2_{L^2(\Omega\times[0,1])} + \alpha \left|\vec{u}\right|^2_{H^1(\Omega\times[0,1])}. \tag{2.7}$$

The flow calculated from (2.7) is more robust because it uses all the information available and not only the one of each pair of frames. In this way, the authors of [142] are able to ensure time coherence of the resulting flow. The drawback is a algorithm more demanding from a computationally point of view. There, minimization was done by applying a semi (Euler forward) scheme to the associated steepest descent equations.

Let us conclude this chapter with a few remarks. The above discussed functionals are straightforward to minimize, however they present several drawbacks. Many attempts have been proposed over the last years [122, 140], in order to manage discontinuities in the flow field, or different data terms. For example in [27, 149] they assume gradient constancy and not only brightness constancy. Indeed, this last assumption for certain data is often violated. Moreover, in [27, 28, 59, 88, 149] they propose other extensions of the original formulation (2.6) in order to use color data, to guarantee contrast invariance and to implement efficiently the optical flow estimation method, but to name a few.

Chapter 3

Contributions of the thesis

Modeling visual attention is a challenging problem. We gave a brief overview about that in Chapter 1. In this thesis we propose solutions to few of these challenges, which aim to simulate the human behavior. In order to do so, we therefore need to comprehend this behavior and undertake the following steps. Initially, we analyze the behavior of human while they are watching a dynamic sequence from a cognitive perspective. In particular, we propose a cognitive model and perform eye-tracking experiments. Secondly, we move from the proposed cognitive model to a mathematical model. Finally, we apply the studied theory to the modeling of motion features in order to estimate saliency of dynamic sequences and dejitter static images.
In this thesis we will try to answer the following questions:

- Are repeated features in edited video sequences of interest?

  If yes, how do we model repeated features, like repeated movement?

- How can we model motion features in order to estimate saliency of a dynamic sequence?

- Is it possible to reconstruct corrupted (dejittered) images?

3.1 Visual attention analysis

Let us answer the first question: **Are repeated features in edited video sequences of interest?**
In [7] we proposed a two-step architecture for the modeling of visual attention in edited dynamical images. The type of editing considered is cinematic cut. As already explained in chapter 1, visual attention models can be divided in bottom-up and top-down models. *Bottom-up* models assume that the focus of attention is fully determined by the characteristics of the visual stimuli [63]. *Top-down* models emphasize the importance of past experiences, goals, intentions, interpretations and interests of the viewer as predictors of visual attention [128].
We proposed an architecture that combines one bottom-up feature and several top-down principles. The bottom-up feature is the motion information. The top-down principles describe the relevance of new information and of repeated one. The model is driven by the temporal coherence of the optical flow across

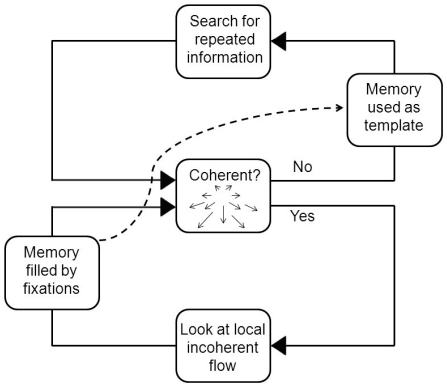

Figure 3.1: Whenever the global optical flow is temporally coherent, the human gaze is steered towards novel information within movie takes. In this setting, when new objects appear in the visual field, this is characterized by locally incoherent flow. The situation is different in the case of a cut. Cuts are characterized by incoherence (or temporal discontinuities) of the global optical flow. In this situation, the human gaze is steered towards repeated information.

subsequent images. The two-step architecture proposed is shown in Figure 3.1. We indicated that the human gaze is steered toward novel information within movie takes. In this case, attraction to novelty is achieved by down-weighting optical flow and up-weighting temporally locally incoherent flow for the selection of gaze directions. The situation changes in the case of a cut. *Cuts* are characterized by incoherence of the global optical flow. In this situation, the human gaze is steered towards repeated information. We motivated our model by the following evidence:

- The visual attention is attracted by human action in general [54] and human faces in particular [46]. In these cases, actions and facial movements are defined by local motion patterns that have regularities differing (*local incoherent flow* in Figure 3.1) from the global temporally coherent optical flow. Similarly, a motion singleton among coherently moving objects captures human attention. Let us think for example of an object entering in the scene. It will describe a motion singleton with respect to time, more precisely a locally incoherent optical flow.

- Humans have the ability to selectively look for particular visual shapes or combinations of shapes and colors [129]. Similarly, viewers could also search for landmarks seen in the past and use them, after a cinematic cut, to decide whether a visual scene continues or has changed. The ability of viewers to learn from known stimuli or part of them is shown in Brooks et al. [26].

- As we are going to show below in this section, in recent experiments, participants preferentially have looked at videos bearing a high similarity of pre-cut and post-cut images [131, 133]. There, we used two videos presented side by side and asked participants to keep their eyes on only

one of the videos. We applied two types of cut to the videos: cuts with a high pre- and post-cut feature similarity and cuts with a lower pre- and post-cut feature similarity. Moreover, the images could switch position during the cut. In this setup, the participants showed a clear preference for looking at images with a higher pre- and post-cut similarity.

- Another research [138] chooses to rearrange an otherwise coherent take by introducing cuts. The resulting video was temporally incoherent and this lead to a drastically reduced reliability of the gaze pattern.

In [131] we tested the hypothesis that visual attention is attracted in videos by repeated information after a cut takes place. This hypothesis was tested through an eye-tracking experiment, which consisted of participants keeping their gaze focused on a video that was shown next to another irrelevant video. Cuts were applied to both videos. This resulted in a low correlation of object locations pre- and post-cut. The cuts were of two types: cuts with high pre- and post-cut feature similarity (within scene cut, abbreviated WSC), and with low pre- and post-cut feature similarity (between scenes cut, abbreviated BSC). An example is shown in Figure 3.2. Moreover, the location of the videos on the screen was randomly switched. The measurement used for the analysis of this experiment was the saccadic reaction time (SRT). This is defined as the latency of the first saccade towards the target video after the videos switched position.

In Figure 3.3 we show the results of this experiment. First, we sought confirmation of the fact that more visual information is repeated after WSCs than BSCs and validated it. This was done by calculating the mean Euclidean dis-

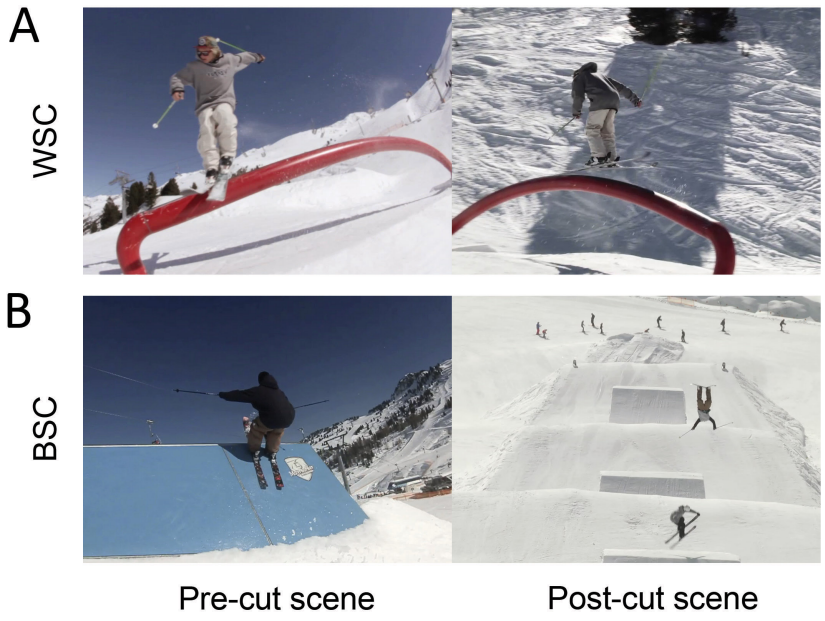

Figure 3.2: This are cuts examples. (A) Within scene cut (WSC). (B) Between scenes cut (BSC).

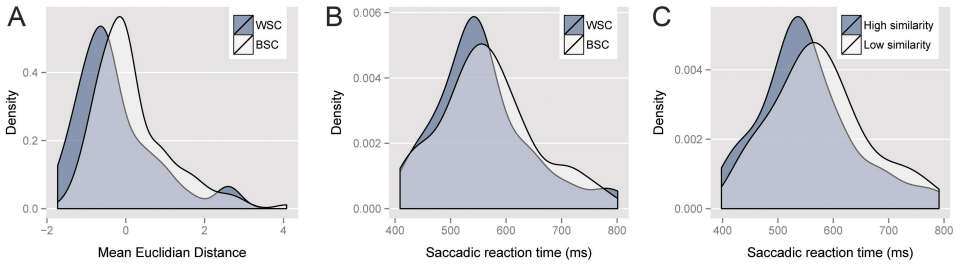

Figure 3.3: Results. (A) Distribution of mean Euclidean distances (z-transformed) of RGB color histograms of the last pre-cut and the first post-cut frame as a function of cut category. Values below 0 represent higher similarity, values above 0 represent lower similarity. (B) Distribution of individual median SRT as a function of cut category. (C) Distribution of individual median SRT as a function of color histogram similarity across the cut.

tance of the red-green-blue (RGB) color histograms between the final pre-cut and the first post-cut frame. In order to clarify the result, we transformed these values, so that values below 0 represented higher similarity and low similarity otherwise. Figure 3.3 (A) confirms that in WSCs more visual content is repeated than BSCs. Indeed, the majority of the values of the mean euclidean distance for WSCs is below 0, therefore indicating high similarity. The opposite is true for BSCs. Secondly, we tested the connection between WSCs, BSCs and SRT. We noticed in Figure 3.3(B) that median SRT is on average 9 milliseconds (ms) shorter in WSCs than BSCs. Finally, we categorized the cuts into *high similarity* cuts if their value of mean distance, as reported in 3.3(A), was below 0 or *low similarity* cuts otherwise. Also this analysis, in Figure 3.3(C) confirmed that SRTs were 23 ms shorter after *High similarity* than after *Low similarity cuts*.
These experiments and evidence suggest that in the case of cuts, viewers benefit from repeated information and are attracted by it.

3.2 On a spatial temporal decomposition of optical flow

Let us answer the second question: **How do we model repeated features, like repeated movement?**
The two-step architecture depicted in Figure 3.1 suggests the need for an algorithm able to decompose the motion information. In particular, the algorithm should divide the repeated motion in time (not coherent in time in Figure 3.1), from the one smooth (or coherent) in space and time. In [101], we proposed a variational approach able to decompose the optical flow.
A first step was to analyze the equation (2.3) in case of cinematic cut. We considered in [101] a particular case of cut called *flickering*. In a standard flickering experiment, the visual attention is investigated by inclusion of blank images in a repetitive image sequence. Although, in general, these blank images are not deliberately recognized, they change the awareness of the test persons.
We analyzed simple examples that explain the properties of the solution of

the optical flow equation (2.3) under conditions of illumination changes, also known as *flickering* and motivated the scientific approach we used below. First, we studied the 1D optical flow equation

$$\partial_x f(x,t)u(x,t) + \partial_t f(x,t) = 0 \text{ in } (0,1)\times(0,1) \tag{3.1}$$

to solve for u for the specific data

$$f(x,t) = \tilde{f}(x)g(t) \text{ for } (x,t) \in (0,1)\times(0,1). \tag{3.2}$$

f represents a dynamic sequence with brightness variation g over time. We were more specific and took:

$$\tilde{f}(x) = x(1-x) \text{ and } g(t) = \exp\left\{-\frac{1}{\beta}(1-t)^\beta\right\} \text{ with some } 0<\beta<1 \tag{3.3}$$

for $(x,t) \in \hat{\Omega} := (0,1/4)\times(0,1)$. The optical flow is given by

$$u(x,t) = -\frac{x(1-x)}{1-2x}(1-t)^{\beta-1}.$$

We showed in [101] that for $0<\beta<1/2$, $u \notin L^2(\hat{\Omega})$ but $\widehat{u}(\cdot,t) = \int_0^t u(\cdot,\tau)\,d\tau \in L^2(\hat{\Omega})$. The bottom line is that changes of illumination, such as flickering, may result in singularities of the optical flow and a violation of standard smoothness assumptions of the optical flow field. The potential appearance of the singularities motivated us to consider regularization terms for optical flow computations, which allow for singularities over time, such as negative Sobolev norms or G-norms.

From now on, let us consider functions on a two-dimensional domain $\vec{x} \in \Omega = (0,1)^2$ and in a time-interval $t \in (0,1)$. We assumed in [101] that the optical flow field is composed of two flow field components:

$$\vec{u}(\vec{x},t) = \vec{u}^{(1)}(\vec{x},t) + \vec{u}^{(2)}(\vec{x},t) = \begin{pmatrix} u_1^{(1)}(\vec{x},t) \\ u_2^{(1)}(\vec{x},t) \end{pmatrix} + \begin{pmatrix} u_1^{(2)}(\vec{x},t) \\ u_2^{(2)}(\vec{x},t) \end{pmatrix}.$$

In this case, the equation (2.3) contains four unknown (real valued) functions $u_j^{(i)}, i,j = 1,2$ and thus is highly under-determined. We proposed to minimize the unconstrained regularization functional:

$$\begin{aligned} \mathcal{F}(\vec{u}^{(1)},\vec{u}^{(2)}) &:= \mathcal{E}(\vec{u}^{(1)},\vec{u}^{(2)}) + \sum_{i=1}^{2}\alpha^{(i)}\mathcal{R}^{(i)}(\vec{u}^{(i)}), \\ \mathcal{E}(\vec{u}^{(1)},\vec{u}^{(2)}) &:= \int_{\Omega\times(0,1)} (\nabla f\cdot(\vec{u}^{(1)}+\vec{u}^{(2)}) + \partial_t f)^2\,\mathrm{d}\vec{x}\mathrm{d}t \text{ with } \alpha^{(i)} > 0. \end{aligned} \tag{3.4}$$

We introduced two regularizers $\mathcal{R}^{(i)}$ in order to obtain a unique solution for (3.4). For the sake of simplicity of presentation, we omitted the space and time arguments of the functions $u_j^{(i)}$ and f, whenever it simplifies the formulas without possible misinterpretations. For the two regularizers we chose:

- a spatial-temporal regularizer [142] for $\vec{u}^{(1)}$

$$\mathcal{R}^{(1)}(\vec{u}^{(1)}) := \int\limits_{\Omega\times(0,1)} \nu\left(\left|\nabla_3 u_1^{(1)}\right|^2 + \left|\nabla_3 u_2^{(1)}\right|^2\right) \mathrm{d}\vec{x}\mathrm{d}t, \tag{3.5}$$

  where $\nabla_3 = (\partial_{x_1}, \partial_{x_2}, \partial_t)$ and $\nu : [0,\infty) \to [0,\infty)$ is a monotonically increasing, differentiable function. For the choice of ν we followed [142] and took

$$\nu(r) = \epsilon r + (1-\epsilon)\lambda^2\sqrt{1+\frac{r}{\lambda^2}}\,, \forall r \in [0,\infty)\,, \tag{3.6}$$

  with $0 < \epsilon \ll 1$ and $\lambda > 0$. The function $r \mapsto \nu(r^2)$ is convex in r and there exist constants $c_1, c_2 > 0$ with $c_1 r^2 \leq \nu(r) \leq c_2 r^2$ for all $r \in \mathbb{R}$.

- a regularizer for $\vec{u}^{(2)}$ that penalizes for variations in time. In particular the regularizer that we introduced is non-local in time and is motivated by Y. Meyer's book [87]. We showed in this section and in [101] that in 1D, in case of flickering, u may violate the L^2 smoothness, however not the primitive of u in time. Variations of Meyer's G-norm were used in energy functionals for calculating spatial decomposition of the optical flow [1, 69]. It is a challenge to compute the G-norm efficiently, and therefore workarounds were proposed. For instance in [98] they proposed as an alternative of the G-norm the following realization for the H^{-1} norm: For a generalized function $v : (0,1) \to \mathbb{R}$, they defined

$$\|v\|^2_{H^{-1}} = -\int_0^1 v(t)\partial_{tt}^{-1} v(t) dt\,.$$

  Here, we use this workaround for a realization for the *temporal* H^{-1}-norm, which we used as a regularization functional:

$$\mathcal{R}^{(2)}(\vec{u}^{(2)}) = \sum_{j=1}^{2} \int_\Omega \left\|u_j^{(2)}(\vec{x},\cdot)\right\|^2_{H^{-1}} \mathrm{d}\vec{x}. \tag{3.7}$$

In [101] we minimized the functional (3.4) and we tested the proposed decomposition on different sequences.

3.3 Dynamical optical flow of saliency maps for predicting visual attention

Let us answer the third question: **How can we model motion features in order to estimate saliency of a dynamic sequence?**
The aim of visual attention models is the identification of salient areas or objects. A location in an image is considered *salient* if it stands out compared to its local surroundings. Usually, a visual attention model [64] assigns a value of saliency for each pixel of an image resulting in a *saliency map*. The literature regarding calculation of saliency maps for static images is rich and well-established [3, 53, 58, 64, 108]. However, no consensus exists on how to compute a saliency map of a dynamic sequence. Recent work has been focused

on saliency maps for dynamic sequences [42, 60, 62, 63, 77, 109]. These *spatial-temporal saliency maps* are modeled as the weighted sum of motion features and of static saliency maps [42, 60, 62, 63, 76, 77, 79, 81, 104, 109, 119, 148]. The motion features are usually [99] obtained from optical flow features (see Figure 7.1) of two consecutive frames. In [102] we proposed a new *dynamic saliency map* representing the motion features in a dynamic sequence. In detail:

- we calculated the flow of a high-dimensional image sequence, which consisted of (i) intensity and (ii) color channels, complemented by saliency maps, respectively;
- we also considered the complete movie (consisting of all frames) for the computation of a dynamic saliency map.

Using the notation as in section 3.2, if the frames composing a movie consisted of gray-valued images, then we described each by a function $f : \Omega \to \mathbb{R}$. If the frame was completed by a spatial saliency map, then $\vec{f} := (f_1, f_2)^t : \Omega \to \mathbb{R}^2$, where f_1 was the recorded movie and f_2 was the corresponding saliency map. For a colored frame, $\vec{f} := (f_1, f_2, f_3)^t : \Omega \to \mathbb{R}^3$, where each component represented a channel of the color images; typically RGB (red-green-blue) or HSV (hue-saturation-value) channels. A colored frame, which was complemented by a saliency map was described by $\vec{f} := (f_1, f_2, f_3, f_4)^t : \Omega \to \mathbb{R}^4$, where the first three components were the color channels and the fourth component was the corresponding saliency map. In [102] we showed that taking into account saliency maps for calculating the optical flow allows for overcoming the aperture problem (see 7.2).
In this setting the optical flow equation (2.3) reads as follow:

$$J_{\vec{f}}(\vec{x}, t) \cdot \vec{u}(\vec{x}, t) + \frac{\partial \vec{f}}{\partial t}(\vec{x}, t) = 0 \text{ for all } \vec{x} \in \mathbb{R}^2 \text{ and } t \in (0, 1)\,, \tag{3.8}$$

where $\vec{u}(\vec{x}, t) = (u_1(\vec{x}, t), u_2(\vec{x}, t))$ is the optical flow and $J_{\vec{f}}$, $\frac{\partial \vec{f}}{\partial t}$ are the partial derivatives in space and time of the function $\vec{f}$, respectively. Note that the *Jacobian* $J_{\vec{f}} = \nabla \vec{f} = \left(\frac{\partial}{\partial x_1}\vec{f}, \frac{\partial}{\partial x_2}\vec{f}\right)$ is a (2×2)-dimensional matrix if it is a saliency complemented gray-valued image, and a (4×2)-dimensional matrix if a color image is complemented by a saliency map.
The complemented data allows us to overcome the aperture problem for part of the pixels, but we still need a regularizer for the uniqueness of the solution. We suggested to use:

$$\int_{\Omega \times [0,1]} \Psi(|\nabla_3 u_1(\vec{x}, t)|^2 + |\nabla_3 u_2(\vec{x}, t)|^2) d\vec{x} dt$$

as in [142], with the difference that the function $\Psi(r^2) = \epsilon r^2 + (1-\epsilon)\lambda^2\sqrt{1 + \frac{r^2}{\lambda^2}}$ was replaced by

$$\Psi(r^2) = \sqrt{r^2 + \epsilon^2} \text{ with } \epsilon = 10^{-6} \tag{3.9}$$

as in [27]. Moreover, we substituted the spatial-temporal gradient operator in [142] with $\nabla_3 = (\nabla, \lambda\frac{\partial}{\partial t})$ considering $\lambda > 1$. This was done in order to accentuate the smoothness of the solution more over time than in space. This

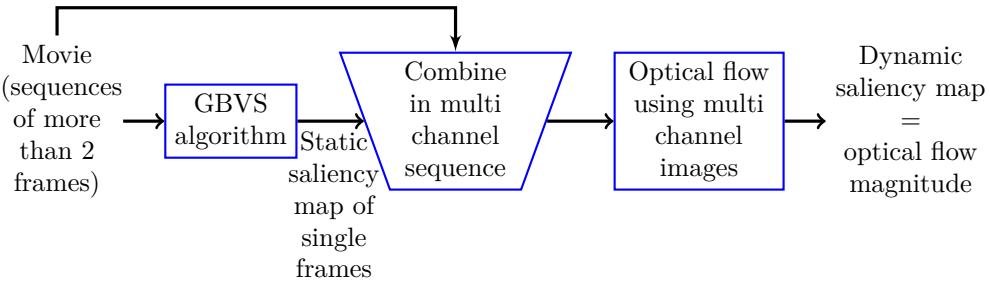

Figure 3.4: The proposed approach for calculating a dynamic saliency map.

is particularly important during occlusions. Psychological evidence [4, 45, 118] suggests that humans follow motion of an object through occlusions. From the point of view of the optical flow scientific community, having a flow through an occlusion is a wrong result [122]. Instead, as described above, such a result is desired for a visual attention model.
The resulting model for optical flow computations consisted furthermore in minimization of the functional:

$$\int_{\Omega\times[0,1]} \left\| J_f(\vec{x},t)\cdot\vec{u}(\vec{x},t)+\frac{\partial\vec{f}}{\partial t}(\vec{x},t)\right\|_{\mathcal{B}}^2 d\vec{x}dt + \alpha\int_{\Omega\times[0,1]}\Psi(|\nabla_3 u_1(\vec{x},t)|^2+|\nabla_3 u_2(\vec{x},t)|^2)d\vec{x}dt. \quad (3.10)$$

where $\|\cdot\|_{\mathcal{B}}^2$ was a semi-norm introduced in order to restore the contrast invariance for the minimizer of the optical flow. The magnitude of the optical flow calculated as minimizer of (3.10) is the dynamic saliency map. The algorithm proposed is schematically depicted in Figure 3.4.

3.4 Dejittering Models

Let us answer the fourth and last question: **Is it possible to reconstruct corrupted (*dejittered*) images?**
Humans are able to recognize objects up to a certain level of distortion [40, 139]. In [36] we simulated the ability to reconstruct distorted images for a particular type of distortion, *jitter*. Jitter arises often in signal processing, when the distance (time) between sampling points varies and leads to signal errors. In case of jitter applied to images, let $u^\delta : \Omega \to \mathbb{R}$ denote a continuous jittered image, u the original image without jittering and $\eta(\vec{x}) : \Omega \to \mathbb{R}$ noise. We defined different types of jitter:

- *line jitter* that consists of horizontal shifts of each row (line) of an image. In this case the shift is the same for all pixels of the same line. Formally:

$$u^\delta(\vec{x}) = u(x_1 + \mathtt{d}(x_2), x_2) + \eta(\vec{x})\,, \quad (3.11)$$

  respectively, where $\mathtt{d} : [0,1] \to \mathbb{R}$ denotes the jitter function of the y-th component.

- *line pixel jitter* that is similar to line jitter, but is characterized by different shifts in the same line. Formally:

$$u^\delta(\vec{x}) = u(x_1 + \mathtt{d}(\vec{x}), x_2) + \eta(\vec{x})\,, \tag{3.12}$$

  respectively, where $\mathtt{d} : \Omega \to \mathbb{R}$ denotes the jitter function of the point $(\vec{x})$ in x_1-direction.

- *pixel jitter* when for each pixel the shift is not limited to be in a line, but can be in each direction. Formally:

$$u^\delta(\vec{x}) = u(\vec{x}) + \mathtt{d}(\vec{x}) + \eta(\vec{x})\,, \tag{3.13}$$

  respectively, where $\mathtt{d} : \Omega \to \mathbb{R}^2$ denotes the jitter vector field at the point $(\vec{x})$.

Our interest in these types of problems is not only motivated by psychological reasons. Indeed, the algorithms that correct these jitters, or *dejittering* algorithms, result in calculating a flow, similarly to what we did in the above sections.
For line pixel jitter, we proposed in [36] to recover the dejittered image u by minimizing the functional:

$$\mathcal{N}(u) := \alpha\frac{1}{2}\int_\Omega \left|\frac{u^\delta(\vec{x}) - u(\vec{x})}{\partial_{x_1} u(\vec{x})}\right|^2 d(\vec{x}) + \frac{1}{p}\int_\Omega \left|\partial_{x_2}^k u(\vec{x})\right|^p \, d\vec{x}\,. \tag{3.14}$$

where ∂_{x_1} is the partial derivatives in the first component, $\partial_{x_2}^k$ denotes the k-th derivative with respect to the second component, and p is a parameter. When we use this approach to correct for line jitter, we have to consider the fact that each pixel in a line has the same shift, which leads to

$$0 = \partial_{x_1}\mathtt{d}(x_2) \approx \partial_{x_1}\left(\frac{u^\delta(\vec{x}) - u(\vec{x})}{\partial_{x_1} u(\vec{x})}\right)\,.$$

Thus, line jitter correction can be rephrased as an unconstrained minimization of the functional:

$$\mathcal{N}(u) + \beta\int_\Omega \left(\partial_{x_1}\left(\frac{u^\delta(\vec{x}) - u(\vec{x})}{\partial_{x_1} u(\vec{x})}\right)\right)^2 d(\vec{x})\,, \tag{3.15}$$

where β is a penalty parameter. Instead, for pixel jitter we end up with the functional [74, 75]:

$$\hat{\mathcal{N}}(u) := \alpha\frac{1}{2}\left\|(\nabla u)^\dagger(u^\delta - u)\right\|^2_{L^2(\Omega)} + \int_\Omega |\nabla u(\vec{x})| \, d\vec{x}\,. \tag{3.16}$$

where $(\nabla u)^\dagger$ denotes the Moore-Penrose pseudo-inverse of ∇u. Note that in comparison with (3.14), $\int_\Omega \left|\partial_{x_2}^k u(\vec{x})\right|^p \, d\vec{x}$ was replaced with the TV-semi norm $\int_\Omega |\nabla u(\vec{x})| \, d\vec{x}$.

3.5 Discussion and further research

We started from psychological assumptions and moved to mathematical models for solving the challenging task of visual attention simulation. We deployed variational optical flow as a basis for our models, that were used in order to simulate attention in the case of repeated movement (oscillating pattern), occlusions and jitter.
First, we proposed a general architecture to model human visual attention. We presented evidence that confirms our architecture. Moreover, we proved the importance of repeated information for the modeling of visual attention, through eye-tracking experiments.
Second, we suggested a model able to decompose the optical flow. Our model decomposes the flow in a repeated component, or oscillating patterns, and one coherent over time. This model allows the analysis of dynamic sequences with illumination changes, or flickering, which is also a particular type of cut.
Third, we addressed the problem of modeling the saliency of motion information for dynamic sequences. We suggested an algorithm based on optical flow of high dimensional data. Such data was composed of the video sequence and of the saliency map of each frame. The high dimensional data allowed us to overcome the aperture problem. Our experiments proved that the proposed model is particularly suitable in case of occlusions. Indeed, it is highly similar to human behavior. In practice, humans would continue to look at a moving object, also during a temporary occlusion.
Finally, we proposed an algorithm for reconstructing images affected by jitter. We gave a continuous formulation usable for reconstructing images affected by line jitter, line pixel jitter and pixel jitter.
The proposed algorithms allow cognitive scientists to test theories and perform quantitative evaluation. Eye-tracking experiments should be designed for testing the response of human visual attention compared to the one of our algorithms. For example, in the case of oscillating patterns, eye-tracking experiments are needed to test quantitatively (in terms of time response and eye patterns) how repeated movements affect the viewers. It is to be noted that an eye-tracking experiment was designed and the data collection started at the time of writing of this thesis.
A further step of mathematical interest could constitute the extension of our models towards a general model, able to simulate the visual attention when all the above-mentioned situations (oscillating patterns, occlusions and jitter) take place. Such a mathematical models would be applied to e.g. an average movie transmitted by online movie-streaming. An average movie may easily contain oscillating pattern, cuts and frames affected by jitter.
Finally, let us point out that the visualization of oscillating pattern and of saliency maps regarding videos is challenging. Usually in literature, only one frame is shown, which discards completely the time evolution. However, we think that further research should be undertaken, in order to provide cognitive scientists with satisfactory results. To summarize, we claim that an appropriate formulation of the optical flow can deliver quantitative methods for the estimation of visual attention.

Part II

Publications

Chapter 4

Visual attention in edited dynamical images

Authors & Contributions The authors are Ulrich Ansorge, Shelley Buchinger, Christian Valuch, Aniello Raffaele Patrone and Otmar Scherzer. The development of this article was a collaborative process, and each of the authors made significant contributions to every aspect of the paper.

Publication Status Published [7]: U. Ansorge, S. Buchinger, C. Valuch, A. R. Patrone and O. Scherzer. Visual Attention in Edited Dynamical Images. In *SIGMAP-2014: Proceedings of the 11th International Conference on Signal Processing and Multimedia Applications*, pages 198-205, SciTePress 2014

The final publication is available at `http://www.scitepress.org/`.

