Detecting Social Signals with Spatiotemporal Gabor Filters

Joosten, Bart

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Detecting Social Signals with Spatiotemporal Gabor Filters

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof.dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 29 juni 2018 om 14.00 uur door Bart Joosten, geboren te Tegelen.

Promotores:	Prof. dr. E.J. Krahmer
	Prof. dr. E.O. Postma
Promotiecommissie:	dr. H. Dibeklioğlu
	Prof. dr. D.K.J. Heylen
	Prof. dr. W. Kraaij
	Prof. dr. J.-C. Martin
	Prof. dr. P.H.M. Spronck

SIKS Dissertation series No. 2018-14

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

TiCC Ph.D. Series No. 62

ISBN 978-94-6295-972-

Cover design: Mats Wilke
Printed by: ProefschriftMaken | | www.proefschriftmaken.nl

CONTENTS

1 INTRODUCTION

When computers interact with each other, as happens for instance in multi-agent systems or via the Internet, they typically follow a strict protocol of message exchanges. Naturally, if a message is somehow damaged during transmission or is not otherwise according to the predetermined interaction protocol, the receiving computer will most likely not be able to interpret it. Moreover, messages received in perfect order will only be interpreted literally, the receiving computer will not draw inferences about, let us say, the underlying intentional state of the sending computer.

Human-human interaction is clearly very different in these respects. When we interact with each other, we do not need to adhere to a strict, fully specified protocol; when a message is 'damaged' (e.g., because an utterance is ungrammatical, or produced in a very noisy setting), we often are still able to interpret substantial parts of it; and we are very adapt at drawing inferences based on the input (for example, about how the sender is feeling), even if the sender did not explicitly signal this or even explicitly intended for it *not* to be perceivable (for example, when lying).

Often, the signals from which such inferences are derived are not explicit in the verbal part of our messages (i.e., in the words we use), but rather in the non-verbal part. For example, during a conversation, we can see that our conversational partners understand what we are saying, based on visual feedback signals which we may perceive from their facial behavior (like a nod, for example) and similarly we can indicate that we pay attention to what they have to say, for instance by the occasional smile. In general, we display a wide array of non-verbal behavioral cues, sometimes not even consciously produced, that are somehow indicative of our social attitude, mental and affective state, personality, or another personal characteristic. In the literature, short-spanned temporal sequences of such non-verbal cues are also called social signals (Vinciarelli, Pantic, and Bourlard, 2009; Vinciarelli et al. 2012).

That computers traditionally lack the ability to send or receive social signals, is a problem when computers and humans start to interact. For many human-computer interaction applications, ranging from health care robotics to automatic tutoring systems, it would be beneficial if computers would be able to understand or express social cues. In this way, computers could become more empathic when interacting with patients or more adaptive when interacting with pupils.

It is for these reasons that researchers in recent years have started exploring the possibilities of automatically producing and interpreting social signals, and as a result the new field of social signal processing (SSP) emerged, which tries to channel efforts towards equipping computers with human-like social sensing abilities; the work by Vinciarelli et al. (2009) provides a recent survey. SSP is a multi-disciplinary field that primarily combines insights from psychology, cognitive science, human physiology and computer science. It is closely related to the field of affective computing (Picard, 1997), which

studies the automatic processing and simulation of human affect, which is also often signaled through behavioral cues, such as facial expressions or tone of voice.

In this thesis we will contribute to SSP by systematically comparing the performance of two different techniques, known as spatial and spatiotemporal Gabor filters respectively, on a range of human social signals.

1.1 HUMAN SOCIAL SIGNALS

Psychologists have long studied the different kinds of non-verbal cues that humans produce during interactions (see e.g., Knapp, Hall, and Horgan, 2013, for a survey), with a focus on facial expressions, vocal cues, posture and manual gestures. Typically, such non-verbal cues can be characterized as temporal changes in physiological and muscular activity, which take place during short stretches of time (ranging from milliseconds to minutes), to distinguish them from behaviors such as politeness, or traits such as personality, which typically have a much longer time-span.

Non-verbal cues form a "repertoire of non-verbal behaviors" (Ekman and Friesen, 1969). In their work, Ekman and Friesen have identified five types of non-verbal behavior. These include *illustrators*, which are non-verbal actions that accompany speech, such as eyebrow movements or manual gestures; *regulators*, which are signals that help structure an ongoing interaction, such as eye gaze and head nods, *manipulators*, which are actions on objects in the environment (like touching) or on the speaker themselves (like scratching), and *emblems*, which are culturally-defined signals, like the waving-hand-next-to-cheek gesture in the Netherlands (to signal tasty food).

The fifth, and for this thesis most important type of non-verbal behaviors discussed by Ekman and Friesen are *affect displays*. These refer to the expression of emotion, which is primarily signaled through facial expressions and tone of voice, but may also be discernible from gestures or specific cues such as laughter or tears. People can deliberately transmit affective cues, for instance when the sender wants to emphasize a certain feeling to the receiver, but affective displays are also often produced in a non-conscious manner.

Much research has focused on the display of so-called *basic emotions*, that is to say: the set of emotions shared across all cultures in the world, in the sense that in every culture these emotions are produced and recognized. Ekman and Friesen (1975) take the set of basic emotions to consist of the following six emotions: joy, surprise, fear, sadness, anger, and disgust, but other candidates have been proposed as well, including affective states like contempt, wonder, and anxiety (Frijda, 1986; Gray, 1982; Izard, 1977; Ortony and Turner, 1990). It is also worth emphasizing that for human-computer interaction basic emotions are perhaps less relevant than "social emotions" (Adolphs, 2002a), such as uncertainty or frustration, which arguably occur more often in interactions than a basic emotion such as, say, disgust.

Adequately detecting and responding to these kinds of emotions can potentially have a big influence on human-computer interaction. For example, if computer systems can automatically detect that the user is getting frustrated with the interaction, they can adapt their interaction style to try and ease this

frustration. The same applies to "cognitive states" such as disagreement, ambivalence and inattention, which like "social emotions" may be signaled using non-verbal cues, and whose (automatic) detection can potentially improve human-computer interaction.

In general, the repertoire of non-verbal behaviors is large: many different kinds of cues can occur, ranging from physical appearance and posture to facial expressions or vocal cues. Moreover, they often occur in tandem with verbal cues, yielding one, multi-modal signal. It has been estimated that as much as 90% of non-verbal behaviors are associated with speech (McNeill, 1996). In addition, even though the non-verbal cues are often produced and picked up in an unconscious manner, they can have a substantial influence on how we interpret someone's words. For example, when someone utters "God I feel great" with a smile, this is perceived rather different from when (s)he utters the same sentence with a sad face (Wilting, Krahmer, and Swerts, 2006).

It is interesting to observe that affective states influence a person's non-verbal behavior (Coulson, 2004a; Gross, Crane, and Fredrickson, 2007; Pollick, Paterson, Bruderlin, and Sanford, 2001; Van den Stock, Righart, and De Gelder, 2007), and hence that by picking up these non-verbal cues, a system can try to determine the affective state of the user. It is generally assumed that these cues are "honest", and hence a reliable and important target for social signal processing. For example, when a child, interacting with an automatic tutoring system appears to be bored (based on non-verbal cues such as yawning and looking away), the system could adapt its strategy by making learning material more challenging.

Facial Expression Analysis

Of all the different non-verbal behaviors, facial expressions have perhaps received most scholarly attention. The human face can express a wide range of signals, that are crucial for interpersonal interaction. Perhaps this is because the visual outlet of the speech system (the mouth) is located in the face (and recall that most non-verbal cues are related to speech), but additionally the face also plays a crucial role in, for example, structuring interactions (by regulating turn taking via gaze and nodding behavior) and for highlighting important information (via eyebrow movements). Additionally, the face provides relatively stable information about someone's gender, age and personality, and more dynamic information about someones emotional state. As a result, much work in SSP has concentrated on facial analyses.

Vinciarelli et al. (2009) note a distinction between so-called *message* and *sign* judgments. A message judgment is made when trying to determine what triggers a certain facial expression, while a sign judgment aims to describe actual facial movement or appearance change. As a message, a raised eyebrow, for example, can be interpreted as (part of) a surprise display, but as a sign it is merely described as a raise followed by a lowering of the eyebrows. Put differently, sign, as opposed to message, judgments aim to present an objective description of actual facial movements, without further interpretation.

One of the best-known systems for sign judgments in facial expressions is the facial action coding scheme (FACS) (Ekman, Friesen, and Hager, 1978). FACS describes expressions in terms of underlying muscle movements, deconstructing them in terms of basic Action Units (AUs). Typically, researchers manually code facial expressions in terms of their AUs, without interpreting the facial expression as such (i.e., without making message judgments). Different facial expressions are described as consisting of different AUs. In this way, for example, a distinction can be made between "insincere", social smiles (only involving AUs around the mouth) and "sincere', Duchenne smiles, also involving AUs around the eyes (Ekman, Davidson, and Friesen, 1990). Dynamics of expressions are coded by marking the onset, apex and offset of AUs. In recent years, various researchers have tried to develop automatic FACS coding systems (Cohn, 2010; Cootes, Edwards, and Taylor, 2001; De la Torre et al. 2015; Littlewort et al. 2011b).

It is worth noting, that other sign judgments systems of facial expressions exist as well. For example, the widely-used Active Appearance Models (AAMs) (Cootes et al. 2001; Matthews and Baker, 2004), essentially a generic method to model appearances of non-rigid objects in images, can also be used to track the location and movement of facial landmarks over time.

When developing SSP techniques for facial expressions (based on FACS, AAMs, or another technique), researchers typically rely on a number of standard steps, as we will also do in this thesis. First of all, obviously, recordings of people are required. These can be collected under semi-controlled, experimental settings, but researchers can also rely on existing, spontaneous fragments that may have been recorded for different purposes. Next, the persons (and their faces) need to be located in the fragments. For this, various techniques have been developed, including the Viola-Jones method (Viola and Jones, 2001). Finally, the social signals of interest in the face need to be detected. In other words, first the faces are found in the scene, after which the facial features and their movements are found in the faces (e.g., is there movement of the mouth or the eyes?). Finally, there may be a subsequent classification of the detected facial behavior (e.g., is this person talking?, or sincerely smiling?, to give two examples).

Vinciarelli et al. (2009) note that many approaches to facial expression recognition work on static, 2D facial feature extraction, see for example the works of Pantic and Bartlett (2007) and Tian, Kanade, and Cohn (2005). However, these approaches are limited in at least two respects. First of all, as noted above, social signals may also be detectable from gestures and body postures. Indeed, various researchers have started exploring this (Coulson, 2004a; Gross et al. 2007; Pollick et al. 2001; Van den Stock et al. 2007). A main challenge here is automatically detecting the relevant body parts and selecting good visual features that represent the body parts. Second, and especially relevant for the current thesis, human social signals, both facial expressions as well as gestures and other bodily cues, are not static, but change over time. Dynamic social signals influence both the sign and the message.

In this thesis, we will investigate whether automatic SSP techniques can benefit from explicitly taking dynamic information into account. We will study this in the context of a well-established technique that has been used

in SSP and in many other visual tasks, i.e., Gabor filters. In the next section, we will start with an informal introduction of Gabor filters as a method to study visual perception, followed by a formal description in terms of Gabor equations.

1.2 VISUAL PERCEPTION

As we have discussed above, when we communicate with someone, we perceive their non-verbal social signals, such as gestures and facial expressions, though vision. So, how does human vision work in the context of non-verbal communication? The established view is that human vision relies on the interplay of bottom-up and top-down processing (Bar et al. 2006; Itti and Koch, 2001, e.g.,). Bottom-up processing refers to the processing of incoming visual information. For instance, when we look at our communication partner (and the surrounding visual scene) rays of light that are reflected on the persons and objects in the scene enter our eyes and are projected through the lenses onto the retinal receptors. Subsequently, the visual information is encoded in neural activity and propagated (via intermediate stations) towards the back of the brain where the left and right visual cortices are located. In the visual cortex, the information is processed in a feed-forward way through multiple cortical stages up to the level where object and scene representations reside. The problem of visual recognition is under-constrained and can not be solved by bottom-up information only (Palmer, 1999). The brain deals with this problem by combining bottom-up processing with top-down processing. Top-down processing refers to prior knowledge and the generation of expectations which are generally assumed to work in the direction opposite to feed-forward processing. Activation of object or scene representations, gives rise to top-down processing that activates cortical stages downstream.

Visual illusions provide apt illustrations of the complex interplay between bottom-up and top-down processing. Visual illusions may arise when our top-down knowledge is biased with respect to the visual information. For example, our brain "expects" to see convex faces (this is how we normally see faces), rather than concave ones (which we rarely observe). When we are presented with a two-dimensional image of a hollow face (Gregory, 1970), i.e., a concave mask, we still perceive the face as normal (i.e., convex), as illustrated in Figure 1. The limited experiences with concave faces gives rise to a situation where top-down processing (the expectation of a convex face) supersedes the bottom-up information (a concave face).

Another illustration is the puzzle face illusion shown in Figure 2. The picture contains little bottom-up information (i.e., object contours are deliberately obscured). Therefore, the perceiver has to rely on top-down processing by generating hypotheses about the depicted object. Initially, these hypotheses may be guided by bottom-up cues. For example, the black and white regions may suggest that the depicted object is a spotted cow. After prolonged viewing, the correct hypothesis is generated (a bearded man) and matched successfully with the contents of the image.

1 Stills taken from https://www.youtube.com/watch?v=G_Qwp2GdB1Mt

Figure 1: The Hollow Face Illusion[1](Gregory, 1970) as an illustration of how top-down processing supersedes bottom-up information. The left picture shows the front-side of a mask, that we correctly interpret as a face. The right picture shows the back-side of the same mask, that we incorrectly perceive as a convex face, rather than a concave face.

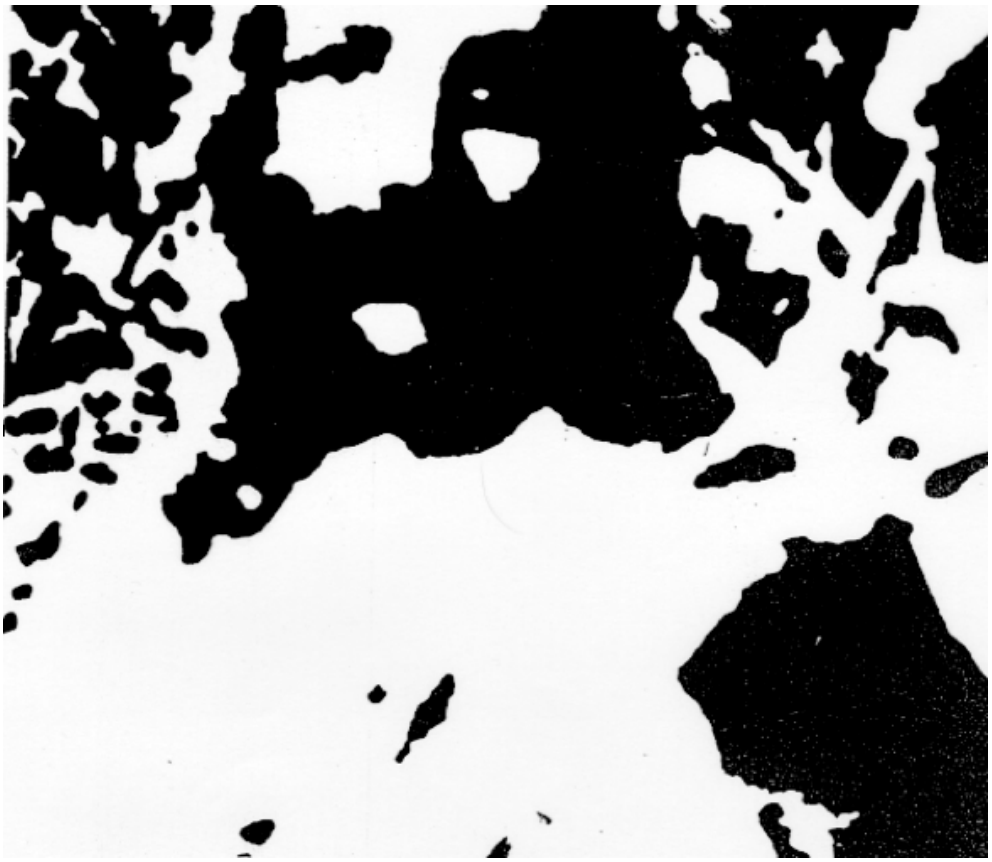


Figure 2: Puzzle face illusion reproduced from Porter (1954). Typically, prolonged viewing is required to recognize the image.

Although in the case of illusions we are fooled into perceiving something different from reality, in most other cases the top-down processing of the visual information helps us to efficiently understand and respond to the world. Helmholtz referred to top-down processing as "unconscious inferences" (Von Helmholtz, 1924).

The challenge when developing a computer vision system for social signal processing is essentially to simulate the various information processing components in the human visual system. According to Marr (1982) the visual system consists of three stages: (i) the primal sketch (e.g., detection of colors, edges and contours), (ii) the $2\frac{1}{2}$D sketch (e.g., local surface orientation and discontinuities), and (iii) 3D models (e.g., object representations that are isomorphic to their real-world counterparts). In more recent computational approaches to vision (Li and Allinson, 2008; Szeliski, 2010), the first stage consists of a global filtering operation, using for example a Gabor filter (Fischer, Šroubek, Perrinet, Redondo, and Cristóbal, 2007) or SIFT descriptor (Lowe, 1999), followed by a second stage consisting of the aggregation of (selected) filter responses. The third stage consists of classification by means of a machine learning algorithm.

In this thesis, we investigate to what extent dynamic information contributes to the performance on social signal processing tasks. In doing so, we adopt the three-stage computational approach sketched above. Social signals have static and dynamic components. For instance, a static smile can be recognized as a joyful expression, whereas the smiling dynamics could facilitate its social interpretation. Throughout the thesis we will study the contribution of static and dynamic information to social signal processing. To this end we will use static and dynamic filters known as spatial and spatiotemporal Gabor filters, respectively. These filters decompose visual images and video sequences into building blocks of visual shapes and movements. There exist many introductions to the theory and application of Gabor filters (Derpanis, 2007; Grigorescu, Petkov, and Kruizinga, 2002; Jain and Farrokhnia, 1991; MacLennan, 1991; Movellan, 2005), below we summarize the most important points by relying partly on MacLennan (1991).

1.3 A BRIEF INTRODUCTION TO GABOR FILTERS

Gabor filters originate from the work of Dennis Gabor on communication theory, an area of research that combines elements from information theory (e.g., signal processing) and mathematics in order to formalize human communication (Gabor, 1946). Before the development of Gabor filters, Fourier analysis was the method of choice for signal processing and in particular for signal analysis. Fourier analysis decomposes a signal into its constituent oscillatory parts. For instance, an auditory signal, such as speech, can be represented as a time-varying variable representing changes in air pressure. This representation allows for the exact temporal localization of the auditory information, i.e., the air pressure at a given time. However, the essence of auditory signals is in its constituent frequencies. The air pressure at a specific time conveys little information about the frequency of a signal. Fourier analysis extracts the frequency, amplitude, and phase from an auditory (or

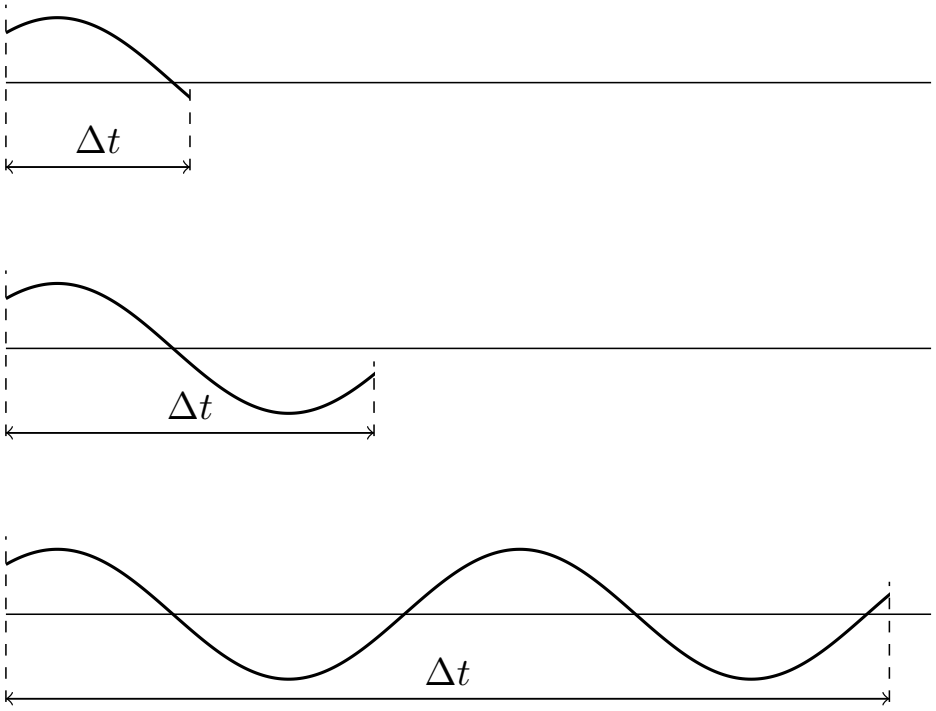


Figure 3: Illustration of a signal sampled over time intervals Δt of increasing length. Illustration after MacLennan (1991).

any other) signal. For a given time interval Δt, Fourier analysis decomposes the signal into its sinusoidal components. This analysis results in describing the signal as a function of frequencies, their associated amplitudes, and their phases. The time interval over which the analysis is performed should be sufficiently large to reliably estimate the presence of sinusoidal components. Clearly, a time interval consisting of a single discrete sample ($\Delta t = 1$) can not be decomposed into sinusoidal components. To detect the presence of a sinusoidal component of a certain frequency requires at least two samples and preferably much more.

Figure 3 illustrates that the time interval Δt should be sufficiently large to discover the periodicity of a signal. Assuming that the signal is sinusoidal, we are able to assess the signal's periodicity, by, for instance, counting the number of maximums over the interval yielding the signal's frequency. The three rows in Figure 3 show three time intervals of increasing duration. The top two intervals are too small to capture a full cycle of the sinusoidal signal. Counting only one maximum, does not reveal the frequency. Only in the bottom interval two positive maximums can be identified and used to estimate the frequency f of the signal.

The Uncertainty Relation

There is an inversely proportional relation between the time interval Δt and the frequencies that can be determined by counting the maximums of the sinusoidal signal within the interval.

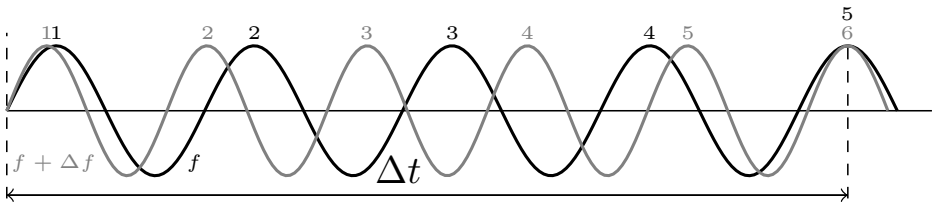


Figure 4: Distinguishing the frequencies of two signals by counting their maximums requires a sufficiently large time interval Δt. Illustration after MacLennan (1991).

Figure 4 shows a plot with two sinusoidal signals with two different frequencies, f_1 and f_2 with $f_2 = f_1 + \Delta f$. If we want to tell the two signals apart by using the maximums localization strategy, we need a Δt that obeys the following inequality:

$$\Delta t \geq 1/\Delta f \tag{1}$$

Following this inequality, the interval Δt must be at least of length $1/\Delta f$ time samples, in order to tell the two signals apart. The inequality (Equation 1) can be rewritten as:

$$\Delta f \Delta t \geq 1. \tag{2}$$

The constant 1 may be smaller or larger depending on the method of determining the frequency of the signal. What is important is that the product of Δf and Δt is a constant. Hence, the inequality (Equation 2) implies that there is a limit to the degree of certainty to which we can simultaneously measure both frequency and (temporal) location. Improving the temporal resolution by making Δt smaller, leads to a less adequate estimation of the frequency. Improving the frequency resolution can only be achieved by making Δt larger. This is analogous to the well-known *Uncertainty Relation* in quantum mechanics that applies to all wave-like systems. In fact, showing that the uncertainty principle also applies to communicative signals is one of the core contributions of the work of Gabor. In his seminal work Gabor (1946), he derived a function that provides the best combination of temporal and frequency resolution; the *Gabor function*. Filters designed according to Gabor's function are called *Gabor filters*. When applied to a temporal signal, these filters perform a localized measurement of the signal's frequency.

Inspired by Gabor's classic work on one-dimensional signals, other researchers (Daugman, 1985; Heeger, 1987; Petkov and Subramanian, 2007) extended his ideas to two-dimensional signals, including the visual ones studied in this thesis.

Formal Description of Gabor Filters

The aforementioned Gabor filters are one-dimensional and are typically associated with the analysis of temporal signals and hence referred to as *temporal*

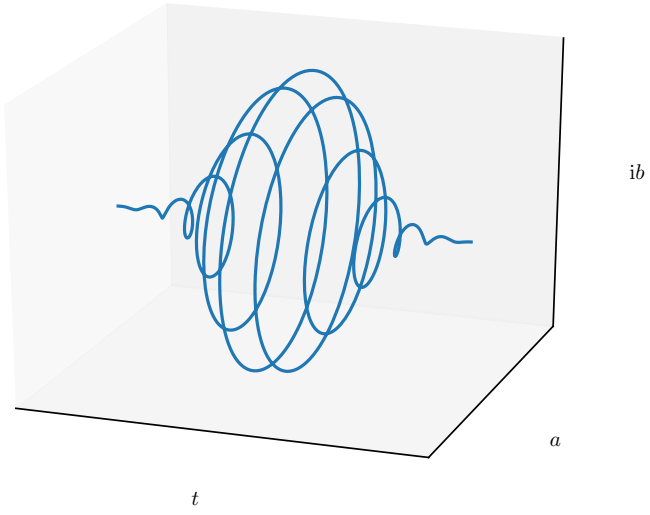


Figure 5: Illustration of the elementary Gabor function (Gaussian-modulated complex exponential) defined in the complex space (a, ib) as a function of time (t). After MacLennan (1991).