Visual Attention in Edited Dynamical Images

Ulrich Ansorge[4], Shelley Buchinger[3], Christian Valuch[3], Aniello Raffaele Patrone[1], and Otmar Scherzer[1,2]

[1]Computational Science Center, University of Vienna,
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[2]Radon Institute of Computational and Applied Mathematics,
Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria
[3]Cognitive Science Research Platform,
University of Vienna Liebigg. 5, A-1010 Wien
[4]Faculty of Psychology,
University of Vienna Liebigg. 5, A-1010 Wien
[5]Faculty of Computer Science,
University of Vienna Währinger Str. 29, 1090 Wien

Abstract

Edited (or cut) dynamical images are created by changing perspectives in imaging devices, such as videos, or graphical animations. They are abundant in everyday and working life. However little is known about how attention is steered with regard to this material. Here we propose a simple two-step architecture of gaze control for this situation. This model relies on (1) a down-weighting of repeated information contained in optic flow within takes (between cuts), and (2) an up-weighting of repeated information between takes (across cuts). This architecture is both parsimonious and realistic. We outline the evidence speaking for this architecture and also identify the outstanding questions.

Keywords: Attention, Eye Movements, Visual Motion, Video, Editing, Saliency.

4.1 Introduction

Our visual world is complex and rich in detail but the human mind has a finite cognitive capacity. This is one of the reasons why humans pick up only a fraction of the visual information from their environment. At each instance in time, humans select only some visual information for purposes such as in-depth recognition, action control, or later retrieval from memory, whereas other visual information is ignored in varying degrees. This fact is called selective visual attention.

One particularly widespread source of visual information is technical dynamic visual displays. These displays depict images of visual motion and are used in computers, mobile telephones, or diverse professional imaging devices (e.g., in devices for medical diagnosis). Importantly, the widespread use of

technical dynamic visual displays in human daily life during entertainment (e.g. video), communication (e.g. smart phones), and at work (e.g. computer screens) significantly adds to the visual complexity of our world. An accurate and ecologically valid model of human visual attention is essential for the optimization of technical visual displays, so that relevant information can be displayed in the place and at the right time in order to be effectively and reliably recognized by the user.

One important characteristic of videos and other technical motion images that contrasts with the dynamics of 3-D vision under more natural conditions is the fact that this material is highly edited (or cut). Videos consist of takes and cuts between takes. In this context, takes denote the phases of spatio-temporally continuous image sequences. By contrast, cuts are the spatio-temporal discontinuities by which two different takes (e.g., taken on different days, at different locations, or from different camera angles at the same location) can be temporally juxtaposed at the very same image location. Despite the fact that edited material conveys a substantial part of the visual information that competes for human selective attention, little is known about the way that attention operates in this situation. Specifically, attention research in this domain has almost exclusively focused on the impact of image motion per se [23, 30, 89], without paying too much attention to the very different cognitive requirements imposed by extracting information from takes versus cuts. Here, we propose a two-step model in response to this demand. In this model, within takes (between cuts) viewers would attend to novel information and would down-weigh repeated visual input.

In the following, we will develop our arguments for this model. We start with the simplest conceivable bottom-up model, and proceed by a brief discussion of top-down factors as one additional important factor. We then introduce our two-step model as a more realistic and yet parsimonious extension of existing bottom-up and top-down models. Next, we turn to review the evidence that is in line with our model. Finally, we conclude with a discussion of the outstanding questions.

4.2 Related work

To understand how selective visual attention works in humans, one can investigate gaze direction, visual search performance, and visual recognition. The relationship between these three measures will be explained next. To start with, we know that gaze direction is tightly linked to interest, attention, and recognition. Eye movements are an objective index of the direction of visual attention. This assumption is well supported by research on recognition during saccade programming [35].It is therefore not surprising that eye movements provide important cues to the personal intentions and interests of another person.When observing another individual, we use direction of fixation (when the eyes are still), of saccades (when the eyes move quickly from one location to another), or of smooth pursuit eye movements (when the eyes track a moving object in the environment) as a window into the other individual's mind.

Of course, gaze direction is not perfectly aligned with attention and does not always tell us what another person sees [107]. For this reason, in attention research, one cannot rely on fixation directions alone. If one wants to under-

stand, where attention is directed, one has to equally draw on conclusions from visual search and visual recognition performance [129].

The Bottom-Up Model

What is true of attention in general is also true of the so called bottom-up model of visual attention. The bottom-up model is supported by both visual search behavior [125] and eye-tracking [64], and its charms lie in its simplicity and parsimony. Bottom-up models rely on one simple principle: "the strength of the visual signal" to explain where humans direct their visual attention. These models disregard different human goals, interests, and other top-down influences, such as prior experiences of an individual, or also task- and situation-specific factors. Instead, bottom-up models define the principles of visual attention in simple objective terms and assume that the focus of attention is fully determined by the characteristics of the momentary visual stimuli in the environment [47, 63].

Beyond Bottom-Up Influences

Despite the evidence supporting the bottom-up model, this model is not satisfying because humans do not all look in a task-unspecific way at the same locations [128]. But how could individual goals influence visual attention? Top-down models explain this. They emphasize past experiences, goals, intentions, interpretations, and interests of the viewer as predictors of visual attention [128, 143]. Top-down principles can influence seeing and looking in two ways: They either boost the subjectively interesting image features or they deemphasize the subjectively uninteresting image features for the summed salience. Top-down models assign different weights to specific features [143] or locations [128]. Thus, top-down models are suited for accommodating the influence of subjective interests and goals. They can bridge the gap between model behavior and subjective influences for an improved prediction of eye movements and visual recognition into more realistic predictions of visual attention.

What is lacking so far is a convincing top-down model of visual attention for edited dynamically changing visual displays. Given the fact that humans spend much of their time viewing edited videos (on the Internet, television, or in the cinema), it is unfortunate that even the approaches that tried to model top-down influences mostly operated on static images without considering visual changes over time. Progress in this direction has been made in the form of a surprise-capture or novelty-preference model. Researchers observed that during watching of movies, human attention is captured by surprising or novel visual information [61]. In the surprise model, stimulus information that repeats over time is deemphasized as an attractor of attention. The surprise model is also parsimonious because it requires just one principle of visual memory of what has been seen in the recent past for an explanation of the creation of goal templates.

However, the surprise model is too rigid. It is incorrect to consider visual feature repetition as always being disadvantageous for the attraction of attention. Many experiments have shown that repeated features attract attention [16, 80].

4.3 Two-step model

We suggest that a two-step model of visual attention offers a realistic description of how attention is allocated in videos and other edited dynamic images (e.g. animated computer graphics, or even medical imaging devices). In the two-step model, surprise capture towards novel information and feature priming towards repeated visual information as the two major top-down principles driving visual attention will take turns as a function of one shared steering variable: the temporal coherence of the optic flow across subsequent images (see Figure 4.1). With the two-step model we thereby seek to overcome existing limitations of (1) bottom-up models that fail to account for inter-individual variability of visual attention, (2) too rigid forms of top-down models of attention that incorporate only one of the two top-down principles, and (3) models that fail to consider the specificities of edited dynamically changing visual images at all. This two-step model is based on empirical observations. It also allows deriving new testable hypotheses that can be investigated with the help of psychological experiments.

To start with, attraction of attention by repeated features (as in feature priming) conflicts with the finding of Itti and Baldi (2009) that repeated features do not attract attention. Attraction of attention by feature repetition can, however, be reconciled with the findings of Itti and Baldi by the two-step model. Itti and Baldi based their conclusions on gaze directions recorded during the viewing of edited video clips and video games. How this could have masked repetition priming across cuts can be understood if one takes into account the specificities of the high temporal resolution of the surprise model that was set to the level of single frames. For each frame, a prior and a new probability distribution were computed and their difference was tested for its potential to attract the eyes. This resulted in a higher number of model tests between cuts (or within takes) of the videos than model tests across cuts (or between takes), even in the highly edited video clips with relatively many cuts. Between-cuts events encompassed 30 frames/second because monitor frequency was set to 60.27 Hz (and assumed that videos were displayed in half frames). However, by definition, each across-cut event consisted of only two frames. Therefore, between-cuts events by far outweighed across-cut events in the test of the model of Itti and Baldi (2009).

Importantly, between cuts (or within takes), the correlation between successive feature or stimulus positions is high, whereas across cuts (or between takes) it is lower. To understand this, think firstly of an example of a take (i.e., a between-cuts event), such as the filming of a moving object in front of a static background. Here the background objects and locations are correlated for all frames of the take. In fact, they would be the same (see Figure 4.2, for a related example). Now secondly think of what happens across a cut (or between takes). Here, the correlation between successive features or stimulus positions must be lower, simply because of occasional cutting between takes of completely different scenes (or at least different camera angles within the same scene). With temporal juxtaposition of different scenes by a filmic cut, no stimulus contained in the take preceding the cut needs to be repeated after the cut. Basically, this low take-by-take correlation across cuts in videos is exactly what corresponds best to the conditions of the experiments demonstrating feature priming: In psychological experiments, a low correlation between positions and

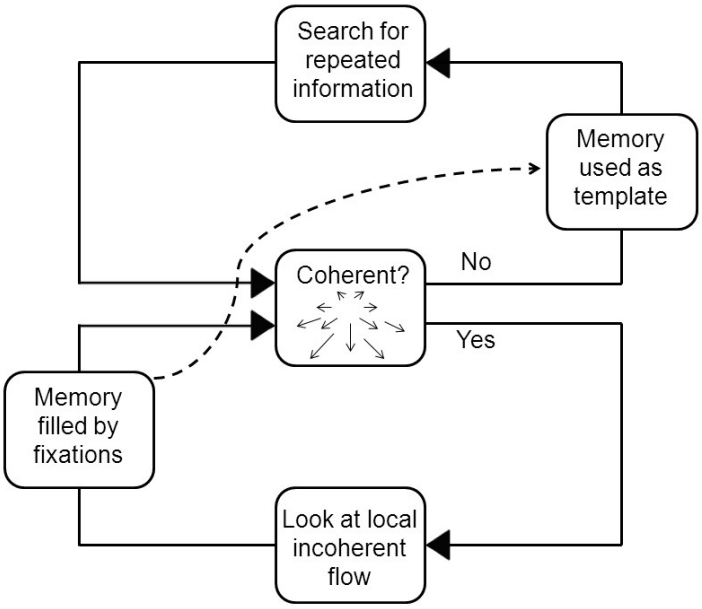

Figure 4.1: According to the model, within takes the human gaze is steered toward novel information. This mode is supported by the presence of temporally coherent global optic flow (see center of Figure) and an attraction to novelty is achieved by down-weighting global optic flow and up-weighting local incoherent flow for the selection of gaze directions because per definition, the information contained in the global flow field relates present to past information whereas local incoherencies form new features themselves and are diagnostic of the appearance of new objects in the visual field. The situation changes if a cut is encountered. Cuts are signaled by incoherencies of the global flow field. In this situation, the human gaze is steered towards repeated information. For further information, refer to the text.

even colors of relevant to-be searched-for target stimuli between trials has been the way to prevent anticipation of target positions and target features [80]). This low correlation corresponds much better to effects across cuts. Basically, in our two-step model we will therefore assume that across cuts, the surprise model of attention would be falsified and a feature priming model would be confirmed, whereas between cuts a preference for novel information holds.

The two-step model comprises three components: one spatially organized representation of the visual image as its input and two internal top-down representations of visual features. The input representation is the same as in standard bottom-up models [63]. The two alternative internal top-down representations of the two-step model are (1) search templates of scene- or take-specific object-feature matrices that a viewer can retrieve from visual memory, and (2) a track record of the temporal coherence of the optic flow within the image that the viewer applies online while watching a video.

The two-step model's visual memory contains representations of visual fea-

Figure 4.2: An image from a sequence of a man shutting the back of his car on the left and a schematic representation of the regions of the highest movement (in black) on the right. As compared to the coherent null vector of optic flow in the background, the optic flow of the moving man would be less coherent and, within a take, should capture human attention.

ture combinations (e.g. edge representations) for objects and for scenes (or takes). If such a representation is retrieved, this memory representation can be used as a template to up-weigh repeated feature combinations as relevant during visual search. This conception of a retrieved search template is very similar to that of other top-down models of attention [143, 147]. In contrast to past feature-search template models, however, as in the surprise model, in the two-step model the content of the visual memory will be empirically specified: What a particular person looks at is stored in visual memory [82]. The two-step model thus uses gaze direction for segmentation and stores objects as a vector of visual features at a fixated position, and each scene or take within a video as a matrix of the vectors of the looked-at objects within a take. Each take-specific matrix will be concluded when a minimum of the temporal coherence of optic flow indicates a change of the scene (see below), and matrices will be successively stored in the order of their storage until a capacity limit of visual working memory has been reached [78]. In this manner, the two-step model adapts to interindividual variation of looking preferences and keeps track of them, without having to make additional assumptions.

Related but operating on a different time scale, for the two-step model optic flow will be continually calculated as a mathematical function that connects one and the same individual features or objects at subsequent locations in space and time by one joint spatio-temporal transformation rule that is characteristic of the change of the larger part of the image for a minimal duration [100]. Moreover, the temporal coherence of the optic flow will be continuously tracked. We calculate the temporal coherence of optic flow as the similarity of the optic flow across time. In the two-step model, an increasing temporal coherence of optic flow will thus be reflected in a descending differential function. This coherence signal can be topographically represented in image coordinates and directly feeds into one visual filter down-weighing those image areas characterized by the temporally coherent optic flow. In this way, the two-step model instantiates the surprise-capture principle and filters out the repeated visual features proportional to the duration and area of uniform optic-flow (see Figure 4.2).

By contrast to this, the local minima of the coherence of optic flow (or the maxima of the differential function) are used as signals indicating cuts that

trigger the retrieval of a search template, and the resultant up-weighing of the repeated features of the image representation resembling the search template.

The two-step model is more realistic than the surprise model because it incorporates feature priming of attention, too. Yet, the two-step model is parsimonious because it couples the two top-down principles of attention to the same shared steering value of optic flow coherence, and, as in the surprise model, most of two-step models free parameters (the content of the visual memory) will not be arbitrarily chosen or have to be specified by task instructions as in standard top-down models (e.g., [92]) but will be specified on the basis of empirical observation (i.e., will be measured as the feature values at fixated positions).

4.4 Evidence

Weighting Coherent Optic Flow

The surprise-capture principle outperforms the bottom-up model when predicting fixations within animated video games and movies [61]. According to the two-step model, this surprise-capture effect reflects the suppression of coherent optic flow. Optic flow denotes the global commonality or unifying mathematical rule of the global visual motion signal across the image that is frequently due to the cameras (or the observers) self-motion. Optic flow is tied to visual feature repetition because across time, like other types of visual motion, too, optic flow reflects a track record of repeated features and objects found at different places.

In line with the assumed down-weighting of coherent optic flow, visual search for a stationary object is facilitated if it is presented in an optic flow field as compared to its presentation among randomly moving distractors [111]. Likewise, objects moving relative to the flow field pop out from the background [112]. A tendency to discard optic flow as a function of its coherence over time and space in dynamic visual scenes also accounts for many instances of attention towards human action in general [54] and human faces in particular [46]. In these situations, actions and facial movements are defined by local motion patterns that have regularities differing from the larger background's coherent flow field. Equally in line with and more instructive for the present hypothesis are the cases in which one motion singleton among coherently moving distractors captures human attention [2, 21]. To perform further analysis in this direction we are developing a decomposition procedure of the motion in dynamic image sequences. For example, on the web-page `http://www.csc.univie.ac.at/index.php?page=visualattention` a movie presenting the projection of a cube moving over an oscillating background can be viewed. The optical flow computed between the sixteenth and the seventeenth frame of the sequence is visualized in Figure 4.3. The motion can be decomposed in global movement of the cube depicted in $U1$ and in the background movement in $U2$.

Templates Combining Features

Selectively attending to the relevant visual features for directing the eyes and for visual recognition is one way by which humans select visual information in

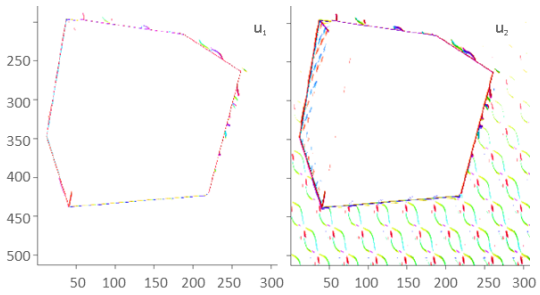

Figure 4.3: Flow visualization of a dynamic image sequence showing a projection of a cube moving over an oscillating background. $U1$ and $U2$ depict the global and the background movement respectively (`http://www.csc.univie.ac.at/index.php?page=visualattention`).

a top-down fashion [143]. For example, informing the participants prior to a computer experiment about the color of a relevant searched-for target helps the participants in setting up a goal template representation to find the relevantly colored target object and ignoring irrelevantly colored distractors (e.g., to find red berries in green foliage during foraging; [38]. Equally important and well established is the human ability to selectively look-for particular visual shapes or for specific combinations of shapes and colors [129]. In this way human viewers could also search for landmarks that they have seen in the past to re-orient after a cinematic cut, and to decide whether a visual scene continues or has changed.

In line with this assumption, participants learn to adjust the search templates to the visual search displays that they have seen in the past. During contextual cueing, for example, participants benefit from the repetition of specific search displays later in a visual search experiment [26]. Similar advantages have been demonstrated in the context of visual recognition under more natural conditions, with static photographs of natural scenes [82, 132].

In the study of [132], for example, participants first viewed a variety of photographs for later recognition of the learned photographs among novel pictures. Critically, during recognition participants only saw cutouts from scene images. Cutouts from the learned scenes were either from a previously fixated area(see Figure 4.4) or they were from an at least equally salient non-fixated area of the learned images(see Figure 4.5). In line with an active role of fixations for encoding and successful recognition, the participants only recognized cutouts that they fixated during learning. In contrast, the participants were unable to recognize cutouts showing areas that were not fixated during learning with better than chance accuracy (see Figure 4.6).

Reorienting After Cinematic Cuts

We consider re-orienting between subsequent visual images as one of the most fundamental tasks for the human viewer. Under ecological conditions, orienting is required in new environments, as well as when time has passed between successive explorations of known environments. During the viewing of edited videos, orienting is required to make sense of temporally juxtaposed images

Figure 4.4: Cutouts from old images were selected contingent upon the participants gaze pattern. Old/fixated cutouts showed the location of longest fixation. Old/control cutouts showed a nonfixated but highly salient location. Copyright by AVRO [132]

Figure 4.5: Cutouts from new images showed highly salient scene regions or were randomly chosen. Copyright by AVRO [132]

with a low correlation of objects or their locations. The latter situation is typical of all technical imaging devices for dynamically changing visual images. Think of video cuts in which the image before the cut does not have to bear any resemblance to the image after the cut.

In line with the assumed role of repetition priming on eye movements, Valuch et al. observed that participants preferentially looked at videos baring a high similarity of pre-cut and post-cut images [131]. These authors used two videos presented side by side and asked participants to keep their eyes on only one of the videos. Critically, during two kinds of cuts, the images could switch positions: cuts with a high pre- and post-cut feature similarity and cuts with a lower pre- and post-cut feature similarity. For example, participants were asked to look at a ski video and to quickly saccade to the ski video if the video switched from the left to the right side. In this situation, the participants showed a clear preference to look at the more similar images. Saccade latency was much lower in the similar than in the dissimilar condition (see Figure 4.7).

Repetition priming would also explain why participants fail to notice so-

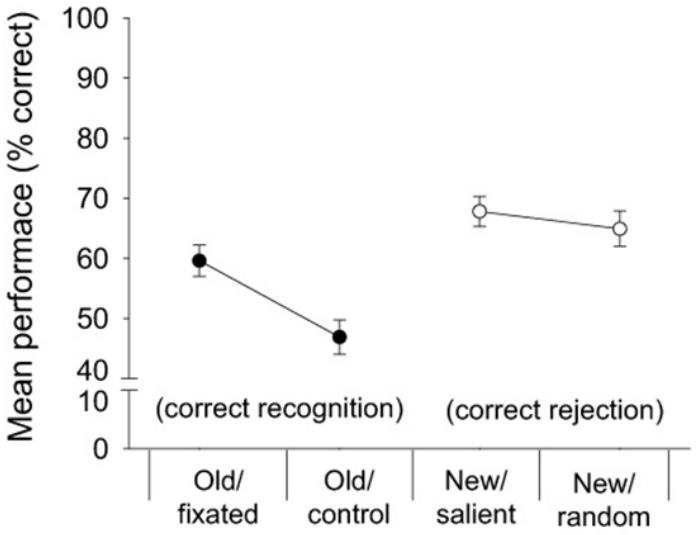

Figure 4.6: Rate of correct responses in percent as a function of cutout type in the transfer block. Copyright by AVRO [132]

called matching cuts. Participants fail to register matching cuts, such as cuts within actions (with an action starting before the cut and being continued after the cut), as compared to non-matching cuts from one scene to a different scene [121]. This is because with matching cuts, that is, cuts within the same scene, the overall changes in visual image features between two images are smaller than with cuts that connect two different scenes [34].

Repeated vs. Novel Features

When researchers rearranged an otherwise coherent take by cutting it and re-arranging the take into a new and incoherent temporal sequence, the reliability

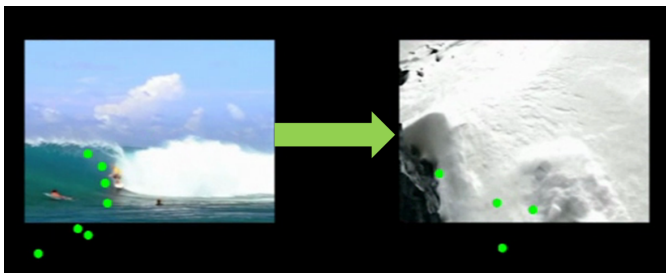

Figure 4.7: Two videos showing different content were presented side by side and participants were asked to follow only one content with their eyes. Videos could switch position during two kinds of cuts: (1) high and (2) low pre- and post-cut feature similarity cuts. The participants showed a clear preference to look at the more similar images.

of the gaze pattern was drastically reduced [138]. These authors argued that their participants kept track of objects within takes and reset this search tendency after a cut. This interpretation is in line with our view that participants apply different strategies within than between takes.

4.5 Open questions and potential Applications

Regarding our two-step architecture many open questions remain. Most critically, it is unclear whether the coherence of optic flow is indeed down-weighted for attention. To address this question, one would need to correlate the decomposition of optical flow [1] into bounded variation and an oscillating component with viewing behavior in natural images. One would also have to test whether changes for novel versus repeated features are characteristic of phases of low global coherence of the flow pattern.

Another open question concerns the impact of top-down search templates for features in natural scenes. This influence is relatively uncertain. Most of the evidence for the use of color during the top-down search for targets stems from laboratory experiments with monochromatic stimuli [29]. This is very different from the situation with more natural images, such as movies, where each color stimulus is polychromatic and consists of a spectrum of colors. In addition, a lot more questions than answers arise with regard to the storage and usage of different take-specific topdown templates.

Among the open questions, the potential applications of the model are maybe the most interesting ones. The model should be useful for improving the prediction of visual attention in more applied contexts, such as clinical diagnosis based on visual motion (e.g., in ultrasound imaging), QoE (quality of experience) assessment and videos coding in entertainment videos.

In medical imaging, much as with cuts, the optical flow of an image sequence can be interrupted by noise or by changes of perspective. For example in case of angiography, a new perspective of the vessels can be suddenly shown.

Also, due to a lack of contact between imaging devices and body (e.g., during ultrasound diagnosis), noise or blank screens can interrupt medical image sequences. These examples illustrate that the two-step model is applicable to medical imaging and that it captures a new angle on these problems. During medical imaging, pervasive eye-tracking could be used to extract visual feature vectors at the looked-at image positions. After an interruption of the imaging sequence, these vectors could then be convolved with post-interruption images for a highlighting of those regions baring the closest resemblance with the input extracted before the interruption. Likewise, in the area of video coding and compression, scene cuts represent an important challenge.

4.6 Conclusion and comparative evaluation

With a simple two-step model of down-weighting redundant information contained in optic flow versus up-weighting repeated information contained in two images divided by a cut, we proposed a framework for studying attention in edited dynamic images. This model is very parsimonious because it does not require many assumptions and it can be empirically falsified. In comparison to a bottom-up model the two-step model is able to accommodate inter-individual

viewing differences but has more free parameters and is therefore less economical. In comparison to existing top-down models the two-step model takes the particularities of visual dynamics into account and used empirical observations to specify many of its free parameters. Although this model nicely explains a variety of different findings, future studies need to address many outstanding questions concerning the model.

Chapter 5

The effect of cinematic cuts on human attention

Authors & Contributions The authors are Christian Valuch, Ulrich Ansorge, Shelley Buchinger, Aniello Raffaele Patrone and Otmar Scherzer. The development of this article was a collaborative process, and each of the authors made significant contributions to every aspect of the paper.

Publication Status Published [131]: C. Valuch, U. Ansorge, S. Buchinger, A. R. Patrone and O. Scherzer. The Effect of Cinematic Cuts on Human Attention. In P. Olivier, P. Wright and T. Bartindale editors, *TVX'14: Proceedings of the 2014 ACM international conference on Interactive experiences for TV and online video*, pages 119–122, New York, NY, USA. ACM.

The final publication is available at `http://www.acm.org/`.

The Effect of Cinematic Cuts on Human Attention

Christian Valuch[3], Ulrich Ansorge[4], Shelley Buchinger[5], Aniello Raffaele Patrone[1], and Otmar Scherzer[1,2]

[1]Computational Science Center, University of Vienna,
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[2]Radon Institute of Computational and Applied Mathematics,
Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria
[3]Cognitive Science Research Platform,
University of Vienna Liebigg. 5, A-1010 Wien
[4]Faculty of Psychology,
University of Vienna Liebigg. 5, A-1010 Wien
[5]Faculty of Computer Science,
University of Vienna Währinger Str. 29, 1090 Wien

Abstract

Understanding the factors that determine human attention in videos is important for many applications, such as user interface design in interactive television (iTV), continuity editing, or data compression techniques. In this article, we identify the demands that cinematic cuts impose on human attention. We hypothesize, test, and confirm that after cuts the viewers' attention is quickly attracted by repeated visual content. We conclude with a recommendation for future models of visual attention in videos and make suggestions how the present results could inspire designers of second screen iTV applications to optimise their interfaces with regard to a maximally smooth viewing experience.

Keywords: Attention, Eye Movements, Visual Motion, Video, Editing, Saliency.

5.1 Introduction

Designing successful applications for online and interactive television (iTV) requires a proper understanding of the factors that determine the user's experience. Working towards this objective, HCI research has been using eye tracking as a means of evaluating user interfaces [106]. For instance, in multiple screen applications [18] users frequently shift their gaze between at least two locations [56]. The presence of the second screen can distract the viewer from the main content of the show [18]. Understanding which factors determine the viewer's attention in such situations would allow designers to optimize their applications in favor of a maximally smooth viewing experience. To investigate these questions on a more general level – with a broad range of applications, such as video coding [113], or continuity editing [121] – we looked at gaze shifts

after cinematic cuts. Human attention is closely related to eye movements. *Saccades* -– abrupt gaze shifts between two locations — are a direct consequence of shifting attention to a new location [71]. Accordingly, by looking at the properties of saccades, it is possible to formulate and test theories about attention.

Current models of human attention and gaze behavior in videos emphasize the role of novelty, or *Bayesian surprise*. They assume that visual content that is maximally dissimilar from the viewer's prior visual experience is the best predictor of human attention and gaze direction. Indeed, eye tracking confirmed that human gaze direction in continuous videos is better explained by Bayesian surprise than by alternative models [61]. However, this is not necessarily true for cuts within edited videos. Existing evidence suggests that attention is attracted by repeated visual features in situations where location correlations between two successive images are low [20, 80].

Edited videos frequently contain hard cuts, i.e. visual discontinuities that require shifting attention from one location to another because object locations are uncorrelated across the cut. Moreover, making sense of narratives and content across cuts implicitly requires deciding whether the post-cut scene is a continuation of the pre-cut scene [121]. Here, *within scene cuts (WSCs)* continue with the same scene from a different angle; *between scenes cuts (BSCs)* continue with a different scene (see Figure 5.1). Orienting attention to repeated visual features could enable viewers' quick and efficient recognition of content that connects the cut images (in the case of WSCs).

The Present Study

We tested the hypothesis that after cuts, attention is more strongly attracted by repeated visual content than by novel, or surprising content. We conducted an eye tracking experiment, in which participants had to watch and keep their gaze on a video that was shown next to another, irrelevant video. Both videos contained hard cuts and unforeseeably kept or switched their locations at the cuts. This manipulation created a low correlation of object locations as is typical of cuts. Presenting two videos side by side also allowed us to measure influences of repeated versus novel content on saccades, during which attention and eye movements are tightly coupled [71]. If locations switched, participants had to saccade to the new location of the video they were instructed to follow (similar to shifting gaze between two screens).

We analyzed *saccadic reaction time (SRT)* as a measure of viewers' reorienting of attention to the post-cut scene after a location switch. Following our hypothesis, we predicted shorter SRT after cuts where much visual content was repeated (WSCs, or cuts with high image-image similarity) and longer SRT after cuts where less visual content was repeated (BSCs, or cuts with low image-image similarity).[1] In the remaining sections of this paper we give details on our method, results, and discuss implications for further research and improvements of iTV applications.

[1]This prediction is the opposite of that of the Bayesian surprise model which generally predicts a shorter SRT for less similar than for more similar image content.

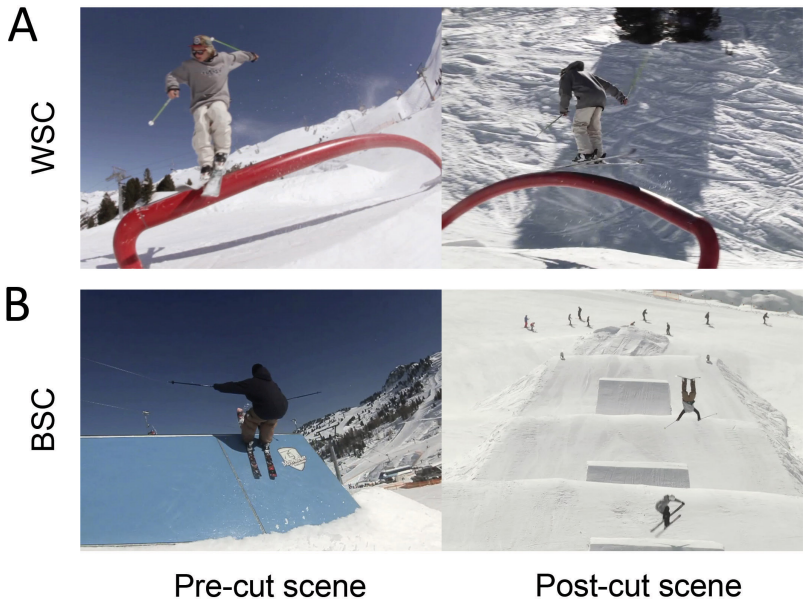

Figure 5.1: Example cuts. (A) Within scene cut (WSC). (B) Between scenes cut (BSC). Screenshots derived from videos by QParks.com, available under CC BY 3.0 at vimeo.com/89901459 and vimeo.com/89248621.