Gabor filters. Two-dimensional Gabor filters are often applied in image analysis and called *spatial* Gabor filters. Adding the temporal dimension to spatial Gabor filters leads to (three dimensional) *spatiotemporal* Gabor filters (SGFs). In what follows, we provide a formal description of each of these three types of Gabor filters.

Temporal Gabor filters

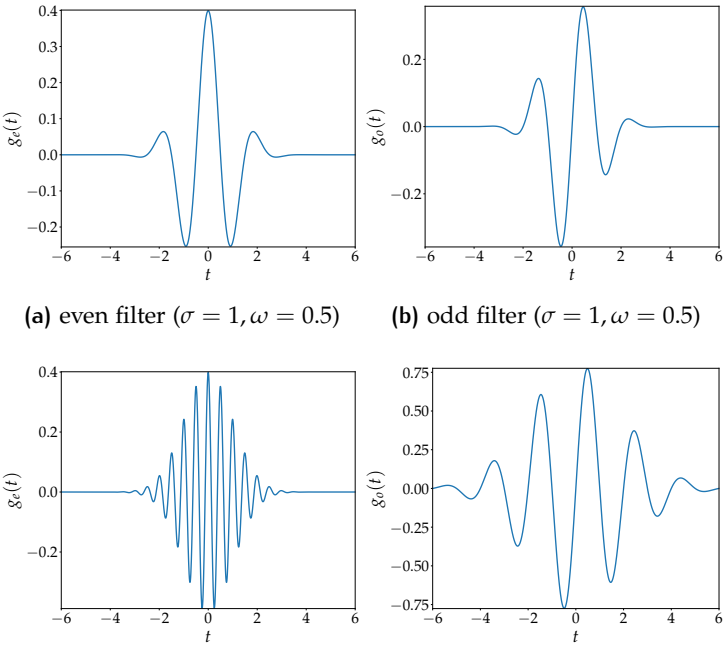
The elementary Gabor function can be defined in terms of complex numbers, consisting of a real number a and an imaginary number ib, where the norm of the complex number represents the amplitude of the signal and the complex angle represents the phase of the signal. Figure 5 illustrates the elementary Gabor function in the three dimensional space spanned by time t, and the real and imaginary numbers a and ib. The elementary Gabor function is also referred to as a Gaussian-modulated complex exponential, because a complex number z can also be written as a complex exponential, i.e. $z = a + ib = r\exp^{i\theta}$, where $a = r\cos\theta$, $b = r\sin\theta$, $r = \sqrt{a^2 + b^2}$ and $\theta = \arctan\frac{b}{a}$.

In the one-dimensional case a Gabor filter is derived from Gabor's elementary signal. When the complex sinusoidal function is decomposed into a real ("even", g_e) and an imaginary ("odd", g_o) part this yields the following equations (Heeger, 1987):

$$g_e(t) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp -\frac{t^2}{2\sigma^2} \cos(2\pi\omega t) \tag{3}$$

$$g_o(t) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp -\frac{t^2}{2\sigma^2} \sin(2\pi\omega t) \tag{4}$$

where ω denotes the center frequency with the highest energy (i.e., filter response), and σ represents the spread of the Gaussian envelope. Figure 6 is a visualization of even and odd one-dimensional Gabor filters for four different values of σ and ω.



(a) even filter ($\sigma = 1, \omega = 0.5$) **(b)** odd filter ($\sigma = 1, \omega = 0.5$)

(c) even filter ($\sigma = 1, \omega = 2$) **(d)** odd filter ($\sigma = 0.5, \omega = 0.5$)

Figure 6: One-dimensional Gabor filters with different parameters.

Spatial Gabor Filters

With his work on cells in the primary visual cortex Daugman (1985) extended the one-dimensional Gabor filter to two spatial dimensions x and y. Two-dimensional Gabor filter responses for even and odd filters are defined as follows:

$$g_e(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp -\frac{1}{2}\left(\frac{x^2}{\sigma_x} + \frac{y^2}{\sigma_y}\right) \times \cos(2\pi\omega_x x + 2\pi\omega_y y) \tag{5}$$

$$g_o(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp -\frac{1}{2}\left(\frac{x^2}{\sigma_x} + \frac{y^2}{\sigma_y}\right) \times \sin(2\pi\omega_x x + 2\pi\omega_y y), \tag{6}$$

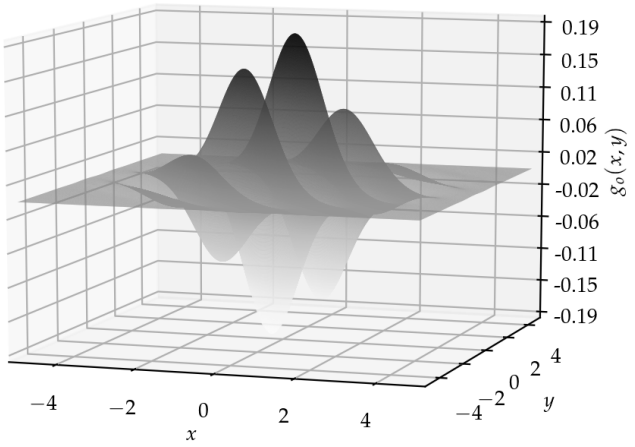


Figure 7: Two-dimensional odd Gabor filter for the parameters values $\sigma = 2$, $\omega = 0.5$, and $\theta = 45°$.

where (ω_x, ω_y) denote the maximum response center frequencies and (σ_x, σ_y) the spread of the Gaussian envelope in the x and y direction respectively. To make the filter sensitive to any arbitrary orientation, we can substitute rotation functions x_r and y_r for x and y, respectively:

$$x_r = x\cos(\alpha) - y\sin(\alpha) \tag{7}$$
$$y_r = -x\sin(\alpha) + y\cos(\alpha) \tag{8}$$

where α is the desired orientation. Figure 7 shows an illustration of a two-dimensional Gabor filter.

Spatiotemporal Gabor Filters

A spatiotemporal Gabor filter extend the spatial Gabor filter with a temporal component. A formal definition due to Heeger (1987) is as follows:

$$g_e(x,y,t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t}\exp -\frac{1}{2}\left(\frac{x^2}{\sigma_x} + \frac{y^2}{\sigma_y} + \frac{t^2}{\sigma_t}\right) \times \cos(2\pi\omega_x x + 2\pi\omega_y y + 2\pi\omega_t t) \tag{9}$$

$$g_o(x,y,t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t}\exp -\frac{1}{2}\left(\frac{x^2}{\sigma_x} + \frac{y^2}{\sigma_y} + \frac{t^2}{\sigma_t}\right) \times \sin(2\pi\omega_x x + 2\pi\omega_y y + 2\pi\omega_t t) \tag{10}$$

The added variable t denotes time. Analogous to the one- and two-dimensional cases, σ and ω denote the spread of the Gaussian and the center frequency for their corresponding axes, respectively. Increasing σ for any axis, results in filters that are less localized for that axis, in other words the spread over the

axis is larger. This results in narrower frequency responses measured over a wider area. Conversely, by decreasing σ we can improve the localization of the measurement, with the trade-off that the filter becomes less selective for frequency. With ω we can specify the filter's center frequency, i.e., the frequency for which the filter has the highest response. Similarly to the two-dimensional case, we can substitute rotation functions for x and y given by Equation 7. This makes the filter steerable to orientation α.

Spatiotemporal Gabor filters provide good models for the functional properties of cells in the primary visual cortex (Petkov and Subramanian, 2007). These cells have a sharp tuning to motion with a certain speed and direction. Based on this biological perspective, an alternative formalization of spatiotemporal Gabor filters was proposed by Petkov and Subramanian (2007):[2]

$$g(x, y, t, v, \theta, \phi) = \frac{\gamma}{2\pi\sigma^2} \exp \frac{-\left((x_r + v_c t)^2 + \gamma^2 y_r^2\right)}{2\sigma^2} \times \cos\left(\frac{2\pi}{\lambda}(x_r + vt) + \phi\right). \tag{11}$$

Where ϕ is the phase of the filter which determines the symmetry of the filter. Values of 0 and π correspond to even filters, whereas values of 0.5π and 1.5π generate odd filters. Here, γ controls the ellipticity of the Gaussian envelope in the spatial domain. This basically controls the selectivity to the amplitude of the signal. Parameters v and v_c control the speed preferences of the filters, where v denotes the preferred speed in pixels per frame (PPF), v_c determines whether the Gaussian envelope moves along the x-axis at a certain speed ($v_c > 0$) or remains stationary ($v_c = 0$). The primary visual cortex hosts both cells that are selective to the temporal frequency as well as to the speed of movement. The model of Petkov and Subramanian (2007) can accommodate both variants, giving rise to velocity tuning ($v_c \neq 0$) and frequency tuning ($v_c = 0$). We discuss both variants in more detail below. The λ parameter corresponds to the preferred wavelength of the periodic part of the filter which corresponds to spatial frequency $1/\lambda$. This value is determined by the relation with the preferred speed v: $\lambda = \lambda_0\sqrt{1 + v^2}$, where λ_0 is a constant denoting the duration of one cycle. If we keep t at a fixed value, we can plot the profile of the spatiotemporal filter in the (x, y) plane at time t. The result is shown in Figure 8 for three subsequent time steps t where we kept the envelop stationary. From left to right, the images show the Gaussian-weighted grating moving from the upper right to the lower left.

Two Implementations of Spatiotemporal Gabor Filters

In this thesis, we will experiment with two implementations of the spatiotemporal Gabor filters, one due to Heeger (1987) [3] and one due to Petkov and Subramanian (2007) [4]. Heeger was among the first who developed a computational implementation of the idea of spatiotemporal filters, highlighting the

2 We omitted two terms from Petkov and Subramanian's equation: a surround inhibition term and a causality constraining term. Both terms were included in the original work to enhance the biological plausibility.

3 Our implementation is partly based on the code found here: http://www.bu.edu/vip/files/pubs/reports/EZLR10-04buece.pdf

4 http://www.cs.rug.nl/~imaging/spatiotemporal_Gabor_function/GaborApp.html

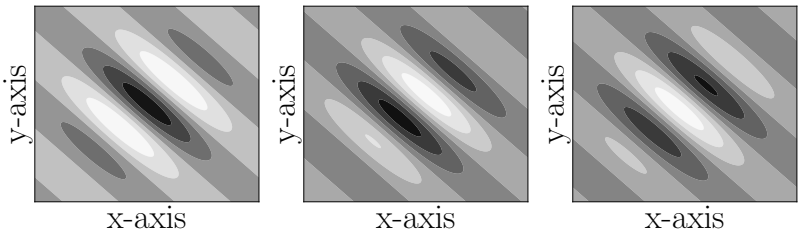


Figure 8: Three (x, y) contour plots for different t.

importance of motion information in visual perception. While Heeger does not make explicit claims about the biological realism of the method, Petkov and Subramanian study the spatiotemporal filters as models of dynamic receptive fields of cells in the primary visual cortex (V1). Both implementations originate from the same mathematical model of Gabor filters. The main difference between the implementations is in the choice of parameters and parameter constraints. Table 1 specifies the main parameters of both implementations. In our experiments we will explore different values of these parameters. The left column lists the parameters of Heeger's implementation, given by Equation 9 and Equation 10. These are the center frequencies and standard deviations of the spatial coordinates and temporal coordinate and parameter θ to control the selectivity to the direction of motion. The right column lists the parameters of Petkov and Subramanian's implementation.

Selectivity to spatial frequency is controlled by the λ parameter in the Petkov and Subramanian's implementation, which corresponds to the first two listed parameters in the Heeger implementation, i.e., ω_x and ω_y. The Heeger implementation does not have an explicit parameter to tune for a specific speed, in contrast to Petkov and Subramanian's v and v_c parameters. Instead, selectivity to a preferred speed is controlled by ω_t. Both implementations use the Gaussian envelope's standard deviation to control the selectivity to frequency, however Petkov and Subramanian use one parameter for all axes, whereas for Heeger they are specified separately. The ellipticity parameter γ for Petkov and Subramanian does not have a counterpart in the Heeger implementation. This is also the case for phase parameter ϕ, which determines the construction of even or odd filters. The Heeger implementation simply considers the real and the imaginary part of the filter as even and odd respectively.

Velocity Tuning versus Frequency Tuning

As mentioned above, the human primary visual cortex has two types of cells that are sensitive to motion, i.e., cells that respond to a certain temporal frequency of moving contours and cells that respond to a specific velocity of the moving contour. These cells can be modeled by applying a stationary envelope to the temporal Gaussian component (i.e., frequency tuned) or by letting the envelope move along the temporal axis (i.e., velocity tuned). In the primary visual cortex most neurons are frequency tuned Wu, Bartlett, and Movellan (2010). Petkov and Subramanian's implementation is able to model both types of cells, whereas Heeger's implementation can only construct filters that are sensitive to a specific frequency of movement. In

our experiments we experiment with both types of filters and switch them between experiments. We will explicitly mention whether we used frequency tuned or velocity tuned filters. Preliminary results showed little to hardly any difference in performance between the two types of filters for Petkov and Subramanian's implementation when we applied them in a social signal processing context.

Table 1: List of the different parameters for two implementations of spatiotemporal Gabor filters.

Heeger		*Petkov and Subramanian*	
symbol	definition	symbol	definition
ω_x	center frequency x-axis	λ	Spatial wavelength
ω_y	center frequency y-axis	v	Preferred speed
ω_t	center frequency t-axis	v_c	Gaussian's center velocity
θ	Direction of motion	θ	Direction of motion
σ_x	Standard deviation	σ	Gaussian's standard deviation
σ_y	Standard deviation	γ	Spatial aspect ratio
σ_t	Standard deviation	ϕ	Phase

Throughout this thesis, we will often use both the Heeger and the Petkov and Subramanian implementations, to see whether it matters for our application — social signal processing — whether the spatiotemporal Gabor filters are an explicitly biologically inspired (and hence more constrained) model or not.

1.4 THE CURRENT THESIS

1.4.1 Methodology

In every study in this thesis we follow a comparable deductive research methodology. In general, we expect that adding temporal information to spatial Gabor filters is beneficial for the performance of automated social signal processing tasks. We systematically assess this expectation by (1) acquiring data that display the social signal we are interested in, (2) applying image processing and analysis techniques to represent the phenomenon in a static (i.e., spatial Gabor filters) and dynamic (i.e., spatiotemporal Gabor filters) manner, and (3) evaluating the performance of the two representations by means of classification experiments. In all of our studies we use video sequences of human behavior that were collected in an experimental setup. The purpose of the work in this thesis is not to obtain state-of-the-art classification results but to systematically compare spatial Gabor filters to spatiotemporal Gabor filters.

We have chosen four areas to investigate our claims, viz., (1) human speech, (2) question answering, (3) smiling, and (4) human gait. With these areas of focus we cover a wide range of human non-verbal behavior, both in terms of how easily they are perceived ("is this person talking?" vs. "how hard was this question?") and in terms of physiological scale (from mouth to face to human body).

1.4.2 Outline

In **Chapter 2** we start exploring the benefits of adding spatiotemporal information to Gabor filters, by looking at voice activity detection (VAD) based on facial movements. VAD is the task of detecting human speech in an audio signal, and most earlier approaches to this problem have typically *only* looked at the auditory channel. However, when speakers talk, they also produce visual cues: they move their lips and often also other parts of their head, including, for example, their jaws or eyebrows. *Visual* VAD (VVAD) tries to detect voice activity based on solely visual cues, which can be helpful, for instance, in noisy environments. Moreover, it has been argued that visual speech cues (e.g., the opening of the mouth) often precede the onset of speech, so that Visual VAD can help for early detection of speech as well. In Chapter 2 we rely on two existing datasets: one is the publicly available CUAVE dataset (Patterson, Gurbuz, Tufekci, and Gowdy, 2002) in which different speakers utter digits, while being filmed both frontally and from the side, the other dataset is the so-called LIVER dataset (Joosten, Postma, Krahmer, Swerts, and Kim, 2012), in which participants utter a single word ("liver"). As a result the two datasets differ substantially in the ratio between speech and non-speech. We systematically compare a standard Gabor filter approach with a dynamic, spatiotemporal variant (which we call STem-VVAD) relying on different speeds (based on the implementation of Petkov and Subramanian, 2007 using velocity tuned filters), also including a baseline merely relying on frame differencing. In addition, we systematically compare the performances of the methods at different levels of detail: looking only at the mouth region, at the whole head, and at the entire clip.

Next, in **Chapter 3**, we move to a more complex non-verbal social signal, namely detecting learning difficulties based on automatic facial expression analysis, asking what the benefits of dynamic information is in this particular task. Being able to automatically detect whether a child considers a task, like for example an arithmetic problem, easy or difficult to solve, is an important prerequisite for developing adaptive learning environments. To study this, we collected our own dataset of children from two age groups solving easy and hard arithmetic problems using a game-like interface. In this study, we compared static, spatial Gabor filters and dynamic, spatiotemporal ones and compared the performances of the implementations of Petkov and Subramanian (2007) (using velocity tuned filters) and Heeger (1987) (using frequency tuned filters). In addition, we compared the performances of the Gabor filter methods with the performance of a method that models the children's facial dynamics explicitly: an Active Appearance Model (AAM)

(Cootes et al. 2001; Matthews and Baker, 2004; Van der Maaten and Hendriks, 2010).

Then, in **Chapter 4**, we continue our explorations by considering yet another social signal: smiles. It is well-known that people can smile in at least two different ways, either because they are truly happy (the so-called Duchenne smile) or as a social response (the non-Duchenne smile) (Niedenthal and Mermillod, 2010). Being able to detect "genuine" smiles is important for automatic emotion recognition systems, but has various practical application as well. For example, it can be used by photo camera's to decide automatically when a picture is best taken. Various factors play a role when trying to distinguish "genuine" from "posed" smiles. For example, it has been suggested that a Duchenne smile is accompanied by a narrowing of the eyes (Ekman and Friesen, 1976; Niedenthal and Mermillod, 2010), causing wrinkles to appear at the outside corners of the eyes. More recently, and particularly relevant for the current thesis, it has been claimed that genuine, Duchenne smiles can also be detected based on the speed with which they appear on the face (with Duchenne smiles appearing slower than non-Duchenne ones) (Krumhuber, Manstead, and Cosker, 2009; Schmidt, Ambadar, Cohn, and Reed, 2006). In this chapter we study the added value of dynamic, spatiotemporal Gabor filters for smile classification (once again in two different implementations: Petkov and Subramanian, and Heeger, and both implementations tuned to the frequency of movement), based on a publicly available dataset of spontaneous and posed smiles: the UVA-NEMO Smile database (Dibeklioğlu, Salah, and Gevers, 2015). We once again compare the benefits of having different speeds in the spatiotemporal Gabor filters. In addition, given the potential impact of head movements on smile classification, we compare results for both "raw" (unprocessed) faces and automatically "fixed" ones.

The preceding chapters all look at *facial* signals, but, of course, it is also possible to consider the body as a whole, which clearly impacts the size of the movements to be considered. Therefore, in **Chapter 5** we consider a basic, full-body task, namely gender classification based on a person's movements while walking (their gait). Again, this task has potential practical applications: shops, for example, may want to automatically track the number of male and female shoppers in particular shop areas. It is well established that humans are rather good at predicting someone's gender based on general movement characteristics, as has been demonstrated, for instance, by means of point-light displays, in which only the movements of key joints are represented against an otherwise dark background (Kozlowski and Cutting, 1977). Additionally, good computational techniques for this task have been developed, including one based on Gait Energy Images (GEIs), which essentially capture all movement in a single image (Han and Bhanu, 2006). In this final empirical chapter we study how Gabor filters fare on this task, once again comparing static, spatial Gabor filters and dynamic, spatiotemporal ones, with Petkov and Subramanian's frequency tuned implementation. For this purpose, we use the CASIA Gait Dataset B (Yu, Tan, and Tan, 2006), a benchmark for comparing gait recognition methods. We compare both Gabor filter methods with a state-of-the-art GEI-based method, looking both at frontal and sideways clips of people walking, at different levels of detail: the head as well as the upper and the lower body.

Finally, in **Chapter 6**, we summarize and discuss the findings, asking whether there are indeed general benefits of using spatiotemporal Gabor filters over static, spatial ones, and discussing to what extent this depends on the nature of the task and possibly the specific implementation.

2 VISUAL VOICE ACTIVITY DETECTION

2.1 INTRODUCTION

Human speech comprises two modalities: the auditory and the visual one. Many researchers have emphasized the close connection between the two (e.g., McGurk and MacDonald, 1976; Stekelenburg and Vroomen, 2012). A speaker cannot produce auditory speech without also displaying visual cues such as lip, head or eyebrow movements, and these may provide additional information to various applications involving speech, ranging from speech recognition to speaker identification. For many of these applications it is important to be able to detect *when* a person is speaking. Voice Activity Detection (VAD) is usually defined as a technique that automatically detects human speech in an auditory signal. Using VAD enables speech processing techniques to focus on the speech parts in the signal, thereby reducing the required processing power. This is, for example, applied in digital speech transmission techniques (e.g., GSM or VoIP), where VAD helps to transmit speech and not silence segments (Beritelli, Casale, and Cavallaero, 1998; Lee, Kwon, and Cho, 2005).

Arguably, the straightforward approach to VAD would be to look into the auditory channel to see when speech starts. This is indeed what various researchers have done, and what is required for situations in which only the auditory signal is available (Chang, Kim, and Mitra, 2006; Ghosh, Tsiartas, and Narayanan, 2011; Ramírez, Segura, Benítez, Torre, and Rubio, 2004; Sohn, Kim, and Sung, 1999). However, this approach suffers from a number of complications. For instance, when background noise is present it becomes more difficult to differentiate between noise and speech, because they are entwined in one signal. Moreover, when multiple speakers are present, recognizing speech onset also becomes more difficult (because the speech signals are overlapping). Even though solutions for these problems have been proposed (e.g, Furui, 1997; Kinnunen and Li, 2010; Reynolds, 2002), various researchers have argued that taking the visual signal into account (if available) can help in addressing these issues, e.g. because the presence or absence of lip movements can help in distinguishing noise from speech (Sodoyer, Rivet, Girin, Schwartz, and Jutten, 2006), and because visual cues can help for speech segmentation. Moreover, importantly, visual cues such as mouth and head movements typically precede the actual onset of speech (Wassenhove, Grant, and Poeppel, 2005), allowing for an earlier detection of speech events, which in turn may be beneficial for the robustness of speech recognition systems. For this reason, various researchers have concentrated on Visual Voice Activity Detection (VVAD).

This chapter is a slightly extended version of Joosten, B., Postma, E., & Krahmer, E. (2015). Voice activity detection based on facial movement. *Journal of Multimodal User Interfaces*, *9* 183-193.

Previously proposed VVAD methods mostly relied on lip tracking (Aubrey et al. 2007; Liu, Wang, and Jackson, 2011; Sodoyer et al. 2009). While these approaches have been successful, both in detecting voice activity based on visual cues and in combination with auditory VAD approaches, we know that there are more visual cues during speech in the face beyond the movement of the lips (Krahmer and Swerts, 2005). Besides, evidently (extracting features from) lip tracking is challenging when a speaker turns their head sideways. In their overview on *audiovisual automatic speech recognition*, Potamianos, Neti, Luettin, and Matthews (2012) point out that robust visual features for speech recognition should be able to handle changing speaker, pose, camera and environment conditions, and they have identified three types of visual features that apply to VVAD as well: 1) appearance-based features using pixel information extracted from a region of interest (typically the mouth region), 2) shape based features derived from tracking or extracting the lips, and 3) a combination of the aforementioned types of features. Potamianos et al. note that extensive research comparing these features is still missing.

2.1.1 Related Work

Previous work on VVAD methods can be distinguished into two classes of models: lip-based approaches and appearance-based approaches. Below, we review examples of each of these classes.

Lip-Based Approaches

Lip-based approaches employ geometrical models based on the shape of lips. The geometrical models typically consist of a flexible mesh formed by landmarks, or connected fiducial points surrounding the lips, flexible active contours that are automatically fitted to the lip region. In what follows, we describe three examples of lip-based approaches and the features extracted to perform VVAD.

Aubrey et al. (2007) employed a geometrical lip model for VVAD that consisted of landmarks. Given a video sequence of a speaking and silent person, the task was to distinguish speech from non-speech. Their landmarks (constituting the lip model) were fitted to the video data of a speaking person by means of an Active Appearance Model (AAM) (Cootes et al. 2001). For each frame, the two standard geometric features, i.e., the width and height of the mouth, were extracted from the positions of the landmarks and submitted to a Hidden Markov Model.

Using an Active Contour Model (Kass, Witkin, and Terzopoulos, 1988), also called "snakes", Liu, Wang, and Jackson (Liu et al. 2011) computed the two standard geometric features as well an appearance feature, i.e., the mean pixel values of a rectangular patch aligned with the lip corners and centered at the center of the mouth. For each frame, these three features form the basis of their classification vector, which is extended with dynamic features. To classify a frame as VOICE or SILENT, AdaBoost (Freund and Schapire, 1995) was used, a technique that incrementally builds a (stronger) classifier by adding a new feature from the classification vector to the previous classifier

at each consecutive step of the training process. The snake-based VVAD method was evaluated on a selected YouTube video of a single speaker.

The Sodoyer et al. (2009) study relied on segmented lips, which were obtained by painting the lips of recorded speakers in order to be able to extract them from the rest of the face (like in the chroma key technique used in movies). In their study, they employed the chroma key technique to build a 40 minute long audiovisual corpus of two speakers, each in a separate room, having a spontaneous conversation. In spontaneous conversation speech events are generally followed up by silence or non-speech audible events such as laughing and coughing. Such events are characterized by specific lip motion (even in silence parts). The aim of the study was to find a relationship between lip movements during speech and non-speech audible events on the one hand and silence on the other. The two standard geometrical features were extracted from the segmented lips of both speakers and used to define a single dynamic feature based on the sum of their absolute partial derivatives.

Appearance-Based Approaches

Appearance-based VVAD approaches go beyond the lips by taking into consideration the surrounding visual information. We describe three examples of appearance-based methods, each of which emphasizes another visual feature: color, texture, and optical flow.

Scott, Jung, Bins, Said, and Kalker (2009) propose a VVAD that relies on a comparison of the pixel colors of the mouth region and the skin regions just below the eyes. They defined a *mouth openness* measure, which corresponds to the proportion of non-skin pixels in the mouth region. The regions were extracted with automatic face-detection and facial geometry heuristics. Their manually annotated VVAD dataset consisted of three videos.

Navarathna, Dean, Sridharan, Fookes, and Lucey (2011) measured textural patterns in the mouth region using the Discrete Cosine Transform (DCT). Their dataset consisted of frontal and profile faces of the CUAVE dataset (Patterson et al. 2002). They classified the DCT coefficients by means of a Gaussian Mixture Model using speaker-independent models. This was realized by training and testing on different subsets of groups of speakers.

Tiawongsombat, Jeong, Yun, You, and Oh (2012) measured the optical flow in the mouth region using the pyramidal Lucas-Kanade algorithm (Bouguet, 2000). They recorded 21 image sequences of 7 speakers to evaluate and 7 individual mouth image sequences to train their method. Classification was done using a two-layered HMM that considers the states *moving* and *stationary* lips at the lower level and *speaking* and *non-speaking* at the higher level simultaneously.

Evaluation of Existing Approaches

Directly comparing results between the different studies is complex, since they all vary in certain dimensions, e.g., the datasets used differ in size and complexity, different evaluation metrics are employed, and generalizability is often not tested (i.e., evaluations tend to be speaker-dependent). With the exception of the CUAVE dataset, there are no publicly available datasets to enable a comparison across different situations and speakers. However,

in general these methods all perform well in comparison to their specific task and in a comparable range. Typically, scores between 70 and 90% are reported.

2.1.2 Current Studies

Since many VVAD studies acknowledge the importance of modeling movement during speech, we choose to explicitly examine movement information at an early stage, an approach called *Early Temporal Integration* (Wu et al. 2010), by designing a VVAD that incorporates features that represent spatiotemporal information. In this chapter, we propose an appearance-based approach to VVAD, representing images in terms of movement, without explicitly tracking the lips. Our novel method, which we call STem-VVAD (STem abbreviates **S**patio**Tem**poral, but also happens to mean "voice" in Dutch) is based on spatiotemporal Gabor filters (STGF), a type of filter which is sensitive to movement at a certain direction and speed (Petkov and Subramanian, 2007), as explained in Chapter 1, which have, to the best of the author's knowledge, never been applied to VVAD. Intuitively, lip movements during speech have a specific spatiotemporal signature which may be different from those associated with non-speech (e.g., couching, laughing). In a similar vein, the orientation of movements may show different patterns for speech and non-speech, facilitating VAD.

Spatial Gabor filters (SGF) have been frequently used for automatic visual tasks, ranging from texture segmentation (Jain and Farrokhnia, 1990) to coding of facial expressions (e.g., Littlewort et al. 2011b; Lyons, Akamatsu, Kamachi, and Gyoba, 1998) and automatic speech recognition (Kleinschmidt and Gelbart, 2002). The use of SGFs in computer vision is inspired by biological findings on the neural responses of cells in the primary visual cortex (e.g., Daugman, 1985; Field, 1987; Jones and Palmer, 1987), as the 2D Gabor function is able to model these responses. This makes them biologically plausible for use in automatic vision systems. Moreover, Lyons et al. (1998) argue that the use of SGFs for facial expression recognition is also psychologically plausible, since the properties of the neurons that they are modeled on allow neurons in the higher visual cortex to be able to distinguish between different facial expressions.

As explained in Chapter 1, STGFs are the dynamic variants of their spatial counterparts. Whereas SGFs respond to visual contours or bars of a certain orientation and thickness, STGFs respond to moving visual contours or bars. The responses of motion-sensitive cells in primary visual cortex can be modeled by STGFs and have been shown to be the independent components of natural image sequences (Hateren and Ruderman, 1998). In this chapter, we apply Spatiotemperal Gabor filters to Visual VAD, in our STem-VVAD approach.

To examine the extent to which our approach is successful in detecting voice activity, we have conducted a series of experiments on two different datasets, i.e., the CUAVE dataset (Patterson et al. 2002), and our LIVER dataset (Joosten et al. 2012). The CUAVE dataset contains multiple speakers uttering digits, with frontal as well as profile recordings, whereas our LIVER

dataset consists of frontally recorded speakers each with a single speech event, i.e., the uttering of the Dutch word for “liver". In the CUAVE set, the ratio between speech and non-speech is approximately balanced, this in contrast to the LIVER set where the majority of frames is non-speech.

For each dataset we assess the voice activity detection capabilities of our STem-VVAD method as well as for two reference VVADs: a VVAD based on frame differencing and a method based on standard, spatial Gabor filters. In addition, we determine the contribution of various visual speeds to VVAD performance, to determine if certain speeds of, for instance, lip motion contribute more to VVAD than others. As a third evaluation, three regions in the clips are examined, to determine if zooming in on the mouth region leads to better VVAD performance, or that other dynamic facial characteristics contribute as well to the performance as suggested by Krahmer and Swerts (2005).

Since human speech is inextricably connected to the idiosyncratic characteristics of its speaker (Dellwo, Leemann, and Kolly, 2012) and, moreover, since the location with respect to the camera varies among the subjects, we will evaluate STem-VVAD on a speaker-dependent and a speaker-independent basis. By using these two evaluations we focus on the applicability of STGF in VVAD (speaker dependent) versus the generalizability of our method (speaker independent). In the area of speech recognition, systems tailored towards one specific speaker generally outperform systems that are able to handle multiple speakers. We therefore expect to see better results with our speaker-dependent scheme than with our speaker-independent scheme. It will be interesting to see how this distinction affects our different VVADs.

In the next Sections, we present our own appearance-based method (STem-VVAD), which is inspired by the biological example of early spatial-temporal integration in the brain. In addition, to get a better understanding of the problem, and in view of the complex, difficult to compare pattern of results in related work, here we systematically compare analyses of the mouth area with full facial analyses as well as analyses of the entire frame, and we look at different speeds of movement, both in isolation and combined into one feature vector. We evaluate the method on two different datasets (including CUAVE (Patterson et al. 2002)), and look at both speaker-dependent and speaker-independent models.

2.2 METHOD

The Spatiotemporal Visual Voice Activity Detection (STem-VVAD) method is based on two stages: (i) the preprocessing stage consisting of spatiotemporal Gabor filters to determine the energy values at certain speeds, and (ii) the aggregation and classification stage employing summation and a classifier to summarize and map the aggregated energy values onto the binary classes SPEECH and NON-SPEECH.

Preprocessing Stage

The preprocessing stage transforms video sequences with spatiotemporal Gabor Filters into a so-called energy representation (Heeger, 1987; Petkov and Subramanian, 2007; Wu et al. 2010). As described in Chapter 1, the spatiotemporal Gabor filters may be considered to be dynamic templates, i.e., oriented bars or gratings of a certain thickness that move with a certain speed and in a certain direction. The transformation of a video sequence by means of STGFs proceeds by means of convolution, in which each STGF (dynamic template) is compared with the contents of the video sequence at all pixel locations and at all frames. The presence of a moving elongated object in the video that matches the STGF in terms of orientation, thickness, speed and direction, results in a large “energy value" at the location and time of the elongated object. A better match results in a larger energy value. Each STGF results in one energy value for each pixel per frame of the video. Hence, the result of convolving a video sequence with a single filter, yields an energy representation that can be interpreted as an “energy video sequence" in which the pixel values represent energies. Large energy values indicate the presence of the filter’s template at the spatial and temporal location of the value.

In order to capture all possible orientations, a suitable range of sizes (spatial frequencies), and appropriate speeds and directions, a spatiotemporal Gabor filter bank is used which consists of filters whose parameters (orientation, spatial frequency, speed and direction) are evenly distributed over the relevant part of the parameter space. Each of these filters generates an “energy movie" and hence convolving a video sequence with a filter bank gives rise to an enormous expansion of the data. Given a video of F frames and N pixels per frame (PPF), convolution with a filter bank of G filters results in $G \times F \times N$ energy values. The number of filters, G, is determined by the range and number of parameter values selected. In the STem-VVAD method the direction of movement is always perpendicular to the orientation. Hence, the number of filters is defined as $G = k \times d \times s$, where k is the number of spatial frequencies, d the number of orientations and s the number of speeds.

Aggregation and Classification Stage

The applied filter bank of G filters (that vary in three dimensions of parameter space, i.e., spatial frequency, orientation, and speed) result in G energy videos obtained from the convolution in the preprocessing stage. In what follows, we refer to a Gabor filter tuned to a specific combination of spatial frequency, orientation, and speed, as a *Gabor feature*. Representing the energy value for Gabor feature g, frame f, and pixel n by $E_g(f,n)$, the aggregated features $A_g(f)$ are computed by summing the energy values for feature g for each frame, which results in, $A_g(f) = \sum_{n=1}^{N} E_g(f,n)$. The aggregation generates one G-dimensional vector $A(f)$ per frame, the elements of which signal the presence of a filter-like visual pattern in the video frame under consideration. Since the G filters represent different combinations of spatial frequencies, orientations, and speeds, the summed energy values signal the presence of moving contours with these frequencies, orientations, and speeds.

2.3 EXPERIMENTAL EVALUATION

As stated in the introduction, the experimental evaluation of the STem-VVAD method consist of three parts. First, its performance is evaluated on two video datasets. Second, it is compared to two reference VVADs: (1) to determine the contribution of using a sophisticated spatiotemporal filtering method, the STem-VVAD method's performance is compared to the simplest method of change detection called frame differencing, and (2) to assess the contribution of dynamic information, a comparison is made with a version of the method in which the speed is set to zero, thereby effectively creating static, spatial Gabor filters. Third, the VVAD performances obtained for three visual regions of analysis are compared. These regions are: the entire frame, the face, and the mouth.

2.3.1 Datasets

As stated in the introduction, the two datasets used to evaluate the VVAD method are the publicly available CUAVE dataset[1] (Patterson et al. 2002) and our own *LIVER* dataset[2] (Joosten et al. 2012). Both datasets were recorded for different purposes and have different characteristics.

CUAVE

The CUAVE dataset is an audio-visual speech corpus of more than 7000 utterances. It was created to facilitate multimodal speech recognition research and consists of video recorded speakers uttering digits. The dataset contains both individual speaker recordings as well as speaker-pair recordings. We used the individual speaker recordings only. The set contains 36 different speaker video recordings (19 male and 17 female) in MPEG-2, 5000 kbps, 44 KHz stereo, 720×480 pixels, at 29.97 fps. All speech parts are annotated at millisecond precision. The speakers vary in appearance, skin tones, accents, glasses, facial hair and therefore represent a diverse sample. Speakers were recorded under four conditions of which we used the following two: stationary frontal view and stationary profile view. In both cases speakers were successively pronouncing the digits. In these clips, the frontal face videos have an average length of 52 seconds ($sd = 14s.$) compared to 24 seconds ($sd = 6s.$) for the profile videos.

LIVER

Our LIVER dataset was constructed in the context of a surprise elicitation experiment (Joosten et al. 2012). This experiment yielded a dataset of 54 video sequences of 28 participants (7 male and 21 female) uttering the Dutch word for liver ("lever") in a neutral and in a surprised situation resulting in two recordings per person. The participants all sit in front of the camera but are allowed to move their heads and upper body freely. The videos are in WMV format, 7000 kbps, 48 KHz stereo, 29.97 fps, at 640 by 480 pixels

1 http://www.clemson.edu/ces/speech/cuave.htm
2 The dataset was created by our colleague prof. Swerts, and is available upon request.

and were automatically annotated for speech using a VAD based solely on the audio channel. By means of visual inspection we checked the correctness of annotations. The recordings are cropped at approximately four seconds (i.e. around 120 frames) and start when the participants are about to speak. Contrary to in the CUAVE database, where speakers produce speech about half of the time, speakers in the LIVER dataset produce just one word in a 4 second interval, resulting in a dataset that is unbalanced for speech and non-speech frames (1053 to 6524, respectively).

2.3.2 Implementation Details

For the preprocessing stage of the STem-VVAD method, we used the STGF implementation of Petkov and Subramanian (2007)[3] with velocity tuned filters as mentioned in Chapter 1. We created a filter bank of $G = 6 \times 8 \times 2$ filters sensitive to 6 different speeds ($v = \{0.5, 1, 1.5, 2, 2.5, 3\}$ PPF), 8 orientations ($\theta = \{0, 0.25\pi, 0.50\pi, 0.75\pi, \ldots, 1.75\pi\}$ radians) covering the range of speeds and orientations in our datasets, and two constant spatial periods, defined by the parameter λ_0^{-1}, where $\lambda_0^{-1} = \{1/2, 1/4\}$ (recall the relation $\lambda = \lambda_0\sqrt{1+v^2}$). The dimensionality of the resulting STem-VVAD feature vector for frame f, $A(f)$, is equal to $G_{STem-VVAD} = 6 \times 8 \times 2 = 96$. A separate version with the same parameters, but with $v = 0$ was used for comparison. In this version, the dimensionality of feature vector $A(f)$ is equal to $G_{zero-speed} = 2 \times 8 = 16$. This is the same dimensionality as the STem-VVADs where we take only one speed into consideration. We implemented frame differencing by taking the absolute differences of the pixel intensities of two consecutive frames and computing their sum, average and standard deviation, yielding three values per frame.

The video sequences in the datasets were convolved with the STGFs. The resulting energy values were aggregated as specified in Section 2.2. For the three regions of analysis, i.e., frame, face, and mouth, the aggregation was performed over the entire frame, the rectangle enclosing the face, and the rectangle enclosing the lower half of the face, respectively. The lower half of the face was defined as the half of the bounding box enclosing the face region. The face region was detected automatically using the OpenCV implementation of the Viola-Jones face detector with Local Binary Pattern features (Liao, Zhu, Lei, Zhang, and Li, 2007). Since we used face detection in each frame instead of face tracking, we had to deal with false positives and frames in which the detector failed to find a face. By manually ascertaining that the face in the first frame of each video sequence was correctly detected by the face detector, we could automatically remove false positives in subsequent frames by stipulating that a bounding box' size and location should not differ more than a fixed number of pixels, 50 pixels in our setup, from the face detected in the previous frame. We used a simple heuristic to account for the missing detections by interpolating between the previous and upcoming detected face's bounding boxes. Visual inspection of the detected face regions throughout the video sequences confirmed that this procedure worked for almost all videos. Eight video sequences in total (i.e., two in the CUAVE

3 http://www.cs.rug.nl/~imaging/spatiotemporal_Gabor_function/GaborApp.html

frontal condition, one in the CUAVE profile condition, and five in the LIVER dataset) yielded too little face detections and were excluded from the experiments. This amounts to 5% of the total data, which suggests that any biases introduced by face detection failures are minimal.

A support vector machine was used to classify each frame as SPEECH or NON-SPEECH using feature vectors of the aggregated values as input. Feature vectors were classified with a linear Support Vector Machine, for which we used the LIBLINEAR SVM library (Fan, Chang, Hsieh, Wang, and Lin, 2008).

2.3.3 Evaluation Procedure

The generalization performance is an estimate of how well the VVAD performs on unseen videos. To estimate the generalization performance we used two validation procedures: 10 fold cross validation for the speaker-dependent evaluation and Leaving One Speaker Out (LOSO) cross validation for the speaker-independent evaluation. The LOSO cross validation measures the performance on speakers not included in the training set. The resulting generalization performances obtained for (1) frame differencing, (2) the zero-speed version, (3) separate speed versions, and (4) the full-fledged STem-VVAD, are reported in terms of F1-scores. The F1-score, which originates from Information Retrieval, is the harmonic mean of precision and recall (Rijsbergen, 1979). The use of F1-scores is motivated by the unequal distributions of our two datasets (i.e., the CUAVE dataset is approximately balanced, while the liver dataset contains more non-speech frames than speech frames). In contrast to accuracy, the F1-score is insensitive to the unbalance of the two classes. In our tables and figures in the next section we also report the F1-score of the chance classifier, i.e., the classifier that randomly picks between the classes SPEECH and NON-SPEECH. The final F1-score at chance level is the average F1-score between all folds for the specific evaluation procedure.

2.4 RESULTS

Our results are divided over two sections, i.e., speaker-dependent results, and speaker-independent results. In each section we start by presenting the results of the frontal-view speakers in both the CUAVE and the LIVER dataset, followed by the results of the profile-view speakers, obtained only on the CUAVE dataset.

Speaker-Dependent Results

The upper part of Table 2 summarizes the overall results obtained on the frontal faces of the CUAVE dataset. Inspection of this table reveals that, as expected, the best results (for all three detector types, FD, zero-speed and STem-VVAD) are obtained for the mouth region. Looking closer at the results for the mouth region, we can see that, importantly, the STem-VVADs outperform the two reference methods (FD and zero-speed). Of the six nonzero speeds examined, the STem-VVAD with 0.5 PPF performs best, with an F1-

Table 2: Average speaker-dependent F1-scores obtained on all three datasets. The left part of the table shows the results for the frame differencing (FD) and the zero-speed (0) version VVADs and the right part of the table lists the F1-scores for the STem-VVAD method. The columns labeled $0.5 - 3$ contain the scores of the associated speeds, the rightmost column labeled *All*, lists the result for the full-fledged STem-VVAD in which all speeds are included. The three rows for each dataset show the results for the three regions of analysis: frame, face, and mouth. The best scores are printed in bold-face. Chance level F1-scores for the three datasets are 0.47, 0.23 and 0.49 respectively. All scores are significantly different from chance level scores as determined by a two-sample Kolmogorov-Smirnov test at the 1% significance level.

		References		STem-VVADs						
Dataset	Region	FD	0	0.5	1	1.5	2	2.5	3	All
CUAVE Frontal	Frame	0.50	0.50	0.67	0.64	0.60	0.58	0.58	0.57	**0.72**
	Head	0.51	0.53	0.67	0.66	0.64	0.63	0.62	0.62	**0.75**
	Mouth	0.56	0.55	0.70	0.68	0.67	0.66	0.65	0.65	**0.78**
LIVER	Frame	0.34	0.55	0.51	0.43	0.41	0.42	0.44	0.44	**0.70**
	Head	0.40	0.56	0.63	0.56	0.51	0.54	0.55	0.53	**0.80**
	Mouth	0.40	0.57	0.68	0.58	0.57	0.60	0.62	0.60	**0.86**
CUAVE Profile	Frame	0.48	0.53	0.63	0.61	0.58	0.56	0.55	0.54	**0.71**
	Head	0.52	0.59	0.66	0.66	0.64	0.62	0.61	0.61	**0.78**
	Mouth	0.54	0.63	0.70	0.69	0.68	0.65	0.65	0.64	**0.80**

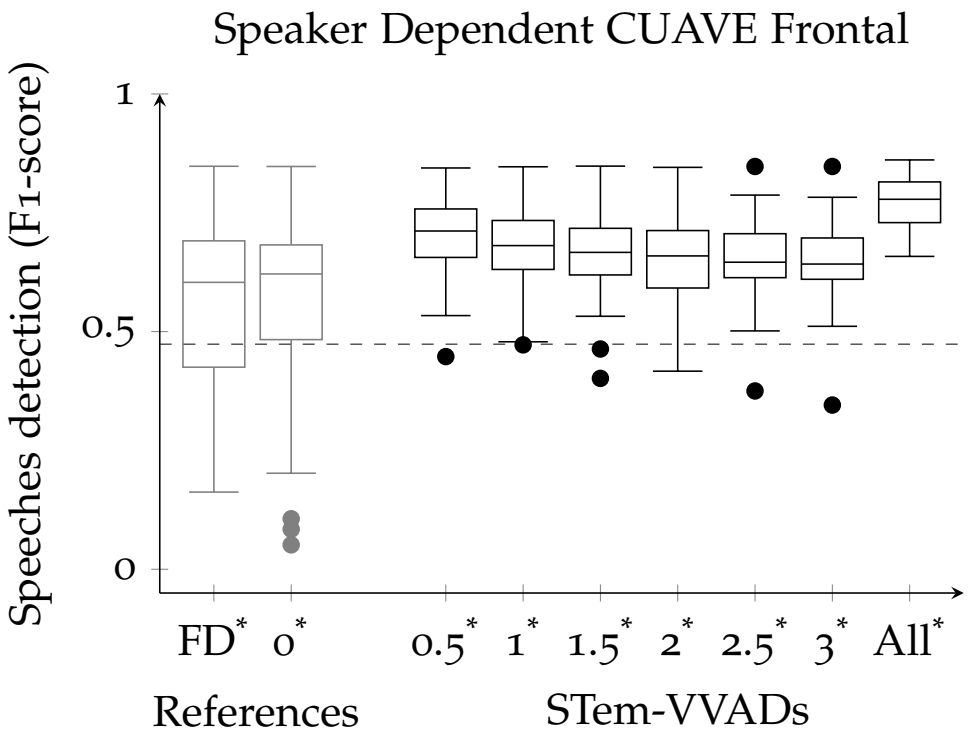


Figure 9: Boxplots of speaker-dependent F1-scores obtained on the CUAVE frontal dataset. The boxes correspond to the **Mouth** results in the upper part of Table 2. The left part of the Figure shows the distribution for the frame differencing (FD) and the zero-speed (0) version VVADs and the right part of the Figure displays box plots of F1-scores for the STem-VVAD method. The boxes labeled $0.5 - 3$ represent the F1-scores of the associated speeds, the rightmost box labeled *All*, shows the F1-scores for the full-fledged STem-VVAD in which all speeds are included. The dashed line indicates performance at chance level.

score of 0.7, which is almost 0.15 above the reference methods. Performance of the single-speed STem-VVADs decreases slightly with increasing speed. The best result is obtained for the full-fledged STem-VVAD in which all speeds are combined: an F1-score of 0.78. This result is comprised of a precision of 0.76 and a recall of 0.79.

Figure 9 visualizes the distributions over speakers of the results for the mouth region with box-whisker-plots as a function of VVAD. Each plot visualizes the distribution of the mean F1-scores per speaker. The horizontal line in the middle of each box represents the median of the data, while the top and bottom horizontal lines of the box represents the upper and lower quartile of the data, respectively. The upper whisker depicts the largest data value which is smaller than the upper quartile plus $1.5\times$ inter-quartile-range (i.e., absolute difference between upper and lower quartile). The reverse holds for the lower whisker, i.e., the smallest data value larger than the lower quartile minus $1.5\times$ inter-quartile-range. Any data larger or smaller than the upper and lower whisker respectively is considered an outlier and is depicted by a dot. The spread of the STem-VVADs is considerably smaller than those of the reference methods, implicating a more robust detection performance for the STem-VVADs. The positions of the box plots' medians are in line

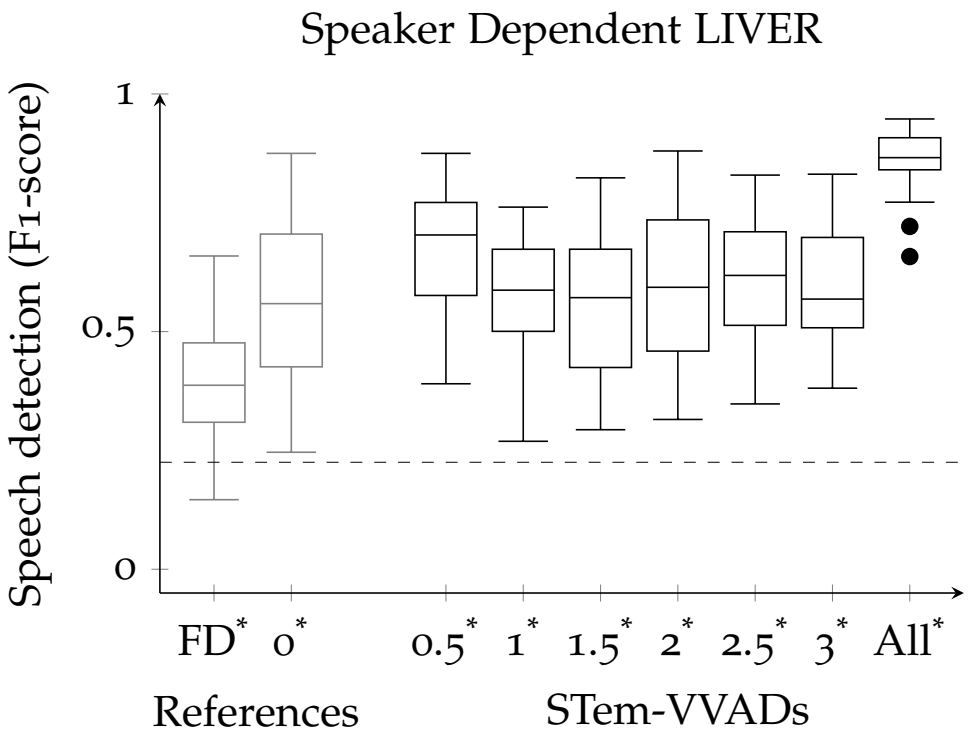


Figure 10: Boxplots of speaker-dependent F1-scores obtained on the LIVER dataset. The boxes correspond to the *Mouth* results in the middle part of Table 2. For explanation see Figure 9.

with the mean values reported on the last line of the upper part of Table 2, showing a gradual descent for increasing speeds and a best performance when combining all speeds.

The results of our VVADs on the LIVER dataset evaluated with ten-fold CV are summarized in the middle part of Table 2. The overall pattern of results is similar to those obtained on the CUAVE dataset. The performances improve with smaller regions, with the best performance obtained for the mouth region. For the mouth region, the single-speed STem-VVADs outperform the reference methods (best single-speed performance is obtained for speed 0.5 (0.68). Again, the full-fledged STem-VVAD yields the best overall performance on all three regions of analysis (0.86 on the mouth region). When we zoom in on this result, we see that the recall here is higher, i.e., 0.93, than the precision, which is 0.8.