5.2 Method

Participants

Forty-two students (34 female) with a mean age of 23 years took part in an eye tracking experiment. Informed consent was obtained from all participants.

Stimuli

We used 20 sports videos in which we deliberately inserted new cuts. Each video showed the same sport throughout (e.g., *skiing*). Videos were edited in pairs, resulting in ten sets of two videos. The sport in the first video was always different from the sport in the second video (e.g., *skiing* vs. *surfing*). Cuts always occurred simultaneously in both videos. Average video duration was 2.5 minutes and the complete set contained 212 cuts. Cuts were assigned to either a WSC or a BSC condition. Whenever major visual changes, e.g. in scenery, actors, or ongoing actions occurred with the cut, the cut was coded as a BSC. In contrast, cuts that connected two images showing the same scene, action, and actors were coded as WSCs. Figure 5.1 shows examples. We assumed that more visual content is repeated after WSCs than after BSCs.

To validate this, we compared the similarity of color histograms of the last pre-cut and the first post-cut frame and, based on this measure, assigned each cut to a *High similarity* or a *Low similarity* condition. We used color similarity because color contributes to gaze and attention preferences for repeated information [80], allows visual recognition after location and/or perspective shifts

[123], and conveys information useful for cut detection [49].

Apparatus

Gaze data were recorded using an EyeLink 1000 Desktop Mount eye tracker (SR Research Ltd., Kanata, Ontario, Canada) at a sampling rate of 1000 Hz. The eye tracker was calibrated to each viewer's dominant eye using a 5-point calibration. Every time the videos switched locations, the exact timestamp was saved to the eye tracking data file, which allowed analyzing the latency of the first saccade to the target video with millisecond precision. Stimuli were displayed on a 19-in. color CRT monitor (Sony Multiscan G400) with a resolution of 1,280 × 1,024 pixels and a refresh rate of 60 Hz. The experimental procedure was implemented in MATLAB (MathWorks, Natick, MA, USA) using the Psychophysics Toolbox [25, 103] and the Eyelink toolbox [31]. Viewing distance to the monitor was 72 cm supported by chin and forehead rests. The viewable screen area subtended 28° × 21°. The apparent size of the 400 × 300 pixel videos was 8.75° × 6.15° and they were shown vertically centered at a horizontal eccentricity of 6.56°.

Procedure and Design

The experiment consisted of 20 blocks in which two videos were presented on the screen. Importantly, participants were instructed to view only one of the videos (the target video) while ignoring the other (the distractor video). At the beginning of each block, the starting location of the target video was announced by a green rectangle. Participants were informed that the videos switched locations at random intervals, and instructed to relocate their gaze as fast as possible to the target video's new location once the videos had switched locations. Throughout the experiment, each block was presented twice so that either of the videos was serving as the target in the first and as the distractor in the second half of the experiment (or vice versa).

Data Analysis

Saccades were identified as sample periods where the change in gaze direction was larger than 0.1°, eye movement velocity exceeded 30°/s, and acceleration exceeded 8000°/s^2. The main dependent variable was SRT, defined as the latency of the first saccade towards the target video after the videos switched locations. SRT was analyzed as a function of the type of cut (WSC vs. BSC) in the target video, and the similarity of RGB color histograms across cuts (High similarity vs. Low similarity) – for further details see Stimuli and Results. We expected shorter SRTs (faster gaze relocation) after WSCs than BSCs. Similarly, we expected shorter SRTs after High similarity than after Low similarity cuts.

Gaze data were preprocessed in MATLAB and statistical tests were run in R [124]. Out of 8,904 collected data points (i.e. 212 cuts for each of the 42 participants), 8,397 (94.3 %) contained valid SRTs and were subjected to statistical analyses. Data were excluded if no saccade to the target video was identified within a time-window of 3 s after the location switch or if gaze was already at the new location shortly ahead of the switch. Figure 5.2 depicts the

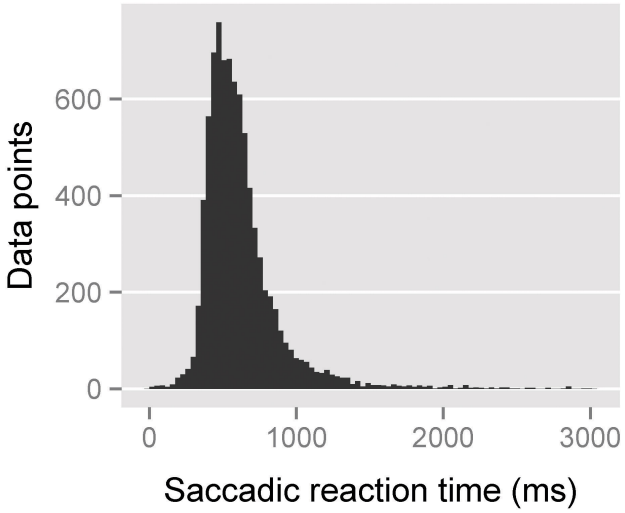

Figure 5.2: Distribution of valid saccadic reaction times (i.e. the latencies of the first saccade to the target video after a location switch).

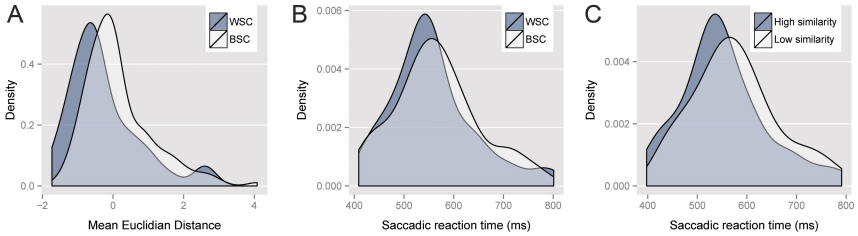

Figure 5.3: Results. (A) Distribution of mean Euclidian distances (z-transformed) of RGB color histograms of the last pre-cut and the first post-cut frame as a function of cut category. Values below 0 represent higher similarity, values above 0 represent lower similarity. (B) Distribution of individual median SRT as a function of cut category. (C) Distribution of individual median SRT as a function of color histogram similarity across the cut.

distribution of valid SRTs. Individual median SRTs per condition were tested for within-participant differences by t-tests. We report Pearson correlation coefficients as measures of effect sizes. For all statistical tests, we set α at 0.05.

5.3 Results

Image Similarity Across Cuts

To validate that more visual content is repeated after WSCs than BSCs, we calculated the mean Euclidian distance of the RGB color histograms between the final pre-cut and the first post-cut frame. For better interpretability, we z-transformed these values, so that values below 0 represent higher similarity (indicated by the smaller Euclidian distance), and values above 0 represent

lower similarity (indicated by the greater Euclidian distance). A Welch two sample t-test indicated significantly higher color similarity in WSCs than BSCs, $t(148.3) = 2.86$, $p<.01$, $r = .23$ (see also Figure 5.3A).

Saccadic Reaction Time After Location Switches

In a first analysis, we tested whether the a priori categories of WSCs and BSCs could explain any variance in SRTs. Using a paired t-test, we found that median SRT was on average 9 ms shorter in WSCs than BSCs, $t(41) = -2.03$, $p<.05$, resulting in a medium-sized effect of $r = .30$ (see also Figure 5.3B).

For a second analysis, we categorized the cuts into either a *High similarity* or a *Low similarity* condition, depending on whether the z-transformed similarity measure for these cuts was below or above 0. Again, we tested for significant differences in SRTs between these conditions. A paired t-test of median SRTs confirmed that on average SRTs were 23 ms shorter after High similarity than after Low similarity cuts, $t(41) = -6.83$, $p<.001$, representing a large effect of $r = .73$ (see also Figure 5.3C).

5.4 Discussion

Our data suggest that after cuts viewers are able to re-orient their attention more quickly if visual content is repeated from the pre-cut scene: Following WSCs or High similarity cuts, saccades to the target video were initiated significantly faster than after BSCs, or Low similarity cuts. Results confirmed viewers' preference for repeated features during reorienting after cuts with low object-position correlations. The following limitations apply.

First, our results seem to conflict with the assumption that novel or surprising information is the best predictor of attention and gaze direction in videos [61]. However, we argue that an advantage for repeated information characterizes only a short time frame following cuts. During this period, viewers search for familiar visual content for deciding whether the previous scene continues, or not. Soon after, a preference for novel or surprising information should take over but future models should account for the effect of cuts on attention, too.

Second, in an effort to precisely measure the speed of attentional orienting after cuts we presented two videos simultaneously. This enabled us to elicit and record saccades of comparable start/end points for each cut. This is good because saccades are valid reflections of attention. However, the surprise model was supported during viewing of single videos. Viewing single videos is a situation that we ultimately also want to understand. Therefore, future research should aim to replicate our findings under single video viewing conditions.

Third, motivated by previous research [20, 80, 49, 123], we validated the stronger repetition of visual content across WSCs as compared to BSCs based on color similarity only. However, other descriptors that do not rely on color might also sufficiently explain the observed differences in SRTs. Also, we are unable to isolate color-repetition effects operating on a short timescale from the viewers' long-term knowledge about object-associated colors that possibly contributed to the color repetition effect (e.g., the knowledge that *snow* is *white*). These questions are open to debate and should be studied in future experiments, possibly by including control conditions with black and white videos.

Implications for iTV Applications

To conclude, we think a preference for repeated visual content applies in all situations in which the location of objects is uncorrelated across successive views. This is relevant for improving user interfaces in iTV. To give just one example, with second screen applications a second screen showing information that is visually unrelated to the main screen might distract the viewer [18]. Following from the present study, we would recommend that designers of second screen applications should include visual elements that repeat across both screens to minimize the time necessary for shifting attention between the two screens and assure a maximally smooth user experience. Even more interesting applications could become possible once eye tracking becomes widely available in consumer electronics. Then, it will be possible to dynamically adapt the content on a second device based on what was just looked at on the primary screen. Finally, we would like to stress that the methods presented in this paper can be easily adapted to study the effects of particular second screen iTV applications on human attention.

5.5 Conclusion

Our paper presents evidence that after cinematic cuts viewers quickly re-orient their attention to visual content that is repeated from the pre-cut scene. A preference for repeated visual content after cuts should be incorporated into models of human attention which currently assume that novelty or Bayesian surprise is the best predictor of human attention and gaze direction in videos. We also discussed implications of our results for the improvement of iTV applications.

Acknowledgments

We thank the reviewers for their excellent comments on a previous version of the paper, as well as Melanie Szoldatics and Heide Maria Weißenböck for assistance with data collection, and gratefully acknowledge grant CS11-009 from the Wiener Wissenschafts-, Forschungs-, und Technologiefonds (WWTF, Vienna Science and Technology Fund) to Ulrich Ansorge, Shelley Buchinger, and Otmar Scherzer. We also wish to express our special gratitude to Dr. Anton Luger.

Chapter 6

On a spatial-temporal decomposition of optical flow

Authors & Contributions The authors are Aniello Raffaele Patrone and Otmar Scherzer. The development of this article was a collaborative process, and each of the authors made significant contributions to every aspect of the paper.

Publication Status Submitted to *Inverse Probl. Imaging* in 2nd round of revision. Date of acknowledgement of receipt: June, 22, 2016. Preprint available in [101].

On a spatial-Temporal Decomposition of Optical Flow

Aniello Raffaele Patrone[1], Otmar Scherzer[1,2]

[1]Computational Science Center, University of Vienna,
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[2]Radon Institute of Computational and Applied Mathematics,
Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria

Abstract

In this paper we present a spatial-temporal variational decomposition algorithm for computation of the optical flow of a dynamic image sequence. We consider several applications, such as the extraction of temporal motion patterns of different scales and motion detection in dynamic sequences under varying illumination conditions, such as they appear for instance in psychological flickering experiments. For the numerical implementation we are solving an *integro-differential* equation by a fixed point iteration. For comparison purposes we use the standard time dependent optical flow algorithm, which in contrast to our method, constitutes in solving a spatial-temporal *differential* equation.

Keywords: Optical Flow, Decomposition, Oscillating patterns.

6.1 Introduction

Analyzing the motion in a dynamic sequence is of interest in many fields of applications, like human computer interaction, medical imaging, psychology, to mention but a few. In this paper we study the extraction of motion in dynamic sequences by means of the optical flow, which is the apparent motion of objects, surfaces, and edges in a dynamic visual scene caused by the relative motion between an observer and the scene. There have been proposed several computational approaches for optical flow computations in the literature. In this paper we emphasize on variational methods. In this research area the first method is due to Horn & Schunck [57]. Like many alternatives and generalizations, this methods calculates the flow from two consecutive frames. Here, we are calculating the optical flow from all frames simultaneously. Spatial-temporal optical flow methods were previously studied by Weickert & Schnörr [141, 142], [24], [137] and [6], to name but a few. However, in contrast to these paper we emphasize on the *decomposition* of the optical flow into appropriate components.

Variational modeling of patterns in *stationary* images has been initialized with the seminal book of Y. Meyer [87]. In the context of total variation

regularization, reconstructions of patterns was studied first in [134]. Here, we are implementing similar ideas as have been used before for variational image denoising [8, 11, 13, 10, 12, 39, 135] and optical flow decomposition [1, 70, 144, 146, 145]. However, a conceptual difference is that we aim for extracting *temporal* patterns, and in all the mentioned papers the decomposition was with respect to the space component. We emphasize that the proposed method is one of very few variational optical flow algorithms in a space-time regime. Within this class, this algorithm is the only spatial-temporal *decomposition* algorithm.

The outline of this paper is as follows: In Section 6.2 we review the optical flow equation. In Section 6.3 we present analytical examples of the optical flow equation in case of illumination disturbances. In Section 6.4 we introduce the new model for spatial-temporal optical flow decomposition. We formulate it as a minimization problem and derive the optimality conditions for a minimizer. In Section 6.5 we make calculations, which help to understand the decomposition algorithm from an analytical point of view. In Section 6.6 we derive a fixed point algorithm for numerical minimization of the energy functional. Finally in Section 6.7 and Section 6.8 we present experiments, results and a discussion of them.

6.2 Registration and optical flow

The problem of aligning dynamic sequences $f(\cdot, t)$ can be formulated as the operator equation, of finding a flow Ψ of diffeomorphisms,

$$\Psi(\cdot, t) : \Omega \to \Omega, \quad \forall t \in [0, T],$$

such that

$$f(\Psi(\vec{x}, t), t) = f(\vec{x}, 0), \qquad \forall \vec{x} \in \Omega \text{ and } t \in [0, T]. \tag{6.1}$$

For natural images, in general, it is not possible to solve equation (6.1) subject to the constraint that Ψ is a diffeomorphism for every t, because of occlusions, illumination changes, noise, and information gain/loss in the movie over time. Thus the optical flow and image registration community typically do not consider this constraint, in contrast to the shape registration community (see for instance [19, 65]).

Differentiation of (6.1) with respect to t for a fixed $\vec{x}$ gives

$$\nabla f(\Psi(\vec{x}, t), t) \cdot \partial_t \Psi(\vec{x}, t) + \partial_t f(\Psi(\vec{x}, t), t) = 0, \qquad \forall \vec{x} \in \Omega \text{ and } t \in [0, T]. \tag{6.2}$$

Switching from a Lagrangian to an Eulerian description allows to define the *optical flow equation (OFE)* on Ω:

$$\nabla f(\vec{x}, t) \cdot \vec{u}(\vec{x}, t) + \partial_t f(\vec{x}, t) = 0, \qquad \forall \vec{x} \in \Omega \text{ and } t \in [0, T]. \tag{6.3}$$

In particular this means that (6.3) is not well-motivated on subsets of Ω which are not met by a characteristic curve in space and time starting from $t = 0$. This problem is less relevant if the optical flow equation is considered for just two consecutive frames, which is the standard optical flow approach in the literature. In this case, a characteristic originates always at some point Ω through the re-initialization at each pair of frames. Instead of solving (6.3)

usually the relaxed problem is considered, which consists in minimizing the functional

$$S(\vec{u}) := \int_\Omega (\nabla f(\vec{x},t)\cdot\vec{u}(\vec{x},t) + \partial_t f(\vec{x},t))^2\,\mathrm{d}\vec{x} \to \min,, \qquad \forall t\in[0,T] \tag{6.4}$$

subject to appropriate constraints.

6.3 The optical flow equation in case of illumination disturbances

In this section we are showing simple motivating examples explaining properties of the solution of the optical flow equation (6.3) under changing illumination conditions.

Example 1. *We consider the 1D optical flow equation, to solve for u in*

$$\partial_x f(x,t)u(x,t) + \partial_t f(x,t) = 0 \text{ in } (0,1)\times(0,1) \tag{6.5}$$

for the specific data

$$f(x,t) = \tilde{f}(x)g(t) \text{ for } (x,t)\in(0,1)\times(0,1). \tag{6.6}$$

f represents a dynamic sequence with brightness variation, g over time. We are more specific and take:

$$\tilde{f}(x) = x(1-x) \text{ and } g(t) = 1-t. \tag{6.7}$$

The function f and the level lines are plotted in Figure 6.1 and the optical flow can be explicitly calculated:

$$u(x,t) = \frac{x(1-x)}{1-2x}\frac{1}{1-t}$$

indicates a transport of intensities from outside to the center $1/2$. We observe that $u(1/2,t)$ and $u(x,1)$ are singularities of the optical flow. From the definition of u it follows that

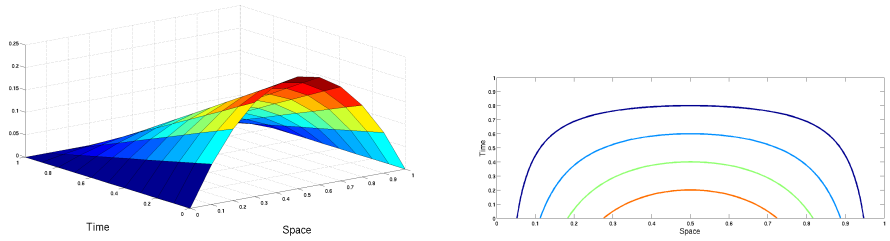

Figure 6.1: $f(x,t) = x(1-x)(1-t)$ from (6.7). Level lines of f are parametrized by $(\Psi(x,t),t)$.

$$\hat{u}(x,t) := \int_0^t u(x,\tau)\,d\tau = -\frac{x(1-x)}{1-2x}\log(1-t),$$

and thus

$$\begin{aligned}\|\hat{u}\|^2_{L^2((0,1)^2)} =& \int_0^1 \frac{x^2(1-x)^2}{(1-2x)^2}\,dx \int_0^1 \log^2(1-t)\,dt \\ =& 2\int_0^1 \frac{x^2(1-x)^2}{(1-2x)^2}\,dx = \infty,\end{aligned}$$

or in other words $u \notin L^2((0,1)^2)$. *In contrast, constraining the equation to a compact domain* C *of* $(0,1/2)$ *in space* $u \in L^2(C \times (0,1))$.

The deformation Ψ *of the non-linear optical flow equation is given by*

$$\Psi(x,t) = \frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{x(1-x)}{1-t}} \text{ for } t \leq 4\left(x - \frac{1}{2}\right)^2,$$

where the branch of Ψ *with* $+$ *is active if* $x > 1/2$ *and the other branch holds for* $x < 1/2$. *Moreover, we have*

$$\partial_t \Psi(x,t) = \mp \frac{x(1-x)}{\sqrt{1-t-4x(1-x)}} \frac{1}{(1-t)^{3/2}}.$$

This shows that the flow has a singularity (endpoint) at $t = 1 - 4x(1-x)$.

In Figure 6.2 there are shown u *and* $\partial_t \Psi$. *Along characteristics (initiated at* $t = 0$*) the optical flow equation produces the same results as the registration equation. However, note that* $\Psi(x,t)$ *satisfies the equation*

$$\partial_x f(\Psi(x,t),t)\partial_t \Psi(x,t) + \partial_t f(\Psi(x,t),t) = 0,$$

which in comparison to (6.3) *evaluates* f *at space locations* $\Psi(x,t)$ *instead of* x *and* $\partial_t \Psi$ *and* u *have to be compared at different space positions in the domain, which are covered by characteristics.*

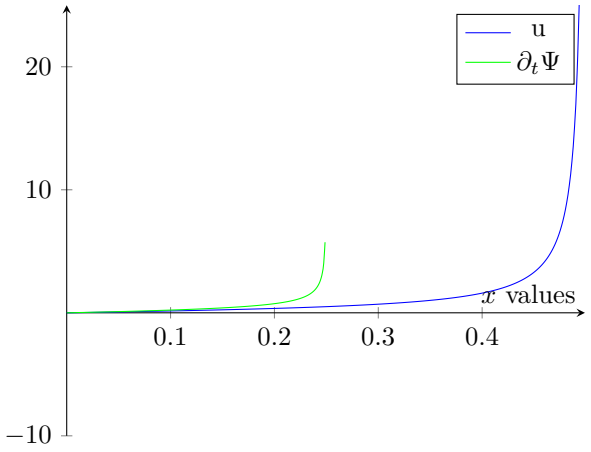

Figure 6.2: Linear versus non-linear optical flow: u and $\partial_t \Psi$ at $t = 1/4$. Note that $\partial_t \Psi$ is only defined in the interval $[0, 1/4]$, which is plotted, while u is defined for $[0, 1/2)$.

We expect that a collapse of characteristics manifest itself in less smoothness of the optical flow. In fact such situations appear in practical situations when occlusions are recorded.

Example 2. *We consider input data f of the form* (6.6) *with*

$$\tilde{f}(x) = x(1-x) \text{ and } g(t) = \exp\left\{-\frac{1}{\beta}(1-t)^\beta\right\} \text{ with some } 0 < \beta < 1 \quad (6.8)$$

for $(x,t) \in \hat{\Omega} := (0, 1/4) \times (0,1)$. The optical flow is given by

$$u(x,t) = -\frac{x(1-x)}{1-2x}(1-t)^{\beta-1}.$$

Integrating this function over time gives

$$\hat{u}(x,t) := \int_0^t u(x,\tau)\, d\tau = \frac{x(1-x)}{1-2x}\frac{1}{\beta}((1-t)^\beta - 1),$$

and consequently with

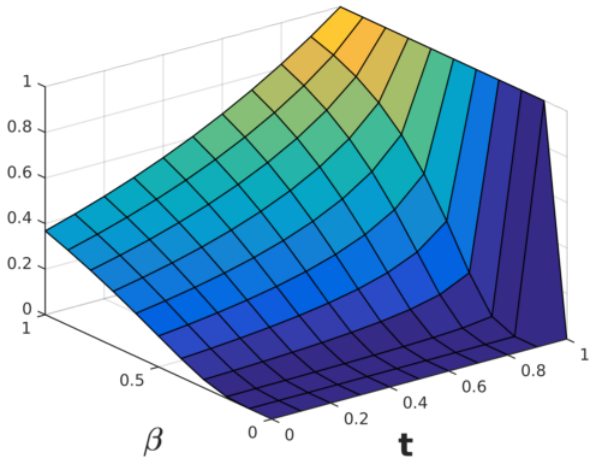

Figure 6.3: $g(t) = \exp\left\{-\frac{1}{\beta}(1-t)^\beta\right\}$

$$C := \frac{1}{\beta^2}\int_0^{1/4} \frac{x^2(1-x)^2}{(1-2x)^2}\, dx < \infty\,,$$

we get

$$\|u\|^2_{L^2(\hat{\Omega})} = C\int_0^1 t^{2\beta-2}\, dt = \begin{cases} C\frac{1}{2\beta-1} & \text{if } \beta > \frac{1}{2}\,, \\ \infty & \text{else}\,. \end{cases}$$

$$\|\hat{u}\|^2_{L^2(\hat{\Omega})} = C\int_0^1 t^{2\beta} - 2t^\beta + 1\, dt = \begin{cases} C\left(\frac{1}{2\beta+1} - \frac{2}{\beta+1} + 1\right) & \text{if } \beta > -\frac{1}{2}\,, \\ \infty & \text{else}\,. \end{cases}$$

This shows that for $0 < \beta < 1/2$ $u \notin L^2(\hat{\Omega})$ but $\hat{u} \in L^2(\hat{\Omega})$.

The bottom line of these examples is that illumination changes, such as flickering, may result in singularities of the optical flow and a violation of standard smoothness assumptions of the optical flow field. The potential appearance of the singularities motivates us to consider regularization terms for optical flow computations, which allow for singularities over time, such as negative Sobolev norms or G-norms.

6.4 Optical flow decomposition: basic setup and formalism

In this paper we derive an optical flow model for decomposing the flow field into spatial and temporal components. We consider the frames defined on a two-dimensional domain and in order to minimize the number of constants we take in (6.3) the time-interval $(0,1)$ and $\Omega = (0,1)^2$. We assume that the optical flow field to be a compound of two flow field components

$$\vec{u}(\vec{x},t) = \vec{u}^{(1)}(\vec{x},t) + \vec{u}^{(2)}(\vec{x},t) = \begin{pmatrix} u_1^{(1)}(\vec{x},t) \\ u_2^{(1)}(\vec{x},t) \end{pmatrix} + \begin{pmatrix} u_1^{(2)}(\vec{x},t) \\ u_2^{(2)}(\vec{x},t) \end{pmatrix}.$$

Because there appears a series of indices and variables it is convenient to specify the notation first:

$\vec{x} = (x_1, x_2)$	Euclidean space
$\partial_k = \frac{\partial}{\partial x_k}$	Differentiation with respect to spatial variable x_k
$\partial_t = \frac{\partial}{\partial t}$	Differentiation with respect to time
$\nabla = (\partial_1, \partial_2)^T$	Gradient in space
$\nabla_3 = (\partial_1, \partial_2, \partial_t)^T$	Gradient in space and time
$\nabla\cdot = \partial_1 + \partial_2$	Divergence in space
$\nabla_3\cdot = \partial_1 + \partial_2 + \partial_t$	Divergence in space and time
$\vec{n}$	normal vector to Ω
f	input sequence
$f(\cdot,t)$	movie frame
$\vec{u}^{(i)}$	optical flow module, $i = 1,2$
$\vec{u} = \vec{u}^{(1)} + \vec{u}^{(2)}$	optical flow
$u_j^{(i)}$	j-th optical flow component of the i-th module
$\Psi^{(i)}$	components of deformation,
$\Psi = \Psi^{(1)} + \Psi^{(2)}$	total deformation
$\widehat{u}(\cdot,t) = \int_0^t u(\cdot,\tau)\,d\tau$	Primitive of u
$\widehat{\widehat{u}}(\cdot,t) = -\int_t^1 \widehat{u}(\cdot,\tau)\,d\tau$	2nd Primitive of u - note $\partial_t\widehat{\widehat{u}}(\cdot,t) = \widehat{u}(\cdot,t)$

The OFE-equation (6.3) contains four unknown (real valued) functions $u_j^{(i)}$, $i,j = 1,2$, and thus is highly under-determined. To overcome the lack of equations, the problem is formulated as a constrained optimization problem, to determine, for some fixed $\alpha > 0$,

$$\operatorname{argmin}\left(\mathcal{R}^{(1)}(\vec{u}^{(1)}) + \alpha\mathcal{R}^{(2)}(\vec{u}^{(2)})\right) \tag{6.9}$$

subject to (6.3). Here $\mathcal{R}^{(i)}$, $i = 1,2$ are convex, non-negative functionals, such that $\mathcal{R}^{(1)} + \alpha\mathcal{R}^{(2)}$ is strictly convex. Instead of solving the constrained optimization problem, we use a soft variant and minimize the unconstrained regularization functional:

$$\begin{aligned} \mathcal{F}(\vec{u}^{(1)},\vec{u}^{(2)}) &:= \mathcal{E}(\vec{u}^{(1)},\vec{u}^{(2)}) + \sum_{i=1}^{2}\alpha^{(i)}\mathcal{R}^{(i)}(\vec{u}^{(i)}), \\ \mathcal{E}(\vec{u}^{(1)},\vec{u}^{(2)}) &:= \int_{\Omega\times(0,1)} (\nabla f\cdot(\vec{u}^{(1)} + \vec{u}^{(2)}) + \partial_t f)^2\,\mathrm{d}\vec{x}\mathrm{d}t \text{ with } \alpha = \frac{\alpha^{(2)}}{\alpha^{(1)}}. \end{aligned} \tag{6.10}$$

In the following we design the regularizers $\mathcal{R}^{(i)}$. Moreover, for the sake of simplicity of presentation, we omit the space and time arguments of the functions $u_j^{(i)}$ and f, whenever it simplifies the formulas without possible misinterpretations.

- For $\mathcal{R}^{(1)}$ we use a common spatial-temporal regularization functional for optical flow regularization (see for instance [142]):

$$\mathcal{R}^{(1)}(\vec{u}^{(1)}) := \int\limits_{\Omega\times(0,1)} \nu\left(\left|\nabla_3 u_1^{(1)}\right|^2 + \left|\nabla_3 u_2^{(1)}\right|^2\right) \mathrm{d}\vec{x}\mathrm{d}\tau, \tag{6.11}$$

  where $\nu : [0,\infty) \to [0,\infty)$ is a monotonically increasing, differentiable function. For the choice of ν we follow [142] and take

$$\nu(r) = \epsilon r + (1-\epsilon)\lambda^2\sqrt{1+\frac{r}{\lambda^2}}\,, \forall r \in [0,\infty)\,, \tag{6.12}$$

  with $0 < \epsilon \ll 1$ and $\lambda > 0$. The function $r \to \nu \circ (r \to r^2)$ is convex in r and there exist constants $c_1, c_2 > 0$ with $c_1 r^2 \leq \nu(r) \leq c_2 r^2$ for all $r \in \mathbb{R}$. Moreover, we denote by ν' the derivative of ν.