The corresponding box-whisker plots for the mouth region in Figure 10 show a similar pattern of results as obtained for the CUAVE dataset. The most striking result is the superior performance obtained for the STem-VVAD.

The lower part of Table 2 shows the speaker-dependent results obtained on the subset of profile faces in the CUAVE dataset. A comparison with the results obtained for the frontal faces in the upper part of Table 2, reveals that the STem-VVAD method can deal with profile faces very well. The mouth-region results are displayed in Figure 11.

Table 3: Speaker-independent F1-scores obtained on all three datasets. For explanation, see Table 2. Chance level F1-scores are 0.48, 0.24 and 0.49 respectively. Light gray values indicate F1-scores which or *not* significantly different from the chance level F1-scores as determined by a two-sample Kolmogorov-Smirnov test at the 1% significance level.

		References		STem-VVADs						
Dataset	Region	FD	0	0.5	1	1.5	2	2.5	3	All
CUAVE Frontal	Frame	**0.53**	0.38	0.51	0.45	0.44	0.45	0.45	0.42	0.50
	Head	0.51	0.39	0.51	0.50	0.49	0.50	0.52	0.51	**0.53**
	Mouth	0.53	0.38	0.55	0.54	0.54	0.53	0.51	0.50	**0.58**
LIVER	Frame	0.38	0.34	0.36	0.29	0.27	0.30	0.31	0.29	**0.46**
	Head	0.37	0.31	0.43	0.38	0.30	0.30	0.32	0.34	**0.44**
	Mouth	0.40	0.22	0.40	0.38	0.35	0.35	0.33	0.32	**0.55**
CUAVE Profile	Frame	**0.49**	0.42	0.41	0.42	0.41	0.42	0.40	0.39	0.42
	Head	0.50	0.49	0.49	0.51	0.51	0.51	0.50	0.49	**0.53**
	Mouth	0.51	0.53	0.52	0.55	**0.56**	0.55	0.54	0.54	**0.56**

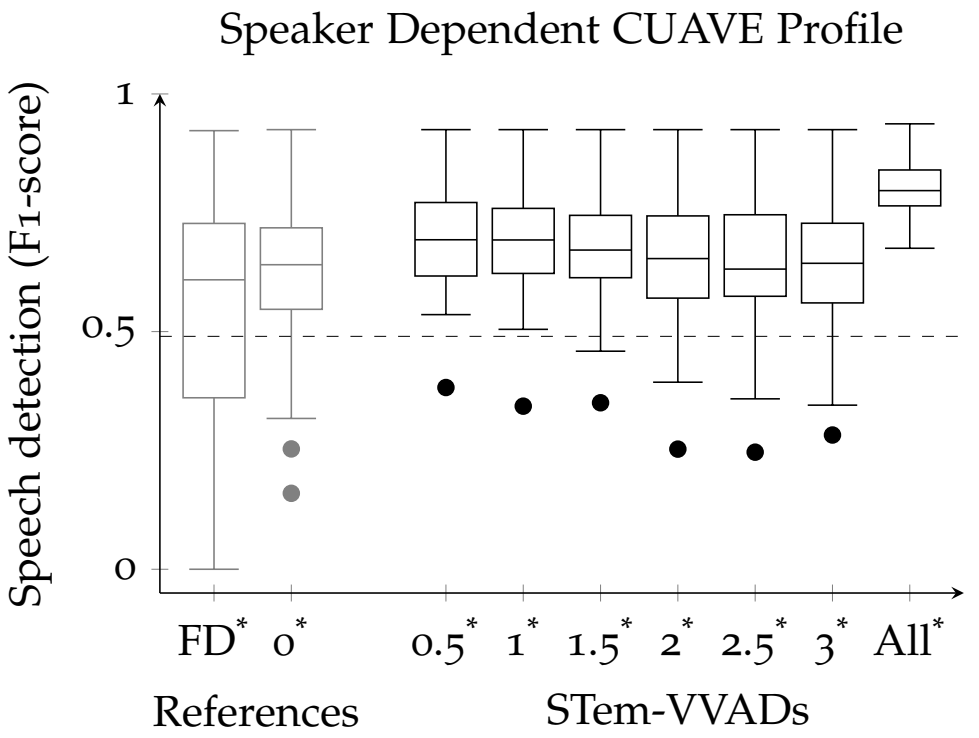


Figure 11: Boxplots of speaker-dependent F1-scores obtained on the CUAVE profile dataset. The boxes correspond to the *Mouth* results in the lower part of Table 2. For explanation see Figure 9.

Speaker-Independent Results

The upper part of Table 3 gives the results for the CUAVE database with the Leave One Speaker Out validation method, which tests the generalizability of our VVAD methods across speakers. Inspection of this table reveals a similar pattern of results as in the upper part of Table 2, although with a lower overall performance. In particular, results for the mouth region are generally better overall than those for the head and the mouth region. Moreover, the best performing individual method is the STem-VVAD with speed 0.5 PPF, although the difference with the FD reference VVAD is much less pronounced than in the ten-fold cross validation results in the upper part of Table 2. Interestingly to remark here is the performance of the FD reference method (0.53%) for the entire frame compared to all the other detectors applied to the same region, since it is the best performing VVAD. Moreover, this VVAD also has a higher score than it's equivalent applied to the head region. In general the FD's performances here are only slightly below the best performing VVADs, i.e., the 0.5 PPF and the combined speeds, whereas the zero-speed's performance here is considerably less.

Again, we zoomed in on the results for the mouth region and visualized them using a box-whisker-plot, as depicted in Figure 12. Compared to Figure 9 the boxes generated from the LOSO experiment are less compressed, corresponding to a wider spread of the individual results, it does however, show roughly the same pattern of performance as the previous plot when comparing them individually.

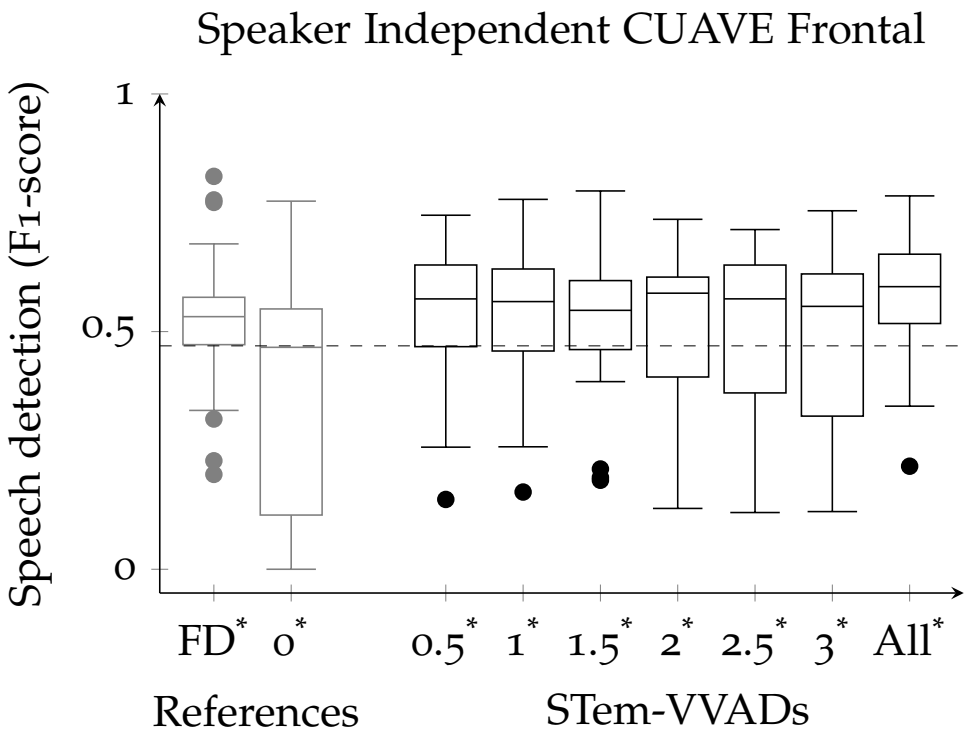


Figure 12: Boxplots of speaker-independent F1-scores obtained on the CUAVE frontal dataset. The boxes correspond to the *Mouth* results in the upper part of Table 3. For explanation see Figure 9.

The middle part of Table 3 shows the speaker-independent results of our VVADs applied to the LIVER dataset, using a Leave One Speaker Out CV. The speaker-independent results are clearly inferior to the speaker-dependent results listed in the middle part of Table 2. Interestingly, simple frame differencing often outperforms single-speed STem-VVADs. The full-fledged STem-VVAD shows the best performance at all three regions of analysis with the best result (0.55) obtained for the mouth region. The box plots in Figure 13 illustrate the corresponding results for the mouth region.

The lower part of Table 3 lists our VVAD results obtained on the profile faces of CUAVE dataset. Compared to the lower part of Table 2 the full-fledged STem-VVADs here do not show a clear prevailing performance. Although the performance tends to improve when zooming in from frame to head to mouth, at each level the results for all VVADs are very similar. The small difference in results is visualized by Figure 14 which contain the results for the mouth area.

2.5 DISCUSSION

In this chapter, we studied whether it is possible to detect voice activity based on facial movements, which has various potential applications when auditory voice detection is difficult (e.g., when there is background noise or when there are multiple speakers). Obviously, movement is an essential ingredient of visual voice activity detection (VVAD), and hence we studied whether

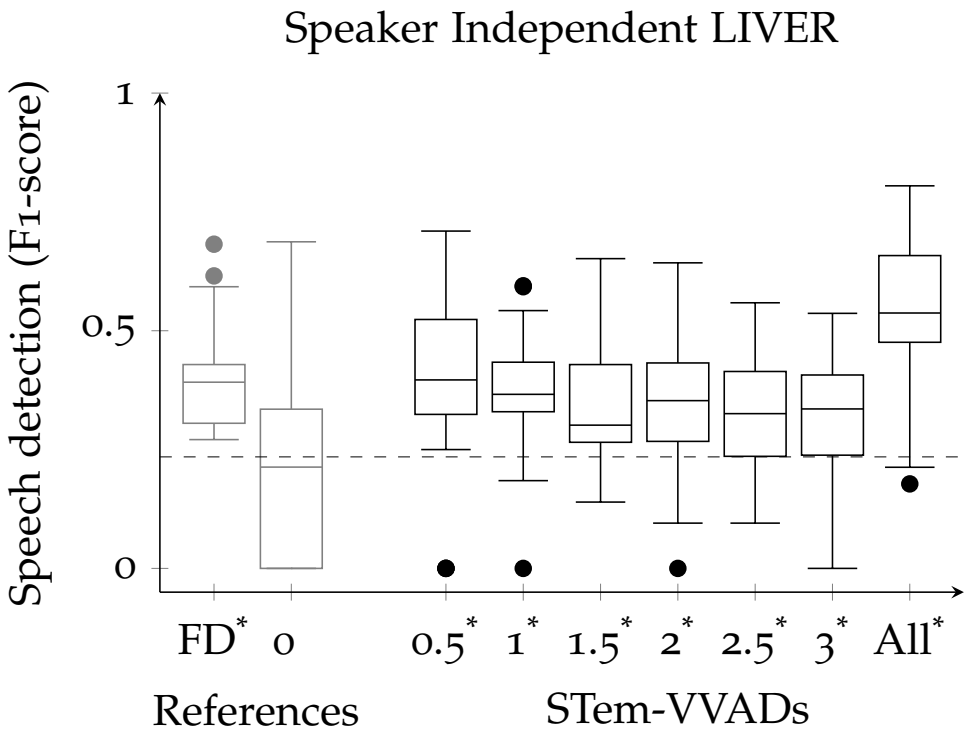


Figure 13: Boxplots of speaker-independent F1-scores obtained on the LIVER dataset. The boxes correspond to the *Mouth* results in the middle part of Table 3. For explanation see Figure 9.

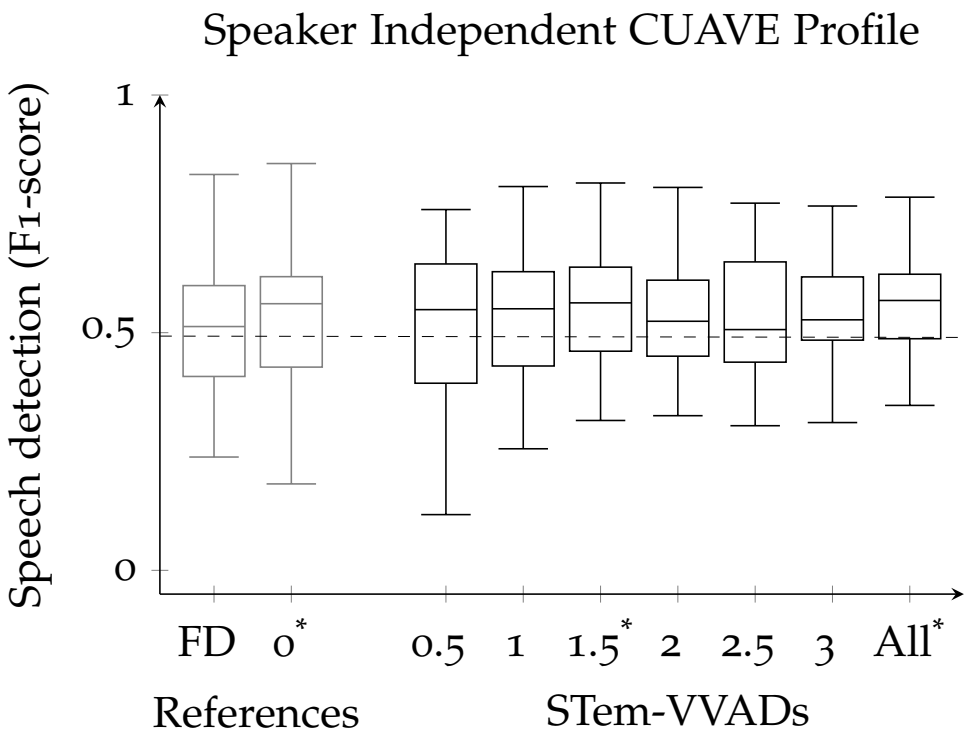


Figure 14: Boxplots of speaker-independent F1-scores obtained on the CUAVE profile dataset. The boxes correspond to the *Mouth* results in the lower part of Table 3. For explanation see Figure 9.

spatiotemporal Gabor filters could be used successfully for this task. Our set-up was as follows: we compared the performance of spatiotemporal Gabor filters in our Stem-VVAD approach with two reference methods, namely a straightforward frame differencing method and a static Gabor filter method (i.e., zero-speed STem-VVAD), allowing us to capture the added value of both Spatial and Temporal information. We compared results on two different datasets (representing two extremes in the speech to silence ratio, which is low in the LIVER and high in the CUAVE dataset). We looked at both frontal and profile recorded faces, and compared performance at three levels of granularity (entire frame, entire face, mouth only). Finally, we evaluated the performance of the VVADs with both speaker-dependent models (where each speaker is used both for training and testing) and speaker-independent models (where we train and test on separate speakers).

The results present a clear picture. In almost all comparisons, the STem-VVAD (combining all speeds) yields the best performance, outperforming both the two baseline systems (and the chance performance level), sometimes by a wide margin.

Our STem-VVAD does not suffer from unbalanced training and test data. The results obtained from the LIVER dataset appear to be slightly better than those obtained on the CUAVE dataset for both individual and generic models. This suggests that the information extracted from this single-speech event data is informative enough to distinguish between speech and non-speech, even though the model is trained with an abundance of non-speech frames. As we pointed out above, the LIVER dataset was originally collected to study verbal and non-verbal expressions of surprise. It is interesting to point out that apparently the facial movements associated with speech differ from the ones associated with surprise, since our STem-VVAD approach picks up on the former but not the latter.

Given the similar results obtained on the frontal and profile conditions of the CUAVE dataset we argue that our STem-VVAD is robust to turning faces (most notably in the speaker-dependent version). STem-VVAD does not rely on advanced lip models, which makes it potentially well suited for automatic speech detection in conference systems, where speakers tend to move their heads freely.

VVAD performance increases when focusing on the mouth; for all three techniques (FD, zero-speed, STem-VVAD), better results are usually obtained when taking only the head into account rather than considering the entire frame, and better results still when zooming in on just the mouth. Even though it has been argued that information from the upper part of the face (e.g., eyebrows) can be a useful cue for VVAD, this turned out not to help for the techniques we studied, perhaps because when considering a larger region of interest the chance of picking up speech irrelevant movements increase, and the movement cues that could be informative are more likely to be lost in the noise.

In addition, the speaker-dependent models (10-fold) perform (substantially) better than the generic models (LOSO), even though all three methods usually perform better than chance. This is perhaps not surprising because the speaker-dependent models capture some of the idiosyncratic properties of each speaker, which is not case for the generic models.

Perhaps more importantly for our current purposes, we find that adding temporal information, as we do in the spatiotemporal Gabor filters, does pay off for VVAD. Zooming in on the mouth (where VVAD works best in our set-up) the best performing STem-VVAD, which combines different speeds, outperforms both reference VVADs, in both datasets, both frontal and profile, and in both individual models as well as for the generic LIVER models. Although the full-fledged mouth results for the generic CUAVE models are better than the reference methods, the differences are negligible.

Looking at the experimental data for the mouth region we can see that our STem-VVAD approach with all speeds could be a valuable addition to traditional auditory VAD systems, especially in the speaker-dependent case were a system is trained on an individual speaker basis. Achieving average F1-scores of 0.78, 0.86 and 0.8, respectively for the three datasets, a reasonable performance by itself. In the speaker-independent case the average F1-scores obtained for the mouth region of our full fledged STem-VVAD appear to be inaccurate enough for useful VVAD applications.

Our current method does not generalize very well, looking at the considerable differences between the speaker-dependent and the speaker-independent results. Apparently, idiosyncratic speech characteristics are prevailing over general speech patterns, considering the high F1-scores in the speaker dependent case. Another possibility could be the non-linearity of the feature space, to which we applied a linear SVM. In Wu et al. (2010) the authors used spatiotemporal Gabor filters to classify facial expressions. Although they report that using a non-linear SVM instead of a linear SVM yielded no significant performance increase, they state that their considerably large feature space (i.e., more than $2.2M$ per video sequence) generated by the non-linear spatiotemporal Gabor filter responses might have made their problem linearly separable. In our case the dimensionality of the feature space was never greater than 96. Not being able to generalize very well is a disadvantage for practical application where you would want to use these techniques out-of-the box, for new speakers. It is conceivable that better results for the generic model can be obtained when more data from more different speakers become available. In addition, in future work we plan to experiment with techniques that have the potential to make our STem-VVAD method generalize better to unseen speakers. For instance by scaling the mouth's bounding box to a fixed size, or by taking the complete (normalized) STGF transformed mouth area (after dimensionality reduction) as input to a classifier.

2.6 CONCLUSION

In general, we can conclude that STGFs offer a promising method for visual voice activity detection. In particular, we have shown that adding temporal information to the widely used spatial Gabor filters yields substantially better results, than can be obtained with Frame Differencing or "standard" Gabor filters, since STGFs make better use of the inherent visual dynamics of speech production.

In the next chapters, we will study whether STGFs also outperform static, spatial Gabor filters (SGFs) when applied to other social signal processing

tasks, beginning, in Chapter 3, with assessing how difficult children find a learning assignment, based on their non-verbal behaviour.

3 LEARNING DIFFICULTY ASSESSMENT

3.1 INTRODUCTION

In a tutoring environment (e.g., classroom or online course) affective states of students play an important role in learning (Grafsgaard, Wiggins, Boyer, Wiebe, and Lester, 2013; Kort, Reilly, and Picard, 2001; Lehman, Matthews, D'Mello, and Person, 2008; Masters, Barden, and Ford, 1979; Meyer and Turner, 2006). For children it has been established that their positive affective states stimulate learning whereas negative states inhibit it (Masters et al. 1979; Meyer and Turner, 2006). Kort et al. (2001) have extended this finding and suggest a broad range of valencies of affective states that influence learning. The authors have examined various emotional states, including anxiety-confidence and frustration-euphoria, that possibly play an important role in learning. Arguably, identifying and properly acting upon those affective states distinguishes expert tutors from novice ones. In this chapter, we will study whether Gabor filters can be used to detect these kinds of affective states, and whether, as we predict, the dynamic versions (STGFs) will do so better than their static counterparts (SGFs). Additionally, in this chapter we will compare Gabor filters with an alternative method, which more explicitly models the face of the learner.

So what, in general, are cues for students' affective states? Often, these will be closely related to their skill level. If students are presented with material far above their skill level this leads to frustration or anxiety which is counter-productive to learning, whereas material that is too easy leads to boredom or disappointment. According to Flow Theory (Csikszentmihalyi, 1990) one of eight mayor components of an optimal (learning) experience is the balance between the challenges of the presented problem and the student's skills to solve it. Put differently: problems should be just challenging enough. Meyer and Turner (2006) found evidence for this. They observed students in classrooms in student teacher situations. Students that were confronted with challenges that by far exceeded their skills, reported low experiences of flow, whereas students who were highly involved in challenging tasks reported more experiences of flow. Therefore, a proper balance between skills and challenges to get students motivated is essential for their learning process.

Ideally, for course material to be just challenging enough, a tutor would adapt the challenges presented to the skill level of the individual students. In (human) teacher-student situations this can be achieved through adjusted or personalized challenges, however for large groups this can be infeasible. For

This chapter is partly based on Joosten, B., van Amelsvoort, M., Krahmer, E., & Postma, E. (2011). Thin slices of head movements during problem solving reveal level of difficulty. *Proceedings of the International Conference on Audio-Visual Speech Processing 2011 (AVSP 2011)*. Aug 31 – Sep 3, 2011 Volterra, Italy, pp. 87-92. and Amelsvoort, M., Krahmer, E., Joosten, B., & Postma, E. (2013). Using non-verbal cues to (automatically) assess children's performance difficulties with arithmetic problems. *Computers in Human Behavior*, 29, 654-664.

these cases computer-aided learning systems may be provided. Since the 60's, computer-aided learning systems have been developed that determine the level of challenges presented partly based on the student's past input (Suppes, 1966). For example, when a student appears to be struggling, the system lowers the level of difficulty of the exercises or simplifies the instructions. The success of these early computer-aided learning systems was hampered by their inability to reliably assess the skill levels of students. The reason was that these systems determined skill level by only analyzing the answers provided by students. By only taking into account the answers given by students and neglect their affective states (e.g., boredom, frustration, happiness), these computer-aided learning systems are bound to fail.

Affective states are important, because, for instance a student can give a correct answer (e.g., by guessing) and still be insecure about it. In addition, compared to classical assessments such as written or oral exams, affective state assessments potentially spot problems earlier by monitoring the student's attitude (e.g., bored, frustrated, anxious) towards the material (Rothblum, Solomon, and Murakami, 1986). Looking inside students' minds to assess their affective state is generally not feasible. Potentially, non-verbal cues may be informative of the underlying affective state of the learner. Therefore, taking non-verbal cues into consideration could support computer-aided learning system's ability to evaluate students. This raises two questions: (1) which non-verbal cues do learners actually display? And, (2) can we detect these automatically?

Non-verbal cues associated with learning

As discussed in the general introduction (Chapter 1), non-verbal cues can be defined as any physically observable human type of behavior that is not directly derived from the spoken words yet may convey a certain message. These cues can be either intentional or unintentional and range from hand or bodily gestures to tone of voice and facial expressions (Knapp et al. 2013). For instance, they can amplify or emphasize the spoken message (e.g., with hand gestures and tone of voice) or they can signal a cognitive or emotional state (such as thinking, curiosity or happiness, using facial expressions or a certain body pose).

As in all everyday situations non-verbal cues occur in learning situations. Generally, a distinction can be made between non-verbal cues associated with affective (e.g., happiness, boredom) states and those associated with cognitive (e.g., concentration, puzzlement) or physiological (e.g., fatigue or pain) states.

A long line of research has looked into the non-verbal expression of emotion (e.g., Ekman, 1973). Traditionally, much of this work has concentrated on so-called basic emotions (e.g., Ekman, 1992a; Ekman and Friesen, 1975), some of which are prima facie relevant in the context of learning (e.g., surprise, because the learner is confronted with unexpected material, happiness, because the learner successfully understands material, or sadness, because he or she does not understand the material). However, it can be argued that other, more social emotions are at least as relevant (e.g., Adolphs, 2002b). Lehman et al. (2008) have identified four emotional states that also have (clear) non-verbal behavior associated with them, occurring significantly in one-on-one learn-

ing sessions with a human expert tutor: confusion, happiness, anxiousness and frustration. Other research suggest the presence of boredom, interest, surprise, curiosity, anger or satisfaction in learning (Craig, Graesser, Sullins, and Gholson, 2004; Sidney et al. 2005).

Besides affective states, cognitive states may also play a central role. Examples of cognitive states of relevance to the educational setting are: comprehension and contemplation (D'Mello et al. 2008; Hart, 2008; Howell and Shepperd, 2013). Inferring cognitive states from facial expressions is an active area of research (El Kaliouby and Robinson, 2005; Gatica-Perez, 2009; Littlewort, Bartlett, Salamanca, and Reilly, 2011a). An obvious source for these cues to originate from is the presented material and the student's level of knowledge towards it. Novel or complex topics are likely to provoke different non-verbal cues than well-known or simple ones do. In our earlier work, we have also found that children's learning states (i.e., whether they experience an arithmetic problem as easy or hard) are reflected in their non-verbal behavior of the children (Amelsvoort, Joosten, Krahmer, and Postma, 2013). Moreover, it was found that adult judges are capable of correctly interpreting this non-verbal behavior. More in particular, it was found that adults are able to determine above chance whether a child is experiencing learning problems based on children's faces and by listening to their voice, both in isolation and in combination. The visual cues generally were most predictive for participants. Pausing information was one of the strongest cues, but even when clips were presented without pauses, adult judges were able to determine whether a child found the arithmetic problem easy or difficult, based on just the answer. For more details, we refer to the paper.