- $\mathcal{R}^{(2)}$ is designed to penalize for variations of the second component in time. Motivated by Y. Meyer's book [87], we introduce a regularization term, which is non-local in *time*. We have seen in Example 1 that u may violate L^2-smoothness in case of changing illumination conditions. Variations of Meyer's G-norm where used in energy functionals for calculating *spatial* decompositions of the optical flow [1, 69]. It is a challenge to compute the G-norm efficiently, and therefore workarounds have been proposed. For instance [134] proposed as an alternative at the G-norm the following realization for the H^{-1} norm: For a generalized function $u : (0,1) \to \mathbb{R}$, they defined

$$\|u\|_{H^{-1}}^2 = -\int_0^1 u(t)\partial_{tt}^{-1}u(t)dt\,.$$

  Here, we use this workaround for a realization for the *temporal* H^{-1}-norm, which we use as a regularization functional:

$$\mathcal{R}^{(2)}(\vec{u}^{(2)}) := \int\limits_{\Omega\times(0,1)} \left|\int_0^t \vec{u}^{(2)}(\vec{x},\tau)\mathrm{d}\tau\right|^2 \mathrm{d}\vec{x}\mathrm{d}t = \sum_{j=1}^{2}\int\limits_{\Omega\times(0,1)} \left(\widehat{u}_j^{(2)}(\vec{x},t)\right)^2 \mathrm{d}\vec{x}\mathrm{d}t. \tag{6.13}$$

  To see the analogy with the $\|\cdot\|_{H^{-1}}$-norm from [134] we note that the second primitive of the optical flow component $\vec{u}^{(2)}$, satisfies for $j = 1, 2$ and $\vec{x} \in \Omega$

$$\widehat{\widehat{u}}_j^{(2)}(\vec{x},1) = 0, \quad \partial_t\widehat{\widehat{u}}_j^{(2)}(\vec{x},0) = \widehat{u}_j^{(2)}(\vec{x},0) = 0. \tag{6.14}$$

  Then, by integration by parts it follows that

$$-\int_0^1 \underbrace{\widehat{\widehat{u}}_j^{(2)}(t)}_{=\partial_{tt}^{-1}u_j^{(2)}} u_j^{(2)}(t)\,dt = \int_0^1 \left(\widehat{u}_j^{(2)}(t)\right)^2 dt$$

and therefore

$$\mathcal{R}^{(2)}(\vec{u}^{(2)}) = \sum_{j=1}^{2} \int_{\Omega} \left\| u_j^{(2)}(\vec{x},\cdot) \right\|_{H^{-1}}^2 \mathrm{d}\vec{x}. \tag{6.15}$$

Energy functional and minimization

We are determining the optimality conditions for minimizers of $\mathcal{F}$ introduced in (6.10). Necessary conditions for a minimizer are that the directional derivatives vanish for all 2-dimensional vector valued functions $\vec{h}^{(j)} : \Omega \times (0,1) \to \mathbb{R}^2$, $j = 1, 2$. To formulate these conditions we use the simplifying notation:

$(\mathcal{E}, \mathcal{F}) := (\mathcal{E}, \mathcal{F})(\vec{u}^{(1)}, \vec{u}^{(2)}), \mathcal{R}^{(i)} := \mathcal{R}^{(i)}(\vec{u}^{(i)})$ and $\mathrm{res} = \nabla f \cdot (\vec{u}^{(1)} + \vec{u}^{(2)}) + \partial_t f$.

Therefore the directional derivative of $\mathcal{F}$ in direction $\vec{u}^{(j)}$ is given by:

$$(\partial_{\vec{u}^{(j)}} \mathcal{F})\vec{h}^{(j)} = \lim_{s \to 0} \frac{\mathcal{F}(\vec{u}^{(1)} + s\delta_{1j}\vec{h}^{(1)}, \vec{u}^{(2)} + s\delta_{2j}\vec{h}^{(2)}) - \mathcal{F}}{s} = 0$$

where $\delta_{ij} = 1$ for $i = j$ and zero else is the Kronecker symbol. The gradient of the functional $\mathcal{F}$ from (6.10) can be determined by calculating the directional derivatives of $\mathcal{E}$ and $\mathcal{R}^{(i)}$, separately.

- The directional derivative of $\mathcal{E}$ in direction $\vec{h}^{(j)}$ is given by

$$(\partial_{\vec{u}^{(j)}} \mathcal{E})\vec{h}^{(j)} = 2 \int\limits_{\Omega \times (0,1)} \mathrm{res} \nabla f \cdot \vec{h}^{(j)} \, \mathrm{d}\vec{x}\mathrm{d}t. \tag{6.16}$$

- The directional derivative of $\mathcal{R}^{(1)}$ at $\vec{u}^{(1)}$ in direction $\vec{h}^{(1)}$ is determined as follows: Let us abbreviate for simplicity of presentation

$$\nu := \nu \left(\left| \nabla_3 u_1^{(1)} \right|^2 + \left| \nabla_3 u_2^{(1)} \right|^2 \right), \quad \nu' := \nu' \left(\left| \nabla_3 u_1^{(1)} \right|^2 + \left| \nabla_3 u_2^{(1)} \right|^2 \right),$$

then the directional derivative of $\mathcal{R}^{(1)}$ in direction $\vec{h}^{(1)}$ at $\vec{u}^{(1)}$ is given by

$$\begin{aligned}
&(\partial_{\vec{u}^{(1)}} \mathcal{R}^{(1)})\vec{h}^{(1)} = \lim_{s \to 0} \frac{\mathcal{R}^{(1)}(\vec{u}^{(1)} + s\vec{h}^{(1)}) - \mathcal{R}^{(1)}}{s} \\
&= \lim_{s \to 0} \frac{1}{s} \int_{\Omega \times (0,1)} \nu \left(\left| \nabla_3 (u_1^{(1)} + s h_1^{(1)}) \right|^2 + \left| \nabla_3 (u_2^{(1)} + s h_2^{(1)}) \right|^2 \right) - \nu \, \mathrm{d}\vec{x}\mathrm{d}t \\
&= -2 \int_{\Omega \times (0,1)} \nabla_3 \cdot \left(\nu' \nabla_3 u_1^{(1)} \right) h_1^{(1)} + \nabla_3 \cdot \left(\nu' \nabla_3 u_2^{(1)} \right) h_2^{(1)} \mathrm{d}\vec{x}\mathrm{d}t,
\end{aligned} \tag{6.17}$$

where integration by parts is used to prove the final identity.

- The directional derivative of $\mathcal{R}^{(2)}$ is derived as follows:

$$\begin{aligned}
&(\partial_{\vec{u}^{(2)}} \mathcal{R}^{(2)})\vec{h}^{(2)} = \lim_{s \to 0} \frac{\mathcal{R}^{(2)}(\vec{u}^{(2)} + s\vec{h}^{(2)}) - \mathcal{R}^{(2)}}{s} \\
&= \lim_{s \to 0} \frac{1}{s} \int_{\Omega \times (0,1)} \left(\left| \int_0^t \vec{u}^{(2)} + s\vec{h}^{(2)} \mathrm{d}\tau \right|^2 - \left| \int_0^t \vec{u}^{(2)} \mathrm{d}\tau \right|^2 \right) \mathrm{d}\vec{x}\mathrm{d}t \\
&= 2 \sum_{j=1}^{2} \int_{\Omega \times (0,1)} \widehat{u}_j^{(2)} \widehat{h}_j^{(2)} \mathrm{d}\vec{x}\mathrm{d}t.
\end{aligned} \tag{6.18}$$

Moreover, it follows by integration by parts of the last line of (6.18) with respect to t that

$$(\partial_{\vec{u}^{(2)}}\mathcal{R}^{(2)})\vec{h}^{(2)} = -2\sum_{j=1}^{2}\int_{\Omega\times(0,1)}\widehat{\widehat{u}}_j^{(2)}(\vec{x},t)h_j^{(2)}(\vec{x},t)\mathrm{d}\vec{x}\mathrm{d}t. \tag{6.19}$$

Now, because of (6.17) and (6.16) it follows that the minimizer $\vec{u}^{(i)}$, $i = 1, 2$ has to satisfy for every $j = 1, 2$,

$$\begin{aligned}\partial_j f(\nabla f\cdot(\vec{u}^{(1)}+\vec{u}^{(2)})+\partial_t f)-\alpha^{(1)}\nabla_3\cdot\left(\nu'\nabla_3 u_j^{(1)}\right)&=0 \text{ in } \Omega\times(0,1),\\ \partial_{\vec{n}}u_j^{(1)}&=0 \text{ in } \partial\Omega\times(0,1),\\ \partial_t u_j^{(1)}&=0 \text{ in } \Omega\times\{0,1\}.\end{aligned} \tag{6.20}$$

Because of (6.16) and (6.19) hold for all $\vec{h}_j^{(2)}$, it follows that for every $j = 1, 2$,

$$\partial_j f(\nabla f\cdot(\vec{u}^{(1)}+\vec{u}^{(2)})+\partial_t f)-\alpha^{(2)}\widehat{\widehat{u}}_j^{(2)}=0 \text{ in } \Omega\times(0,1). \tag{6.21}$$

Thus the optimality conditions for a minimizer are (6.20) and (6.21).

6.5 Optical flow decomposition in 1D

In order to make transparent the features of our decomposition we study exemplary the 1D case again. From regularization theory (see e.g. [116]) we know that the minimizers $(u_{\vec{\alpha}}^{(1)}, u_{\vec{\alpha}}^{(2)})$, for $\vec{\alpha} = (\alpha^{(1)}, \alpha^{(2)}) \to 0$, are converging to a solution of the optical flow equation which minimizes

$$\mathcal{R} = \mathcal{R}^{(1)} + \alpha\mathcal{R}^{(2)} \text{ for } \alpha = \lim_{\vec{\alpha}\to 0}\frac{\alpha^{(2)}}{\alpha^{(1)}} > 0\,.$$

Such a solution is called $\mathcal{R}$ minimizing solution. Note that by the 1D simplification the modules $u^{(i)}$, $i = 1, 2$ are single valued functions.

We calculate the decomposition for the optical flow equation (6.5), for the specific test data (6.6). The regularized solutions approximate the $\mathcal{R}$ minimizing solution, and thus these calculations can be viewed representative for the properties of the minimizer of the regularization method. For these particular kind of data the solution of the optical flow equation is given by :

$$u(x,t) = -\frac{\tilde{f}(x)}{\partial_x\tilde{f}(x)}\frac{\partial_t g(t)}{g(t)} = -\frac{\partial_t(\log g)(t)}{\partial_x(\log\tilde{f})(x)}\,. \tag{6.22}$$

Let us assume that $(\log g)(t) - (\log g)(0)$ can be expanded into a Fourier sin-series:

$$(\log g)(t) - (\log g)(0) = \int_0^t \partial_t(\log g)(\tau)\,d\tau = \sum_{n=1}^{\infty}\hat{g}_n\sin(n\pi t). \tag{6.23}$$

Moreover, we assume that $1/\partial_x(\log\tilde{f})(x)$ can be expanded in a cos-series:

$$\frac{1}{\partial_x(\log\tilde{f})(x)} = \sum_{m=0}^{\infty} f_m\cos(m\pi x). \tag{6.24}$$

Then

$$-\frac{(\log g)(t)-(\log g)(0)}{\partial_x(\log \tilde{f})(x)} = \widehat{u}(x,t) = \widehat{u}_1(x,t) + \widehat{u}_2(x,t). \tag{6.25}$$

Inserting this identity in the regularization functional

$$\mathcal{R}(u^{(1)},u^{(2)}) = \int_{(0,1)\times(0,1)} (\partial_x u^{(1)})^2 + (\partial_t u^{(1)})^2 + \alpha \left(\hat{u}^{(2)}\right)^2 \mathrm{d}x\mathrm{d}t\,,$$

we remove the $u^{(2)}$ dependence, and we get

$$\mathcal{R}(\hat{u}^{(1)}) := \int_{(0,1)\times(0,1)} (\partial_{xt}\hat{u}^{(1)})^2 + (\partial_{tt}\hat{u}^{(1)})^2 + \alpha \left(\frac{(\log g)(t)-(\log g)(0)}{\partial_x(\log \tilde{f})(x)} + \hat{u}^{(1)}\right)^2 \mathrm{d}x\mathrm{d}t,$$

where we enforce the following boundary conditions on $\hat{u}^{(1)}$: From the definition of $\hat{u}^{(1)}$ it follows that $\hat{u}^{(1)}(x,0) = 0$. Secondly, we enforce $\hat{u}^{(1)}(x,1) = 0$. In fact, the assumption is reasonable because of the choice of $\hat{g}$, when the series $\sum_{n=0}^{\infty} \hat{g}_n$ is absolutely convergent, $\hat{g}(1) = 0$, which implies that $\hat{u}^{(1)}(x,1) + \hat{u}^{(2)}(x,1) = 0$, which is guaranteed in particular by $\hat{u}^{(1)}(x,1) = \hat{u}^{(2)}(x,1) = 0$.

By substituting the relation between $\hat{u}^{(2)}$ and $\hat{u}^{(1)}$ we reduce the constraint optimization problem to an unconstrained optimization problem for $\hat{u}^{(1)}$, and the minimizer solves the partial differential equation

$$\partial_{ttxx}\hat{u}^{(1)} + \partial_{tttt}\hat{u}^{(1)} + \alpha\left(\frac{(\log g)(t)-(\log g)(0)}{\partial_x(\log \tilde{f})(x)} + \hat{u}^{(1)}\right) = 0 \text{ in } (0,1)\times(0,1),$$

together with the boundary conditions:

$$\begin{aligned} \partial_{tt}\hat{u}^{(1)} = \hat{u}^{(1)} = 0 \text{ on } (0,1)\times\{0,1\}\,,\\ \partial_x\partial_{tt}\hat{u}^{(1)} = 0 \text{ on } \{0,1\}\times(0,1). \end{aligned} \tag{6.26}$$

Now, we substitute $\hat{w} := \partial_{tt}\hat{u}^{(1)}$, and we get the following system of equations

$$\begin{aligned} \partial_{xx}\hat{w} + \partial_{tt}\hat{w} = -\alpha\left(\frac{(\log g)(t)-(\log g)(0)}{\partial_x(\log \tilde{f})(x)} + \hat{u}^{(1)}\right) \text{ in } (0,1)\times(0,1),\\ \hat{w} = 0 \text{ on } (0,1)\times\{0,1\}\,,\\ \partial_x\hat{w} = 0 \text{ on } \{0,1\}\times(0,1). \end{aligned} \tag{6.27}$$

and

$$\hat{u}^{(1)}(x,t) = \int_0^t\int_0^\tau \hat{w}(x,\hat{\tau})\mathrm{d}\hat{\tau}\mathrm{d}\tau - t\int_0^1\int_0^\tau \hat{w}(x,\hat{\tau})\mathrm{d}\hat{\tau}\mathrm{d}\tau.$$

$\hat{w}$ can be expanded as follows:

$$\hat{w}(x,t) = \sum_{m,n=0}^{\infty} \hat{w}_{mn}\cos(m\pi x)\sin(n\pi t),$$

and we expand $\hat{u}^{(1)}$ in an analogous manner:

$$\hat{u}^{(1)}(x,t) = \sum_{m,n=0}^{\infty} \hat{u}^{(1)}_{mn}\cos(m\pi x)\sin(n\pi t)\,,$$

such that

$$\hat{w}_{mn} = -n^2\pi^2\hat{u}^{(1)}_{mn}, \quad \forall m, n \in \mathbb{N}_0. \tag{6.28}$$

Thus it follows from (6.27) that

$$\hat{w}_{mn}(m^2+n^2)\pi^2 = \alpha\left(\hat{u}^{(1)}_{mn} + f_m\hat{g}_n\right), \quad \forall m, n \in \mathbb{N}_0. \tag{6.29}$$

(6.28) and (6.29) imply that

$$\hat{u}^{(1)}_{mn} = -\frac{\alpha}{\alpha + \pi^4(m^2+n^2)n^2} f_m\hat{g}_n, \quad \forall m, n \in \mathbb{N}_0. \tag{6.30}$$

Now, consider a specific test example $g(t) = \exp\left\{\frac{\sin(n_0\pi t)}{n_0\pi}\right\}$ for some $n_0 \in \mathbb{N}$. Then, from (6.22) it follows that $u(x,t) = -\frac{\cos(n_0\pi t)}{\partial_x(\log f)(x)}$, and correspondingly we have

$$(\log g)(t) - (\log g)(0) = \frac{\sin(n_0\pi t)}{n_0\pi} = \sum_{n=1}^{\infty} \frac{\delta_{nn_0}}{n_0\pi}\sin(n\pi t).$$

In this case it follows from (6.30) that

$$\hat{u}^{(1)}_{mn} = -\frac{\alpha}{\alpha + \pi^4(m^2+n_0^2)n_0^2}\frac{\delta_{nn_0}}{n_0\pi} f_m.$$

For flickering $u^{(2)}$ is pronounced (if n_0 is large $\hat{u}^{(1)}_{mn} \approx 0$) while in the quasi-static case $u^{(1)}$ is dominant. Moreover, we also see that spatial components belonging to Fourier-cos coefficients with large m are more pronounced in the $u^{(2)}$ component, and the spatial and temporal coefficients are mixed.

6.6 Numerics

In this section we discuss the numerical minimization of the energy functional $\mathcal{F}$ defined in (6.10). Our approach is based on solving the optimality conditions for the minimizer $u^{(i)}_j$, $i,j = 1,2$ from (6.20), (6.21) with a fixed point iteration. We call the iterates of the fixed point iteration $u^{(i)}_j(\vec{x},t;k)$, for $k = 1,2,\ldots,K$, where K denotes the maximal number of iterations. We summarize all the iterates of the components of flow functions $u^{(i)}_j$ in a tensor of size $M \times N \times T \times K$. In this section we use the notation as reported in table 6.1. For every tensor $H = (H(r,s,t)) \in \mathbb{R}^{M\times N\times T}$ (representing a complete movie) we define the discrete gradient

$$\nabla^h_3 H(r,s,t) = (\partial^h_1 H(r,s,t), \partial^h_2 H(r,s,t), \partial^h_t H(r,s,t))^T \text{ for } (r,s,t) \in \mathbb{R}^{M\times N\times T},$$

where

$$\begin{aligned}
\partial^h_1 H(r,s,t) &= \frac{H(r+1,s,t) - H(r-1,s,t)}{2\Delta_x} && \text{if } 1 < r < M \\
\partial^h_2 H(r,s,t) &= \frac{H(r,s+1,t) - H(r,s-1,t)}{2\Delta_y} && \text{if } 1 < s < N \\
\partial^h_t H(r,s,t) &= \frac{H(r,s,t+1) - H(r,s,t-1)}{2\Delta_t} && \text{if } 1 < t < T
\end{aligned} \tag{6.31}$$

$f = f(r,s,t) \in \mathbb{R}^{M\times N\times T}$	Input sequence
$\vec{u}^{(i)} = \vec{u}^{(i)}(r,s,t;k) \in \mathbb{R}^{M\times N\times T\times K\times 2}$	artificial optical flow module
$\vec{u}^{(i)} = \vec{u}^{(i)}(r,s,t) = \vec{u}^{(i)}(r,s,t;K) \in \mathbb{R}^{M\times N\times T\times 2}$	formal relation between artificial and optical flow module
$u_j^{(i)} = u_j^{(i)}(r,s,t;k) \in \mathbb{R}^{M\times N\times T\times K}$	component of artificial optical flow module
$u_j^{(i)} = u_j^{(i)}(r,s,t) = u_j^{(i)}(r,s,t;K) \in \mathbb{R}^{M\times N\times T}$	formal relation between components of artificial and optical flow module
∂_k^h	Finite difference approximation in direction x_k
∂_t^h	Finite difference approximation in direction t

Table 6.1: Discrete Notation

and $\Delta_x = \frac{1}{M-1}$, $\Delta_y = \frac{1}{N-1}$ and $\Delta_t = \frac{1}{T-1}$. Again, whenever possible, we leave out the indices. Moreover, we define the discrete divergence, which is the adjoint of the discrete gradient: Let $(H_1, H_2, H_3)^T(r,s,t)$, then

$$\nabla_3^h \cdot (H_1, H_2, H_3)^T = \partial_1^h H_1 + \partial_2^h H_2 + \partial_t^h H_3. \tag{6.32}$$

The realization of the fixed point iteration for solving the discretized equations (6.20) and (6.21) reads as follows:

- $k = 0$: we initialize two flow components $\vec{u}^{(1)}(\cdot;0), \vec{u}^{(2)}(\cdot;0) \in \mathbb{R}^{M\times N\times K\times 2}$.
- $k \to k+1$: let $\nu'^{(k)} := \nu'(|\nabla_3^h u_1^{(1)}(\cdot;k)|^2 + |\nabla_3^h u_1^{(2)}(\cdot;k)|^2)$, then

$$\begin{aligned}\frac{u_1^{(1)}(\cdot;k+1) - u_1^{(1)}(\cdot;k)}{\Delta_\tau} &= \nabla_3^h \cdot \left(\nu'^{(k)} \nabla_3^h u_1^{(1)}(\cdot;k)\right) \\ &\quad - \frac{\partial_1^h f}{\alpha^{(1)}} \Big[\partial_1^h f \left(u_1^{(1)}(\cdot;k+1) + u_1^{(2)}(\cdot;k)\right) \\ &\quad + \partial_2^h f \left(u_2^{(1)}(\cdot;k) + u_2^{(2)}(\cdot;k)\right) + \partial_t^h f\Big],\end{aligned} \tag{6.33}$$

$$\begin{aligned}\frac{u_2^{(1)}(\cdot;k+1) - u_2^{(1)}(\cdot;k)}{\Delta_\tau} &= \nabla_3^h \cdot (\nu'^{(k)} \nabla_3^h u_2^{(1)}(\cdot;k)) \\ &\quad - \frac{\partial_2^h f}{\alpha^{(1)}} \Big[\partial_1^h f \left(u_1^{(1)}(\cdot;k+1) + u_1^{(2)}(\cdot;k)\right) \\ &\quad + \partial_2^h f \left(u_2^{(1)}(\cdot;k+1) + u_2^{(2)}(\cdot;k)\right) + \partial_t^h f\Big],\end{aligned} \tag{6.34}$$

$$\frac{u_1^{(2)}(\cdot;k+1)-u_1^{(2)}(\cdot;k)}{\Delta_\tau} = \widehat{\widehat{u}}_1^{(2)}(\cdot;k) - \frac{\partial_1^h f}{\alpha^{(2)}}\Big[\partial_1^h f\left(u_1^{(1)}(\cdot;k+1)+u_1^{(2)}(\cdot;k+1)\right) + \partial_2^h f\left(u_2^{(1)}(\cdot;k+1)+u_2^{(2)}(\cdot;k)\right)+\partial_t^h f\Big], \tag{6.35}$$

and

$$\frac{u_2^{(2)}(\cdot;k+1)-u_2^{(2)}(\cdot;k)}{\Delta_\tau} = \widehat{\widehat{u}}_2^{(2)}(\cdot;k) - \frac{\partial_2^h f}{\alpha^{(2)}}\Big[\partial_1^h f\left(u_1^{(1)}(\cdot;k+1)+u_1^{(2)}(\cdot;k+1)\right) + \partial_2^h f\left(u_2^{(1)}(\cdot;k+1)+u_2^{(2)}(\cdot;k+1)\right)+\partial_t^h f\Big], \tag{6.36}$$

where

$$\widehat{\widehat{u}}_j^{(2)}(r,s,t;k) = -\sum_{\tau=t}^{1}\sum_{\tilde{\tau}=0}^{\tau} u_j^{(2)}(r,s,\tilde{\tau};k), \qquad j=1,2.$$

and Δ_τ is the step size parameter, which has been set to 10^{-4}.

The system (6.33),(6.34),(6.35),(6.36) can be solved efficiently using the special structure of the matrix equation similarly to [141, 142].

The iterations are stopped when the Euclidean norm of the relative error

$$\frac{|u_j^{(i)}(\cdot,k)-u_j^{(i)}(\cdot,k+1)|}{|u_j^{(i)}(\cdot,k)|}, \qquad j=1,2$$

drops below the precision tolerance value $tol = 0.05$ for at least one of the component of $\vec{u}^{(1)}$ and one of $\vec{u}^{(2)}$. The typical number of iterations is much below 100.

6.7 Experiments

In this section we present numerical experiments to demonstrate the potential of the proposed optical flow decomposition model. In the first two experiments we use for visualization of the computed flow fields the standard flow color coding [14]. The flow vectors are represented in color space via the color wheel illustrated in Figure 6.4. For the third and fourth experiment we use a black and white visualization technique. There black is assigned to pixels where no flow is present and a gray-shade elsewhere, which is proportional to the flow magnitude. In order to compare frequencies of the sequences used for testing all the intensity values of f are scaled in the range $(0,1)$. The used parameters are reported for each experiment except for $\Delta_x, \Delta_y, \Delta_t$ defined as in Section 6.6. In this work we consider the following four dynamic image sequences:

- The first experiment is performed with the video sequence from [83] (available at `http://of-eval.sourceforge.net/`) which consists of forty-six frames showing a rotating sphere with some overlaid patterns. The

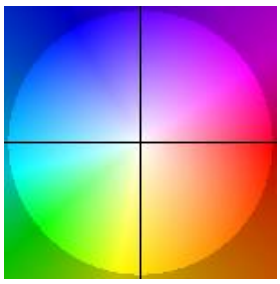

Figure 6.4: Color Wheel.

analytical results from Section 6.5 in 1D show that the intensity of the $\vec{u}^{(2)}$ component increases monotonically with increasing frequency over time. We verify this hypothesis numerically in higher dimensions. We simulate in particular two, four and eight times the original motion frequency. In order to do so, we duplicate the sequence periodically, however consider it to be in the same time interval $(0, 1)$. The flow visualized in Figure 6.5 is the one between the 16th and the 17th frame of every sequence. We study the behavior of the sphere at different motion frequencies with the same parameter setting $\alpha^{(1)} = 1$, $\alpha^{(2)} = \frac{1}{4}$. The numerical results confirm the $1D$ observation that for high frequency movement $\vec{u}^{(2)}$ is dominant (cf. Figures 6.5) and $\vec{u}^{(1)}$ is always 20% of $\vec{u}^{(2)}$.

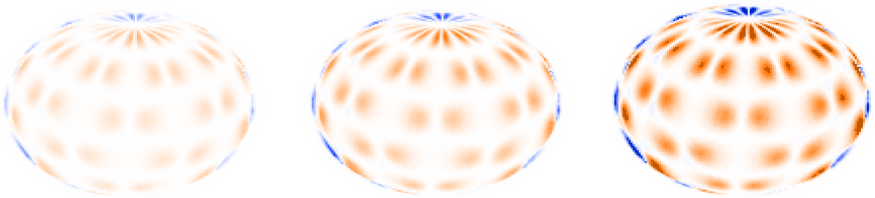

Figure 6.5: $\vec{u}^{(2)}$ at different frequencies of rotations: 2, 4 and 8× faster than the original motion frequency. $\alpha^{(1)} = 1$, $\alpha^{(2)} = \frac{1}{4}$. The intensity of $\vec{u}^{(2)}$ increases when the frequency of rotation is increased.

- The second experiment concerns the decomposition of the motion in a dynamic image sequence showing a projection of a cube moving over an oscillating background. The movie consists of sixty frames and can be viewed on the web-page `http://www.csc.univie.ac.at/index.php?page=visualattention`.
  The background is oscillating in diagonal direction, from the bottom left to the top right, with a periodicity of four frames. In each frame the oscillation has a rate of 5% of the frame size. The flow visualized in Figure 6.6 is the one between the 20th and the 21st frame of the sequence. Applying the proposed method with a parameter setting $\alpha^{(1)} = 10^3$, $\alpha^{(2)} = 10^3$, $\Delta_\tau = 10^{-5}$, and precision tolerance $tol = 0.001$, we notice that the background movement appears almost solely in $\vec{u}^{(2)}$ and the global movement of the cube appears in $\vec{u}^{(1)}$. In Figure 6.6 we represent only flow vectors

with magnitude larger than 0.05 and omit in $\vec{u}^{(2)}$ the part in common with $\vec{u}^{(1)}$ for better visibility.

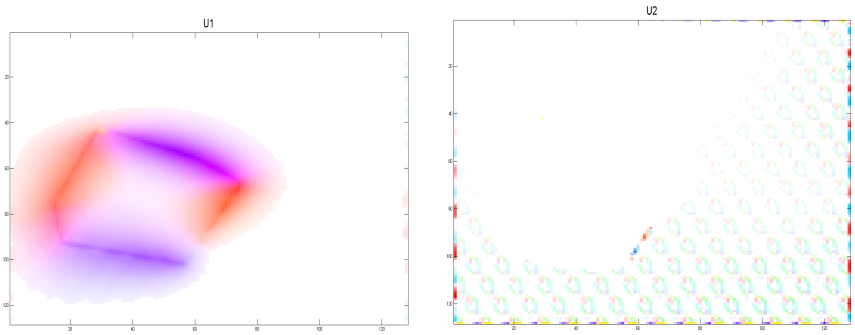

Figure 6.6: The dynamic sequence consists of the smooth (translation like) motion of a cube and an oscillating background. The oscillation has a periodicity of four frames and takes place along the diagonal direction from the bottom left to the top right, moving at a rate of 5% of the frame size in each frame. The proposed model decomposes the motion, obtaining the global movement of the cube in $\vec{u}^{(1)}$ (left) and the background movement solely in $\vec{u}^{(2)}$ (right).

- In the third experiment the original movie consists of thirty-two frames and can be viewed together with the decomposition result on the webpage `http://www.csc.univie.ac.at/index.php?page=visualattention`. The flow is decomposed into two components. The first part shows the movement of a Ferris wheel and people walking. The second part shows blinking lights and the reflection of the wheel. The flow visualized in Figure 6.7 is the one between the 4th and the 5th frame of the sequence with a parameter setting $\alpha^{(1)} = 1$, $\alpha^{(2)} = \frac{1}{4}$. In order to improve the visibility we represent only flow vectors with magnitude larger than 0.18 and we omit in $\vec{u}^{(2)}$ the part in common with $\vec{u}^{(1)}$.

- The fourth example is flickering. In a standard flickering experiment, the difference in human attention is investigated by inclusion of blank images in a repetitive image sequence. Although, in general, these blank images are not deliberately recognized, they change the awareness of the test persons. J. K. O'Regan [97] states that "*Change blindness is a phenomenon in which a very large change in a picture will not be seen by a viewer, if the change is accompanied by a visual disturbance that prevents attention from going to the change location*". They provided test data `http://nivea.psycho.univ-paris5.fr`, which we used for our simulations. The proposed optical flow decomposition is able to detect regions, which also humans can recognize, but standard optical flow algorithms fail to: To show this the input sequence is composed by four frames consisting of Frame 1, a blank image, Frame 2 and again an identical blank image (see Figure 6.9 (top)). This sequence is then aligned periodically to a movie. We interpret the movie as a linear interpolation between the frames.

Figure 6.7: $\vec{u}^{(1)}$: Movement of a Ferris wheel and people walking in the foreground (top left). $\vec{u}^{(2)}$ consists of blinking lights and the reflections of the wheel (top right). The third image (bottom) is a reference frame.