An important channel of non-verbal cues for determining a student's affective state is the face. Whether students are easing through or struggling with a presented problem is often revealed by their facial expressions. Especially solving difficult problems often yields characteristic facial expressions (Craig, D'Mello, Witherspoon, and Graesser, 2008). The question is if and how such cues can be detected automatically.

3.1.1 Related Work

In the computer-aided learning domain several attempts have been made to incorporate facial expression detection in an automatic system (Banda and Robinson, 2011; Bosch, Chen, and D'Mello, 2014; D'Mello et al. 2008; Dragon et al. 2008; Grafsgaard et al. 2013; Kapoor, Burleson, and Picard, 2007; Kapoor and Picard, 2005; Littlewort et al. 2011a; Whitehill, Bartlett, and Movellan, 2008). In Bosch et al. (2014), Grafsgaard et al. (2013), Littlewort et al. (2011a), and Whitehill et al. (2008) the researchers used the Computer Expression Recognition Toolbox (CERT) (Littlewort et al. 2011b). CERT is a tool that automatically detects specific facial muscle movements, so called Action Units (AU) as well as head pose information and the basic emotions. AUs are the building blocks of the Facial Action Coding Scheme (FACS), a taxonomy to describe all possible facial movement, which we also briefly discussed in Chapter 1. The approximations of AUs and head pose data are explicit features pertaining to the physiological state of the face. Banda and Robinson

(2011) and Dragon et al. (2008) build upon the "mind reading" system developed by El Kaliouby and Robinson (2005) to infer mental states from facial expressions. These systems map the occurrence of geometric and appearance features through a hierarchical model of three layers to a probability for each one of six mental states (agreeing, concentrating, disagreeing, interested, thinking and unsure). The AutoTutor system (D'Mello et al. 2008) uses an array of sensors to detect cognitive states such as boredom, engagement/flow, confusion and surprise and is based on the IBM BlueEyes system developed by Kapoor and Picard (2005). By tracking the pupils of the eyes they can also infer the location of the eyebrows and subsequently measure upper facial action units using template matching.

Different methods for automatic facial expression detection exist, which mainly differ in their use of features and expression models. Generally speaking, we can identify two types of features, i.e., 1) geometrical features, estimated from the locations of fiducial points in the face, and 2) appearance features, which use pixel intensity values at a certain region of interest (ROI) of either the original input image or a (filter) transformed image. Automatic facial expression detection systems exist which employ one or multiple instances of the two different feature types (Ashraf et al. 2009; Dibeklioğlu et al. 2015; Littlewort et al. 2011b). Detecting the expression based on the features involves choosing the right classifier and determining whether to operate on static feature information (i.e., geometrical or appearance information from one frame or image) or also take the dynamical, temporal aspect of the expression into account.

The choice of feature type and whether or not to use temporal information, remains an active area of research and often largely depends on the expression(s) to detect and the available training data. Obviously, when the training data consists of isolated still images, temporal information as from successive frames is absent. Nevertheless, recent advances in automatic facial expression detection have shown to be effective at detecting certain expressions without using temporal information (Chu, De la Torre, and Cohn, 2013; Littlewort et al. 2011b). These advances mainly relate to what is being referred to as message-based expressions (Cohn, 2007), i.e., expressions with a clear intentional message such as the prototypical facial expressions ascribed to the basic emotions (i.e., joy, surprise, sadness, disgust, fear and anger). Training data for message-based expression detection is often obtained from posed or acted facial expressions. Unfortunately for automatic detection, spontaneous expressions, in contrast to their posed ones, often show less intense characteristic traits and may even differ in configuration and timing altogether (Reisenzein, Bordgen, Holtbernd, and Matz, 2006; Visser, Krahmer, and Swerts, 2014). Moreover, facial expressions related to more complex emotional states or cognitive states may even vary from person to person.

3.1.2 Current Studies

The goal of the current study is to compare the effectiveness of different automatic methods to track non-verbal cues as an indication of learning difficulties. Like in the other chapters of this thesis, we will be comparing

static (SGF) versus dynamic features Gabor filters (STGFs), to see whether the addition of dynamic information leads to an improvement of the results. In addition, in this chapter we will also compare the implicit Gabor filter method, with an explicit alternative method, which tracks specific predefined locations in the face.

More specifically, the explicit method is based on Active Appearance Models (AAM) (Cootes et al. 2001; Littlewort et al. 2011b; Matthews and Baker, 2004) that extract geometrical information of fiducial facial landmarks. Obviously, a single sample of an AAM's output contains no dynamics, however, considered over time, the displacements of landmarks represent facial movement. We will refer to this method as our explicit dynamic method, since it explicitly represents speed (i.e., in pixels per frame). The implicit method, as said, is based on spatiotemporal Gabor filters (STGF), as discussed in Chapter 1, and will be compared to SGFs. Responses of spatiotemporal Gabor filters consider pixel intensity changes within a specified window of frames. In our case the spatiotemporal Gabor filters operate on the pixel changes in facial regions, so-called appearance features. Since the filters can be tuned to respond maximally to specific motion, the responses of multiple individual filters generate a numerical signature of the frame's dynamics. The method based on STGF will be considered our implicit dynamic method, since movement is indirectly assessed through the filter's responses.

We will evaluate the two facial expression detection methods on the task of difficulty assessment. This task involves determining whether facial expressions displayed during the answering of questions, as assessed by the two methods, are indicative for the level of difficulty perceived by the surveyed.

In this study we use data of elementary school children (i.e., second and fifth grade) that answer arithmetic questions that are either *easy* or *hard*, based on their pre-assessed skill level. This dataset was collected in the context of a behavioral study (Amelsvoort et al. 2013) in Tilburg. Further details about the dataset are described in Section 3.3.1.

3.2 METHOD

To measure facial expressions we have introduced two methods, the explicit method and the implicit method, that employ two different types of features focusing on specific traits of facial movement. A method in our case is comprised of three stages: face localization, extraction of dynamic features, and evaluating their performance with respect to detecting facial expressions or affective or cognitive states displayed through facial expressions. The explicit method identifies per frame a set of fiducial landmarks on the face and evaluates specific geometrical displacements over time so that facial movement can be derived. The implicit method captures facial dynamics by measuring the intensities of moving contours at different spatial and temporal frequencies (i.e., spatial scales) and spatial orientations. Both methods will be discussed in greater detail in the next two subsections.

Explicit Method

Our explicit method's main component is the well-known Active Appearance Model (AAM, (Cootes et al. 2001; Matthews and Baker, 2004; Van der Maaten and Hendriks, 2010)), which is a statistical method that is able to learn a set of predefined landmarks on any instance of a generic object. In our case it is able to locate a set of fiducial landmarks on frontal faces. By manually specifying a grid of fiducial facial landmarks on a sufficient number of frames, the AAM method is able to create a prototypical model of a face that combines both the geometrical variation in landmark positions and the differences in appearance of the underlying faces. Figure 15 shows the specification of the landmark grid superimposed on the face in one of our frames. The dots (fiducial landmarks) and lines represent the grid. The lines connect dots belonging to the same facial part (e.g., mouth, nose, eyes, or eyebrows). If the selected annotated training frames are representative for most of the facial movement the method learns the correct landmark displacement which corresponds to the visual changes in the face. This makes it possible to automatically fit the grid of fiducial facial landmarks on faces in frames that have not been annotated.

Dynamic Feature

With the AAM applied to all the frames in the data set we have the location information of a substantial number of facial landmarks at our disposal. Given these locations over time we can extract numerous motion features that originated from certain head and facial movements. Our explicit method focuses on head movement. More specifically, we roughly approximate head pose in 2D space by computing a center-of-mass feature that changes with movements of the head. The large dot in the center of the triangle illustrated in Figure 16 represents our center-of-mass feature, computed using the locations of the eyes and chin (i.e., corners of the triangle) and the tip of the nose. From the Cartesian vectors defined by the change of position of our center-of-mass feature we derive polar coordinates which represent the angle (in the 2D plane) and magnitude of the head movement.

Implicit Method

Our implicit method relies on Gabor filters to quantify spatial and temporal changes in the input images. In general, filters applied to images calculate a new value for each pixel by combining filter coefficients with the values of the surrounding pixels. The filter coefficients determine what type of pixel transitions will mostly be affected. In spatial Gabor filters the filter coefficients are defined by a Gabor function as defined in Section 1.3. Using the parameters of the Gabor function we can create filters that respond maximally to pixel transitions at a specific orientation and with a specific spatial frequency, making them well suited to code facial movements.

Spatial Gabor filters have proven to be quite effective in classifying facial expressions (Littlewort et al. 2011b). STGFs extend spatial Gabor filters with the dimension of time, by incorporating the pixel transitions within a given

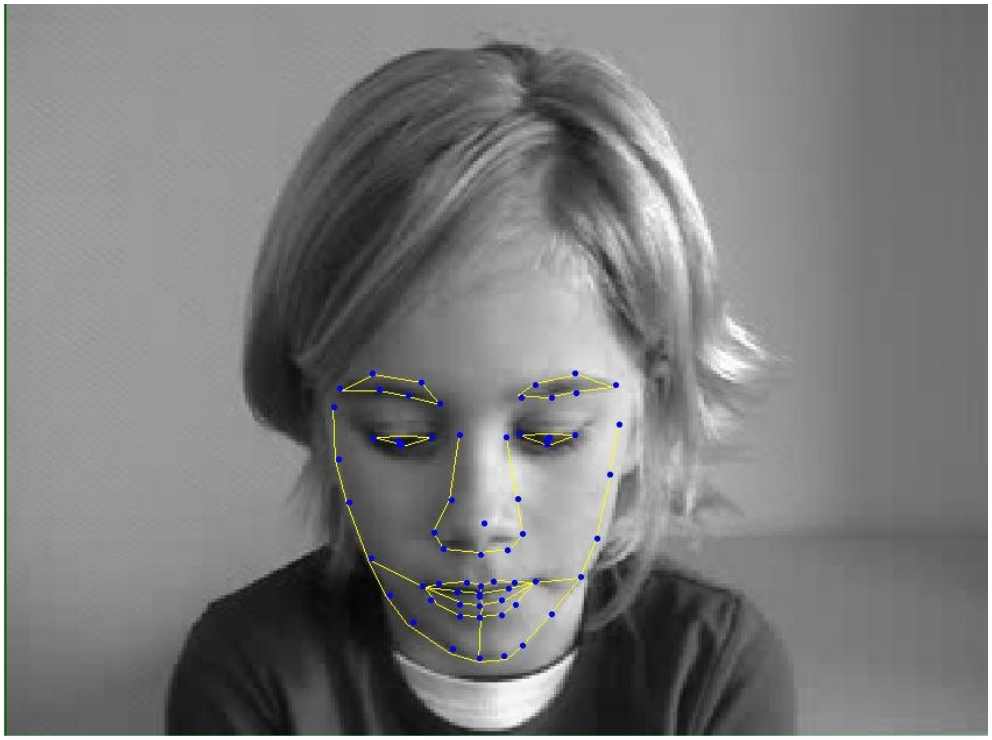


Figure 15: Example of a manually annotated landmarks (dots) with the lines connecting them representing the grid projected on a single frame in the data set.

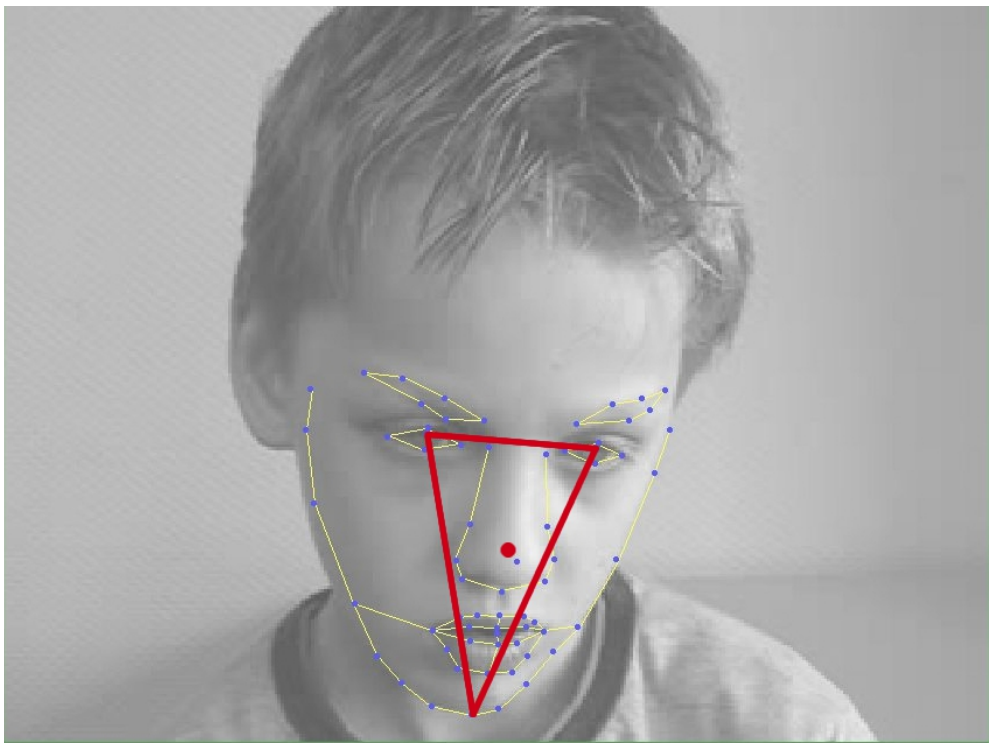


Figure 16: Illustration of our center-of-mass feature. The position of the large dot is the geometrical center of the fitted landmark positions of the eyes and the middle of the chin (depicted by the triangle) and the tip of the nose, approximating their center-of-mass.

Figure 17: Illustration of two frames in our data set (upper row) and two corresponding spatiotemporal Gabor filter responses (lower row). The left filter was tuned to respond maximally to upward movement ($\theta = \pi/2$) at speed is 1 pixel per frame (PPF) and spatial frequency of 0.35 pixels per cycle and the right filter is tuned to rightward movement ($\theta = 0$) at a speed of 2 PPF and spatial frequency of 0.22 pixels per cycle. The darkness of a pixel is proportional to the filter response.

window. Effectively, this means that we can tune filters to respond to moving contours with a specific direction, speed, and spatial frequency. Figure 17 illustrates the transformation of two pairs of subsequent frames in our data set using two specific STGFs. The left filter was tuned for upward motion at the speed of 1 pixel per frame, whereas the right filter responds maximally to rightward motion at 2 pixels per frame (PPF). Filter responses are represented by shades of gray. Darker shades represent higher filter responses. Note that the filters responds to contours that are perpendicular to the direction of movement that the STGF is tuned to. With a filterbank of STGFs that covers the range of movements in our data set we "measure" the responses in terms of movement generated by the facial expressions.

Dynamic Feature

Our implicit method convolves a filter bank of G STGFs on every video sequence in the data set, resulting in G transformed sequences. Using a face detection algorithm we determine the location of the head from which we extract the filters' responses. For each filter in our filter bank we aggregate the responses of the identified head region by summation yielding one value per filter per frame. Each value signals the presence of visual structure that matches the filter's properties. For instance, a large value for a vertical filter tuned to rightward motion indicates the presence of rightward moving

vertical contours in the facial region. Our data vector *preliminary* to our final feature vector thus consists of $G \times F$ aggregated STGF values for a given window of time (where F represents the number of frames). Similar to our spatial aggregation, we also apply a temporal aggregation of the responses. Per region of interest we average each spatially aggregated value over the number of frames in the time window, which results in our *final* feature vector. Predicting which type of question was posed amounts to training a support vector machine on labeled feature vectors.

3.3 EXPERIMENTAL EVALUATION

We evaluated our two aforementioned dynamic facial expression detection methods, viz., the explicit method, and the implicit method on the task of difficulty assessment. For this task we used data acquired during an expression eliciting experiment (Amelsvoort et al. 2013). The next subsection will give more details about the data collecting procedure and the characteristics of our data set. Then we will discuss our experimental approach to difficulty assessment, which we based on the notion of *thin slices* (Ambady and Rosenthal, 1992) and we will provide relevant implementation details for both the explicit method as well as the implicit method. Finally, we report the evaluation procedure we employed to test the effectiveness of both methods.

3.3.1 Dataset

Our dataset consists of video recordings of children in second (group 4 in the Dutch school system) and fifth grade (group 7) of elementary school. Each video clip captures a child's response to an arithmetic problem up until the given answer. For each child there are two recordings corresponding to two levels of difficulty of the problem, i.e., easy and difficult (based on level of skill expected for children of their respective grades), which also comprise our two labels. By varying the level of difficulty the purpose of the data acquisition was to elicit different facial expressions in both conditions. Given the straightforward answers they evoke and the structured ways to determine what type of arithmetic level a child should have, these problems are well suited for this expression eliciting task.

During the data collection the children were instructed that they would participate in the evaluation of a new video brain game, where they would have to answer arithmetic problems a quickly as possible. In the recording setup children were presented with a PowerPoint presentation that resembled a game interface that showed them arithmetic problems. Examples of the slides are shown in Figure 18. These problems were taken from an official test in the Dutch School System, one with which problems could be easily mapped to the easy or hard category, based on the level a particularly child should have. On the laptop showing the PowerPoint presentation, a camera was placed that recorded children's facial expressions to the presented problems.

Consented data were collected from 55 children, resulting in a data set of 110 videos (i.e., one video per category). The set is almost balanced for grade,

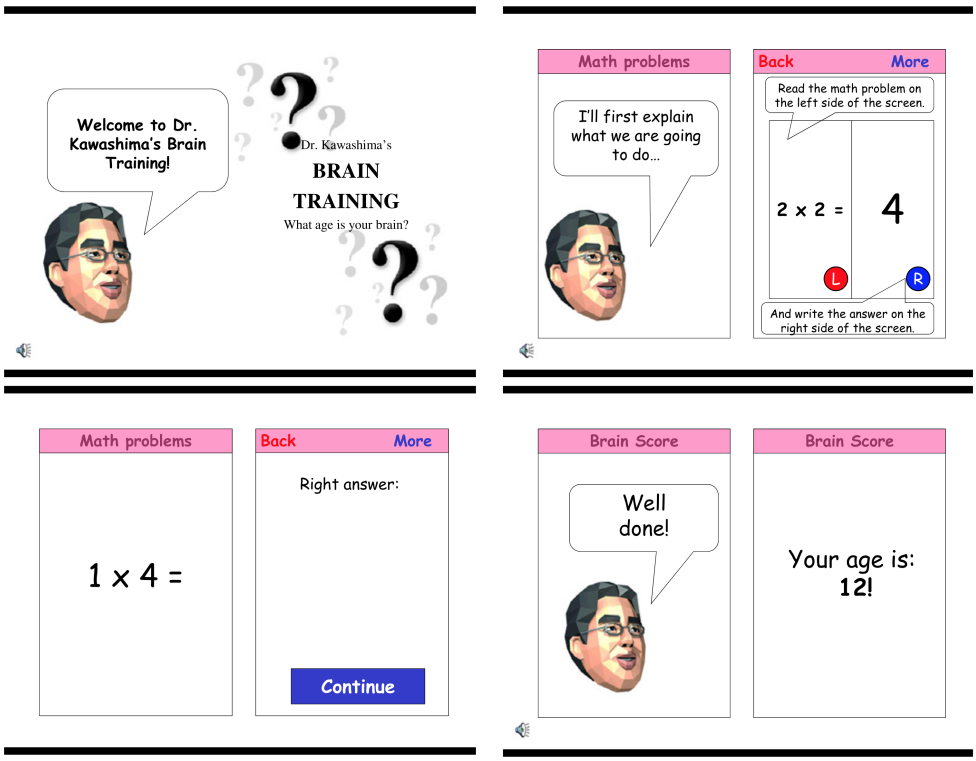


Figure 18: Slides from the game interface (Amelsvoort, Joosten, Krahmer, and Postma, 2013), extracted from the Nintendo game: Dr Kawashima's Brain Training How old is your Brain?

i.e., 27 second grade children (14 boys, 13 girls) and 28 fifth grade children (15 boys, 13 girls). Videos are recorded with a frame rate of 25 frames per second with video lengths varying from 37 frames in the shortest clip and 1114 frames in the longest clip. Representative stills are shown in Figure 19.

There are small variations of position and pose of the head between and within videos. However, with respect to the camera, children tend to look slightly tilted downwards and a little to the left. Partial facial occlusions due to glasses and hair occur in some of the clips and there are moderate variations in illumination since the videos were recorded throughout the day using a natural light source.

3.3.2 Implementation Details

The human judgment experiment of Amelsvoort et al. (2013) revealed that pause information (e.g., the duration of pauses) is one of the strongest cues for difficulty, hence including this cue in an automatic method would be easy but arguably also somewhat trivial and uninformative. Our goal here is to see whether automatic detection is also feasible on short fragments. Inspired by the notion of *thin slicing* (Ambady & Rosenthal, 1992; Gladwell, 2005), we decided to restrict our computational analysis to the first second (i.e., 25 frames) of each fragment. The "thin slice paradigm" conjectures that certain behavior can already be spotted by observing only a small window of time.

Figure 19: Representative stills from our dataset with varying head poses and facial expressions. The top row shows children performing easy problems, the bottom hard ones.

Note also that this approach is helpful for intelligent tutoring systems, because they would benefit from detecting relevant cues as quickly as possible.

Both facial expression detection methods rely on the correct discovery of a person's face in each video frame in order to correctly measure expressions. For face detection we used the OpenCV implementation of the Viola and Jones (2001) face detector with Local Binary Pattern features. This detector is very fast and has a high detection rate. We estimate the locations of faces where the detector was unsuccessful by interpolation. Unfortunately for three video sequences reliable face detection was not possible, therefore these sequences were omitted from the analysis. Next we will describe the experimental setup for both of the expression detection methods using the face locations as input.

Explicit Method: AAM

Active Appearance Models do not work out of the box. They need at least a number of representative annotated images to fit the landmark grid to unseen instances. Although pre-annotated datasets exist, these usually consist of images of adults. Thus, for AAMs to work on our dataset we had to manually annotate a substantial representative portion of the frames before we could apply them to the rest. Per participant, we manually selected at least 8 frames that varied as much as possible in pose and expression and adopted a landmark location configuration of 66 points inspired by Gross, Matthews, Cohn, Kanade, and Baker (2010). Specifying the positions of the landmarks was performed with the publicly available AAM annotation tool developed by Tim Cootes [1].

Following the annotation phase is the creation of the actual shape and appearance model that constitute the AAM, for which we used the MATLAB® implementation provided by Laurens van der Maaten [2] (Van der Maaten and Hendriks, 2010). With the instantiated AAM the facial landmark grid is fitted to each frame, including the ones previously annotated.

We extract the center-of-mass feature Figure 16 for each frame in the thin slice and convert its displacement between consecutive frames to polar coordinates. We aggregate the coordinates of each slice in a 2D histogram which divides the angles over 10 bins (ranging from $-\pi$ tot π radians) and the magnitudes over 6 bins (ranging from 0 to 6 pixels). One such histogram per participant is considered our final feature for classification.

Implicit Method: Petkov and Subramanian STGF

Our first implicit method uses the implementation of Petkov and Subramanian (2007) as discussed in Section 1.3 and will henceforth be referenced as PS-STGF. Their approach allows us to construct velocity tuned STGFs. We constructed a filter bank of $G = 6 \times 8$ filters sensitive to 6 different speeds ($v = \{0.5, 1, 1.5, 2, 2.5, 3\}$ PPF), 8 orientations ($\theta = \{0, 0.25\pi, 0.50\pi, 0.75\pi, \ldots, 1.75\pi\}$ radians), to cover as much as possible the variation in speeds and orientations of movement present in our dataset. The remaining parameters are set to their standard values as reported by Petkov and Subramanian (2007). We set the

1 http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/software/am_tools_doc/index.html

2 http://sspnet.eu/2011/03/active-appearance-models/

parameter for the Gaussian envelope to "moving", which results in velocity tuned filters (setting the parameter to "stationary" yields temporal frequency-tuned filters). The video sequences in the dataset were convolved with the PS-STGFs. The energy responses of the area enclosed by the bounding box indicating the location of the face is aggregated as specified in Figure 3.2. The dimensionality of the resulting PS-STGF feature vector for frame f, $A(f)$, is equal to $G = 6 \times 8 = 48$. We also construct a static variant (PS-SGF) to compare the relative contribution of the dynamic filter. For this variant we set $v = 0$, i.e., the zero-speed filter responses, which yields an eight-dimensional feature vector.

Implicit Method: Heeger STGF

The second implicit method we evaluated is frequency-tuned in the temporal plane, i.e., the envelope of the spatial Gaussian is stationary. This Heeger method was introduced in Section 1.3 and will be referred to as H-STGF. Here we opted for a straightforward implementation (Zorn and Lokesh, 2010) inspired by the work of Heeger (1987), who used the Gabor motion energies of multiple filters to estimate velocity in image sequences, as explained in Section 1.3. Each filter is constructed using 6 input variables, viz., 2 spatial center frequencies (x_0, y_0) expressed in cycles per pixel, one temporal center frequency (t_0) expressed in cycles per frame, and 3 standard deviations $(\sigma_x, \sigma_y, \sigma_t)$ for the Gaussian envelope in each axis (i.e., x,y and t). Table 4 lists the center frequencies we used that are sensitive to 8 different orientations. We also varied the spread of the Gaussian envelope in the 2D spatial domain, by restricting $\sigma_x = \sigma_y = \{2, 4, 8, 16\}$ and keeping σ_t at 1. This resulted in a spatiotemporal filterbank of $G = 8 \times 4$ filters. Also here, we construct a static counterpart (H-SGF) of the dynamic features, by setting the temporal Gaussian in the filter's equation to 1. Since the static variants of two perpendicular oriented directional dynamic filters (e.g., filters sensitive to left and right movement) are identical, our static feature vector is half the length of its dynamic counterpart: $G = 4 \times 4$.

Table 4: Center frequencies for each of the eight orientations of our Heeger STGFs. Spatial frequency ($\omega_x and \omega_y$) is expressed in cycles per pixel and temporal frequency (ω_t) in cycles per frame (Zorn and Lokesh, 2010).

orientation	ω_x	ω_y	ω_t
↙	1/4	−1/4	−1/4
↓	1/4	0	−1/4
↘	1/4	1/4	−1/4
→	0	1/4	−1/4
↗	1/4	−1/4	1/4
↑	1/4	0	1/4
↖	1/4	1/4	1/4
←	0	1/4	1/4

3.3.3 Evaluation Procedure

The comparative evaluation of the implicit and explicit dynamic features relies on a single classification procedure. Often in visual pattern recognition experiments the performance greatly depends on some sort of dimensionality reduction scheme or other feature selection procedure to optimize results. For most classifiers there are generally various parameters that can be tweaked to improve results of classification experiments. In this experiment we adopt a straightforward feature selection procedure that we apply to the data of both the explicit as well as the implicit methods in order to make a fair comparison between all methods. Our goal is not to construct an optimized system that is robust in assessing a person's perceived level of difficulty.

The optimization scheme we apply is able to sift through the important features and can learn fairly complex decision boundaries but with only *one* model parameter to vary. For our purposes we use the Random Decision Forest (RDF) classifier in MATLAB® to 1) find the most relevant subset of features, and 2) to determine the number of trees that yields the highest Out-Of-Bag Accuracy (OOBA). Our procedure is consists of three steps. In the first step we grow an RDF with an arbitrarily chosen number of 50 trees, and all other settings at their default values, except for the parameter `OobVarImp`, which is set to '`on`'. This parameter allows us to evaluate afterwards the importance of each feature for classifying the out-of-bag samples during the creation of the RDF. Feature importance is determined by observing the classification error after permuting the values of a specific feature in the out-of-bag samples. If the error increases the feature is considered relevant (after all, the original values achieved a better classification). The number of relevant features varies per method and if the procedure does not find any informative features we fall back to using all available features. The number of features selected following this approach typically ranged from 5 to 9 features. In the second step we take a subset of the data using the identified "important" features from the previous step. Then, we once more construct an RDF, and let it use up to 500 trees. Again, we use all the default values, except for `OOBPred` which is turned on to check the classifier's performance on the OOB samples at all intermediate number of trees. Finally, in the third step we report the OOBA of the model with the number of trees that resulted in the highest performance. These three steps give rise to optimized models that may have been overfitted. However, our intention is to find the optimal model to exploit both STGF variants maximally, in order to establish which method is the best. Since RDFs rely on random re-sampling of the data, different partitions of the data give different performance results. To account for the probabilistic behavior of the classifier, both aforementioned steps are repeated a hundred times and their results are averaged. The motivation for describing performance in terms of accuracy is that our dataset is almost balanced, this in contrast to our datasets in Chapter 2.

Table 5: Average out-of-bag accuracy scores plus standard deviation for five types of non-verbal cues assessment methods.

Method	Accuracy	Standard Deviation
AAM	0.66	0.02
PS-STGF	0.53	0.03
H-STGF	0.56	0.02
PS-SGF	0.49	0.04
H-SGF	0.51	0.04

3.4 RESULTS

We start by illustrating the results of the explicit methods. This distinction is clearly visible in Figure 20, where for each participant the raw coordinates (i.e., with respect to the coordinates of the frame) of the center-of-mass features are plotted as a trace for the easy (left) and hard (right) questions during one second (25 frames), directly following the presentation of the stimulus. The traces for the easy questions tend to be vertically oriented, whereas the traces for the hard questions seem to be diagonally oriented. These traces correspond to vertical and diagonal movements of the head (nose), respectively. The movements are also apparent in some individual fragments. Figure 21 provides an illustration by showing the movements of the same participant in both the easy and hard conditions. The photo's show two color-coded frame differences between the first and the last frame of our thin slice to emphasize the movement. Magenta-colored regions denote the position of the participant at frame $t = 1$, whereas the green parts correspond to the position at frame $t = 25$. Furthermore, we superimposed the center-of-mass trajectory for reference. By examining the displacements of the eyes in the photo's, we clearly see a vertical (top image) and a diagonal (bottom image) direction for the easy and hard question respectively. Apparently, this distinction in movement is also picked up by the RDF classifier which is summarized in Table 5. The first row in this table indicates a 66% average OOB classification accuracy for the explicit (AAM) method. The AAM result clearly illustrate that the explicit method is able to detect behavioral differences in head movements for easy and hard questions, at least to some extent.

We now turn to a quantitative comparative evaluation of our implicit methods with the explicit method. Inspection of Table 5 reveals that the explicit method outperforms all implicit methods, with a 66% accuracy for the AAM method versus a 56% accuracy for the best achieving implicit method, H-STGF. Importantly, for both Gabor implementations we find that whereas the static (SGF) variants perform near chance level, their dynamic counterparts (STGF) result in a (somewhat) better performance. This suggests that the more relevant information for the task at hand is in the facial dynamics.

The results of the comparative evaluation of the different methods is illustrated in Figure 22. The box-whisker plots depict the distribution of

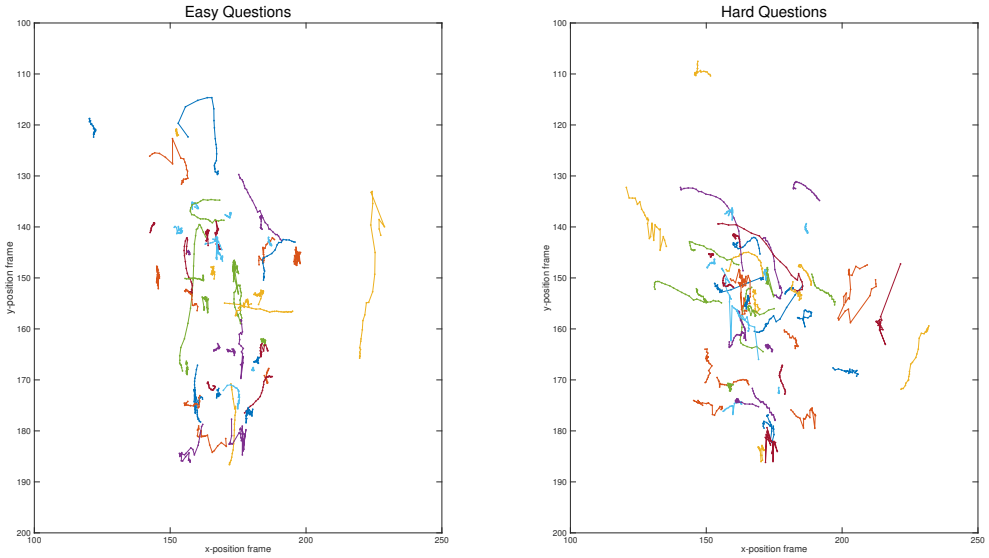


Figure 20: Individual trajectories of the center-of-mass feature plotted on a zoomed in region of the frame for all video clips. Different colors represent individual participants although multiple participants share the same color and are therefore merely added as a visual clarification.

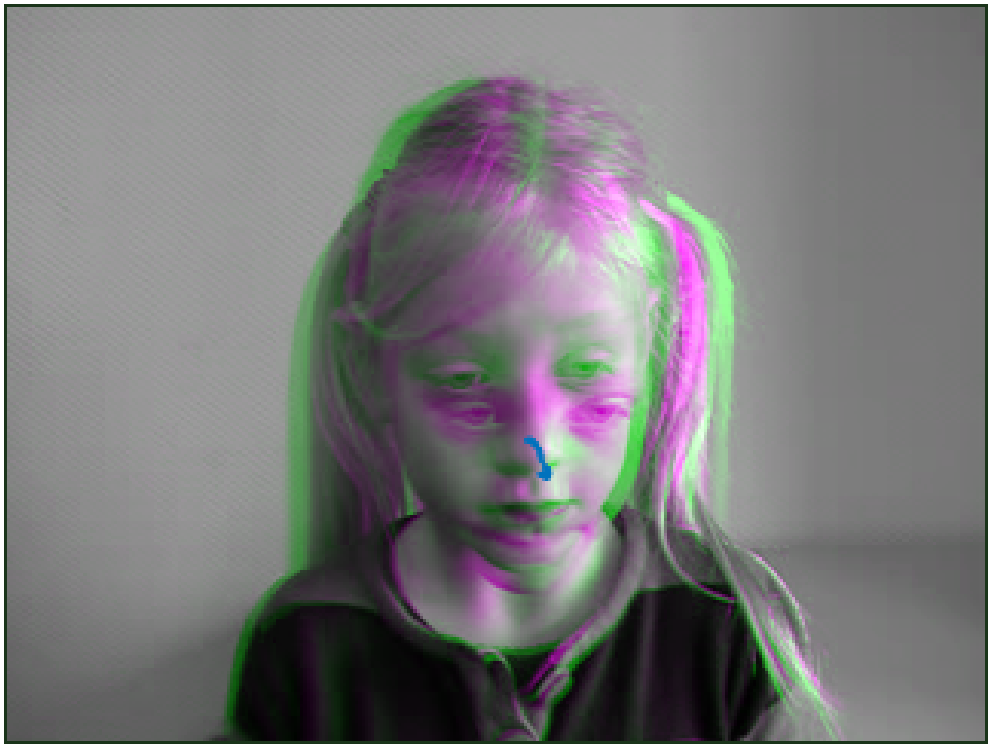
the out-of-bag accuracy (OOBA) for an RDF classification experiment that was repeated a hundred times. We refer to Section 2.4 for an explanation of the representation of boxes and whiskers. From left to right, the five plots show the distribution of accuracies for the explicit method (AAM) and the two dynamic implicit methods (PS-STGF and H-STGF) and the two static implicit methods (PS-SGF and H-SGF). The dashed line crossing all boxes corresponds to classification at chance level. Methods for which the box-whiskers extend below the chance-level line perform on a par with random guessing (i.e., tossing a coin). From this plot it is apparent that the explicit dynamic AAM feature outperforms all implicit features based on Gabor filters, since the whole box-whisker of the AAM extends above all other box-whiskers. Furthermore, it is noteworthy to mention that only for the AAM method and the H-STGF method each subject has a higher average accuracy than chance.

3.5 DISCUSSION

The goal of this study was to compare different approaches to detect non-verbal cues for learning. As in the other chapters, we compare static (SGF) and dynamic (STGF) Gabor filters for this task. Additionally, we compare the implicit Gabor method with the explicit AAM method.

These methods differ first and foremost in how explicitly they model movement. The explicit method used AAMs to determine facial landmarks and subsequently derived an explicit movement feature that represents rigid head movement. The implicit method is based on S(T)GFs, whose features model facial movement implicitly, for which we examined two implementations,

(a) Easy question



(b) Hard question

Figure 21: Composite images of the first and last frame of two thin slices pertaining to (a) an easy question and (b) a hard question. Superimposed are the trace plots of our center-of-mass feature for all subsequent frames which shows the (subtle) difference in movement orientation, i.e., vertical for easy questions and diagonal for hard questions. The gray overlay represents the areas where the two frames have the same intensities. The magenta (first frame) and green (last frame) regions show where they differ.

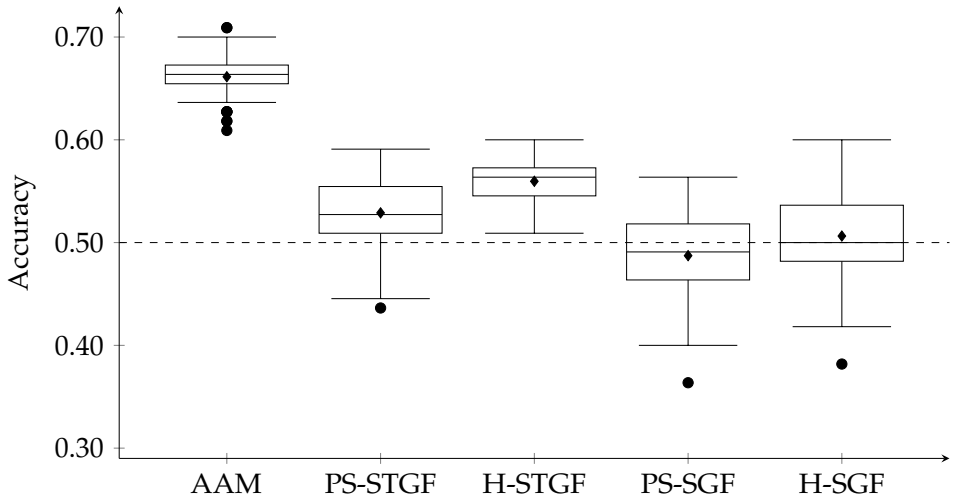


Figure 22: Estimation of the classification performance of difficulty assessment for five methods methods of movement detection: 1) AAM, 2) PS-STGF, 3) H-STGF, 4) P-SGF, and 5) H-SGF. The box-whiskers represent the spread of the out-of-bag accuracy (OOBA) for a repeated (100 times) random decision forest (RDF) classification experiment with different random initializations per repetition. For each feature type we chose the number of trees (i.e., $[1, 500]$) in the RDF experiment that maximizes the average accuracy. The diamond in each box represents the averaged classification accuracy for all samples. The dashed line indicates performance at chance level.

i.e., 1) the implementation by Petkov and Subramanian (2007), and 2) an implementation based on Heeger (1987).

Results for the explicit, AAM-based method revealed that head movements as captured by center-of-mass feature help in distinguishing easy from hard problems, where easy problems are associated predominantly with vertical movements and hard problems are more likely to be accompanied by movement in a diagonal direction. It is worth emphasizing that this cue is already apparent in the first 25 frames, and detecting this cue in such a short fragment for human judges is presumably rather difficult. This orientation preference could be in line with the findings of Wells and Petty (1980). In this work, the researchers studied the effect of vertical (nodding) versus horizontal (shaking) head movement on processing persuasive messages. They postulated that bodily movement can either enhance or inhibit specific behavior, depending on a previous positive or negative association, respectively. In this case, head nodding is generally associated with agreeing (positive), while head shaking often means disagreeing (negative). Participants had to listen to either a pro-attitudinal or a counter-attitudinal spoken message whilst simultaneously moving their head either vertically or horizontally. One of the outcomes of the study was that participants who had to perform (incongruent) vertical head movement in the counter-attitudinal condition found this harder to do than participants who were instructed to move their heads horizontally. In the pro-attitudinal condition the reverse effect was observed, thus, suggesting a preference for congruent head movement with respect to the attitude towards the message. It can be argued that arithmetic problems do not evoke a strong opinionated attitudes, however, depending on skill level or level of academic

risk taking, students might rather answer easy questions than hard questions. The predominantly vertical head movement displayed by the children when answering easy questions could be attributed to a "preference" for easier questions, which could facilitate this movement. Following similar reasoning, being presented with a harder problem could also steer their movement, in this case mostly in a diagonal direction.

While the explicit method worked surprisingly well, our two implicit methods fared considerably less well. Based on our earlier work on visual speech detection in the previous chapter, where we showed that STGFs outperformed various other methods, including, most notably, *static* SGFs, we expected the implicit method to work well for this task. Also other studies (Siritanawan, Kotani, and Chen, 2014; Wu et al. 2010) suggested that STGFs are capable of picking up (subtle) signals, but this could not be confirmed for our current data set. However, crucially, we did find that the two STGFs outperformed the corresponding SGFs, indicating that there is a benefit (albeit a small one) of adding dynamic information for detecting learning difficulties in our stimuli.

We identify two possible explanations for the poorer performance of the implicit method, compared to the explicit one. The first possible explanation is inherent to the averaging over many STGFs as happens in our aggregation scheme. When averaging filter responses over an area as crude as the bounding box around a face, subtle changes of task-relevant movements (the "signal") picked up by one or a few filters, could be smoothed out by the task-irrelevant movements picked up by the remaining filters (the "noise"). This limitation could be addressed by abandoning the spatial aggregation or by reducing the spatial extent of aggregation. A second possible explanation could reside in the vast STGF parameter space. It is well-known that the optimal configuration of parameters for a given visual task requires empirical studies Long, Wu, Movellan, and Bartlett (2012). In the most straightforward case STGFs have three components that can be adjusted, i.e., 1) spatial frequency, 2) orientation, and 3) spatiotemporal frequency. Optimizing the parameter space for each specific task is often infeasible. In our work we choose to keep the parameter settings at the suggested values we found in the respective papers (Heeger, 1987; Petkov and Subramanian, 2007). It may be the case that our task requires different parameter settings.

In addition, it is important to add that the AAMs require manual preprocessing, while the S(T)GFs do not require this. Certainly when it comes to the subtle movements under investigation here, it may be that this manual preprocessing step is essential for good performance results.

In general, our results suggest that it is feasible to automatically assess learning difficulty based on facial expressions of children, where the best performance is obtained using an explicit method. Note that these results are based on thin slices, which suggests that this kind of information can already be used in early stages. However, in practice, the best and more robust results are probably obtained by including longer stretches (including pause information as well as other specific non-verbal cues, such as frowning or lip puckering) as well auditory information. Finally, with the present feasibility and state-of-the-art performances in various visual domains, representation learning using deep neural networks (Goodfellow, Bengio, and Courville,

2016) with time-varying filters, could help to automatically develop and configure the best spatiotemporal filter settings through learning.

3.6 CONCLUSION

In this study we set out to compare different methods of measuring non-verbal cues in the face for assessing learning problems. We compared the performance of two implicit methods based on Gabor filters, one static (SGF) and one dynamic (STGF) in two different implementations. In addition, we compared these to an explicit method, relying on AAMs. We found that the explicit method clearly outperformed the implicit methods. However, when comparing the spatial with the spatiotemporal Gabor filters we found that adding dynamic information did improve the classification accuracy, albeit with a small margin. In general, we conclude that dynamically assessing facial cues in the context of answering arithmetic problems is feasible (even within a very small window of time), however that the movement that clearly sets the two conditions in our experiments apart is too subtle for our spatiotemporal Gabor filters to pick up.

4 SMILE CLASSIFICATION

4.1 INTRODUCTION

In the preceding two chapters, we have looked at two social signal processing tasks of different complexities — visual voice activity detection (Chapter 2) and learning problem assessment (Chapter 3) — and compared static, spatial Gabor filters (SGFs) with their dynamic, spatiotemporal counterparts (STGFs). We found that STGFs outperformed SGFs in both cases, although only in the first case (Chapter 2) did we see a substantial improvement of STFGs over SGFs; in the second case, both SGFs and STGFs were outperformed by a method based on Active Appearance Models. In this chapter we continue our study of the added value of STGFs, by considering a new SSP task: automatically classifying smiles as genuine or not.

It is well established in the literature that people can smile in at least two different ways, either because of genuine happiness (the so-called Duchenne smile) or as a social response (the non-Duchenne smile) (Ekman, 1992b; Johnston, Miles, and Macrae, 2010; Kraut and Johnston, 1979). Even though the distinction between the two can be subtle, researchers have argued that the distinction is noticeable in terms of the parts of the face that play a role (the Duchenne smile is primarily noticeable around the eyes) as well as the speed of movement – the Duchenne smile apparently takes longer to fully appear on the face as well as to disappear(Krumhuber et al. 2009; Schmidt et al. 2006). For these reasons it seems an ideal test case for comparing static and spatiotemporal Gabor filters, as we will do in this chapter.

In addition, we compare results for the original recordings with those for variants in which the rigid head movements are automatically controlled (by means of normalizing the position of the head with respect to the location of the eyes), on the assumption that this will make it easier for the spatiotemporal Gabor filters to detect relevant, non-rigid facial and head movements. Moreover, we also compare results for the entire face as well as for movements in more specific facial regions, based on earlier claims that these allow for better classification of smiles as spontaneous or not.

Posed vs. Spontaneous Smiles

The smile has been argued to be the most complex facial expression because it can convey a surprisingly wide range of intentions (Niedenthal and Mermillod, 2010). Morphologically, the smile is primarily a contraction of the *zygomaticus major* muscle causing the corners of the lips to move in an upward direction, the so-called "lip corner puller" (Ekman and Friesen, 1976). The *zygomaticus major* is controlled by two motor pathways: one that produces non-voluntary facial expressions, and one that produces deliberate expressions (Niedenthal and Mermillod, 2010). Extensive research has been

done to distinguish between the non-voluntary (true) and deliberate (false) smile. The Duchenne marker is probably the most well-known result of this research. Named after Guillaume-Benjamin Duchenne it is the contraction of the *orbicularis oculi* which causes the characteristic wrinkles around the eyes when they narrow as well as the uplifting of the cheeks. It was believed that the presence of the Duchenne marker accompanied by a contraction of the *zygomaticus major* indicates a "true" smile (Ekman et al. 1990; Surakka and Hietanen, 1998). According to this belief, true smiles stem from genuine happiness, whereas "false" smiles are merely used in social contexts to deliberately display a positive expression, without experiencing the positive mood (Krumhuber et al. 2007; Niedenthal and Mermillod, 2010). In recent years there has been an increasing interest in the computational analysis of facial expressions. With advanced image coding and machine learning algorithms, the statics and dynamics of smiles are analyzed to automatically recognize different types of smiles (Dibeklioğlu et al. 2015).

4.1.1 Related Work

Recently, a number of studies have emerged on the topic of automatic facial expression classification (Bettadapura, 2012; Fasel and Luettin, 2003; Lyons et al. 1998; Pantic and Rothkrantz, 2000; Sariyanidi, Gunes, and Cavallaro, 2015). Where the focus used to be on (static) posed or prototypical expressions (Cohn and Schmidt, 2004; Du, Tao, and Martinez, 2014; Pantic and Rothkrantz, 2000) more recent work tends to emphasize the role of the facial dynamics of spontaneous expressions (Cohn and Schmidt, 2004; Dibeklioğlu et al. 2015; Valstar, Pantic, Ambadar, and Cohn, 2006). The evolution of smile detection followed a similar path (Cohn and Schmidt, 2004): after an initial period of research using posed single image smiles, new studies focus on the automatic recognition of naturally occurring smiles.

Static Smiles

Although the majority of research in automatic posed versus spontaneous smile detection focuses on temporal patterns of smile dynamics, some work has been done on static images as well (Liu and Wu, 2012; Nakano, Mitsukura, Fukumi, and Akamatsu, 2002; Radlak, Radlak, and Smolka, 2018). One of the earliest works, by Nakano et al. (2002), used Principle Component Analysis (PCA) to represent images of "true" and "false" smiles. Subsequently, a neural network trained on angles between individual images and the PCA represented classes (i.e., true or false smile), classified unseen examples in a leave-one-out manner. On their private dataset with 25 subjects they achieved a score of around 90% correct. However, given the limited size and lacking of formal description of their dataset, the performance is hard to evaluate.

More recently, the work by Radlak et al. 2018 also focused on genuine smile detection by using only static images. They used smile apexes of the 1240-participants large UvA-NEMO dataset to train a Support Vector Machine on Local Binary Pattern features collected at different facial configurations and using two types of face normalizations. Their maximum classification score was around 65%, where the type of normalization only contributed

marginally to the performance. Unfortunately, the authors do not mention the effect of normalization compared to unnormalized images.

While from a theoretical perspective, automatic static smile discrimination is very interesting, especially with regard to human performance, from a practical perspective, however, compared to systems that incorporate smile dynamics, Radlak et al. 2018's approach performs under par. In what follows we discuss three aspects of dynamic automatic smile recognition that play a role in existing research: the measurement of facial dynamics, the localization of smile cues, and the importance of face registration.

Measuring Facial Dynamics

Smile dynamics, like other facial expressions, can be measured using geometric features, appearance features or a derivative thereof. The measurement of facial dynamics generally relies on identifying or detecting fiducial points, or landmarks in the face, like the corners of the eyes and mouth or the tip of the nose, that can be referenced across each frame. Geometry-based aspects of facial movement, such as speed of movement, amplitude of movement, and duration of movement, are typically calculated from displacements of automatically tracked facial landmarks. Various studies have combined geometrical measures and landmarks to classify smiles (Cohn and Schmidt, 2004; Dibeklioğlu et al. 2015; Dibeklioğlu, Salah, and Gevers, 2012; Hoque, McDuff, and Picard, 2012; Trutoiu, Hodgins, and Cohn, 2013; Valstar, Gunes, and Pantic, 2007).

Geometrical measures applied to landmark representations suffer from two main limitations. The first limitation is that when the accuracy of the landmark tracker drops (e.g., due to illumination changes or out of plane rotations of the face), the geometrical measures are distorted, which results in deterioration of the smile detection performance. The second limitation is that appearance cues such as bulges, frowns and wrinkles are not captured by geometrical measures of landmark representations. These limitations can be addressed by using appearance features to detect smiles (Sénéchal, Turcot, and El Kaliouby, 2013; Wu, Liu, and Zhang, 2014).

Dynamic Smiles

The type of feature that is best suited for automatic smile detection depends on the specific context. For example, the Duchenne marker (wrinkles around the eyes) is perhaps best detected with an appearance feature, whereas a geometry-based approach may be the best choice to measure the amplitude of the lip corner pull.

As mentioned in the previous section, smiles can be signaled from multiple visual cues, of which the shape of the mouth is obviously the most prominent one. In addition, it has been shown that head or shoulder movements (Valstar et al. 2007), eyelid and cheek movement (Dibeklioğlu et al. 2012; Trutoiu et al. 2013), and eye blinking (Trutoiu et al. 2013) contribute to distinguishing posed from spontaneous smiles.