We test and compare Horn-Schunck, Weickert-Schnörr and the proposed algorithm. We set the smoothness parameter $\alpha^{(1)}$ to a value of one for

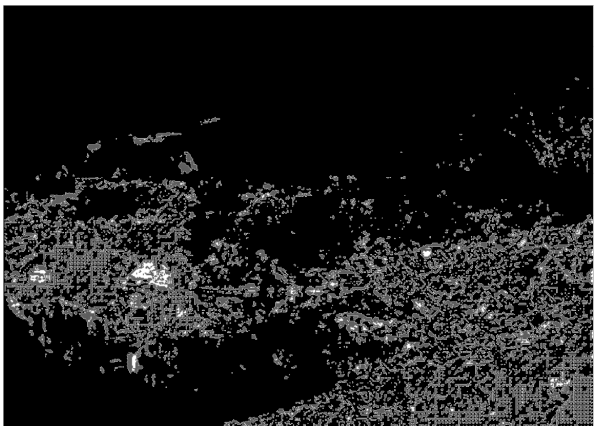

Figure 6.8: Result with Horn-Schunck

all the methods. Moreover, for our approach we set $\alpha^{(2)} = 1$. For Horn-Schunck we visualize the flow field in Figure 6.8. This flow is the one between the blank frame and the slightly changed frame, which exceeds a threshold of 3.9. The results obtained by applying Weickert-Schnörr and the $\vec{u}^{(1)}$ field of our approach, respectively, are small in magnitude. Therefore, we do not visualize them. This behavior is coherent with the motivation of the Weickert-Schnörr method to produce an optical flow that is less sensitive to variations over space and time. Finally, we visualize in Figure 6.9 (down right) the $\vec{u}^{(2)}$ flow field for the proposed

approach. For the visualization we omit all vectors with magnitude lower than 0.18. In order to make transparent the result, we show in Figure 6.9 (down left) the difference between the two frames of the sequence containing information (see Figure 6.9 (top)). In this experiment, we

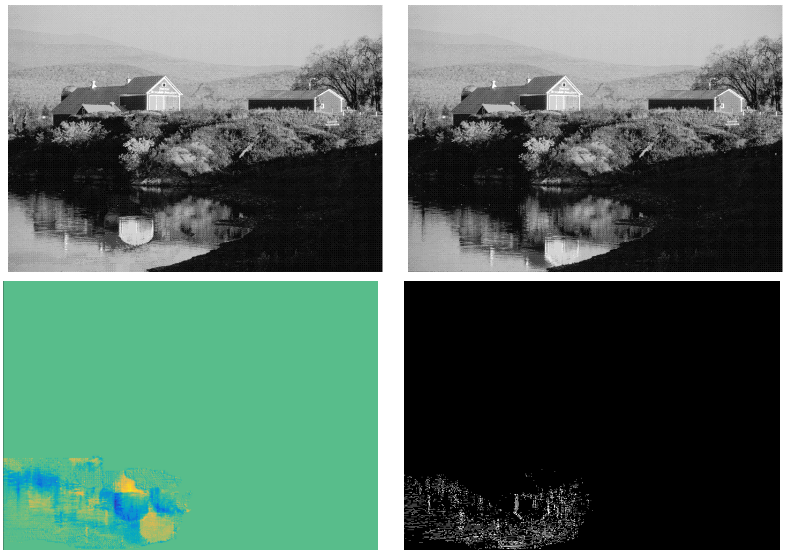

Figure 6.9: The two frames of the flickering sequence containing information (top), the difference between these two frames (down left), and the $\vec{u}^{(2)}$ flow field resulting from the proposed approach (down right). As predicted in sections 6.3 and 6.5 the $\vec{u}^{(1)}$ component is negligible, instead $\vec{u}^{(2)}$ detects the change of intensity across the blank sheet.

notice that the $\vec{u}^{(1)}$ component is negligible, instead $\vec{u}^{(2)}$ detects the areas affected by change of intensities (see Figure 6.9 (down right)).

Additional Information

In the following, we show the capacity of our model to extract more and different information compared to standard optical flow algorithms. The current literature focuses on average angular and endpoint error [14] in order to compare optical flow algorithms. Our model extracts information, that is neglected by standard algorithms. Such difference can be shown through a quantitative comparison of models. For this purpose, we use well-known test sequences from the Middlebury database `http://vision.middlebury.edu/flow/`, and evaluate the residual of the *optical flow constraint*. We compare the residual of our method with the one of the Horn-Schunck method [57]. However, the Horn-Schunck method does not take into account time information, and therefore we calculate for every pair of successive frames and stack the series of flow images into a movie. For calculating the flow for one pair we use the regularization parameter $\alpha = 400$ and 50 iterations for every pair of frames. For the proposed method $\alpha^{(1)} = 400$, $\alpha^{(2)} = 10$ and tolerance value $tol = 0.03$. In this case the whole image sequence is used at once.

The parameters $\alpha^{(1)}, \alpha^{(2)}$ are chosen larger than 1 in order to avoid overfitting. For every pair of successive images f_1 and f_2 we visualize the squared residual

$$\int_\Omega \left(\nabla f_1 \cdot \vec{u} + \frac{f_2 - f_1}{\Delta_t} \right)^2 \mathrm{d}\vec{x},$$

both for Horn-Schunck and the proposed method. Note that for the comparison we omit space dependency of the movie. We notice from Figure 6.10 that the squared residual is larger in every frame for Horn-Schunck than for our decomposition model, meaning that the proposed method is capable to extract more flow information.

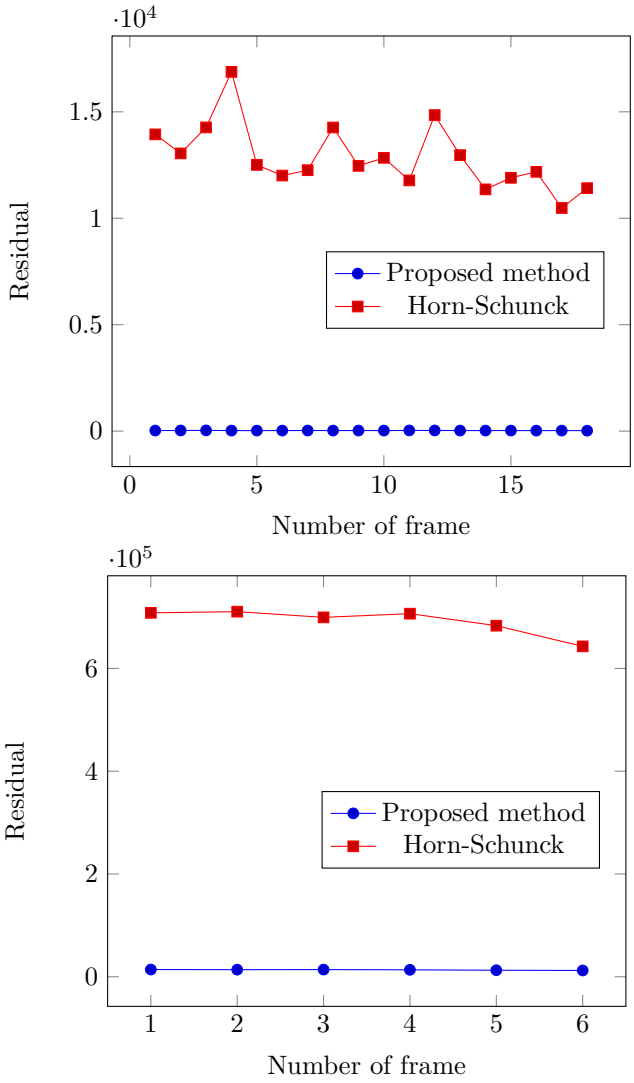

Figure 6.10: Residuals for Hamburg taxi (up) and Minicooper sequence (down) from Middlebury database. Residuals for Horn-Schunck are plotted in red, the proposed method is plotted in blue.

	Weickert-Schnörr	Proposed model
Hamburg Taxi	1374.9	1021
RubberWhale	4459.7	3046.8
Hydrangea	8533.3	7647.2
DogDance	9995.4	8217.6
Walking	8077.5	5944.3

Table 6.2: Comparison of squared residuals over space and time $\mathcal{E}$ between Weickert-Schnörr and the proposed method.

In order to understand how much information our method is capable to extract from an entire dynamic sequence, we also calculate the residual squared over space and time: $\mathcal{E}(\vec{u}^{(1)}, \vec{u}^{(2)})$ as in (6.10) and compare it with the squared residual over space and time of the Weickert-Schnörr method [141, 142]. We use the parameter settings $\alpha^{(1)} = 100$ ($\alpha = \alpha^{(1)}$ in Weickert-Schnörr) and $\alpha^{(2)} = \frac{1}{4}$, tolerance $tol = 0.01$, in order to have a good comparison of the two methods. Again the residual is smaller for the proposed method as shown in table 6.2.

6.8 Conclusion

We present a new optical flow model for decomposing the flow in spatial and temporal components of different scales. A main ingredient of our work is a new variational formulation of the optical flow equation. Finally, many applications are considered both analytically and computationally in case of illumination disturbances.

Acknowledgment

This work is carried out within the project *Modeling Visual Attention as a Key Factor in Visual Recognition and Quality of Experience* funded by the Wiener Wissenschafts und Technologie Fonds - WWTF. OS is also supported by the Austrian Science Fund - FWF, Project S11704 with the national research network, NFN, Geometry and Simulation. The authors would like to thank U. Ansorge, C. Valuch, S. Buchinger, C. Kirisits, P. Elbau, L. Lang, G. Dong, T. Widlak for interesting discussions on the optical flow and José A. Iglesias for discussions and the creation of the cube sequence.

Chapter 7

Dynamical optical flow of saliency maps for predicting visual attention

Authors & Contributions The authors are Aniello Raffaele Patrone, Christian Valuch, Ulrich Ansorge and Otmar Scherzer. The development of this article was a collaborative process, and each of the authors made significant contributions to every aspect of the paper.

Publication Status Submitted to *Computer Vision and Image understanding Journal*. Date of Acknowledgement of receipt: June, 23, 2016. Preprint available in [102].

Dynamical optical flow of saliency maps for predicting visual attention

Aniello Raffaele Patrone[1], Christian Valuch[3,5], Ulrich Ansorge[3], and Otmar Scherzer[1,2]

[1]Computational Science Center, University of Vienna,
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[2]Radon Institute of Computational and Applied Mathematics,
Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria
[3]Faculty of Psychology,
University of Vienna Liebigg. 5, A-1010 Wien
[5]Faculty of Biology and Psychology,
University of Göttingen

Abstract

Saliency maps are used to understand human attention and visual fixation. However, while very well established for static images, there is no general agreement on how to compute a saliency map of dynamic scenes. In this paper we propose a mathematically rigorous approach to this problem, including static saliency maps of each video frame for the calculation of the optical flow. Taking into account static saliency maps for calculating the optical flow allows for overcoming the aperture problem. Our approach is able to explain human fixation behavior in situations which pose challenges to standard approaches, such as when a fixated object disappears behind an occlusion and reappears after several frames. In addition, we quantitatively compare our model against alternative solutions using a large eye tracking data set. Together, our results suggest that assessing optical flow information across a series of saliency maps gives a highly accurate and useful account of human overt attention in dynamic scenes.

Keywords: Saliency, visual attention, eye movements, optical flow, motion, dynamic scenes, variational approach.

7.1 Introduction

Humans and other primates focus their perceptual and cognitive processing on aspects of the visual input. This selectivity is known as *visual attention* and it is closely linked to eye movements: humans rapidly shift their center of gaze multiple times per second from one location of an image to another. These gaze shifts are called saccades, and they are necessary because high acuity vision

is limited to a small central area of the visual field. Fixations are the periods between two saccades, in which the eyes rest relatively still on one location from which information is perceived with high acuity. Consequently, fixations reflect which areas of an image attract the viewer's attention.

Models of visual attention and eye behavior can be roughly categorized into two classes: bottom-up models, which are task-independent and driven mainly by the intrinsic features of the visual stimuli, and top-down models, which are task-dependent and driven by high-level processes [55]. Our focus in this article is on bottom-up models. A central concept in bottom-up models of attention is the saliency map, which is a topographical representation of the original image representing the probability of each location to attract the viewer's attention. Saliency maps are useful for testing hypotheses on the importance of the current image's low-level visual features (such as color, luminance, or orientation). Moreover, saliency models allow for general predictions on the locations that are fixated by human viewers. This is important in many contexts, such as the optimization of video compression algorithms or of graphical user interfaces at the workplace as well as in entertainment environments, to name but a few examples. A location in an image is considered *salient* if it stands out compared to its local surroundings.

Saliency maps are often computed based on low-level visual features and their local contrast strengths in dimensions such as color, luminance, or orientation. A well-known model for *static* scenes is the one of [64]. Other examples are graph-based visual saliency (GBVS)[53], gaze-attentive fixation finding engine (GAFFE)[108], frequency-tuned saliency detection model [3] and models based on phase spectrum, explained by the inverse Fourier transform [52].

Recent work is devoted to develop concepts of saliency maps of *dynamic* sequences [42, 62, 60, 63, 77, 109]. These *spatial-temporal saliency maps* are modeled as the weighted sum of motion features and of static saliency maps [42, 60, 62, 63, 76, 77, 79, 81, 104, 109, 119, 148].

The present work introduces a *novel* dynamic saliency map, which is the optical flow of a high-dimensional dynamic sequence. Extending the concept of saliency to dynamic sequences by including optical flow as an additional source for bottom-up saliency is not new per se [48, 81, 130, 136]. However, while other researchers use the optical flow as a feature of the dynamic saliency map, we define the dynamic saliency map as the optical flow itself. In detail:

1. we calculate the flow of a *virtual*, *high-dimensional* image sequence, which consists of (i) intensity and (ii) color channels, *complemented* by saliency maps, respectively;

2. we also consider the complete movie (consisting of all frames) for the computation of a dynamic saliency map. In contrast, in [42, 60, 62, 63, 76, 77, 79, 81, 104, 109, 119, 148] (as it is standard), dynamic saliency maps are obtained from optical flow features (see Figure 7.1) of *two consecutive* frames. As we show below in section 7.4 this can lead to misinterpretations of visual attention, for instance, in the case of occlusions.

Our algorithm for calculating the dynamic saliency map is schematically depicted in Figure 7.2. From the different methods proposed in the literature to estimate optical flow, we focus on variational methods, which are key methods in computational image processing and computer vision.

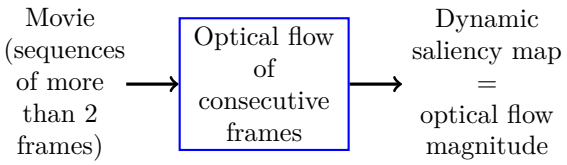

Figure 7.1: The standard approach for calculating a dynamic saliency map

The outline of this paper is as follows: in section 7.2 we review the optical flow, introduce the new model and derive a fixed-point algorithm for the computational realization; in section 7.3 we discuss the acquisition of the eye tracking data; finally, in sections 7.4 and 7.5 we present experiments, results and a discussion.

7.2 Computational methods

The optical flow denotes the pattern of apparent motion of objects and surfaces in a dynamic sequence. The computational model of optical flow is based on the *brightness-constancy assumption*, requiring that for every pixel there exists a path through the movie which conserves brightness.

Basic optical flow calculations

We briefly outline the concept in a continuous mathematical formulation. We consider a *movie* to be a time continuous recording of images, where each image is described by a function defined for $x = (x_1, x_2)^t$ in the Euclidean plane $\mathbb{R}^2$. This function representing the image is called *frame*. Moreover, we assume that the movie to be analyzed has unit-length in time. That is, the movie can be parametrized by a time $t \in [0, 1]$. If the frames composing a movie consist of gray-valued images, then we describe each by a function $f : \mathbb{R}^2 \to \mathbb{R}$. If the frame is completed by a spatial saliency map, then $\vec{f} := (f_1; f_2)^t : \mathbb{R}^2 \to \mathbb{R}^2$, where f_1 is the recorded movie and f_2 is the according saliency map. For a color frame, $\vec{f} := (f_1, f_2, f_3)^t : \mathbb{R}^2 \to \mathbb{R}^3$, where each component represents a channel

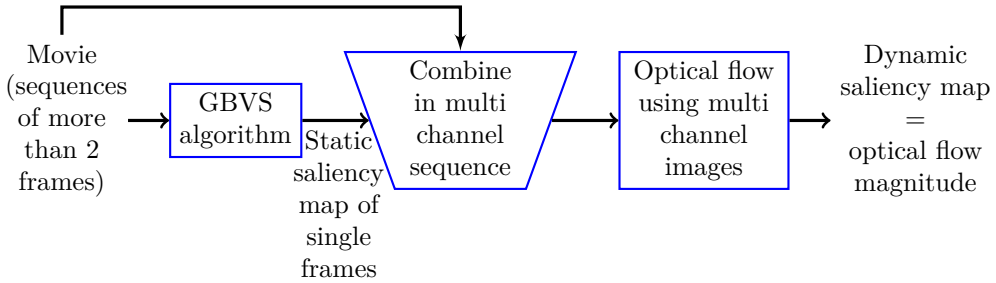

Figure 7.2: The proposed approach for calculating a dynamic saliency map.

of the color images; typically RGB (red-green-blue) or HSV (hue-saturation-value) channels. A color frame, which is complemented by a saliency map is described by $\vec{f} := (f_1, f_2, f_3; f_4)^t : \mathbb{R}^2 \to \mathbb{R}^4$, where the first three components are the color channels and the fourth component is the according saliency map. For the sake of simplicity of notation, from now on, we will always write $\vec{f}$, even if f is a gray-valued image.

The optical flow equation is derived from the *brightness constancy assumption*, which considers paths γ of constant intensity of the movie or the saliency complemented movie, respectively. Note that in our setting we consider brightness in each component. That is for

$$\vec{f}(\gamma(x,t),t) = \vec{c} \tag{7.1}$$

for some constant vector $\vec{c}$. Note that we do not differentiate in our notation between intensity, color, or saliency completed movies anymore and always write $\vec{f}$ for the image data.

A differential formulation of the brightness constancy assumption follows from (7.1) by differentiation with respect to time (see for instance [57]):

$$J_{\vec{f}}(x,t) \cdot \vec{u}(x,t) + \frac{d\vec{f}}{dt}(x,t) = 0 \text{ for all } x \in \mathbb{R}^2 \text{ and } t \in (0,1)\,, \tag{7.2}$$

where $\vec{u}(x,t) = (u_1(x,t), u_2(x,t)) = \frac{d\gamma}{dt}(x,t)$ is the optical flow and $J_{\vec{f}}, \frac{d\vec{f}}{dt}$ are the partial derivatives in space and time of the function $\vec{f}$, respectively. Note that $(u_1(x,t), u_2(x,t))$ can both be vectors, and that the *Jacobian* $J_{\vec{f}} = \nabla\vec{f} = \left(\frac{\partial}{\partial x_1}\vec{f}, \frac{\partial}{\partial x_2}\vec{f}\right)$ is a two-dimensional vector if the movie is gray-valued, a (2×2)-dimensional matrix if it is a saliency complemented gray-valued image, a (3×2)-dimensional matrix if it is in color, and a (4×2)-dimensional matrix if a color image is complemented by a saliency map.

Equation (7.2) is uniquely solvable at points (x,t) where $J_{\vec{f}}$ has full-rank 2. For gray-valued movies the matrix can have at most rank of one, and thus the two unknown functions u_1 and u_2 *cannot* be reconstructed uniquely from this equation. This is known as the *aperture problem* in Computer Vision. The non-uniqueness is taken care of mathematically by restricting attention to the *minimum energy solution* of (7.2) which minimizes, among all solutions, an appropriately chosen energy, such as the one proposed in [57]:

$$\mathcal{E}[t](\vec{u}) := \int_{\mathbb{R}^2} |\nabla u_1(x,t)|^2 + |\nabla u_2(x,t)|^2 dx \text{ for all } t \in [0,1]. \tag{7.3}$$

For every $t \in (0,1)$ the minimum energy solution can be approximated (see for instance [116] for a rigorous mathematical statement) by the minimizer of

$$\begin{aligned} \mathcal{F}[t](\vec{u}) := &\int_{\mathbb{R}^2} \left| J_{\vec{f}}(x,t) \cdot \vec{u}(x,t) + \frac{d\vec{f}}{dt}(x,t) \right|^2 dx \\ &+ \alpha \int_{\mathbb{R}^2} |\nabla u_1(x,t)|^2 + |\nabla u_2(x,t)|^2 dx. \end{aligned} \tag{7.4}$$

Here $\alpha > 0$ is a weight (also called regularization) parameter.

Different optical flow methods have been considered in the literature, which are formulated via different regularization energies. For instance, in [142] it was suggested to minimize

$$\int_{\mathbb{R}^2} \left| J_{\vec{f}}(x,t)\cdot\vec{u}(x,t) + \frac{d\vec{f}}{dt}(x,t)\right|^2 + \alpha\Psi(|\nabla u_1(x,t)|^2 + |\nabla u_2(x,t)|^2)dx \quad \text{for all } t\in[0,1], \tag{7.5}$$

where $\Psi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is a monotonically increasing, differentiable function. Note that in the original work of [142] the function $\vec{f}$ is not saliency complemented and the data $\vec{f}$ only represents intensity images.

We note that in most applications, $\mathcal{F}[t]$ is realized by replacing $J_{\vec{f}}$ by the spatial difference quotients, $\frac{d\vec{f}}{dt}$ by the temporal difference of two consecutive frames $\vec{f}(x,t_i)$, $i=1,2$, respectively, and appropriate scaling. This results in the functional to be minimized:

$$\begin{aligned}&\int_{\mathbb{R}^2} \left| J_{\vec{f}(\cdot,t_1)}(x)\cdot\vec{u}(x) + \vec{f}(x,t_2) - \vec{f}(x,t_1)\right|^2 dx\\ &+\alpha\int_{\mathbb{R}^2}\Psi(|\nabla u_1(x)|^2 + |\nabla u_2(x)|^2)dx.\end{aligned} \tag{7.6}$$

Remark 1. *We emphasize that by the approximation of $\frac{d\vec{f}}{dt}$ with the finite difference $\vec{f}(x,t_2)-\vec{f}(x,t_1)$ (where, we assume that after time scaling we have $t_2-t_1=1$), the equations* (7.2) *and*

$$J_{\vec{f}(\cdot,t_1)}(x)\cdot\vec{u}(x) + \vec{f}(x,t_2) - \vec{f}(x,t_1) = 0 \text{ for all } x\in\mathbb{R}^2, \tag{7.7}$$

might not be correlated anymore. This is especially true for large displacements $\gamma - I$. In order to overcome this discrepancy, researchers in computer vision proposed to use computational coarse-to-fine strategies [5, 22, 28, 84, 85, 86].

We emphasize that in all these approaches the temporal coherence of the movie is neglected and this can lead to rather abruptly changing flow sequences. Therefore, in [142], a *spatial-temporal* regularization was suggested, which consists in minimization of

$$\int_{\mathbb{R}^2\times[0,1]} \left| J_{\vec{f}}(x,t)\cdot\vec{u}(x,t) + \frac{d\vec{f}}{dt}(x,t)\right|^2 + \alpha\Psi(|\nabla_3 u_1(x,t)|^2 + |\nabla_3 u_2(x,t)|^2)dxdt, \tag{7.8}$$

where $\nabla_3 = (\nabla, \frac{\partial}{\partial t})$ denotes the spatial-temporal gradient operator. More sophisticated spatial-temporal regularization approaches have been proposed in [6, 24, 101, 137, 141, 142].

Spatial saliency for optical flow computations

In the following section we investigate the effect of complementing intensity and color images with a spatial saliency map on optical flow computations. We use the GBVS method, introduced by [53], which defines a spatial saliency for each

Sequence	Saliency-Brightness	Saliency-Color	Color
Backyard	3.16%	8.18%	2.45%
Basketball	3.68%	10.06%	1.26%
Beanbags	4.07%	9.73%	0.9%
Dimetrodon	5.91%	7.54%	1.07%
DogDance	3.99%	9.57%	2.50%
Dumptruck	5.20%	10.13%	3.35%
Evergreen	6.12%	12.48%	3.20%
Grove	12.61%	18.94%	4.86%
Grove2	15.48%	24.07%	3.29%
Grove3	16.49%	25.27%	4.20%
Hydrangea	10.30%	29.93%	2.48%
Mequon	11.44%	19.12%	2.38%
MiniCooper	7.35%	16.77%	3.76%
RubberWhale	8.45%	22.34%	1.46%
Schefflera	12.12%	20.02%	1.98%
Teddy	16.65%	32.96%	4.63%
Urban2	3.17%	3.35%	0.31%
Venus	12.86%	17.84%	2.75%
Walking	2.85%	7.41%	1.25%
Wooden	3.38%	7.76%	0.75%

Table 7.1: Percentage of the pixels with condition number smaller than 1000 for different sequences from the Middleburry Dataset [15]. For each sequence we use the third frame and its spatial saliency as input.

frame. Note however, that our approach is not restricted to this particular choice.

We verify the hypothesis that the complemented data uniquely determines the optical flow in selected regions - that is, where $J_{\vec{f}}$ has full rank. A conceptually similar strategy was implemented in [81] where the image data was complemented by Gabor filters as input of (7.2). Here, we verify this thesis by checking the spatially dependent condition number of $J_{\vec{f}}$ in the plane, and by tabbing the area of the points in a region where the condition number is below 1000 (see Table 7.1). In these regions we expect that the flow can be computed accurately from (7.2) without any regularization. We recall that for gray-valued images the optical flow equation is under-determined and the matrix $J_{\vec{f}}$ is singular, or in other words, the condition number is infinite. However, if the intensity data is complemented by a saliency map, 3 to 17% of the pixels have a condition number smaller than 1000, such that the solution of (7.2) can be determined in a numerically stable way. For color images the optical flow equation is already over-determined and 1 to 5% of the pixels have a small condition number. However, if the color information is complemented by a saliency information, 3 to 33% of the pixels have this feature. These results suggest that complementing the original information by saliency is useful for accurate computations of the optical flow.

Contrast invariance

We also aim to recognize motion under varying illumination conditions because humans have this visual capacity. However, a typical optical flow model, like the one in (7.8), would not yield contrast invariant optical flow although this would be necessary for motion recognition under varying illumination conditions.

In order to restore contrast invariance for the minimizer of the optical flow functional, as proposed (for a different reason) by [59] and by [72, 149], we introduce the semi-norm $\|\cdot\|_{\mathcal{B}}^2$ like:

$$\|w\|_{\mathcal{B}}^2 = w^T \mathrm{diag}(b_1, ..., b_n) w\,,$$

where b_i are the components of a vector $\mathcal{B}$. For gray-valued images complemented with saliency $\vec{f} = (f_1; f_2)$ the vector $\mathcal{B}$ is defined as:

$$\mathcal{B} = \left(\tfrac{f_2}{\sqrt{|\nabla f_1|^2+\xi^2}}, 1 \right). \tag{7.9}$$

For saliency complemented color images $\vec{f} = (f_1, f_2, f_3; f_4)$ we define

$$\mathcal{B} = \left(\tfrac{f_4}{\sqrt{|\nabla f_1|^2}}, \tfrac{f_4}{\sqrt{|\nabla f_2|^2+\xi^2}}, \tfrac{f_4}{\sqrt{|\nabla f_3|^2+\xi^2}}, 1 \right). \tag{7.10}$$

The vectors $\mathcal{B}$ are the product of the weighting factors $\frac{1}{\sqrt{|\nabla f_i|^2+\xi^2}}$ and the saliency map, respectively. This means that features with high spatial saliency will be weighted stronger, and thus more emphasis on a precise optical flow calculation is given to these regions. We are then using the weighted semi-norm as an error measure of the residual of (7.2):

$$\int_{\mathbb{R}^2\times[0,1]} \left\| J_{\vec{f}}(x,t)\cdot\vec{u}(x,t) + \frac{d\vec{f}}{dt}(x,t) \right\|_{\mathcal{B}}^2 . \tag{7.11}$$

Finally, one needs to choose the constant ξ for the denominator of each weighting factor. As in [59] we choose $\xi = 0.01$.

The final model

We use as a regularization functional

$$\int_{\mathbb{R}^2\times[0,1]} \Psi(|\nabla_3 u_1(x,t)|^2 + |\nabla_3 u_2(x,t)|^2) dx dt$$

as in [142], with the difference that the function $\Psi(r^2) = \epsilon r^2 + (1-\epsilon)\lambda^2\sqrt{1+\frac{r^2}{\lambda^2}}$ is replaced by

$$\Psi(r^2) = \sqrt{r^2+\epsilon^2} \text{ with } \epsilon = 10^{-6} \tag{7.12}$$

as in [27]. Moreover, we substitute the spatial-temporal gradient operator in [142] with $\nabla_3 = (\nabla, \lambda\frac{\partial}{\partial t})$. In numerical realizations, the weighting parameter λ corresponds to the ratio of the sampling in space squared and the one in time

The resulting model for optical flow computations then consists in minimization of the functional

$$\int_{\mathbb{R}^2\times[0,1]} \left\| J_f(x,t)\cdot\vec{u}(x,t)+\frac{d\vec{f}}{dt}(x,t)\right\|_{\mathcal{B}}^2 + \alpha\Psi(|\nabla_3 u_1(x,t)|^2+|\nabla_3 u_2(x,t)|^2)dxdt. \tag{7.13}$$

.

Numerical solution

The ultimate goal is to find the path γ solving (7.1), that is connecting a movie sequence. By applying Taylor expansion one sees that $\vec{u}\sim\frac{d\gamma}{dt}(x,t)t$ for small t. As we have stated in Remark (1), when trying to find approximations of γ via minimization of the proposed functional (7.13) a coarse-to-fine strategy [5, 22, 28, 84, 85, 86] is useful for the following two reasons:

First, since at a coarse level of discretization large displacements appear relatively small, the optical flow equation (7.2), which is a linearization of the brightness constancy equation (7.1), is a good approximation of that. Indeed, a displacement of one pixel at the coarsest level of a 4-layer pyramid can represent 4 pixels of distance in the finest layer. With reference to [73], we can assume that on the coarsest level, the discretized energy functional has a unique global minimum and that the displacements are still small. We further expect to obtain the global minimum by refining the problem at finer scales and using the outcome of the coarser iteration as an initial guess of the fine level.