In the study by Wu et al. (2014), the authors used the tracked locations of five facial landmarks viz., center of eyes, tip of nose, and corner of lips to divide each face spatially in five blocks. Subsequently they also segmented

each video clip on the temporal axis using their own interpretation of onset, apex and offset (i.e., rise, sustain, decay). Applying a dynamic Local Binary Pattern descriptor to the spatiotemporal segmented blocks achieved a correct classification rate of 91.4% on the UvA-NEMO dataset. Dibeklioğlu et al. (2015) used the tracked location of eleven fiducial facial landmarks around the mouth, eyes and cheeks in their smile classification system. By exploiting the fixed positions of the eyes they align and normalize each facial landmark grid for every frame. With the normalized grids they calculate various landmark displacements between frames, which correspond to eyelid, lip and cheek movement. The landmark displacements are used to automatically segment each recording into the onset, apex and offset phase of the smile and to calculate descriptive movement features such as duration, maximum amplitude and mean of the movement. This generated three (i.e., one for each phase) 25-dimensional vectors of geometrical features for three regions of the face, i.e., eyes, cheeks and mouth. Their best full-fledged system using SVM and feature selection scored an 89.8% correct classification rate on the UvA-NEMO dataset.

Although many of the above studies address the dynamics of smiles, thus far, no study employed spatiotemporal filters for distinguishing real from fake smiles. Since static Gabor energy responses have proven to be valuable features in static smile detection (Whitehill, Littlewort, Fasel, Bartlett, and Movellan, 2009) and given the importance of dynamic information for smile classification, we expect that applying spatiotemporal Gabor filters to posed and spontaneous smile sequences will help distinguish these subtle cases.

Before turning to the current studies, we discuss the potentially important preprocessing step of face registration.

Face Registration

A detailed analysis of facial expressions may be hampered by rigid movements of the face or head. Normalizing the face (face alignment) has been argued to be central to the success of smile recognition (Chew et al. 2012). In almost all research on smile detection a form of face alignment or normalization is applied, but most studies refrain from reporting the effect of normalization (e.g., by specifying the results obtained with and without normalization) and as a result the precise benefits of this normalization remains somewhat unclear. Whitehill et al. (2009) stressed the importance of accurate face registration by comparing the smile detection performance using a fully automatic alignment scheme and one based on human annotations. Their system rotated, cropped and scaled all facial images to a fixed size using the location of the eyes as markers. The fully automatic system used automatic eye detection to register the faces, whereas humans had to label the locations of the eyes in the semi-automatic alignment scheme. The loss in performance of the automatic system compared to the manually aided system ranged from 5% with a dataset of 100 images to 1.7% using 10,000 images. A similar drop in performance was reported by Dibeklioğlu et al. (2012) (1.85%) when comparing performances of a landmark tracker that was manually initialized for the first frames against a fully automatic system.

4.1.2 Current Studies

In the current study we investigate the contribution of adding spatiotemporal pixel information to the task of posed vs. spontaneous smile classification.

The general methodology for our analyses is similar to the one used in Chapter 2, where the STemVAD method was introduced. In particular, we zoom in on the head region and collect Gabor filter responses for the entire face, as well as for the upper part and the lower part, in line with the literature described above, which suggests that the distinction between posed and spontaneous smiles may be visible both in the upper part (around the eyes) and the lower part (around the lips). Crucially, we will compare static (spatial) and dynamic (spatiotemporal) Gabor filters, to find out whether the addition of dynamic information is beneficial for the smile classification, in line with the findings in Chapter 2 and Chapter 3 that revealed that adding dynamic information was beneficial for visual speech detection and learning difficulty assessment, respectively. Similar to the previous chapter, we will compare two implementations of the spatiotemporal Gabor filter; the method due to Petkov and Subramanian (2007) used in Chapter 2 (PS-STGF) and the method due to Heeger (1987) (H-STGF), to see to what extent findings may differ due to specific implementation details. As said above, face registration may be an important factor for facial analyses, which is why we compare results both for the original unaligned recordings and for those were we apply face registration. Since we have seen in Chapter 2 that combining different speeds, as expressed in pixels per frame (PPF), may be beneficial for classification, we will explore these speeds in the current chapter as well.

The classification of posed versus spontaneous smiles is a more subtle task than visual speech detection, where it was shown that zooming in on relevant facial areas (i.e., the mouth) was beneficial to the classification performance. Therefore, in this Chapter, we will perform an additional analysis that inspects appearance changes in the face at an even finer level of detail, i.e., at the level of individual facial landmarks. For a small region around each landmark we will collect Gabor filter responses and evaluate the classification performance with these landmark-based filters. Again, the crucial question is whether spatiotemporal Gabor filters outperform static ones, which we will explore for both the PS-STGF and H-STGF method.

4.2 METHOD

In this section we will briefly discuss our method to quantify facial movement. We will first describe how we measure facial movement. Subsequently, we describe the classification approach.

Measuring Facial Movement with Spatiotemporal Pixel Information

In an almost identical manner as in Chapter 3 we measure facial movement by representing sequences of facial images as a set of STGF transformations. As explained in Chapter 2 and Chapter 3 we can tune an STGF to respond maximally to specific motion (i.e., speed and direction) allowing

us to construct a bank of filters that covers the variability of movements in the video sequences. In the single image case an SGF operates on each pixel, transforming it to the sum of the multiplication between its neighbors and the filter values, a procedure called convolution. Value transitions in the filter are reflected in the residual image after convolution, i.e., a filter can be constructed to respond maximally to specific contour changes. In the case of an STGF, a cubic 3D block of filter values slides over the spatiotemporal image cube defined by the two spatial axes (x and y) and the time axis (t). The result of this spatiotemporal convolution is analogous to the spatial case, except for the incorporation of temporal variations. The values in each filter are calculated using the Gabor functions described in Chapter 1. Similar to Chapter 3, we employ two implementations of STGF, i.e., Petkov and Subramanian's STGF implementation (PS-STGF) and the Heeger inspired implementation (H-STGF) in order to assess implementation dependent effects. Since PS-STGF were modeled to properly reflect the responses of human vision cells, they impose a strict relation on the Gabor function's parameters which was discussed in Section 1.3, whereas H-STGF poses no relation on the parameters of the Gabor function. In both cases the length of the filter block in the third dimension corresponds to the temporal window covered by the filter, which is related to the preferred speed the filter is tuned to (i.e., the size of the temporal window is inversely proportional to the speed the filter is tuned to). By carefully constructing the filters, we can tune their preference for specific temporal contour changes. In our case we construct filter banks where each row in the bank represents a different direction of movement and each column corresponds to a preferred speed in PPF. The combined filters largely cover the distribution of temporal contour changes present in facial image sequences.

Classifying facial movement

Similar to the approach described in Chapter 2 and Chapter 3 aggregation of filter responses constitutes the classification vector. The output of a filter g applied to frame f results in an energy image represented as $E_g(f)$ with an equal number of pixels as the original frame. Each pixel value n in the energy image denoted by $E_g(f, n)$, is a reflection of the filter's specific characteristics, with respect to spatial frequency, speed and orientation, present in the original frame. Summing N pixels in a region of interest in a frame generates the aggregated feature which is computed as follows $A_g(f_{roi}) = \sum_{n=1}^{N} E_g(f_{roi}, n)$. The final G-dimensional (frequencies $\times$ speeds $\times$ orientations) aggregated feature vector $A(f_{roi})$ is computed by applying all the filters in the filter bank to the frame and summing the responses.

4.3 EXPERIMENTAL EVALUATION

We applied PS-STGFs and H-STGFs on the binary task of posed vs. spontaneous smile classification. In our experiments we used the UvA-NEMO Smile Database (Dibeklioğlu et al. 2012). In the next subsections we will discuss

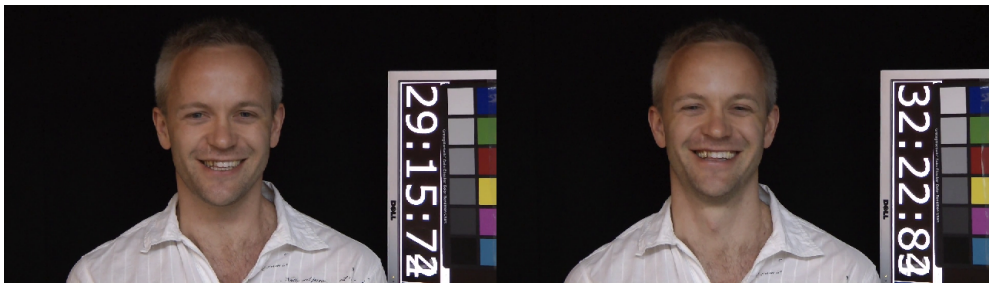


Figure 23: Deliberate and spontaneous example of UvA-NEMO data set.

the characteristics of this dataset, report relevant implementation details for PS-STGF and H-STGF based methods for measuring facial movement and provide information about our evaluation procedure .

4.3.1 Dataset

The UvA-NEMO Smile Database (Dibeklioğlu et al. 2012) was developed to study the facial dynamics of smiles. The database consists of multiple fragments per speaker with an equal number of spontaneous and posed smiles. Figure 23 shows two frames taken from a deliberate and a spontaneous sequence, respectively, of the same participant. Total number of fragments in the data set is 1240 (597 spontaneous and 643 posed), collected from 400 subjects (185 female and 215 male) ranging from 8 to 76 years of age. The fragments start and end with a (near) neutral expression and on average last 3.9 seconds ($\sigma = 1.8$). Frames are RGB recorded in 1920×1080 pixels at 50 frames per second using a Panasonic HDC-HS700 3MOS camcorder under similar illumination conditions.

The smile elicitation procedure was straightforward. Posed smiles were collected by asking subjects to perform an enjoyment smile, in some cases after being shown a proper example. Spontaneous smiles were evoked by showing a funny video compilation. The resulting video clips of the subjects were segmented by two trained annotators to contain only genuine smiles.

We use the same information about the phases of the smiles, i.e., onset, apex and offset, as used by Dibeklioğlu et al. (2012) and cordially provided by the authors. In their work the authors use the distance of the corners of the mouth with respect to the center of the mouth to estimate the different phases. The longest constant increase of distance corresponds to the onset phase and, in a similar vein, the longest decrease indicates the offset phase. The frames between onset and offset are marked as apex.

4.3.2 Implementation Details

We conducted our smile classification experiments on two variants of the data, i.e., the original recorded stimuli, and stimuli where we fixed the head in the middle of each frame using the position of the eyes to normalize the rigid head motion. The frames in the first instance are resized by a factor of 4 to 480×270 pixels which reduces the necessary computational resources in the feature extraction step. By reducing the absolute number of pixels, we

effectively reduce the number of filtering operations and memory required to store the responses. The frames in the second instance are subjected to a more elaborate preprocessing scheme. In order to remove rigid head motion we align any frame to its antecedent frame and re-center the frame around the position of the eyes. By aligning each subsequent head with its previous location we greatly eliminate movement caused by translating and rotating of the head. We use the Enhanced Correlation Coefficient (ECC) algorithm (Evangelidis and Psarakis, 2008) to compute transformation parameters in order to register the heads. The algorithm finds an affine warp that minimizes the residual image of the transformed frame and template frame. We initialize the alignment by scaling and rotating the first frame using the interocular distance and the straight line connecting the inside of the eyes, respectively. From the initialized frame the head is cropped and scaled to result in an 128×128 pixel frame. We choose these dimensions because certain filtering operations are most efficient on images whose dimensions are a power of two (i.e., 2^n) and 128 pixels was the closest to the average of all bounding box sizes of the face detections. Due to (non-rigid) movement of the face an accumulating error can cause the face to gradually move away from its initial position. To account for the so-called drift effect when aligning sequences we also determine the alignment values of the current frame to the first frame. By restricting the translation in the x-y-plane to be the average of the proposed translation with respect to the previous frame and the translation with respect to the initial frame the final adjusted frame can only deviate from the first frame by a small margin, thereby fixating the head over the entire sequence.

Implementation Petkov and Subramanian STGF

For our biologically inspired STGF implementation we rely, like in the previous chapters, on the implementation of Petkov and Subramanian. In a similar vein to the work in Chapter 2 and Chapter 3, we constructed a filterbank with 48 filters varying over 8 orientations and 6 speeds. The speeds we used are $v = \{0.5, 1, 1.5, 2, 2.5, 3\}$ PPF for each of the $\theta = \{0, 0.25\pi, 0.50\pi, 0.75\pi, \ldots, 1.75\pi\}$ radians orientations. By setting $v = 0$ we construct the static counter part of the dynamic filters, i.e, PS-SGFs.

Implementation Heeger STGF

The second, Heeger-inspired implementation, follows a similar approach as in Chapter 3. We constructed a 3D Gabor filter bank of $G = 8 \times 4$ filters sensitive to 8 directions of movement, i.e., 2 vertical 2 horizontal, and 4 diagonal directions, and 4 different values for the spatial Gaussian's envelope width, i.e., $\sqrt{2}$, 2, $2\sqrt{2}$, and 4 with spatial and temporal frequencies set to 1/4 cycles/pixel and a temporal Gaussian envelope width of 1. In a similar vein we constructed a stationary filter bank, containing half of the number of filters of its dynamic counter part, since direction of movement is undefined for stationary filters. The resulting filter bank has $G = 4 \times 4$ filters, with 4 orientations, i.e. vertical, horizontal and 2 diagonals and using the same spatial Gaussian widths as for the spatiotemporal Gabor filter bank. We compute the static variant by removing the temporal Gaussian factor (i.e., setting it to 1).

Also here, we construct a static counterpart (H-SGF) of the dynamic features, by setting the temporal Gaussian in the filter's equation to 1.

Facial Analyses

Our analyses follow a similar approach as the analyses in Chapter 2. Filter responses within the region of interest are aggregated by summing all values, this yields one value per filter per region. We compare the classification results for the entire face, upper part and lower part using the filter responses calculated from the original frames as well as from the face-aligned frames. Since the H-STGF is tuned to one specific preferred velocity of movement, we also select one preferred speed for PS-STGF, i.e., 1 PPF, when we compare both implementations on the whole head and its subparts. For each frame and for all three facial regions, this yields a 8 (filters) $\times$3 (phases) dimensional feature vector for Petkov and Subramanian's implementation (for both the dynamic and static variants), a 32 (filters) $\times$3 (phases) dimensional feature vector for H-STGF, and a 16 (filters) $\times$3 (phases) dimensional feature vector for H-SGF. For each phase, the activation per filter is computed by averaging the scores of all frames within the phase. Since the data set is approximately balanced we report the correct classification rate as our performance measure.

We extend the aforementioned analysis by zooming in on the face and compare results at the level of facial fiducial landmarks. With Intraface (Xiong and De la Torre, 2013) we are able to determine the location of 49 facial landmarks. We decided to include all landmarks in the experiment (and not just, say, the eye and mouth corners). In each frame we measure and sum the filterbank responses of a 5×5 pixel ROI centered at the locations. For each phase, the activation per filter is computed by averaging the scores of all frames within the phase. This adds a factor of 49 (landmarks) to the dimensionality of the above described feature vectors.

Our last comparison evaluates the effect of STGFs tuned to different speeds. For the sake of simplicity, we evaluate PS-STGF on the original frames only, using each of the speeds from 0 PPF (i.e., static filters) to 3 PPF individually and one with all dynamic filters combined. This comparison will be applied to all granularities from the whole head to the individual landmarks.

To facilitate our analyses we developed a response viewer, which allows one to quickly inspect the filter responses for the various settings as illustrated by Figure 24. In this example, we inspected responses around the right corner of the mouth as visualized by the red star on the image of the frame panel and selected in the third list panel (i.e., "Punt 32", point 32). More specifically, we looked at the responses of the PS-STGF with upward orientation (first list panel) and speed 1 PPF (second list panel). In the lower part of the viewer we plot the responses over time and indicate the current response with a similar red star as in the frame panel. The three colors green, blue and red correspond to onset, apex and offset of the smile, respectively.

Similar to our analyses in Chapter 2, we also compare the performances of banks of STGFs that are tuned to one specific preferred speed in eight orientations. Our comparison includes the classification scores of the zero-speed filterbank, six individual single speed filterbanks (ranging from 05 to 3 PPf), and the combined-speeds filterbank.

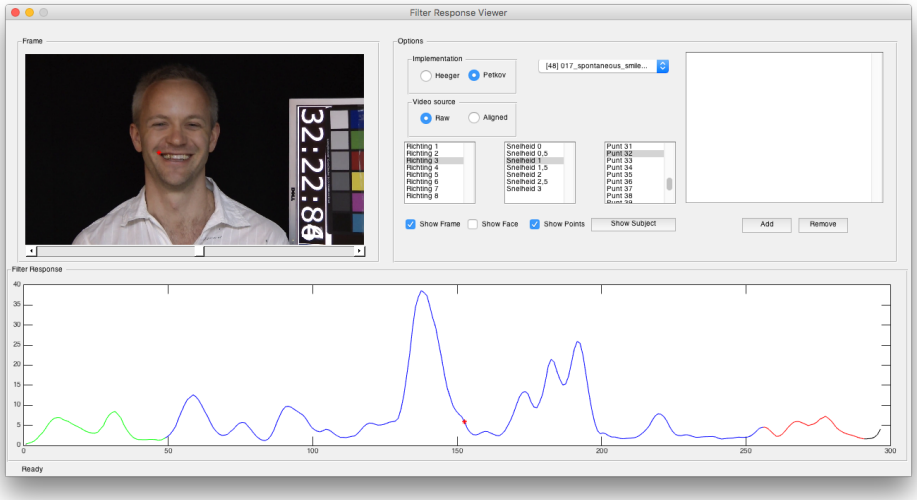


Figure 24: Example of our filter response viewer.

4.3.3 Evaluation procedure

We followed a straightforward 10-fold classification procedure based on the folds proposed by Dibeklioğlu et al. (2012). In contrast to Dibeklioğlu et al. (2012), we do not optimize our classifier's parameters and therefore do not perform a nested cross validation inner loop. We did, however, apply Principal Component Analysis (PCA) to reduce the dimensionality of the feature vectors when we classified at the level of fiducial landmarks. Our resulting transformation matrix should account for at least 99% of the variance of the training fold. Feature vectors were made unit length and were fed to a linear Support Vector Machine as implemented by the LIBLINEAR library (Fan et al. 2008).

4.4 RESULTS

We present our results in three separate sections, i.e., results obtained from different facial parts, from facial landmarks, and from comparing different speeds on all granularities. In the first two sections we present results for both Petkov and Subramanian's and Heeger's implementation, whereas in the last section we only present the results of Petkov and Subramanian's implementation for the sake of simplicity.

Facial parts

Table 6 displays the results from the first experiment, comparing two implementations (Petkov and Subramian, and Heeger) of both STGFs and SGFs applied to the original and aligned data for the entire face, the upper part and the lower part. First and most importantly, examination of this table shows that for all but one cases the STGFs outperform the corresponding SGFs. The overall best result (with an averaged correct classification rate of

Table 6: Correct classification rates (%) of STGFs and SGFs applied to the original frames and to the aligned faces. Performance is evaluated on the whole head, the upper part of the head and the lower part.

Method	Modality	Head	Upper part	Lower part
PS-STGF	Original	**70.2 ± 4.9**	**66.9 ± 4.9**	**72.8 ± 4.2**
	Aligned	68.2 ± 3.6	64.1 ± 2.7	69.2 ± 2.5
PS-SGF	Original	57.5 ± 3.5	55.5 ± 3.9	58.1 ± 3.3
	Aligned	56.2 ± 3.1	55.5 ± 2.7	53.8 ± 2.8
H-STGF	Original	65.4 ± 3.3	62.6 ± 3.3	68.4 ± 4.7
	Aligned	54.0 ± 2.3	53.2 ± 2.2	53.7 ± 3.0
H-SGF	Original	53.5 ± 3.9	55.2 ± 4.5	53.3 ± 2.9
	Aligned	53.9 ± 2.9	53.4 ± 2.9	53.6 ± 2.8

72.8%) is obtained with PS-STGF on the lower part of the face on the original recordings.

Looking closer at the results, various interesting observations can be made. First, for all STGFs, better results are obtained when applied to the lower part of the face then when applied to the upper part of the face. This is in contrast to the findings of Dibeklioğlu, Salah, and Gevers, who found the region of the eyes to be the most informative for distinguishing between posed and spontaneous smiles using their method. Interestingly, although sometimes marginally, the results on the whole head are better than when applied only to the upper part, suggesting that it is indeed, at least for our method, the lower part of the face which is most important for classifying smiles as posed or spontaneous.

Second, PS-STGF achieves the highest results overall on all granularities when applied to the original sequences. In fact, and perhaps surprisingly, the results show that for all cases Petkov and Subramanian's implementation outperforms Heeger's equivalent.

The third and final observation is that when comparing the performance between the original (unaligned) and the face-aligned frames reveals that normalizing the rigid head motion generally leads to a drop in performance. Most notably, in the case of dynamic filters (which, as we have seen, perform best) normalizing head movement almost always leads to a drop in performance, which can amount to close to fifteen percentage points (as in the H-STGF case).

Facial Landmarks

When looking at the entire face, its upper and lower part, our analyses convincingly show that using STGFs is beneficial. However, it has been argued that it is important to look at specific facial landmarks for classifying smiles as spontaneous or posed. It is therefore conceivable that when we would zoom in on such landmarks, the benefits of STGFs disappear. To examine this closer, we have also applied the two implementations of the

Table 7: Correct classification rates (%) of STGFs and SGFs applied to the original frames and to the aligned faces. Filter responses are extracted from small regions around facial landmarks.

Method	Original	Aligned
PS-STGF	**77.4 ± 4.3**	76.5 ± 3.5
PS-SGF	73.1 ± 2.5	71.0 ± 4.5
H-STGF	72.2 ± 4.6	69.1 ± 3.4
H-SGF	65.4 ± 5.1	65.1 ± 5.1

Gabor filters (again at one dimension of velocity) to small regions around facial landmarks. Table 7 summarizes the result of our second experiment.

The most important observation that can be made when inspecting this table is that, again, STGFs systematically outperform their corresponding SGFs, although the differences are less pronounced than in the first experiment. The best result overall is obtained with PS-STGF on the original frames (with a correct classification rate of 77.4%). Zooming in on the results reveals that Petkov and Subramanian's implementation again outperforms Heeger's version in all cases. Furthermore, it can be observed that performance on the original frames is better than on the face-aligned frames, although the differences are often small. Comparing the results with Table 6 shows that the classification rate increases when zooming in on the level of facial landmarks. The effect is most notable for the SGFs, where the increase varies from ten to fifteen percentage points.

Comparing Different Speeds

Table 8: Correct classification rates (%) of PS-STGFs and PS-SGFs applied to the original frames. Performances are compared for various speeds and on all granularities.

Speed	Head	Upper part	Lower part	Points
0	57.5 ± 3.5	55.5 ± 3.9	58.1 ± 3.3	73.1 ± 2.5
0.5	68.0 ± 6.1	66.4 ± 5.7	71.1 ± 6.3	76.6 ± 4.0
1	70.2 ± 4.9	66.9 ± 4.9	72.8 ± 4.2	77.4 ± 4.3
1.5	69.3 ± 4.2	65.8 ± 4.6	**72.9 ± 4.4**	78.1 ± 3.4
2	68.6 ± 4.4	64.2 ± 4.3	71.6 ± 4.9	**79.3 ± 2.7**
2.5	67.7 ± 3.9	63.1 ± 4.1	70.7 ± 4.8	77.5 ± 3.4
3	66.6 ± 4.9	63.0 ± 3.7	70.4 ± 4.5	76.4 ± 3.8
all	**71.0 ± 4.1**	**67.7 ± 4.5**	71.9 ± 4.5	78.0 ± 2.8

So far, we have only looked at STGFs with a one dimensional velocity component, and compared these to their static Gabor counter parts. However, in our earlier work (Joosten, Postma, and Krahmer, 2015, see Chapter 1), we have shown that the comparison of different speeds can have substantial effects.

In Joosten et al. (2015) speed 1 PPF (as we have used in the above experiments for Petkov and Subramanian's implementation) did not systematically yield the best scores. In fact, these were obtained by combining different speeds. Therefore, in our last analyses, we explore the value of looking at different speeds for smile classification. For the sake of simplicity, we only do this for Petkov and Subramanian's implementation. In addition, given the consistently lower scores for head-aligned stimuli, we only look at the unaligned (original) stimuli. The analysis is performed on all granularities, i.e., the whole face and its subparts and the facial landmarks.

Table 8 lists the results of our final experiment. Columns two to four report the scores for the different speeds, looking at the entire head, the lower part of the face and the upper part, respectively. Inspection of this three columns reveals a very consistent pattern: the lowest scores are obtained for the static Gabor filters (speed 0 PPF). This pattern is also reflected in Figure 25, Figure 26 and Figure 27, where we present the spread of the accuracies of the different classification folds (head, upper and lower, respectively) in box-whisker plots. Next, looking at dynamic filters with speeds ranging from 0.5 to 3 PPF, we see that they all outperform the static Gabor filter, with a small upward and then downward trend, such that an increase in speed preference seems to result in a somewhat lower accuracy score. This is also reflected in Table 8, where we see that speed 1, 1, and 1.5 are the optimal speeds for the whole face, the upper part and the lower part respectively. As in Chapter 2 with the study on visual speech detection, best results are obtained by combining all (non-zero) speeds for the whole face and its upper part. Interestingly, this is not the case for the lower part of the face (where 1.5 PPF yields the highest accuracy). The best performance is once again obtained by looking at the lower part of the face (72.9%), although the differences are small.

The fifth column in Table 8 presents the same analysis, applied to facial landmarks, and essentially reveals the same picture. First, the lowest score is obtained with the SGF. Second, of the individual speeds of the landmarks, which all perform better than any other granularity-speed combination, the 2 PPF one yields the best results. This speed also seems to be a peak in performance, both lower and higher speeds seem to deteriorate the performance, clearly depicted in Figure 28. Although all individual speeds and the combination of speeds have better accuracies than the static variant, the differences as less pronounced. And finally, third, the best results are obtained with the filters tuned to speed 2 PPF, which also yields the overall best result of all the classification methods we have described, with an accuracy of 79.3%, which is again in contrast to Chapter 2 and Chapter 3.