Second, this strategy results in a faster algorithm [15]. The optical flow is computed on the coarsest level, where the images are composed by fewest pixels, and then upsampled and used to initialize the next level. The initialization results in far lesser iterations at each level. For this reason, an algorithm using coarse-to-fine strategy tends to be significantly faster than an algorithm using only the finest level. For the coarse-to-fine strategy, we use four pyramid levels and a bicubic interpolation between each level.

In this paper, we combine the coarse-to-fine strategy with two nested fixed-point iterations. We apply a presmoothing at each level, by convolving the images with a Gaussian kernel with standard deviation one, as proposed by [17]. We solve the minimization problem on each pyramid level starting from the coarsest one. There, we initialize the optical flow $\vec{u}$ by 0. The solution of the minimization problem is smoothed applying a median filter, as proposed in [122], and then prolonged to the next finer level. There, we employ it for the initialization of the fixed-point iterations.

For the purpose of numerical realization we call the iterates of the fixed point iteration $u_i^{(k)}$ for $k=1,2,...,K$ where K denotes the maximal number of iterations. We plot the pseudo-code to illustrate the structure of the algorithm in Figure 7.3.

For each level of the pyramid we compute with the fixed point iterations

Given an input sequence f
Initialization
for all frame in the sequence **do**
    create a pyramid of 4 levels
    calculate the saliency map
    smooth the frame
end for
for each level $lev \in 1..4$ **do**
    if lev=1 **then**
        $u_1^{(k)}$=0
        $u_2^{(k)}$=0
    else
        $u_1^{(k)} = u_1^{(k+1)}$
        $u_2^{(k)} = u_2^{(k+1)}$
    end if
    while the precision tolerance ≥ 0.001 **do**
        Calculate an approximation of $u_1^{(k+1)}$ and $u_2^{(k+1)}$
        solving fixed point iterations
    end while
    Apply median filtering to the flow
    Rescale $u_1^{(k+1)}$ and $u_2^{(k+1)}$ with bicubic interpolation
end for

Figure 7.3: Pseudo-code to illustrate the structure of the algorithm

the solution of the optimality condition of the functional (7.13):

$$\begin{aligned} 0 &= \sum_{i=1}^{\sigma} \mathcal{B}_i \frac{\partial f_i}{\partial x_1}\left(\frac{\partial f_i}{\partial x_1}u_1 + \frac{\partial f_i}{\partial x_2}u_2 + \frac{\partial f_i}{\partial t}\right) - \nabla_3 \cdot (\Psi'(|\nabla_3 u_1|^2 + |\nabla_3 u_2|^2)\nabla_3 u_1) \\ 0 &= \sum_{i=1}^{\sigma} \mathcal{B}_i \frac{\partial f_i}{\partial x_2}\left(\frac{\partial f_i}{\partial x_1}u_1 + \frac{\partial f_i}{\partial x_2}u_2 + \frac{\partial f_i}{\partial t}\right) - \nabla_3 \cdot (\Psi'(|\nabla_3 u_1|^2 + |\nabla_3 u_2|^2)\nabla_3 u_2) \end{aligned} \tag{7.14}$$

where $\mathcal{B}_i$ is the *i-component* of the vector $\mathcal{B}$ and $\sigma = 2, 4$, if we consider intensity with complemented saliency data or color with complemented saliency data, respectively.

For the solution of the system of equations we use a semi-implicit Euler method: Let τ be the step size, then the fixed point iterations is defined by

$$\begin{aligned} \frac{u_1^{(k+1)} - u_1^{(k)}}{\tau} &= -\sum_{i=1}^{\sigma} \mathcal{B}_i \frac{\partial f_i}{\partial x_1}\left(\frac{\partial f_i}{\partial x_1}u_1^{(k+1)} + \frac{\partial f_i}{\partial x_2}u_2^{(k)} + \frac{\partial f_i}{\partial t}\right) + \\ &\quad \nabla_3 \cdot (\Psi'(|\nabla_3 u_1^{(k)}|^2 + |\nabla_3 u_2^{(k)}|^2)\nabla_3 u_1^{(k)}) \\ \frac{u_2^{(k+1)} - u_2^{(k)}}{\tau} &= -\sum_{i=1}^{\sigma} \mathcal{B}_i \frac{\partial f_i}{\partial x_2}\left(\frac{\partial f_i}{\partial x_1}u_1^{(k)} + \frac{\partial f_i}{\partial x_2}u_2^{(k+1)} + \frac{\partial f_i}{\partial t}\right) \\ &\quad + \nabla_3 \cdot (\Psi'(|\nabla_3 u_1^{(k)}|^2 + |\nabla_3 u_2^{(k)}|^2)\nabla_3 u_2^{(k)}) \end{aligned} \tag{7.15}$$

where for the discretization of $\nabla_3 \cdot (\Psi'(|\nabla_3 u_1^{(k)}|^2 + |\nabla_3 u_2^{(k)}|^2)\nabla_3 u_1^{(k)})$ and $\nabla_3 \cdot (\Psi'(|\nabla_3 u_1^{(k)}|^2 + |\nabla_3 u_2^{(k)}|^2)\nabla_3 u_2^{(k)})$ we follow [142].

In our experiments we use $\tau = 10^{-3}$. Moreover, we set the regularization parameter $\alpha = 40$ and $\lambda = 1$, unless stated otherwise. The iterations are stopped, when the Euclidean norm of the relative error

$$\frac{|u_j^{(k)} - u_j^{(k+1)}|}{|u_j^{(k)}|}, \qquad j = 1, 2$$

drops below the precision tolerance value of $tol = 0.003$ for both the components u_j. For the discretization of (7.15), we use central difference approximations of $\frac{\partial f_i}{\partial x_1}, \frac{\partial f_i}{\partial x_2}, \frac{\partial f_i}{\partial t}$ for each pixel. We consider that the spacing in the central differences approximations equals a value of one with reference to space and time.

7.3 Eye tracking experiment

We next test our model. This is done by recording human participants' fixations on small video clips of relatively total views of natural scenes, and testing how much of variance of fixation locations could be explained by our model as compared to two alternative models, [52] and [122]. To this end, model performances are compared using the areas under the curves (AUCs) and normalized scanpath saliency

Participants

Twenty-four (five female) human viewers with a mean age of 25 years (range 19–32) volunteered in an eye tracking experiment and received partial course credit in exchange. All were undergraduate Psychology students at the University of Vienna. Viewers were pre-screened for normal or fully corrected eye-sight and intact color vision. Prior to the start of the experiment, written informed consent was obtained from all participants.

Stimuli

The same 71 short video recordings that were used for performing the saliency computations were presented to the sample of human viewers. All videos were presented without sound. Each of the videos contained moving and potentially interesting content at several spatially distinct locations off-center (e.g., moving cars, wind moving trees, people crossing streets etc.). Hence, videos presented viewers with multiple potentially interesting fixation locations and the viewers could visually explore the scene. This distinguished the videos used here from professionally produced footage from TV shows or feature films, which often elicit strong tendencies to keep fixation at the center of the movie scene [37]. The order in which the videos were presented in the experiment was chosen randomly for each participant. All videos were presented in full screen on a CRT monitor.

Apparatus

Throughout each data acquisition session, the viewer's dominant eye position was recorded using an EyeLink 1000 Desktop Mount (SR Research Ltd., Kanata, ON, Canada) video-based eye tracker sampling at 1000 Hz. The eye tracker was calibrated using a standard 9-point calibration sequence. Prior to each individual video, a fixation circle was presented at the center of the screen to perform a drift check. Whenever the acquired gaze position differed by more than 1° from the fixation target's position, the whole calibration sequence was repeated to assure maximal spatial accuracy for each viewer. Video stimuli were delivered in color to a 19-in. CRT monitor (Multiscan G400, Sony Inc.) at a screen resolution of 1280×1024 pixels (85 Hz refresh rate). Viewers sat in front of the monitor and placed their head on a chin and forehead rest, which held viewing distance fixed at 64 cm, resulting in an apparent size of each full-screen video of $31 \times 24.2°$. The presentation procedure was implemented in MATLAB with the PsychToolbox and the Eyelink toolbox functions [25, 32, 103].

Data preprocessing

For the evaluation of the model results we compared the spatial distribution of the human viewers' fixations on each video frame with the computed dynamic saliency maps for each frame. The recorded gaze position vector was parsed into three classes of oculomotor events: blinks, saccades, and fixations. Fixations were defined as the mean horizontal and vertical gaze position during data segments not belonging to a blink, or a saccade (gaze displacement $< 0.1°$, velocity $< 30°/s$, and acceleration $< 8,000°/s^2$). The parsed fixations were mapped onto each video frame depending on their start and end times. For example, if a fixation started 1.25 s after the onset the scene at location (x,y), this location was marked as fixated in a fixation matrix belonging to the 30th frame of the video. If this same fixation ended 2 s after the onset of the video, the corresponding fixation matrix from the 30th through to the 50th frame were set to true (or fixated) at that location. This mapping was done for all viewers, and all videos, resulting in a 3-dimensional fixation matrix with the spatial resolution of one video frame and the temporal extent of the number of analyzed video frames. Each video was presented for 10 s during the data collection.

7.4 Results

Qualitative model evaluation

This section is devoted to the evaluation of the dynamical saliency mapping by comparison with eye tracking data. High spatial saliency should correspond to active visual attention. Particular emphasis is put on the participants' tracking of moving objects which are temporarily occluded, because this is a situation where standard optical flow algorithms fail although humans try to actively track such objects [4, 45, 118] that is, they attend to such temporarily occluded objects. We compare our model (7.13) with saliency complemented data, with a standard optical flow algorithm [122] and with (7.13) without complemented

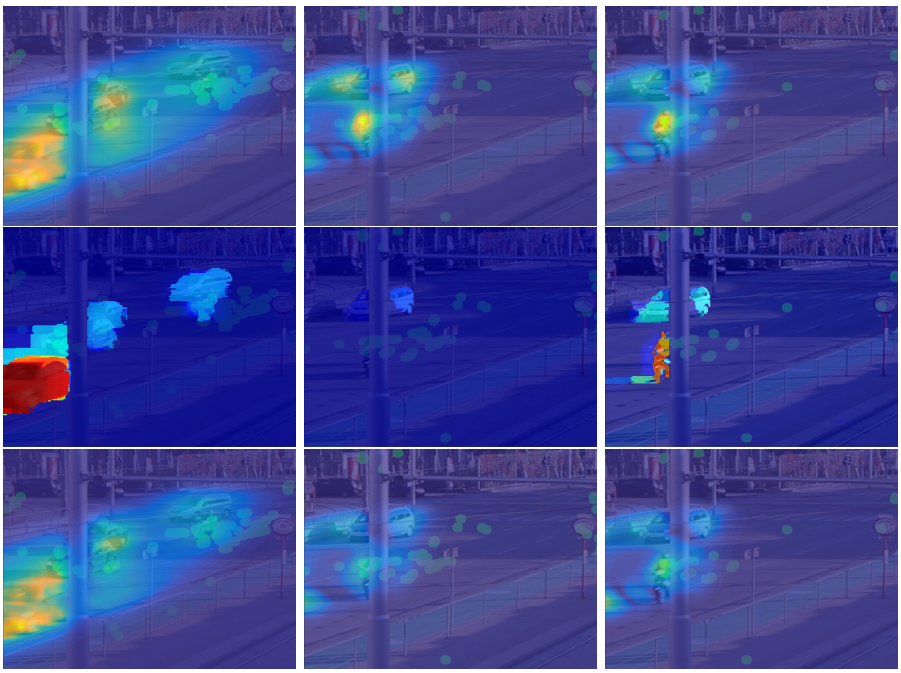

Figure 7.4: The dynamic sequence shows traffic in a public street. A motorcycle is riding behind a pole and it is an object of interest. We notice that when the motorcycle is occluded (central column), only (7.13) with saliency complemented data (top) is able to recognize the occluded part as salient. The method considering only two frames [122] (middle) does not recognize this area as salient. The method (7.13) without complemented data (bottom) results in a lower correlation between the saliency map and the fixation distribution.

data. This last approach is shown to highlight the effect of complementing data with spatial saliency on the calculation of a dynamic saliency map. In order to make a fair comparison between methods, for both our model (7.13) with gray valued images complemented with saliency and the model of [122] we set the amount of regularization (i.e. α) to the value of forty. We set $\alpha = 30$ for (7.13) using color valued images complemented with saliency. Finally, for (7.13) with both types of complemented data, we set the factor enforcing smoothness over time (i.e λ) to the value of ten. The parameters for (7.13) without complemented data are like the one for (7.13) with complemented data. In Figure 7.4, 7.5 and 7.6 we present three sequences with occlusions. For each sequence we show the results of the models in one frame before, one during, and one after the occlusion. On every depicted frame, we superimposed partecipants' fixations, with green dots, of the last five frames before until the last five frames after the depicted frame. For the proposed method (7.13) with complemented data or without complemented data (see Figures 7.4, 7.5 and 7.6 [top and bottom]), the resulting saliency maps are similar for gray valued or color valued images. Therefore, we display only the gray valued version.

The first video in Figure 7.4 shows a motorcycle riding behind a pole. We note that people are looking at the pole in order to follow the motion during

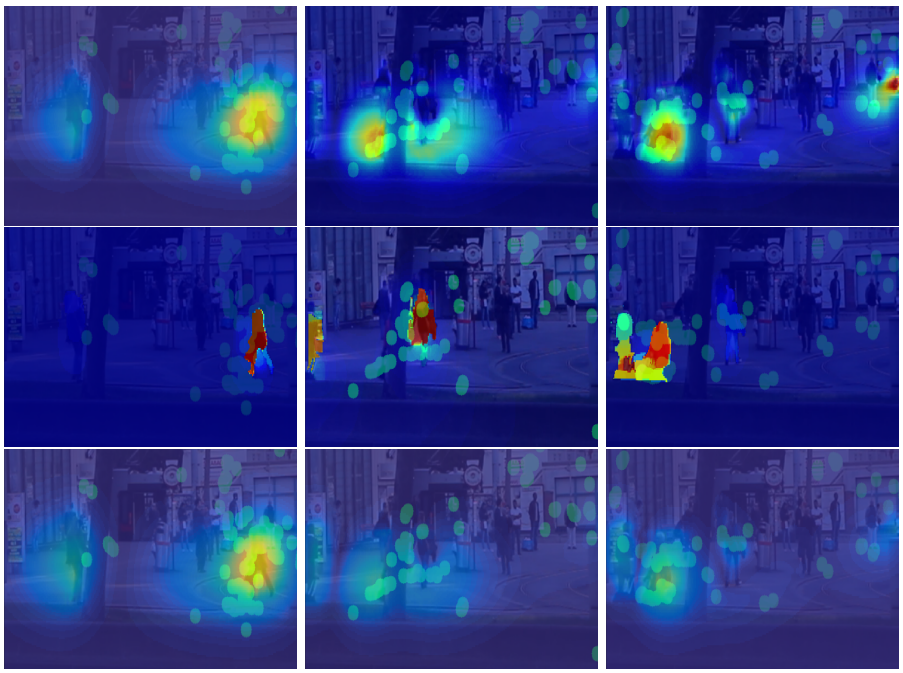

Figure 7.5: The dynamic sequence shows people walking in front of a metro station. A woman in a red coat is walking behind a tree and is salient. We notice that when the woman is occluded (central column), only (7.13) with saliency complemented data (top) is able to recognize the occluded part as salient. The method considering only two frames [122] (middle) does not recognize this area as salient. The method (7.13) without complemented data (bottom) results in a lower correlation between the saliency map and the fixation distribution.

the occlusion. A two-frame method such as [122] does not recognize the pole as salient, while our method (7.13) does so for the occluding parts of the pole (see Figure 7.4 [top]). Moreover, (7.13) without complemented data recognizes the occlusion as salient, but the area is not as strongly marked as attractive for attention compared to (7.13) with complemented data. Indeed, we are not able to predict if the people are looking at the car or at the motorcycle in Figure 7.4. Therefore, the usage of complemented data in (7.13) results in a better fit to the measured fixations compared to (7.13) without complemented data (see Figure 7.5 [top and bottom]) We notice moreover that the method in [122] discards the information regarding the motorcycle in the central column of Figure 7.4.

In the second video in Figure 7.5, we see a woman running which is occluded by a tree for some moments. The woman is highly salient due to the strong color contrast of her red coat. Her saliency is not recognized by [122] while the woman is being occluded by a tree. Using both types of complemented data, our method (7.13) recognizes her saliency correctly. We notice how a two-frame method like [122], in the central column of Figure 7.5, is affected by

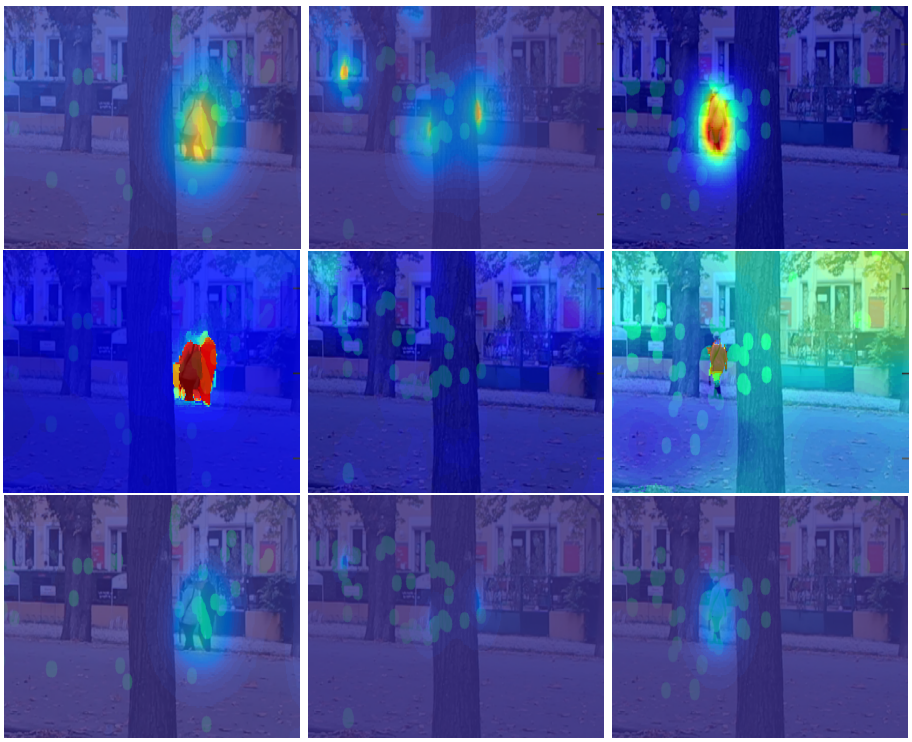

Figure 7.6: The dynamic sequence shows a couple walking in a park. They are walking behind a tree. We notice that when the couple is occluded (central column), only (7.13) with saliency complemented data (top) is able to recognize the occluded part as salient. The method considering only two frames [122] (middle) does not recognize this area as salient. The method (7.13) without complemented data (bottom) results in a lower correlation between the saliency map and the fixation distribution. This sequence is particularly challenging due to light reflections and noise. We notice how the method in [122] (middle) is affected by over-smoothing.

over-smoothing, which results in a wrong interpretation. This image illustrates that the temporal coherence inherent in (7.13) but lacking in [122] makes the method more robust against over-smoothing. Also in this experiment like in the previous one, the usage of complemented data results in a saliency map more correlated to the gaze points compared to the one without complemented data (see Figure 7.5 [top and bottom]).

Finally, the third video shows a couple walking behind a tree. This sequence is particularly challenging because it includes many points with light reflections and noise. We notice that this does not affect the proposed model (7.13), but it influences the result of a two-frame model like [122]. This outcome shows that using temporal coherence, as is done in our model (7.13), results in a robust method, usable for real video sequences. The saliency of the walking couple is correctly recognized by our model (7.13) using complemented data (see Figure 7.6 [top]). Moreover, the model marks the tree section as potentially attractive

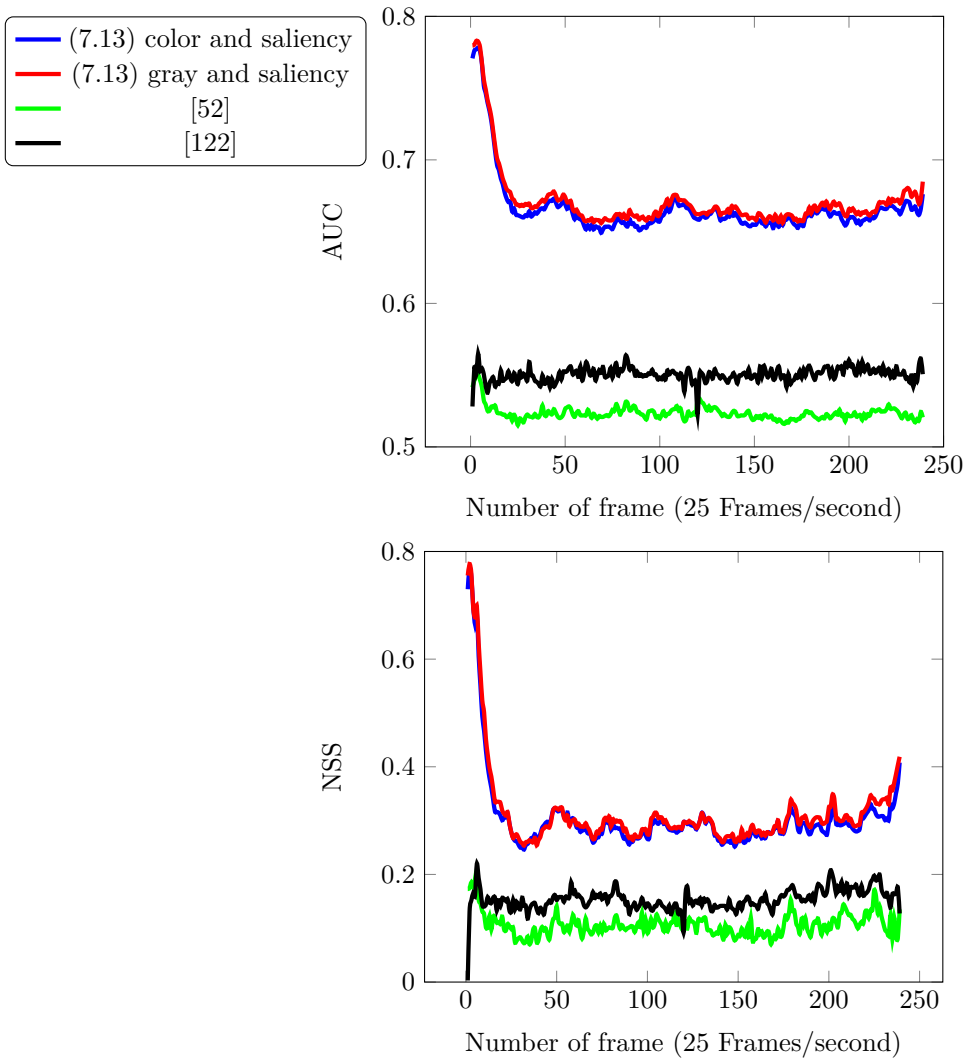

Figure 7.7: Comparison according to AUC (top) and NSS (bottom) for the proposed approach (7.13) with complemented data, the motion as modeled in [52] and a standard optical flow algorithm [122]

for attention, a region that the participants indeed fixate during the occlusion (see Figure 7.6 central column [top]). We notice that in all these experiments the proposed model (7.13) with complemented data recognizes the salient part of the sequences more clearly compared to (7.13) without complemented data and to [122].

Model evaluation measures

As comparison metrics we chose [110]: area under the curve (AUC) of the receiver operating characteristic (ROC) and normalized scanpath saliency (NSS). The AUC treats the saliency map as a classifier. Starting from a saliency map,

the AUC algorithm labels all the pixels over a threshold as True Positive (TP) and all the pixels below as False Positive (FP). Human fixations are used as ground truth. This procedure is repeated one hundred times with different threshold values. Finally, the algorithm can estimate the ROC curve and compute the AUC score. A perfect prediction corresponds to a score of 1.0 while a random classification results in 0.5.
The NSS was introduced by [105]. For each point along a subject scan-path we extract the corresponding position p and a partial value is calculated:

$$NSS(p) = \frac{SM(p) - \mu_{SM}}{\sigma_{SM}} \tag{7.16}$$

where SM is the saliency map. In (7.16) we normalize the saliency map in order to have zero mean and unit standard deviation. The final NSS score is the average of $NSS(p)$:

$$NSS = \frac{1}{N} \sum_{p=1}^{N} NSS(p) \tag{7.17}$$

where N is the total number of fixation points. The NSS, due to the initial normalization of the saliency map, allows comparison across different subjects. For this measurement the perfect prediction corresponds to a score of 1.0.

In this paper, for the implementation of AUC and NSS we follow [66].
These measures are used to test the prediction performances of our *new* dynamic saliency map. In other approaches [81, 52, 60, 99, 104, 148] the dynamic saliency maps are combined through a chosen weighting scheme [81, 99] with spatial saliency maps and then the resulting saliency map tested. Here, we do not discuss the choice of a weighting scheme and we test directly the dynamic saliency map.
In Figure 7.7, we compare the proposed model (7.13) using the two types of complemented data, the motion as modeled in [52] and a standard optical flow algorithm [122]. We set the regularization parameter $\alpha = 40$ for [122]. We notice in Figure 7.7 that the proposed model (7.13) performs similarly with both types of complemented data. Moreover, it performs better than the other models considered. It is worth noticing that two frames models such as [122] and [52] have bad performances. This is coherent with our previous results. The algorithms of [122] and [52] discards more information than (7.13) (as the motorcycle in the central column of Figure 7.4, second row). Moreover, the models of [122] and of [52] fail in particular cases as the occlusion one described in the previous paragraph.

7.5 Conclusion

We have proposed a novel dynamic saliency map based on a variational approach and optical flow computations. The framework is applicable to every type of spatial saliency algorithm and results in significant improvements of model performance with regard to predicting human fixations in videos. We analyzed the possibility to use gray valued images or color valued images complemented with spatial saliency as input of our model. Finally, we studied the contribution of temporal coherence for calculating dynamic saliency maps and

presented an application regarding occlusions. The results underline better performances (AUC and NSS) explaining visual attention compared to other approaches in literature [52, 122].

7.6 Acknowledgements

This work is carried out within the project *"Modeling Visual Attention as a Key Factor in Visual Recognition and Quality of Experience"* funded by the Wiener Wissenschafts und Technologie Fonds - WWTF (Grant no. CS11-009 to UA and OS). OS is also supported by the Austrian Science Fund - FWF, Project S11704 with the national research network, NFN, Geometry and Simulation. The authors would like to thank P. Elbau for interesting discussions on optical flow. The authors also thank R. Seywerth for help with creating the stimulus videos and collecting the eye tracking data.

The use of complemented data results in a more reliable optical flow

We would like to emphasize that the purpose of our model (7.13) is to estimate the dynamic saliency map of a movie sequence and not calculating a precise optical flow. As shown in Table 7.1, if we use saliency complemented data in (7.13), we can calculate the optical flow in imaging regions without regularization. In turn this means that by taking this into account we can reduce the

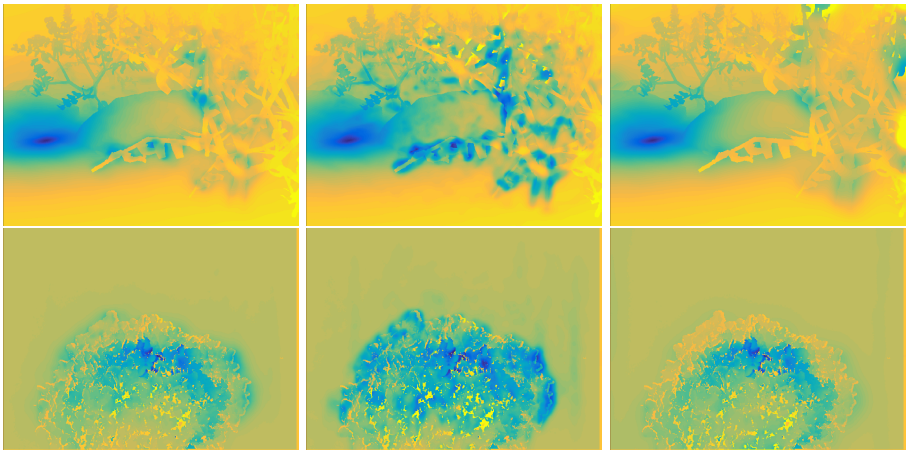

Figure .8: We test the reliability of the optical flow comparing (7.13) using gray valued images complemented with spatial saliency (left column), (7.13) using color valued images complemented with spatial saliency (central column) and (7.13) using color valued images not complemented (right column). The used comparison measure is the average angular error [15] for the Grove 3 (top) and the Hydrangea (bottom) sequences from the Middleburry dataset. Blue shade color indicates points with small average angular error. For these experiments we set $\tau = 0.01$ and $\alpha = 0.8$, $\alpha = 0.01$ and $\alpha = 1.5$, respectively.

amount of regularization (smaller α) in optical flow minimization algorithms. We test the assumptions above for two sequences of the Middlebury dataset [15]. We compare the flow obtained with our complemented data, with the true one, also called *ground truth* for these sequences. We use the average angular error [17] as comparison measure. In Figure .8, areas colored in shades of blue are the ones for which the average angular error is small. This means that in these areas, the flow is close to the true one.

In accordance with the results in Table 7.1, we set α to a value of: 0.8 for gray scale data complemented with saliency, 0.01 for color valued images complemented by saliency, and 1.5 for color valued images. Ideally, we should obtain similar areas with small error.

We notice in Figure .8 that the result we obtain by using color valued images complemented with saliency outperforms the other approaches. Moreover, the model (7.13) that uses gray valued images complemented with saliency performs slightly performs better than the one using only color valued images without saliency. This exemplary test confirms our results in Table 7.1.

Chapter 8

Infinite dimensional optimization models and pdes for dejittering

Authors & Contributions The authors are Guozhi Dong, Aniello Raffaele Patrone, Otmar Scherzer and Ozan Öktem. The development of this article was a collaborative process, and each of the authors made significant contributions to every aspect of the paper.

Publication Status Published [36]: G. Dong, A. R. Patrone, O. Scherzer and O. Öktem. Dimensional Optimization Models and PDEs for Dejittering. In J.F. Aujol, M. Nikolova and N. Papadakis editors, *SSVM'15: Proceedings of the fifth International Conference on Scale Space and Variational Methods in Computer Vision*, volume 9087 of *Lecture Notes in Computer Science*, pages 678–689, Berlin, Heidelberg, 2015. Springer-Verlag.

The final publication is available at `http://link.springer.com/`.

Infinite Dimensional Optimization Models and PDEs for Dejittering

Guozhi Dong[1], Aniello Raffaele Patrone[1], Otmar Scherzer[1,2], and Ozan Öktem[3]

[1]Computational Science Center, University of Vienna,
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[2]Radon Institute of Computational and Applied Mathematics,
Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria
[3]Department of Mathematics,
KTH - Royal Institute of Technology, Lindstedtsvägen 25, SE-10044 Stockholm, Sweden

Abstract

In this paper we do a systematic investigation of continuous methods for pixel, line pixel and line dejittering. The basis for these investigations are the discrete line dejittering algorithm of Nikolova and the partial differential equation of Lenzen et al for pixel dejittering. To put these two different worlds in perspective we find infinite dimensional optimization algorithms linking to the finite dimensional optimization problems and formal flows associated with the infinite dimensional optimization problems. Two different kinds of optimization problems will be considered: Dejittering algorithms for determining the displacement and displacement error correction formulations, which correct the jittered image, without estimating the jitter. As a by-product we find novel variational methods for displacement error regularization and unify them into one family. The second novelty is a comprehensive comparison of the different models for different types of jitter, in terms of efficiency of reconstruction and numerical complexity.

Keywords: Dejittering, Variational methods, Nonlinear evolution PDEs.

8.1 Introduction

A frequent task in image processing is *dejittering*, which is the process of assigning pixel positions to image data recorded with pixel displacements. Jitter is a type of distortions which arises frequently in signal processing, when the distance (time) between sampling points vary rendering signal errors. A specific form of jitter is line jitter that consists of horizontal shifts of each row (line) of an image. The shift is the same for the entire row. This may typically happen when digitizing analog noisy video frames and there are line registration problems due to bad synchronization pulses. The effect is that

the image lines are (randomly) shifted with respect to their original location, so vertical lines become jagged resulting in a disturbing visual effect since all shapes become jagged. One may also have line pixel jitter where pixels in a row are shifted differently. Finally there is pixel jitter where one also experiences vertical shifts.

The main goal of this paper is to establish relations between discrete and continuous models for dejittering. In particular we consider line, line pixel, and pixel jitter. In the literature these problems have been considered in an infinite dimensional continuous and in a finite dimensional discrete setting, resulting in different problem formulations and analysis. To link these approaches and put the theory on solid grounds (based on an infinite dimensional - discretization free - theory) we require to link the approaches.

Presently there exists two kind of algorithms for dejittering which we catalog as follows:

- *Dejittering algorithms* find the displacements by an optimization routine first and then restore the image by composing the jittered image with the displacement.

- *Displacement correction algorithms* compute the image directly without calculating the displacement function first.

The algorithms will be implemented for different purposes: For dejittering we assume a deterministic jitter, while in the later we assume a random perturbation.

Starting point of this paper are publications in different worlds, which deal with dejittering: The discrete optimization formulation of Nikolova [95, 96] and Lenzen et al [74, 75], which deals with displacement correction. We are generalizing Nikolova's algorithm to the infinite dimensional setting and then establish a relation to displacement correction and systems of partial differential equations.

As a consequence we can discuss advantages and shortcuts of the different methods and discretization dependence.

The outline of this paper is as follows: In Section 8.2 we make the basic problem formulation for three types of jittering. Then we explain line dejittering and recall the standard formulation in the field from Nikolova [96] in Section 8.3. After deriving a continuous variant, we put this algorithm in perspective with displacement error regularization [50, 74, 75, 114, 115]. We explain the different philosophies but show the close relation of these areas in the general setting of line pixel dejittering; cf. Section 8.4. Moreover, we review continuous algorithms for pixel dejittering in Section 8.5. In Section 8.6 we formulate partial differential equations, which constitute the flows according to the continuous optimization energies. Finally we present numerical results in Section 8.7. The paper ends with a conclusion, where we outline the novelties of this work.

8.2 Basic Notation and Problem Formulation

In this paper we use the following notations:

- u can either denote a discrete (digital) gray valued image, in which case it is represented as a matrix $u \in \mathbb{R}^{m\times n}$, where m is number of columns, and n is number of rows, or
- u denotes a function $u : \Omega \to \mathbb{R}$ on the unit-square $\Omega = [0,1]^2$. For a continuous image $u : \Omega \to \mathbb{R}$, one way to have the digitized image pixels is
$$u_{ij} = \frac{1}{h_x h_y} \int_{(i-1)/m}^{i/m} \int_{(j-1)/n}^{j/n} u(x,y)\, d(x,y) \; .$$
Here, the pixel size is $h_x \times h_y$, with $h_x = \frac{1}{m}$ and $h_y = \frac{1}{n}$.
- η_{ij} and $\eta : \Omega \to \mathbb{R}$ denote noise. In the discrete setting the lines are horizontally numbered from bottom to top.

Let u^δ denote either a discrete, jittered image - then it is a matrix in $\mathbb{R}^{m\times n}$, or a continuous, jittered image, then it is function $u^\delta : \Omega \to \mathbb{R}$. Assuming that u denotes the original image without jittering, we consider the following discrete and continuous problem formulations:

Line jitter:

$$u^\delta(i,j) = u(i + \mathtt{d}_j, j) + \eta_{ij} \; , \quad u^\delta(x,y) = u(x + \mathtt{d}(y), y) + \eta(x,y) \; , \tag{8.1}$$

respectively, where $\mathtt{d}_j \in \mathbb{Z}$ denotes the discrete jitter of the j-th line, and $\mathtt{d} : [0,1] \to \mathbb{R}$ denotes the jitter function of the y-th component.

Line pixel jitter:

$$u^\delta(i,j) = u(i+\mathtt{d}_{i,j}, j)+\eta_{ij} \; , \quad u^\delta(x,y) = u((x+\mathtt{d}(x,y), y)+\eta(x,y) \; , \tag{8.2}$$

respectively, where $\mathtt{d}_{i,j} \in \mathbb{Z}$ denotes the discrete jitter of the $i-th$ pixel in the j-th line, and $\mathtt{d} : \Omega \to \mathbb{R}$ denotes the jitter function of the point (x,y) in x-direction.

Pixel jitter:

$$u^\delta(i,j) = u((i,j) + \mathtt{d}_{i,j}) + \eta_{ij} \; , \quad u^\delta(x,y) = u((x,y) + \mathtt{d}(x,y)) + \eta(x,y) \; , \tag{8.3}$$

respectively, where $\mathtt{d}_{i,j} \in \mathbb{Z}^2$ denotes the discrete jitter of the $(i,j)-th$ pixel, and $\mathtt{d} : \Omega \to \mathbb{R}^2$ denotes the jitter vector field at the point (x,y).

For those jittered pixels which run out of the domain of the original image u, we define their intensity values as 0.

In the literature, many dejittering algorithms are particularly designed for line jittering, referring to (8.1), see for instance [68, 67, 95, 96, 120]. In these algorithms, the jittering error is considered deterministic, and a probably noisy input image has to be smoothed in an additional step, either before or after dejittering. The problems of line pixel jitter (8.2) and pixel jitter (8.3) have been discussed for instance in [74, 75], where a displacement error correction model has been considered. In this context, it is commonly assumed that noise is significant and jitter is stochastic, and the methods are supposed to dejitter and denoise simultaneously.

8.3 Line Dejittering

In this section we investigate algorithms for line dejittering. After reviewing algorithms from the literature, we will formulate line pixel and pixel dejittering below.

As we have mentioned in the introduction, there are two different kinds of algorithms for dejittering in the literature. The prime example of the first type approach is Nikolova's algorithm [95, 96], which is outlined below. A-priori Nikolova's approach is formulated in a discrete setting. We provide a continuous formulation below, which allows us to put it in perspective with the second approach, and thus in turn to partial differential equation models in the spirit of [74, 75].

Nikolova's Algorithm for Discrete Line Dejittering

Nikolova [95, 96] proposed an efficient algorithm for discrete line dejittering. This algorithm is based on energy minimization and determines in an iterative way, from bottom to top, for each horizontal image line discrete integer values $\mathtt{d}_j, j \in \{1, 2, \cdots, n\}$, which indicate the horizontal displacement of the j-th line, respectively.

The algorithm involves setting values of an exponential parameter p, which Nikolova chooses as $p = 1$ or $p = 0.5$, $p = 0.5$ is better suited for discontinuous images, while $p = 1$ is better suited for smooth images. Moreover, it is assumed that the jitter is bounded, such that there is a parameter σ constraining the maximal line jitter (a typical values is $\sigma = 6$ pixels):

$$|\mathtt{d}_j| \leq \sigma\,, \qquad \forall j = 2, \ldots, n\,.$$

1. The algorithms is initialized by setting $j := 2$, $\mathtt{d}_1 := 0$, $\hat{u}(i,1) := u^\delta(i,1)$ and selecting the parameter $\sigma^* \geq \sigma$. The minimizer $\hat{\mathtt{d}}_2$ of the functional

$$\mathcal{J}_2(\mathtt{d}_2) := \sum_{i=\sigma^*+1}^{m-\sigma^*} \left|u^\delta(i - \mathtt{d}_2, 2) - u^\delta(i,1)\right|^p \tag{8.4}$$

   is used to define $\hat{u}(i,2) := u^\delta(i - \hat{\mathtt{d}}_2, 2)$.

2. For $j = 3, \ldots, n$ determine $\hat{\mathtt{d}}_j$ as the minimizer of the functional

$$\mathcal{J}_j(\mathtt{d}_j) := \sum_{i=\sigma^*+1}^{m-\sigma^*} \left|u^\delta(i - \mathtt{d}_j, j) - 2\hat{u}(i, j-1) + \hat{u}(i, j-2)\right|^p\,, \tag{8.5}$$

   and define $\hat{u}(i,j) = u^\delta(i - \hat{\mathtt{d}}_j, j)$.

A Continuous Optimization Problem for Line Dejittering

We here formulate a continuous variant of Nikolova's algorithm, which also establishes the relation to existing variational methods and partial differential equations for dejittering. Let $u^\delta : \Omega \to \mathbb{R}$ be the line jittered variant of u, so u^δ satisfies (8.1). In order to recover u and $\mathtt{d}$, we minimize (8.6) for each $\hat{y} \in [0,1]$ separately, where $\hat{y}$ indicates the continuum position of the line in the image,

$$\mathcal{J}_c(\mathtt{d})(\hat{y}) := \lim_{\tau \to 0^+} \frac{1}{2\tau} \int_{\max\{\hat{y}-\tau,0\}}^{\min\{\hat{y}+\tau,1\}} \int_{\sigma_*}^{1-\sigma_*} \left|\partial_y^k u^\delta(x - \mathtt{d}(y), y)\right|^p \, d(x,y)\,, \quad (8.6)$$

subject to

$$\|\mathtt{d}\|_{L^\infty([0,1])} \leq \sigma. \quad (8.7)$$

The parameter σ_* is chosen to satisfy $\sigma \leq \sigma_*$. With this choice the integrand in the integral $\int_{\sigma_*}^{1-\sigma_*} \left|\partial_y^k u^\delta(x - \mathtt{d}(y), y)\right|^p \, dx$ is evaluated only for arguments of u^δ in the interior of the image domain $[0,1] \times [0,1]$. This correspond to the discrete sum $\sum_{i=\sigma_*+1}^{m-\sigma_*}$ in the Nikolova algorithm. The term $\partial_y^k u^\delta$ denotes the k-th derivative of u^δ with respect to the second component. Since

$$\frac{u^\delta(i - \mathtt{d}_j, j) - 2\hat{u}(i, j-1) + \hat{u}(i, j-2)}{h_y^2} \approx \partial_y^2 u^\delta(ih_x - \mathtt{d}((j-1)h_y), (j-1)h_y)\,,$$

we propose the following simplified variant of (8.6) and (8.7), namely to minimize

$$\mathcal{J}^{(k)}(\mathtt{d}) := \frac{1}{p} \int_\Omega \left|\partial_y^k u^\delta(x - \mathtt{d}(y), y)\right|^p \, d(x,y) \quad (8.8)$$

subject to

$$\|\mathtt{d}\|_{L^2([0,1])} \leq \hat{\sigma}. \quad (8.9)$$

The main difference to minimizing $\mathcal{J}_c$ is that we consider integration over all of Ω. To make this well-defined, we propose to extend u^δ symmetric across left and right, and top and bottom images boundaries, respectively. Another difference is that we consider an *a joint* approach, which optimizes globally over all pixels, instead of separately for each line. Moreover, from a modelling point of view taking the second derivative ($k = 2$) of u^δ in the functional $\mathcal{J}_c$ is not mandatory, for instance, we may take as well the derivative ($k = 1$) or another integer order. In practice, minimizing the functional with second order derivatives performs better than using first order derivatives in a noise free environment. For the other parameter p in (8.8), in the discrete setting, Nikolova has suggested to use either 0.5 or 1, however, we would propose to choose either $p = 1$ or $p = 2$, in order to keep the convexity of the functional in our continuous model, where $p = 1$ works better with the discontinuities.

8.4 Line Pixel Dejittering

In this section we review line pixel dejittering and displacement regularization: We find that within the continuous setting, formally, the optimization approach for line dejittering from last section can be similarly generalized to the case of line pixel dejittering. However, the formal difference is that for line pixel dejittering $\mathtt{d} : \Omega \to \mathbb{R}$ is a bounded random field over the whole two dimensional domain Ω, while for line jitter $\mathtt{d} : [0,1] \to \mathbb{R}$. Thus, we propose to optimize the functional which is only slightly changed from (8.8)

$$\mathcal{J}_2^{(k)}(\mathtt{d}) := \frac{1}{p} \int_\Omega \left|\partial_y^k u^\delta(x - \mathtt{d}(x,y), y)\right|^p \, d(x,y) \quad (8.10)$$

subject to $\|\mathtt{d}\|_{L^2(\Omega)} \leq \hat{\sigma}$.

Because we assume small displacements $\mathtt{d}$, we also consider approximating the term $\partial_y^k u^\delta(x - \mathtt{d}(x,y), y)$ by its linearisation:

$$\partial_y^k u^\delta(x - \mathtt{d}(x,y), y) \approx \partial_y^k u^\delta(x,y) - \mathtt{d}(x,y)\partial_x \partial_y^k u^\delta(x,y) \, .$$

Replacing the nonlinear term by its linearization, we arrive at the constrained optimization problem, which is to minimize

$$\mathcal{J}_2^{(k)}(\mathtt{d}) := \frac{1}{p} \int_\Omega \left| \partial_y^k u^\delta(x,y) - \mathtt{d}(x,y)\partial_x \partial_y^k u^\delta(x,y) \right|^p \, d(x,y) \, , \quad k = 1,2 \tag{8.11}$$

subject to (8.9).

For $1 < p \leq 2$, $\mathcal{J}_2^{(k)}$ is strictly convex, and for three-times continuously differentiable u^δ also weakly lower semi-continuous. Then, the constrained optimization problem is equivalent to the method of Tikhonov-regularization with parameter choice by Morozov's discrepancy principle, consisting in calculation of

$$\mathtt{d}(\alpha) := \arg\min_{\mathtt{d}} \left\{ \mathcal{J}_2^{(k)}(\mathtt{d}) + \frac{\alpha}{2} \left\| \mathtt{d} \right\|_{L^2(\Omega)}^2 \right\} \, , \tag{8.12}$$

where α is chosen to satisfy $\left\| \mathtt{d}(\alpha) \right\|_{L^2(\Omega)} = \hat{\sigma}$. For further background on the relation between Tikhonov regularization and constrained optimization problems see for instance [41, 93, 51, 90, 91, 116, 126, 127]. For $p \leq 1$ the relation is not obvious, but we ignore this difficulty.

We stress the fact that the minimizer of (8.12) with $p = 2$ can be explicitly calculated: We have

$$\mathtt{d}(\alpha) = \frac{\partial_y^k u^\delta \partial_x \partial_y^k u^\delta}{\alpha + (\partial_x \partial_y^k u^\delta)^2} \, . \tag{8.13}$$

This explicit linearised method provides insufficient results (cf. Figure 8.1).

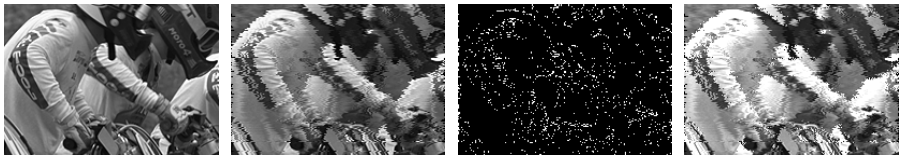

Figure 8.1: Left to right: ground truth, line jittered image, displacement, recovered image

Displacement Error Correction for Line Pixel Dejittering

In the following we outline an approach for dejittering, which does not recover the jitter but the dejittered image directly. We use a first order approximation of the data by assuming that the jitter is only a small disturbance:

$$u^\delta(x,y) \approx u(x + \mathtt{d}(x,y), y) \approx u(x,y) + \partial_x u(x,y)\mathtt{d}(x,y) \, . \tag{8.14}$$

Considering the approximation as an identity we find that

$$\mathtt{d}(x,y) = \frac{u^\delta(x,y) - u(x,y)}{\partial_x u(x,y)} . \tag{8.15}$$

Now, instead of minimizing $\mathcal{J}_2^{(k)}$ with respect to $\mathtt{d}$, we replace in $\mathcal{J}^{(k)}$ the u^δ by $u(x+\mathtt{d}(x,y))$ and use the identity (8.15), and minimize with respect to u. Thus the optimization problem for line pixel dejittering consists in the minimization of the functional:

$$\mathcal{N}(u) := \alpha \frac{1}{2} \int_\Omega \left| \frac{u^\delta(x,y) - u(x,y)}{\partial_x u(x,y)} \right|^2 \, d(x,y) + \frac{1}{p} \underbrace{\int_\Omega \left| \partial_y^k u(x,y) \right|^p \, d(x,y)}_{\mathcal{R}} \, . \tag{8.16}$$

Remark 2. *When we use this approach to correct for line jitter, we have to respect the fact that each line has the same shift, which leads to*

$$0 = \partial_x \mathtt{d}(y) \approx \partial_x \left(\frac{u^\delta(x,y) - u(x,y)}{\partial_x u(x,y)} \right) \, .$$

Thus line jitter correction can be rephrased as an unconstrained minimization of the functional

$$\mathcal{N}(u) + \beta \int_\Omega \left(\partial_x \left(\frac{u^\delta(x,y) - u(x,y)}{\partial_x u(x,y)} \right) \right)^2 d(x,y) \, , \tag{8.17}$$

where β is a penalty parameter.

8.5 Pixel Dejittering

The problem of pixel jitter correction can be formulated again as a constraint optimization problem, consisting in minimization of

$$\mathcal{J}_3^{(k)}(\mathtt{d}) := \frac{1}{p} \int_\Omega \left| \partial_y^k u^\delta((x,y) - \mathtt{d}(x,y)) \right|^p \, d(x,y) \tag{8.18}$$

subject to $\|\mathtt{d}\|_{(L^2(\Omega))^2} \leq \hat{\sigma}$. Note the fundamental difference that $\mathtt{d} : \Omega \to \mathbb{R}^2$, while for line pixel jitter $\mathtt{d} : \Omega \to \mathbb{R}$, and for line jitter $\mathtt{d} : [0,1] \to \mathbb{R}$.

Displacement error regularization for correcting pixel jitter has been considered in [74, 75]. It is again based on Taylor expansion

$$u^\delta(x,y) - u(x,y) \approx \mathtt{d} \cdot \nabla u \, ,$$

which implies that we can choose as a solution $\mathtt{d} \approx (\nabla u)^\dagger (u^\delta - u)$, where $(\nabla u)^\dagger$ denotes the Moore-Penrose pseudo-inverse of ∇u. This choice of an inverse of ∇u considers displacement errors which are orthogonal to level lines of u.

Here, we define

$$\hat{\mathcal{S}}(u) := \frac{1}{2} \left\| (\nabla u)^\dagger (u^\delta - u) \right\|^2_{L^2(\Omega)} \, .$$

Assuming that u is of finite total variation we ended up with the following regularization functional [74, 75]:

$$\hat{\mathcal{N}}(u) := \alpha \hat{\mathcal{S}}(u) + \int_\Omega |\nabla u(x,y)| \, d(x,y) \, . \tag{8.19}$$

Note that in comparison with (8.16), $\int_\Omega \left| \partial_y^k u(x,y) \right|^p \, d(x,y)$ has been replaced by the TV-semi norm $\int_\Omega |\nabla u(x,y)| \, d(x,y)$.

8.6 PDE Models as Formal Energy Flows

Considering $\mathcal{S}$ as a metric, the minimization of functional $\mathcal{N}$ defined in (8.16), can be formally solved as metric flows of $\mathcal{S}$ with energy $\mathcal{R}$. In [75], a PDE according to (8.19) has been derived by considering $\hat{\mathcal{N}}(\alpha, \cdot)$ as an implicit time-step of the associated flow, following that, we state the flows according to (8.16) and (8.19).

- The flow associated with (8.16), for $k = 1, 2$ and $p = 1, 2$ is:

$$\begin{cases} \partial_t u = |\partial_x u|^2 \, \partial_y^k \left(\dfrac{\partial_y^k u}{\left|\partial_y^k u\right|^{2-p}} \right) \; ; \\ u = u^\delta \, , \quad in \; \Omega \times \{0\} \; ; \\ \partial_y^{2l-1} u = 0 \, , \quad on \; \{0,1\} \times [0,1] \, , \quad \forall l = 1, .., k \, . \end{cases} \tag{8.20}$$

- We emphasize that the flow associated to (8.19) is

$$\begin{cases} \partial_t u = |\nabla u|^2 \, \nabla \cdot \left(\dfrac{\nabla u}{|\nabla u|} \right) \; ; \\ u = u^\delta \, , \quad in \; \Omega \times \{0\} \; ; \\ \partial_n u = 0 \, , \quad on \; \partial\Omega \, . \end{cases} \tag{8.21}$$

8.7 Numerical Results

In this section we show the numerical results of our newly developed model (8.20) for different choices of k and p, making comparisons with the approach from [75], that consists in solving (8.21), and with Nikolova's algorithm [96]. In the implementation, for $p = 2$ in (8.20), we use standard finite differences discretization with semi-implicit iteration, but for the case of $p = 1$, the solution of (8.20) is obtained by solving the convex optimization problem (8.22) iteratively, where we generalised the TV denoising algorithm from [43] to approximate the solution.

$$\begin{cases} u^{m+1} := \arg\min_u \left\{ \dfrac{\alpha}{2} \displaystyle\int_\Omega \frac{|u^m(x,y) - u(x,y)|^2}{|\partial_x u^m(x,y)|^2 + \epsilon} + \left|\partial_y^k u(x,y)\right| \, d(x,y) \right\}, \\ u^0 = u^\delta. \end{cases} \tag{8.22}$$

Here α corresponds to the time-stepping and $u^m \approx u(m\alpha)$. In all the experiments, we use as stopping criteria some threshold of $\left\|u^m - u^{m+1}\right\|_{L^2}$. The test data are generated by adding jitter to clean test images. In addition noisy test data are generated by composing the test image with Gaussian noise of mean 0 and standard deviation 10. In order to evaluate the results quantitatively, we consider the *mean square error* (MSE) computed by averaging the intensity difference between the analyzed pixel $\hat{u}(i,j)$ and the reference pixel $u(i,j)$, and the related quantity of *peak signal to noise ration* (PSNR)

$$MSE = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} (\hat{u}(i,j) - u(i,j))^2 \quad \text{and} \quad PSNR = 10 \log_{10} \frac{L^2}{MSE} \, ,$$

Measure	Test data	k=1,p=2	k=2,p=2	k=1,p=1	k=2,p=1	cf.[75]	cf.[96]
Line Jitter Data without Adding Noise							
PSNR	17.814	19.886	20.031	20.109	20.461	19.807	24.818
MSE	1075.7	667.407	645.584	634.035	584.668	679.740	214.408
SSIM	0.622	0.704	0.714	0.709	0.729	0.691	0.998
Line Pixel Jitter Data without Adding Noise							
PSNR	16.608	17.913	17.956	18.193	18.356	19.213	13.999
MSE	1420.0	1051.4	1040.9	985.634	949.517	779.525	2589
SSIM	0.484	0.552	0.558	0.566	0.571	0.618	0.308
Pixel Jitter Data with Adding Noise							
PSNR	15.367	17.460	17.563	17.688	17.891	19.064	-
MSE	1889.8	1167.1	1139.6	1137	1056.6	806.614	-
SSIM	0.316	0.433	0.461	0.457	0.487	0.585	-

Table 8.1: Comparison of noisy and noise free data affected by different jitter types.

where L is the dynamic range of allowable pixel intensities, e.g. for an 8-bit per pixel image $L = 2^8 - 1 = 255$. These quantity are appealing but not well matched to perceived visual quality as reported in [94] and [139]. For that reason we consider also the *structural similarity* (SSIM) index [139] defined as:

$$SSIM(\hat{u}, u) = f(l(\hat{u}, u), c(\hat{u}, u), s(\hat{u}, u)) \,,$$

where the three independent components $l(\hat{u}, u), c(\hat{u}, u), s(\hat{u}, u)$ are the similarity functions of the luminance, the contrast and the structure, respectively, between the reconstructed and test image, and f is a combination function. Quantitatively, the higher *PSNR* value the better similarity between the test data and the original clean image. Moreover, a small value of *MSE* points out a good intensity approximation of the original data, and a larger value of *SSIM* claims that the structure of the original image is better preserved.

Table 8.1 gives a comprehensive evaluation of different methods for image dejittering, which are the algorithm for solving (8.20) presented in this paper,

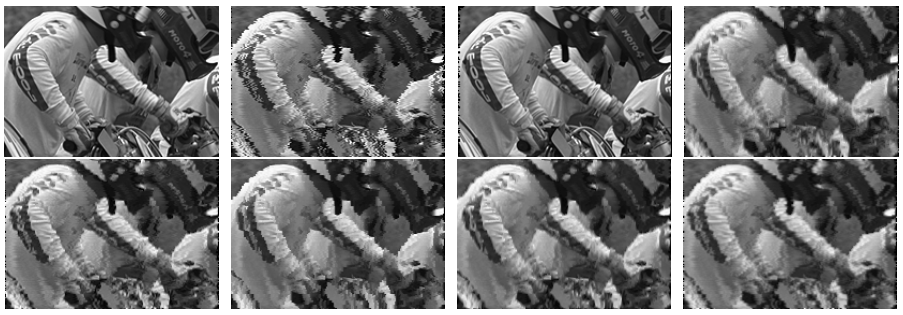

Figure 8.2: Line Dejittering. Top row: The ground truth, the noisy free line jittered image, dejittered with [96], dejittered with (8.20) $k = 1,\ p = 2$. Bottom row: dejittered with (8.20) $k = 2,\ p = 2$, (8.20) $k = 1,\ p = 1$, (8.20) $k = 2,\ p = 1$, approach from [75].

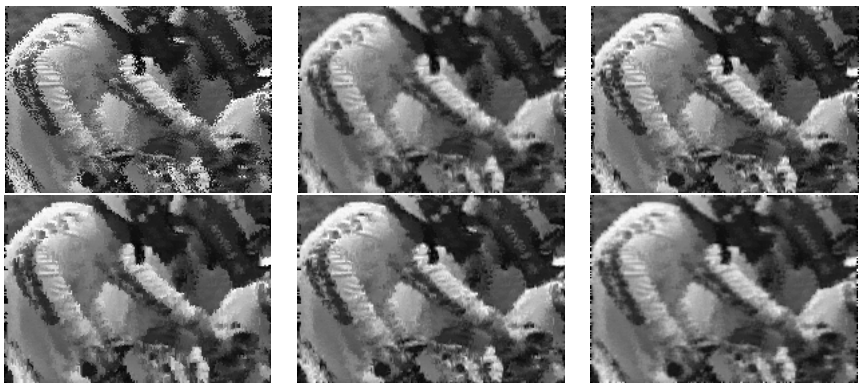

Figure 8.3: Line Pixel Dejittering. Top row: The noisy line pixel jittered image, dejittered with (8.20) $k = 1,\ p = 2$, (8.20) $k = 2,\ p = 2$. Bottom row: (8.20) $k = 1,\ p = 1$, (8.20) $k = 2,\ p = 1$, approach from [75].

and the algorithms from [75] and from [96], respectively. For the test images used for line dejittering and line pixel dejittering, we have not superimposed the data with additive noise. The test data used for pixel dejittered was considered with additive noise. For line dejittering, Nikolova's algorithm [96] gives the most superior results. Evaluating the two different PDE models, we notice that (8.20) performs better than [75] for line dejittering. [96] is not able to handle line pixel dejittering, in contrast with the PDE models. In this case the method in [75] achieves slightly better grades than (8.20); see Table 8.1. However visually, one may find that (8.20) (e.g.with parameter $k = 2, p = 1$) has less blurring of the reconstructed image and keeps more clear details; see Fig 8.3. The highlight of the approach [75] happens in the pixel dejittering task, where it outperforms the others both quantitatively and qualitatively. Over all the tests, it is not hard to find that, for the model (8.20), the choice of parameter $k = 2, p = 1$ gives the most competitive results in compare with

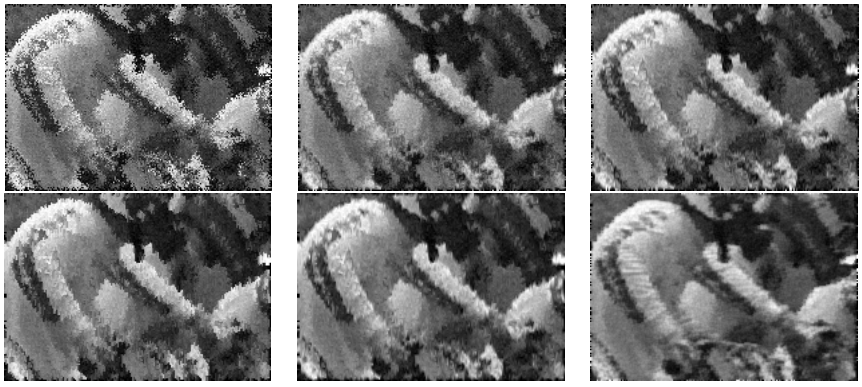

Figure 8.4: Pixel Dejittering. Top row: The noisy line pixel jittered image, dejittered with (8.20) $k = 1,\ p = 2$, (8.20) $k = 2,\ p = 2$. Bottom row: (8.20) $k = 1,\ p = 1$, (8.20) $k = 2,\ p = 1$, approach from [75].

the other parameter choices.

8.8 Conclusion

The novelties of this paper are that we have shown the formal connection of Nikolova's method with variational displacement error correction and PDE methods. To do this, we have unified a family of variational methods for displacement error regularization, which apply for different dejittering applications. The second novelty is a comparison of the different models for different types of jitter. An analysis of the proposed algorithms for minimizing models (8.16) is lacking and this might be a future research topic. Another aspect will be to investigate problems in tomography, which involve reconstruction of objects that show small (unknown) displacements while being imaged.

Acknowledgements

The work of GD and OS has been supported by the Austrian Science Fund (FWF) within the project *Variational Methods for Imaging on Manifolds* within the NFN Geometry and Simulation, project S11704. AP and OS are supported by the project *Modeling Visual Attention as a Key Factor in Visual Recognition and Quality of Experience* funded by the Wiener Wissenschafts und Technologie Funds - WWTF. The authors thank the reviewers for their comments to improve the presentation of the paper.

Part III

Appendix

Bibliography

[1] J. Abhau, Z. Belhachmi, and O. Scherzer. On a decomposition model for optical flow. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 5681 of *Lecture Notes in Computer Science*, pages 126–139. Springer-Verlag, Berlin, Heidelberg, 2009.

[2] R. A. Abrams and S. E. Christ. Motion onset captures attention. *Psychological Science*, 14(5):427–432, 2003.

[3] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, pages 1597–1604. IEEE, 2009.

[4] G. A. Alvarez, T. S. Horowitz, H. C. Arsenio, J. S. DiMase, and J. M. Wolfe. Do multielement visual tracking and visual search draw continuously on the same visual attention resources? *Journal of Experimental Psychology: Human Perception and Performance*, 31(4):643, 2005.

[5] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, 1989.

[6] R. Andreev, O. Scherzer, and W. Zulehner. Simultaneous optical flow and source estimation: Space–time discretization and preconditioning. *Applied Numerical Mathematics*, 96:72–81, 2015.

[7] U. Ansorge, S. Buchinger, C. Valuch, A. R. Patrone, and O. Scherzer. Visual attention in edited dynamical images. In *SIGMAP*, pages 198–205, 2014.

[8] G. Aubert and J.-F. Aujol. Modeling very oscillating signals. Application to image processing. *Applied Mathematics and Optimization. An International Journal with Applications to Stochastics*, 51(2):163–182, 2005.

[9] G. Aubert and P. Kornprobst. *Mathematical problems in image processing: partial differential equations and the calculus of variations*, volume 147. Springer Science & Business Media, 2006.

[10] J.-F. Aujol, G. Aubert, L. Blanc-Féraud, and A. Chambolle. Image decomposition into a bounded variation component and an oscillating component. *Journal of Mathematical Imaging and Vision*, 22(1):71–88, 2005.

[11] J.-F. Aujol and A. Chambolle. Dual norms and image decomposition models. *International Journal of Computer Vision*, 63(1):85–104, 2005.

[12] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher. Structure-texture image decomposition—modeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67(1):111–136, 2006.

[13] J.-F. Aujol and S.H. Kang. Color image decomposition and restoration. *Journal of Visual Communication and Image Representation*, 17(4):916–928, 2006.

[14] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision*, 92(1):1–31, November 2011.

[15] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

[16] M. Bar. The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7):280–289, 2007.

[17] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994.

[18] S. Basapur, H. Mandalia, S. Chaysinh, Y. Lee, N. Venkitaraman, and C. Metcalf. Fanfeeds: evaluation of socially generated information feed on second screen as a tv show companion. In *Proceedings of the 10th European conference on Interactive tv and video*, pages 87–96. ACM, 2012.

[19] M. Bauer, M. Bruveris, and P. W. Michor. Overview of the geometries of shape spaces and diffeomorphism groups. *Journal of Mathematical Imaging and Vision*, 50(1-2):60–97, 2014.

[20] S. I. Becker. Can intertrial effects of features and dimensions be explained by a single theory? *Journal of Experimental Psychology: Human Perception and Performance*, 34(6):1417, 2008.

[21] S. I. Becker and G. Horstmann. Novelty and saliency in attentional capture by unannounced motion singletons. *Acta psychologica*, 136(3):290–299, 2011.

[22] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996.

[23] M. Böhme, M. Dorr, C. Krause, T. Martinetz, and E. Barth. Eye movement predictions on natural videos. *Neurocomputing*, 69(16):1996–2004, 2006.

[24] A. Borzi, K. Ito, and K. Kunisch. Optimal control formulation for determining optical flow. *SIAM journal on scientific computing*, 24(3):818–847, 2003.

[25] D. H. Brainard. The psychophysics toolbox. *Spatial vision*, 10:433–436, 1997.

[26] D. I. Brooks, I. P. Rasmussen, and A. Hollingworth. The nesting of search contexts within natural scenes: evidence from contextual cuing. *Journal of Experimental Psychology: Human Perception and Performance*, 36(6):1406, 2010.

[27] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Computer Vision-ECCV 2004*, pages 25–36. Springer, 2004.

[28] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61(3):211–231, 2005.

[29] B. R. Burnham. Displaywide visual features associated with a search display's appearance can mediate attentional capture. *Psychonomic Bulletin & Review*, 14(3):392–422, 2007.

[30] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision research*, 46(26):4333–4345, 2006.

[31] F. W. Cornelissen, E. M. Peters, and J. Palmer. The eyelink toolbox: eye tracking with matlab and the psychophysics toolbox. *Behavior Research Methods, Instruments, & Computers*, 34(4):613–617, 2002.

[32] F. W. Cornelissen, E. M. Peters, and J. Palmer. The eyelink toolbox: eye tracking with matlab and the psychophysics toolbox. *Behavior Research Methods, Instruments, & Computers*, 34(4):613–617, 2002.

[33] R. Courant and D. Hilbert. *Methods of mathematical physics. Vol. I.* Interscience Publishers, Inc., 1953.

[34] J. E. Cutting, K. L. Brunick, and A. Candan. Perceiving event dynamics and parsing hollywood films. *Journal of experimental psychology: human perception and performance*, 38(6):1476, 2012.

[35] H. Deubel and W. X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*, 36(12):1827–1837, 1996.

[36] G. Dong, A. R. Patrone, O. Scherzer, and O. Öktem. Infinite dimensional optimization models and pdes for dejittering. In *Scale Space and Variational Methods in Computer Vision*, pages 678–689. Springer, 2015.

[37] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, 10(10):28, 2010.

[38] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.

[39] V. Duval, J.-F. Aujol, and L.A. Vese. Mathematical modeling of textures: Application to color image decomposition with a projected gradient algorithm. *J. Math. Imaging Vision*, 37:232–248, 2010.

[40] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya. Visual attention in quality assessment. *IEEE Signal Processing Magazine*, 28(6):50–59, 2011.

[41] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[42] Y. Fang, Z. Wang, W. Lin, and Z. Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *Image Processing, IEEE Transactions on*, 23(9):3910–3921, 2014.

[43] M. A. T. Figueiredo, J. B. Dias, J. P. Oliveira, and R. D. Nowak. On total variation denoising: A new majorization-minimization algorithm and an experimental comparisonwith wavalet denoising. In *Image Processing, 2006 IEEE International Conference on*, pages 2633–2636. IEEE, 2006.

[44] D. Fleet and Y. Weiss. Optical flow estimation. In *Handbook of mathematical models in computer vision*, pages 237–257. Springer, 2006.

[45] J. I. Flombaum, B. J. Scholl, and Z. W. Pylyshyn. Attentional resources in visual tracking through occlusion: The high-beams effect. *Cognition*, 107(3):904–931, 2008.

[46] T. Foulsham, J. T. Cheng, J. L. Tracy, J. Henrich, and A. Kingstone. Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition*, 117(3):319–331, 2010.

[47] S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):6, 2010.

[48] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *Advances in neural information processing systems*, pages 481–488, 2004.

[49] U. Gargi, R. Kasturi, and S. H. Strayer. Performance characterization of video-shot-change detection methods. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(1):1–13, 2000.

[50] M. Grasmair, F. Lenzen, A. Obereder, O. Scherzer, and M. Fuchs. A nonconvex pde scale space. In *Scale Space and PDE Methods in Computer Vision*, pages 303–315. Springer, 2005.

[51] C. W Groetsch. *Inverse problems: activities for undergraduates*. Cambridge University Press, 1999.

[52] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *Image Processing, IEEE Transactions on*, 19(1):185–198, 2010.

[53] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

[54] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640, 2004.

[55] J. M. Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003.

[56] M. E. Holmes, S. Josephson, and R. E. Carney. Visual attention to television programs with a second-screen application. In *Proceedings of the symposium on eye tracking research and applications*, pages 397–400. ACM, 2012.

[57] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[58] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[59] J. A. Iglesias and C. Kirisits. Convective regularization for optical flow. *arXiv preprint arXiv:1505.04938*, 2015.

[60] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 631–637. IEEE, 2005.

[61] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[62] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554, 2005.

[63] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[64] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 20(11):1254–1259, 1998.

[65] A. Jain and L. Younes. A kernel class allowing for fast computations in shape spaces induced by diffeomorphisms. *Journal of Computational and Applied Mathematics*, 245:162–181, 2013.

[66] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. Technical report, MIT tech report, 2012.

[67] S. Kang and J. Shen. Image dejittering based on slicing moments. *Image Processing Based on Partial Differential Equations*, pages 35–55, 2007.

[68] S. H. Kang and J. Shen. Video dejittering by bake and shake. *Image and vision computing*, 24(2):143–152, 2006.

[69] C. Kirisits, L. F. Lang, and O. Scherzer. Decomposition of optical flow on the sphere. *GEM-International Journal on Geomathematics*, 5(1):117–141, 2014.

[70] T. Kohlberger, É. Mémin, and C. Schnörr. Variational dense motion estimation using the helmholtz decomposition. In LewisD. Griffin and Martin Lillholm, editors, *Scale Space Methods in Computer Vision*, volume 2695 of *Lecture Notes in Computer Science*, pages 432–448. Springer Berlin Heidelberg, 2003.

[71] E. Kowler, E. Anderson, B. Dosher, and E. Blaser. The role of attention in the programming of saccades. *Vision research*, 35(13):1897–1916, 1995.

[72] S.-H. Lai and B. C. Vemuri. Reliable and efficient computation of optical flow. *International Journal of Computer Vision*, 29(2):87–105, 1998.

[73] M. Lefébure and L. D. Cohen. Image registration, optical flow and local rigidity. *Journal of Mathematical Imaging and Vision*, 14(2):131–147, 2001.

[74] F. Lenzen and O. Scherzer. A geometric pde for interpolation of m-channel data. In *Scale Space and Variational Methods in Computer Vision*, pages 413–425. Springer, 2009.

[75] F. Lenzen and O. Scherzer. Partial differential equations for zooming, deinterlacing and dejittering. *International Journal of Computer Vision*, 92(2):162–176, 2011.

[76] S. Li and M. C. Lee. Fast visual tracking using motion saliency in video. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–1073. IEEE, 2007.

[77] C. Liu, P. C. Yuen, and G. Qiu. Object motion detection using information theoretic spatio-temporal saliency. *Pattern Recognition*, 42(11):2897–2906, 2009.

[78] S. J. Luck and E. K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, 1997.

[79] Y.-F. Ma and H.-J. Zhang. A model of motion attention for video skimming. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–129. IEEE, 2002.

[80] V. Maljkovic and K. Nakayama. Priming of pop-out: I. role of features. *Memory & cognition*, 22(6):657–672, 1994.

[81] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. Spatio-temporal saliency model to predict eye movements in video free viewing. In *Signal Processing Conference, 2008 16th European*, pages 1–5. IEEE, 2008.

[82] A. M. Maxcey-Richard and A. Hollingworth. The strategic retention of task-relevant objects in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3):760, 2013.

[83] B. McCane, K. Novins, D. Crannitch, and B. Galvin. On benchmarking optical flow. *Computer Vision and Image Understanding*, 84:126–143, 2001.

[84] E. Mémin and P. Pérez. Dense estimation and object-based segmentation of the optical flow with robust techniques. *Image Processing, IEEE Transactions on*, 7(5):703–719, 1998.

[85] E. Mémin and P. Pérez. A multigrid approach for hierarchical motion estimation. In *Computer Vision, 1998. Sixth International Conference on*, pages 933–938. IEEE, 1998.

[86] E. Mémin and P. Pérez. Hierarchical estimation and segmentation of dense motion fields. *International Journal of Computer Vision*, 46(2):129–155, 2002.

[87] Y. Meyer. *Oscillating patterns in image processing and nonlinear evolution equations*, volume 22 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2001. The fifteenth Dean Jacqueline B. Lewis memorial lectures.

[88] Y. Mileva, A. Bruhn, and J. Weickert. Illumination-robust variational optical flow with photometric invariants. In *Pattern Recognition*, pages 152–162. Springer, 2007.

[89] P. Mital, T. J. Smith, S. G. Luke, and J. M. Henderson. Do low-level visual features have a causal influence on gaze during dynamic scene viewing? *Journal of Vision*, 13(9):144–144, 2013.

[90] V. A. Morozov. *Methods for solving incorrectly posed problems*. Springer Science & Business Media, 2012.

[91] V. A. Morozov and M. Stessin. *Regularization methods for ill-posed problems*. Crc Press Boca Raton, FL:, 1993.

[92] J. Najemnik and W. S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005.

[93] M. Z. Nashed. The theory of tikhonov regularization for fredholm equations of the first kind (cw groetsch). *SIAM Review*, 28(1):116–118, 1986.

[94] P. Ndajah, H. Kikuchi, M. Yukawa, H. Watanabe, and S. Muramatsu. An investigation on the quality of denoised images. *Circuits, Systems and Signal Processing*, 5:423–434, 2011.

[95] M. Nikolova. Fast dejittering for digital video frames. In *Scale Space and Variational Methods in Computer Vision*, pages 439–451. Springer, 2009.

[96] M. Nikolova. One-iteration dejittering of digital video images. *Journal of Visual Communication and Image Representation*, 20(4):254–274, 2009.

[97] J. K. O'Regan. Change blindness. *Encyclopedia of cognitive science*, 2007.

[98] S. Osher, A. Solé, and L. Vese. Image decomposition and restoration using total variation minimization and the h 1. *Multiscale Modeling & Simulation*, 1(3):349–370, 2003.

[99] N. Ouerhani. *Visual attention: from bio-inspired modeling to real-time implementation*. PhD thesis, Université de Neuchâtel, 12 2004.

[100] A. R. Patrone. Optical flow decomposition with time regularization. Presented in SIAM-IS14 Conference on IMAGING SCIENCE, Hong Kong, 2014.

[101] A. R. Patrone and O. Scherzer. On a spatial-temporal decomposition of the optical flow. *arXiv preprint arXiv:1505.03505*, 2015.

[102] A. R. Patrone, C. Valuch, U. Ansorge, and O. Scherzer. Dynamical optical flow of saliency maps for predicting visual attention. *arXiv preprint arXiv:1606.07324*, 2016.

[103] D. G. Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10(4):437–442, 1997.

[104] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[105] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005.

[106] A. Poole and L. J. Ball. Eye tracking in hci and usability research. *Encyclopedia of human computer interaction*, 1:211–219, 2006.

[107] M. I. Posner. Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3–25, 1980.

[108] U. Rajashekar, I. Van Der Linde, A. C. Bovik, and L. K. Cormack. Gaffe: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, 17(4):564–573, 2008.

[109] Z. Ren, S. Gao, L.-T. Chia, and D. Rajan. Regularized feature reconstruction for spatio-temporal saliency detection. *Image Processing, IEEE Transactions on*, 22(8):3120–3132, 2013.

[110] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1153–1160. IEEE, 2013.

[111] C. S. Royden, J. M. Wolfe, and N. Klempen. Visual search asymmetries in motion and optic flow fields. *Perception & Psychophysics*, 63(3):436–444, 2001.

[112] S. K. Rushton, M. F. Bradshaw, and P. A. Warren. The pop out of scene-relative object movement against retinal motion due to self-movement. *Cognition*, 105(1):237–245, 2007.

[113] D. Salomon. *Data compression: the complete reference*. Springer Science & Business Media, 2004.

[114] O Scherzer. Explicit versus implicit relative error regularization on the space of functions of bounded variation. *Contemporary Mathematics*, 313:171–198, 2002.

[115] O. Scherzer. Scale-space methods and regularization for denoising and inverse problems. *Advances in imaging and electron physics*, 128:446–530, 2003.

[116] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational methods in imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[117] C. Schnörr. Determining optical flow for irregular domains by minimizing quadratic functionals of a certain class. *International Journal of Computer Vision*, 6(1):25–38, 1991.

[118] B. J. Scholl and Z. W. Pylyshyn. Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive psychology*, 38(2):259–290, 1999.

[119] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner. Object recognition during foveating eye movements. *Vision research*, 49(18):2241–2253, 2009.

[120] J. Shen. Bayesian video dejittering by the bv image model. *SIAM Journal on Applied Mathematics*, 64(5):1691–1708, 2004.

[121] T. J. Smith, D. Levin, and J. E. Cutting. A window on reality perceiving edited moving images. *Current Directions in Psychological Science*, 21(2):107–113, 2012.

[122] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.

[123] M. J. Swain and D. H. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[124] R Core Team. R: A language and environment for statistical computing (r foundation for statistical computing, vienna, 2012). *URL: http:// www. R-project. org*, 2015.

[125] J. Theeuwes. Top–down and bottom–up control of visual selection. *Acta psychologica*, 135(2):77–99, 2010.

[126] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. John Wiley & Sons, Washington, D.C., 1977.

[127] A. N. Tikhonov, A. Goncharsky, V. Stepanov, and A. Yagola. *Numerical Methods for the Solution of Ill-Posed Problems*. Kluwer, Dordrecht, 1995.

[128] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[129] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[130] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1):507–545, 1995.

[131] C. Valuch, U. Ansorge, S. Buchinger, A. R. Patrone, and O. Scherzer. The effect of cinematic cuts on human attention. In *Proceedings of the 2014 ACM international conference on Interactive experiences for TV and online video*, pages 119–122. ACM, 2014.

[132] C. Valuch, S. I. Becker, and U. Ansorge. Priming of fixations during recognition of natural scenes. *Journal of vision*, 13(3):3–3, 2013.

[133] C. Valuch, R. Seywerth, P. König, and U. Ansorge. " why do cuts work?"-implicit memory biases attention and gaze after cuts in edited movies. *Journal of vision*, 15(12):1237–1237, 2015.

[134] L. Vese and S. Osher. Modeling textures with total variation minimization and oscillating patterns in image processing. *Journal of Scientific Computing*, 19(1–3):553–572, 2003. Special issue in honor of the sixtieth birthday of Stanley Osher.

[135] L. Vese and S. Osher. Image denoising and decomposition with total variation minimization and oscillatory functions. *Journal of Mathematical Imaging and Vision*, 20:7–18, 2004.

[136] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal. Overt visual attention for a humanoid robot. In *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*, volume 4, pages 2332–2337. IEEE, 2001.

[137] C. M. Wang, K. C. Fan, and C. T. Wang. Estimating Optical Flow by Integrating Multi-Frame Information. *Journal of Information Science and Engineering*, 24:1719–1731, 2008.

[138] H. X. Wang, J. Freeman, E. P. Merriam, U. Hasson, and D. J. Heeger. Temporal eye movement strategies during naturalistic viewing. *Journal of vision*, 12(1):16–16, 2012.

[139] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[140] J. Weickert, A. Bruhn, T. Brox, and N. Papenberg. *A survey on variational optic flow methods for small displacements.* Springer, 2006.

[141] J. Weickert and C. Schnörr. A theoretical framework for convex regularizers in PDE-based computation of image motion. *International Journal of Computer Vision*, 45(3):245–264, 2001.

[142] J. Weickert and Ch. Schnörr. Variational optic flow computation with a spatio-temporal smoothness constraint. *Journal of Mathematical Imaging and Vision*, 14:245–255, 2001.

[143] J. M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

[144] J. Yuan, C. Schnörr, and G. Steidl. Simultaneous higher-order optical flow estimation and decomposition. *SIAM Journal on Scientific Computing*, 29(6):2283–2304, 2007.

[145] J. Yuan, C. Schnörr, and G. Steidl. Convex hodge decomposition and regularization of image flows. *Journal of Mathematical Imaging and Vision*, 33(2):169–177, 2009.

[146] J. Yuan, G. Steidl, and C. Schnörr. Convex Hodge Decomposition of Image Flows. In *Pattern Recognition – 30th DAGM Symposium*, volume 5096 of *lncs*, pages 416–425. Springer Verlag, 2008.

[147] G. J. Zelinsky. A theory of eye movements during target acquisition. *Psychological review*, 115(4):787, 2008.

[148] L. Zhang, M. H. Tong, and G. W. Cottrell. Sunday: Saliency using natural statistics for dynamic analysis of scenes. In *Proceedings of the 31st Annual Cognitive Science Conference*, pages 2944–2949. AAAI Press Cambridge, MA, 2009.

[149] H. Zimmer, A. Bruhn, and J. Weickert. Optic flow in harmony. *International Journal of Computer Vision*, 93(3):368–388, 2011.

Zusammenfassung

Aufmerksamkeit ist der Prozess, in dem sich unsere geistige Fähigkeiten auf Teile der verfügbaren Informationen konzentrieren. Dies liegt daran, dass Menschen nicht alle verfügbaren Informationen auf einmal verarbeiten können. In dieser Dissertation konzentrieren wir uns auf die visuelle Aufmerksamkeit und versuchen, mathematisch ihr Verhalten zu simulieren.
Die Verbreitung von Informationen durch Videos wird mehr und mehr in der heutigen Gesellschaft, durch TV-On-Demand, Webstreaming, E-Learning und Onlinespiele, um nur einige Beispiele zu nennen, präsent. Die vorliegende Arbeit konzentriert sich auf die folgenden Forschungsgebiete: die Bedeutung von Schnitt in Filmsequenzen für die visuelle Aufmerksamkeit, die Attraktivität einer Region in einem Video und das Verhalten der visuellen Aufmerksamkeit in Gegenwart von Verzerrungen, wie Jitter.

Im Folgenden werden wir uns auf das erste Forschungsgebiet, nämlich auf Schnitte, konzentrieren. Schnitte bezeichnen eine Bearbeitungstechnik, die zu einer starken Veränderung der Filmszene führt. Insbesondere werden Objektorte durch Schnitte unkorreliert. Wir analysieren zunächst das Verhalten der Zuschauer, während sie sich ein Video mit einem Schnitt anschauen, aus der Sicht der Kognitionswissenschaft. Wir schlagen eine zweistufige konzeptuelle Architektur vor und testen sie durch Eyetracking Experimente. Die Architektur wird durch die zeitliche Kohärenz der scheinbaren Bewegung angetrieben, die auch als *optischer Fluss* bekannt ist und sich auf zwei Fälle konzentriert: die Reaktion des Betrachters auf eine Sequenz ohne Schnitte und auf eine mit Schnitten.

Wir schlagen vor, dass die Aufmerksamkeit des Betrachters durch Neuheit in einer Einstellung, die keine Schnitte enthält, angezogen wird. In diesem Fall, während der globale Fluss kohärent ist, weist die lokale Inkohärenz auf die Neuheit hin. Das Verhalten der Zuschauer ändert sich, wenn man auf einen Schnitt trifft. In diesem Fall ist der globale Fluss inkohärent, was den Schnitt signalisiert. Die Aufmerksamkeit des Betrachters wird durch wiederholte Merkmale, wie wiederholte Bewegung, angezogen.
Mathematisch formulieren wir die zweistufige Architektur als Variationsansatz zur Berechnung des optischen Flusses. Wir gehen von der Horn-Schunck Funktional aus und modifizieren es bequem, um den räumlich-zeitlichen Ansatz von Weickert-Schnörr mit einzuschließen. Wir schlagen eine Aufteilung des Flusses in zwei optische Felder vor: eines, das einen zeitlich-kohärenten Fluss charakterisiert und ein anderes, das Bezug auf wiederholte Bewegung, die auch als Schwingungsmuster bekannt ist, nimmt. Um das Schwingungsmuster zu modellieren, schlagen wir ein in Zeit nicht lokalen Regularisator, von Meyers Buch inspiriert, vor.

Wir beschreiben nun das zweite Forschungsgebiet, das sich auf die Attraktivität einer bestimmten Stelle in einem Video bezieht. Das Ziel eines Modells der visuellen Aufmerksamkeit ist, die Attraktivität einer Stelle für den Betrachter, numerisch in einer Wahrscheinlichkeit von Interesse übersetzt, zu schätzen. Eine Karte der Wahrscheinlichkeiten von Interesse für jeden Punkt eines statischen Bildes wird *Salienzkarte* genannt. Um im Standardansatz die Salienz von dynamischen Sequenzen zu berechnen, wird die Salienz jedes Kaders des Videos und die Salienz der Bewegungsmerkmale berechnet, um sie dann durch ein Gewichtungsschema zu kombinieren. Wir schlagen einen Algorithmus zur Berechnung der Salienz der Bewegungsmerkmale in einer dynamischen Sequenz, in einer so genannten *dynamische Salienzkarte* vor. Auch hier formulieren wir die Bewegungsmerkmale als Variationsansatz des optischen Flusses-Problems. Insbesondere berechnen wir den Fluss einer hoch-dimensionalen Sequenz, die durch Intensität- oder Farbkanäle, ergänzt durch die Salienzkarte jedes Kaders, zusammengesetzt ist. Dies ermöglicht uns, das Aperturproblem zu überwinden. Außerdem inkludieren wir eine modifizierte Version des räumlich-zeitlichen Ansatzes von Weickert-Schnörr in unserem Funktional. Dank der vorgeschlagenen Veränderung ist unser Modell besonders wirksam im Falle von Okklusion. In der Tat, in unserer dynamischen Salienzkarte, simulieren wir das menschliche Verhalten, die Bewegung eines Objektes kontinuierlich durch Okklusion zu verfolgen.

Wir sprechen das dritte und letzte Forschungsgebiet, genauer gesagt das Verhalten der visuellen Aufmerksamkeit in Gegenwart von Verzerrungen wie z. B. Jitter, an. Die Menschen sind in der Lage, Formen und Objekte bis hin zu einem gewissen Grad der Verzerrung zu erkennen. Das menschliche Hirn führt eine automatische Rekonstruktion des Originalbildes. Wir simulieren diesen Prozess der Rekonstruktion im Fall von statischen Bildern und konzentrieren uns auf eine bestimmte Art von Verzerrung, so genannt *Jitter*. Jitter entsteht, wenn das Zeitintervall zwischen den Abfragepunkten des Signals nicht korrekt ist. Wir schlagen Variationsansätze des Funktionals, um Bilder, die von Linien-, Linien-Pixel- und Pixeljitter verzehrt werden, zu rekonstruieren, vor.

Die vorgeschlagenen Algorithmen erlauben Kognitionswissenschaftler, Theorien zu testen und quantitative Bewertung durchzuführen. Eyetracking Experimente sollen durchgeführt werden, um die Antwort der menschlichen visuellen Aufmerksamkeit im Vergleich zum Ergebnis unserer Algorithmen zu untersuchen. Ein weiterer Schritt von mathematischem Interesse könnte die Erweiterung unserer Modelle in Richtung eines allgemeinen Modells, das in der Lage ist, die visuelle Aufmerksamkeit in allen oben-genannten Forschungsgebieten gleichzeitig zu simulieren, darstellen. Wir behaupten, dass eine geeignete Formulierung des optischen Flusses, quantitative Methoden zur Abschätzung der visuellen Aufmerksamkeit liefern kann.

Curriculum Vitae

University Education

since 2012	Doctoral studies in *Computer Science* (Informatik, IK: Computational Science), University of Vienna, Austria. Thesis: *Variational models of visual attention with a special focus on dynamic sequences* (in progress), supervised by Univ.-Prof. Dr. Otmar Scherzer.
2008-2012	Master studies (MSc.) in *Computer Science*, University of Naples Federico II, Naples, Italy. Thesis: *Uncalibrated eye-tracking system using Machine Learning for the prediction of the Point of Gaze*, supervised by Dr. Francesco Isgro.
2003-2008	Bachelor studies (BSc.) in *Computer Science*, University of Naples Federico II, Naples, Italy. Thesis: *An approach to streaming real-time for mobile devices*, supervised by Univ.-Prof. Dr. Alfredo Petrosino

Publications

U. Ansorge, S. Buchinger, C. Valuch, A. R. Patrone, and O. Scherzer. Visual attention in edited dynamical images. In SIGMAP, pages 198–205, 2014.

G. Dong, A. R. Patrone, O. Scherzer, and O. Öktem. Infinite dimensional optimization models and pdes for dejittering. In Scale Space and Variational Methods in Computer Vision, pages 678–689. Springer, 2015.

C. Valuch, U. Ansorge, S. Buchinger, A. R. Patrone, and O. Scherzer. The effect of cinematic cuts on human attention. In Proceedings of the 2014 ACM international conference on Interactive experiences for TV and online video, pages 119–122. ACM, 2014.

Preprints

A. R. Patrone and O. Scherzer. On a spatial-temporal decomposition of the optical flow. arXiv preprint arXiv:1505.03505, 2015.

A. R. Patrone, C. Valuch, U. Ansorge, and O. Scherzer. Dynamical optical flow of saliency maps for predicting visual attention. arXiv preprint arXiv:1606.07324, 2016.

Conference Talks

16 – 21 Aug, 2015	*Variational saliency maps for dynamic image sequences*, Published abstract and Poster Presentation at Ecem 2015 XVIII. European Conference on Eye Movements Vienna Austria, `http://ecem2015.univie.ac.at/`
12 – 14 May, 2014	*Optical flow decomposition with time regularization*, SIAM Conference on Imaging Science (SIAM – IS14) Hong Kong, `http://www.math.hkbu.edu.hk/SIAM-IS14/`
23 – 27 Sept. 2013,	*Optical flow with oscillating pattern*, Fgp' 13 – 16th French-German-Polish Conference on Optimization `http://www.fgp13.agh.edu.pl`
10 – 12 July, 2013	*Optical flow with time regularization and oscillating pattern*, Visual Attention Meeting 2013, University of Vienna `http://attention.univie.ac.at`.

Other Activities

10/2014	Attended *Workshop on Variational methods in imaging, Special Semester on New Trends in Calculus of Variations* at Johann Radon Institute for Computational and Applied Mathematics (RICAM), Linz, Austria.
10/2014	Attended *School on Imaging, Special Semester on New Trends in Calculus of Variations* at Johann Radon Institute for Computational and Applied Mathematics (RICAM), Linz, Austria.