4.5 DISCUSSION

In this chapter, we asked whether including dynamic information in Gabor filters is beneficial for classifying smiles as spontaneously happy (Duchenne) or posed, social ones (non-Duchenne). The previous two chapters suggest that adding dynamic information is indeed beneficial, although the added value of STGFs over SGFs was much larger for visual speech detection (Chapter 2)

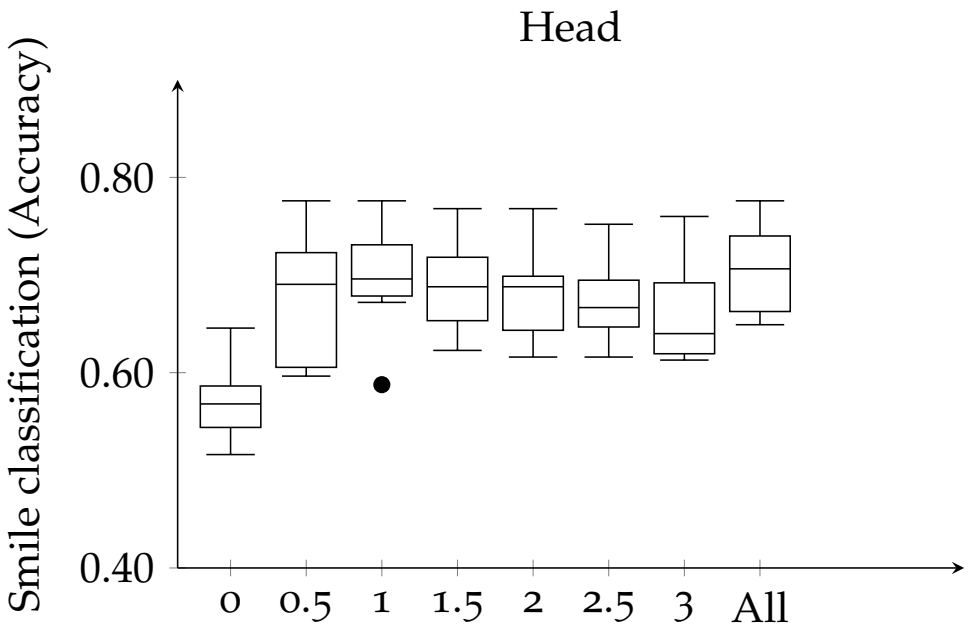


Figure 25: Boxplots of smile classification accuracy scores obtained during cross validation on the UvA-NEMO smile dataset. The boxes correspond to the **Head** results in the second column of Table 8. The x-axis labels correspond to the distribution of scores for each individual speed starting with the zero-speed (0) static version, or SGF and followed by the distribution of accuracy scores of the $0.5 - 3$ speed STGFs. The rightmost box labeled *All*, shows the accuracy scores for the full-fledged STGF in which all speeds are included.

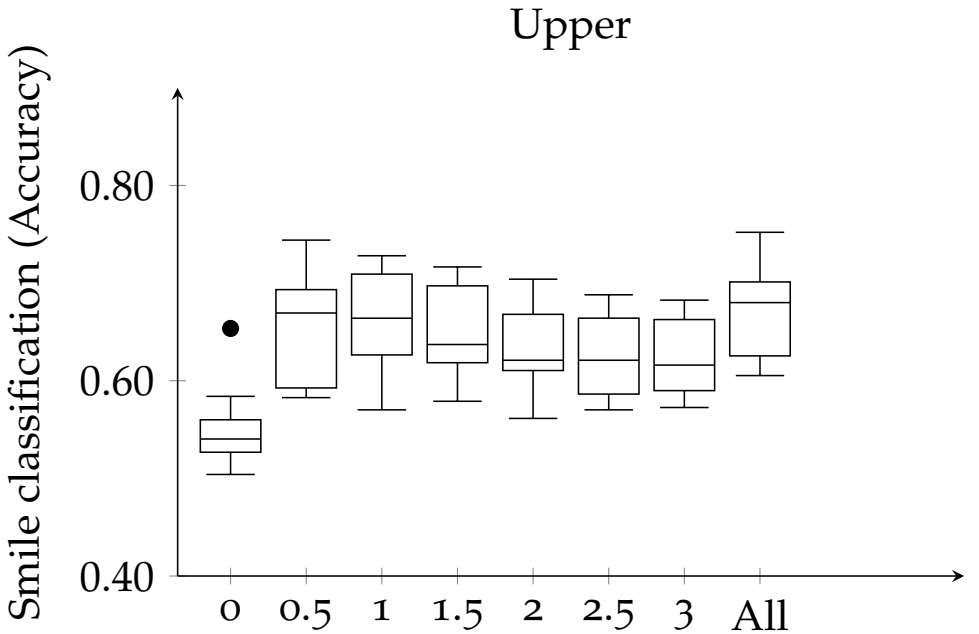


Figure 26: Boxplots of smile classification accuracy scores obtained during cross validation on the UvA-NEMO smile dataset. The boxes correspond to the **Upper part** results in the third column of Table 8. The x-axis labels correspond to the static SGF (0) the individual speeds ($0.5 - 3$) and the combination of speeds (All).

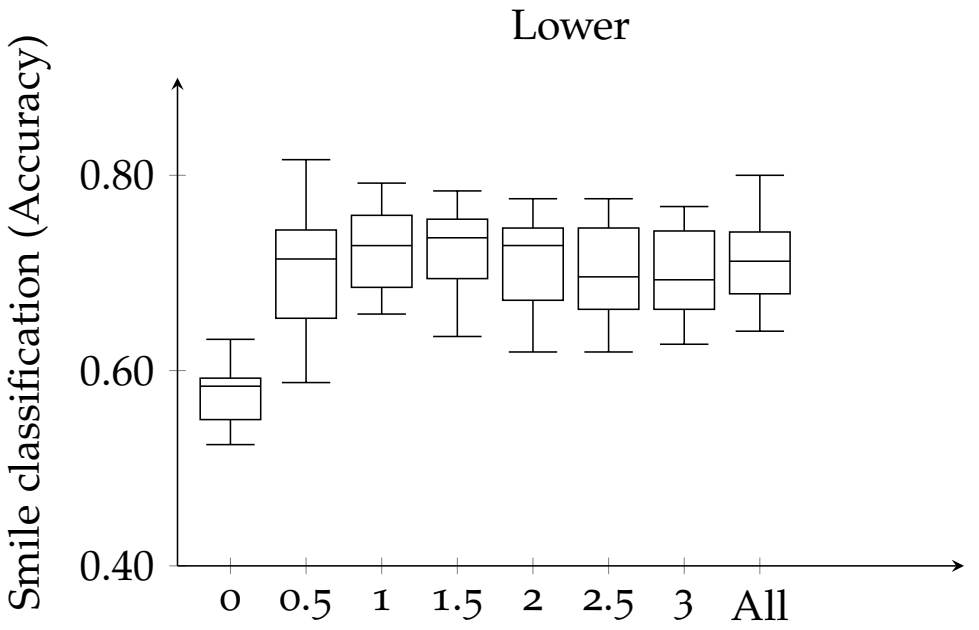


Figure 27: Boxplots of smile classification accuracy scores obtained during cross validation on the UvA-NEMO smile dataset. The boxes correspond to the **Lower part** results in the fourth column of Table 8. The x-axis labels correspond to the static SGF (0) the individual speeds ($0.5 - 3$) and the combination of speeds (All).

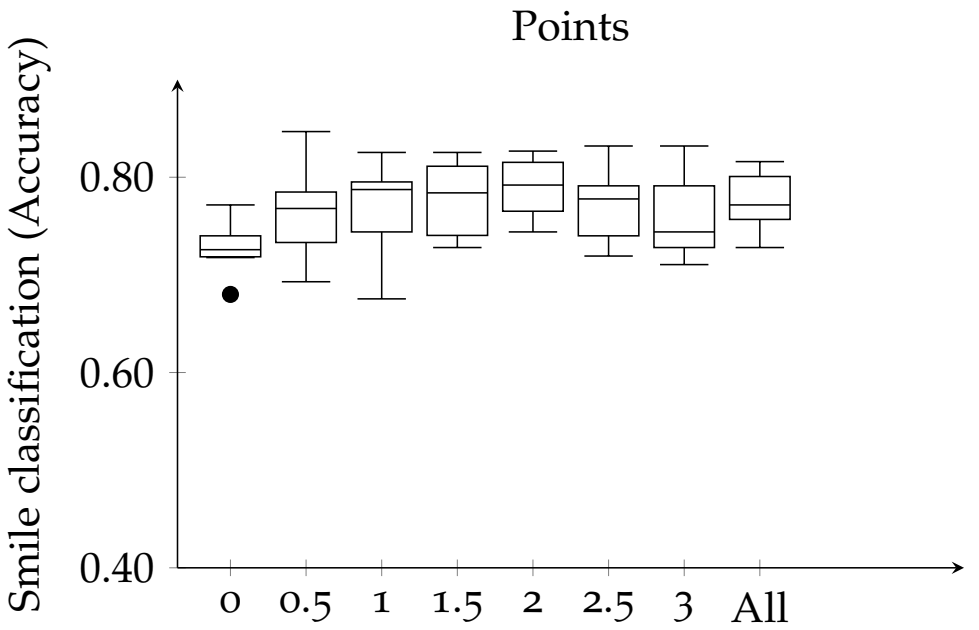


Figure 28: Boxplots of smile classification accuracy scores obtained during cross validation on the UvA-NEMO smile dataset. The boxes correspond to the **Points** results in the fifth column of Table 8. The x-axis labels correspond to the static SGF (0) the individual speeds ($0.5 - 3$) and the combination of speeds (All).

than for classifying difficult assessments (Chapter 3). Smile classification is a good example of a subtle social signal processing task, in the sense that one might think that human judges would find smile classification harder than, for example, visual speech detection. Hence it is interesting to ask, as we did in this chapter, whether dynamic filters are helpful for this more subtle task as well.

Based on a database of posed and spontaneous smiles, the UvA-NEMO smile database, we set-up a number of classification experiments, systematically comparing SGFs and STGFs. As in Chapter 3, we used two different implementations of the STGFs, one due to (Petkov and Subramanian, 2007), PS-STGF, and one to (Heeger, 1987), H-STGF. We applied both methods to the original stimuli as well as their head-aligned, face-registered counterparts, based on the reasoning that rigid head movements might make it harder for the Gabor filters to pick up the "meaningful" movement cues. In addition, we looked at the entire face as well as at the upper and lower part in isolation. Finally, we conducted a separate study where we looked for movement around specific facial landmarks, because the study of Girard, Cohn, and Torre (2014) has argued that extracting appearance information around fiducial points on the face is beneficial for the task of smile intensity classification.

First and foremost, in all our comparisons, we found that dynamic, STGFs outperformed their static, SGF counterparts. This holds true for all granularities that we looked at: the whole face, the upper and lower part of the face, and the facial landmarks. Hence, just as in the previous chapters we found a benefit of dynamic filters over static ones, which in some cases was substantial.

We found that Petkov and Subramanian's implementation always performed better than the one due to Heeger, in some cases with considerable differences in the resulting scores. The speed of 1 PPF seems to result in slightly higher scores for the face and its upper part than the other speeds, whereas speed 1.5 and 2 PPF result in the highest performance for the lower part of the face and the facial landmarks, respectively. The combination of all speeds only results in a higher score for the face and its upper part.

Interestingly, fixing the rigid head motion had a detrimental effect, contrary to our initial expectations. With hindsight, we think that there are a number of possible reasons for this. First of all, it might be that rigid head movements actually are a relevant factor for smile classification (think of moving shoulders in a ha-ha manner during spontaneous laughter), so that removing them hampers classification. Alternatively, it is conceivable that the alignment procedure used to fix the heads might have created movement artifacts (e.g., when two consecutive frames did not align properly causing a small visual perturbation) that were picked up by the Gabor filters. In general, whether eliminating rigid head-movement is a good idea or not for this kind of classification thus warrants future research.

When looking at the results for the face and its subparts, we can say in general that classification for the lower face part is better than for the upper face part. Earlier work, including Dibeklioğlu et al. (2015), has suggested that cues in the upper part of face (e.g., eye blinks or wrinkles around the eye) are important for the classification of smile as spontaneous or not. However, our results did not provide evidence of this. Our best result, 78.2%

classification accuracy, was obtained by zooming in on movements around 49 facial landmarks, using Petkov and Subramanian's STGF tuned to 2 PPF and applied to the original, non-fixed recordings.

The purpose of this study was to compare static and dynamic Gabor filters, and not to achieve state-of-the-art results per se. When comparing the best performance obtained by our STGFs (78.2%), with that of the state-of-the-art method of (Dibeklioğlu et al. 2015) (87.1% and 89.8% without and with feature selection, respectively), we see a considerable difference in favor of the latter. This difference may be explained by the fact that (Dibeklioğlu et al. 2015) employed specific geometrical features that measure displacement and speed for landmarks located at the mouth, cheek and eye regions. Apparently, such explicit features outperform our implicit STGF features. The fact that Dibeklioğlu et al. (2015) found an improved performance for the eye region, whereas we did not, may be explained by their use of explicit features and the fact that we average over the entire upper part of the face. Furthermore, research by Wu, Liu, Zhang, and Gao (2017) suggests that reducing the dimensionality of image descriptor results (like STGF responses) using various schemes such as linear coding, temporal pooling and (whitened) PCA, greatly improve classification results by transforming the large image patches to a compact representation that retains the salient appearance aspects.

Still, it is important to state that our results were obtained without any additional optimization (e.g., feature selection, ensemble methods). Despite this, our system outperforms comparable systems proposed by Cohn and Schmidt (2004), Dibeklioğlu, Valenti, Salah, and Gevers (2010) and Pfister, Li, Zhao, and Pietikäinen (2011) on the UvA-NEMO dataset. This shows that by using local dynamic information, STGFs are quite capable of distinguishing between posed and spontaneous smiles.

Turning back to the main objective of this study, our results support the notion that dynamic information is more informative for smile detection than static information.

4.6 CONCLUSION

We conclude that STGFs outperform their static counterparts on the task of smile detection. Hence, future research could benefit from incorporating STGFs in their feature set.

So far, in this thesis, we have looked facial social signals of different levels of complexity and subtleness. In the final experiment, Chapter 5, we study whether STGFs are also beneficial for tasks at a larger scale: the full body.

5 GAIT-BASED GENDER DETECTION

5.1 INTRODUCTION

Throughout this thesis we have looked at the performance of static and dynamic Gabor filters in the classification of visual communicative behavior. We started, in Chapter 2, with the detection of visual speech, which intuitively is a clear cue occurring in a specific area of interest (mainly around the mouth), and we found that dynamic Gabor filters (STGFs) consistently outperformed static ones (SGFs). In Chapter 3 we asked whether it is possible to automatically assess how a child perceives an exercise (i.e., as easy or hard) based on facial expressions. This is much more subtle (even though human participants can do this above chance) and moreover, it is not a priori clear where in the face the cues (if any) were present. Here the dynamic Gabor filter approach fared less well, and the best results were obtained using an explicit facial modeling approach using AAMs. Nevertheless, despite the relatively poor performance compared to the AAM, we did find STGFs to perform somewhat better than their SGFs variants. Then, in Chapter 4, we turned to the classification of smiles as posed or spontaneous, which arguably is also a rather subtle cue, but which is known to occur in specific area's of interest (most notable around the mouth and the eyes). Again: the dynamic Gabor filters yielded better classification results than the static ones.

So, at the moment it seems that dynamic Gabor filters are indeed beneficial, and work better than their static counterparts, albeit most notably for movements that have a specific temporal signature and occur at specific locations. Of course, we have only looked at movements on a relatively small scale (the face and its parts), and we do not know whether there are benefits for dynamic Gabor filters for movements on a larger scale. This will be addressed in the current chapter, where we will look at full body movement.

In recent years, there has been an increased attention to full-body nonverbal cues (Bouma et al. 2016; Coulson, 2004b; Wallbott, 1998). In the field of emotional expressions, for instance, there is a growing awareness that emotional states are not only expressed using facial cues, but also using the rest of the body. A person that is scared will not only produce a fearful face, but may in addition run for cover, and as a result, other people may be able to recognize this emotional state based on such body movements as well (Gelder, 2006). In fact, more recent work even suggests that the body can be *more* revealing about someone's emotional state than the face (Van den Stock et al. 2007). In a study of tennis players who either won or lost a point, Aviezer, Trope, and Todorov (2012) showed that the valence of the response (positive or negative) could hardly be detected from the tennis player's faces, but judges were substantially better at determining this based on the player's bodies.

In this chapter, we zoom in on a more basic full-body classification task, namely gender detection based on the characteristics of a person's walk, their gait (Hu, Wang, Zhang, and Wang, 2010; Ng, Tay, and Goi, 2012; Yu, Tan, Huang, Jia, and Wu, 2009). This kind of classification task has various potential applications, including, for example, intelligent visual surveillance and customer statistics. Stores may, for instance, want to know how many male or female customers visit specific parts of a shop, which could potentially inform marketing and shop design.

It has been known for some time that humans are good at recognizing gender based on general movement characteristics. This has been studied, for example, using so-called point-light displays (Kozlowski and Cutting, 1977), in which movement is captured using a limited number of moving dots (typically associated with key joints). Based on such minimal movement representations, people are capable of guessing a person's gender well above chance. For example, Pollick, Kay, Heim, and Stringer (2005) were interested in estimating human efficiency in determining gender based on biological motion. To this end they performed a meta-analysis of 21 studies in which human accuracy for gender detection based on point-light motion was estimated. Their meta-analysis estimated that humans are 66% of the time correct in deciding whether the participant was male or female. Human performance increases to 71% correct for frontal or oblique point-light displays.

Of course, based on richer visual representations it might become easier to determine someones gender based on their gait. Yu et al. (2009) presented participants with human silhouette sequences (white against a black background) generated from the sideways recordings of the CASIA Gait Database (to which we will return below), due to (Yu et al. 2006; Zheng, Zhang, Huang, He, and Tan, 2011). Participants either saw the upper part of the body, or the lower part, or both (whole body), and were asked to determine the gender. Their results revealed that participants could do this reasonably well for the lower part of the body (with an accuracy of nearly 68%), while the results for the upper body and the whole body were substantially higher, and very close to each other (94% and 95%, respectively). This shows that gender classification based on gait is feasible for humans, and that the upper part of the body appears to be a more useful cue (at least when looking at walkers from the side) than the lower part. Interestingly, Yu et al. (2009, p. 1906) suggest that when it comes to gender classification from sideways silhouettes "humans are more sensitive to static body shape information than to dynamic information", based on surveys the participants took. In these surveys, they ranked dynamic information (such as movement of the arms and legs) as the lowest informational cue to distinguish gender. In this chapter, we ask whether dynamic information may not be helpful for gender detection after all, by (once again) comparing the performance of static and dynamic Gabor filters.

5.1.1 Related Work

We consider gait-based gender detection as a more abstract form of gait detection: the task of identifying a person based on their gait. After all, by

identifying a person, their gender is implicitly detected as well. The majority of relevant related work involves gait detection, instead of gait-based gender detection. Therefore, we will consider both gait detection and gait-based gender detection research in this section. The computational methods to gender classification or person identification based on gait that have been proposed, generally use one of two approaches, either relying on explicit models or not.

Early approaches to automatic gait classification tend to rely on the skeleton structure of the human body as a model for human motion processing. For example, Niyogi and Adelson (1994) rely on “spatiotemporal snakes" to find the contours of a walking person, after which a simple stick model of the human body is applied. They were able to classify 15 to 21 persons correctly (depending on some parameters) out of 26 sequences (i.e., up to 81% correct with a chance level of 20%). More recent work has extended this in various directions. Yoo, Nixon, and Harris (2002) focus on hip and knee angles, which are used for an extended 2D stick figure, while Bhanu and Han (2002) propose a 3D kinematic approach, which is used to retrieve human gait signatures. Yoo, Nixon, and Harris report a 100% correct classification rate albeit on a set of three participants with two walking sequences each and Yoo, Nixon, and Harris report a 77% correct classification rate on a 30-sequence test set. While these approaches have generally yielded reasonably good results, their results are strongly dependent on the quality of the extracted human contours (Han and Bhanu, 2006), which may be difficult in noisy situations.

To counter this limitation, people have started working on “model-free approaches" (Han and Bhanu, 2006), which do not rely on structural modeling of the full, moving human body but instead focus on the shape and velocity of movements as features for gait-based gender classification (e.g., He and Debrunner, 2000; Little and Boyd, 1998; Shutler, Nixon, and Harris, 2000

One of the most popular model-free approaches for human gait detection is the Gait Energy Image (GEI) representation by Han and Bhanu (2006). Essentially, the GEI represents human motion in a single image, capturing temporal information. The approach starts by extracting the walker from the background in the video recording, which is transformed to a black and white (binary) silhouette image. Subsequent images are size, position, and viewpoint normalized, to make sure that all images of the same walker are of the same size, at the same location, and at the same viewpoint, making it possible to align them. The GEI is now computed by simply taking the average pixel value of the aligned images, where black pixels indicates locations where the human body has not appeared and pixels with more intensity are associated with positions where it has appeared more frequently (see Figure 29). In other words, grey areas represent parts of the body where most movement during walking has occurred, and the claim is that this is beneficial for gender classification. This is indeed what Han and Bhanu show: by extracting features from the GEI representations and using these for learning, they achieve very good classification results, even up to 99% in a 74-participant dataset. Performances decrease when the differences between conditions in the training and test set increase. However, in the toughest cases, their rank 5 classifier still manages to get around 60% of the sequences classified correctly. In general, since the GEI representation is an average of

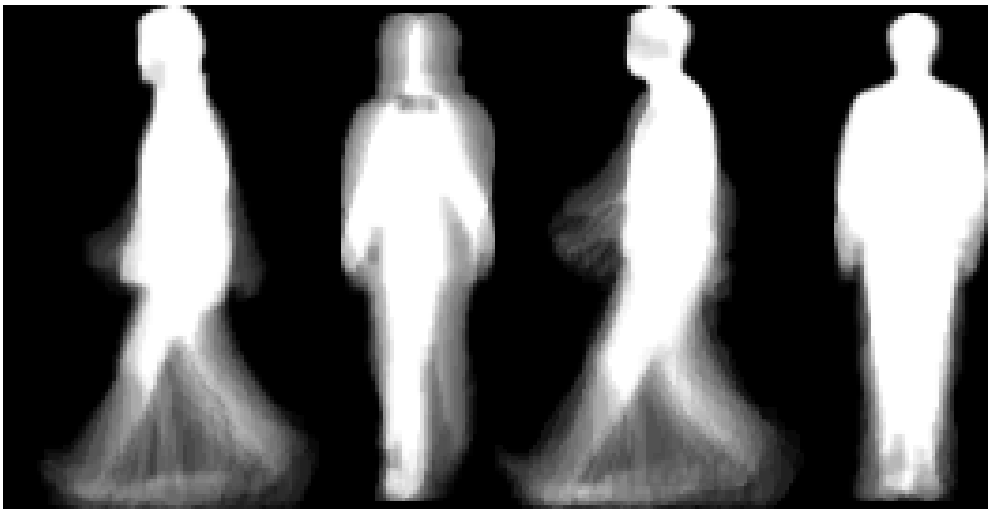


Figure 29: Examples GEI images extracted from a female (first two) and a male participant both from sideways as well as frontal recordings.

multiple binary silhouettes, it is less sensitive to noise in individual silhouette images or to other silhouette errors.

Various researchers have extended the GEI approach. For example, Kusakunniran, Wu, Li, and Zhang (2009) generalize the approach to multiple views.

The main limitation of the GEI approach is that it requires a specific manner of recording movements. In order to maintain size, position, and viewpoint, the camera has to move alongside the walker (as is the case in video recordings of sports(wo)men like skaters or runners). Instead of a moving camera, it is also possible to have multiple camera's along the walking route that take snapshots of the walker in front of the camera. Alternatively, the walkers could walk on a treadmill so that the camera can be fixed to a single position, but this is not very helpful for practical applications.

Despite this limitation, the GEI approach is a popular method for gait-based human analysis, due to its simplicity and elegance, although it is not the only one. Tao, Li, Wu, and Maybank (2007), for example, propose to use spatial Gabor filters like the ones discussed in Section 1.3, given that these have proven to be successful for image understanding and object recognition. They rely on different Gabor filters, either summing over scales, or over directions, or both, and the outcomes of the filters are subjected to a new general tensor discriminant analysis method developed by the authors. They report a 60.6% average correct recognition rate on the task of person identification using their best method with a database of 122 persons. Crucially for our current purposes, all Gabor filters they use are static ones, and do not take movement information into account. This is addressed in the current chapter.

5.1.2 Current Studies

In this chapter, as in previous ones, we compare STGFs with their SGF counterparts, to see how beneficial STGFs are for gait-based gender classification. These two variants of Gabor filters will be compared to the GEI representations proposed by Tao et al. (2007). These three different methods are applied

to recordings from the CASIA Gait dataset, which is one of the standard benchmarks for gait analysis. In our experiments, we concentrate on the frontal and sideways recordings of both male and female walkers. In our first analysis, we compare the overall performance of the three methods (STGF, SGF and GEI). Next, we zoom in on the performance for different body parts, inspired by earlier work (discussed above in Yu et al. (2009)) showing that human judges are better capable of gait-based gender detection from the upper half of the body than from the lower half. To be able to compare the contribution of different body parts across viewing conditions we distinguish between lower body, upper body, and head, and again contrast the performance of the three methods. Finally, as in earlier chapters, we will explore which speeds in the dynamic Gabor approach yield the best results for both the sideways and the frontal recordings.

5.2 METHOD

In this section, we will give a short description of our method to determine gender from human gait. We will first describe how we measure gait movement and end with a section on its classification.

Quantifying Human Gait with Spatiotemporal Pixel Information

We adopt a similar approach that we used in the previous chapters for measuring facial movement to quantify gait-based motion, i.e., gait sequences are transformed into a sets of STGF "energy movies," where each transformed sequence corresponds to a filter that was tuned to give a maximum response in the presence of contours moving at a certain speed and in a certain direction, as was explained in Chapter 2, Chapter 3 and Chapter 4. To capture all motions present in the sequences, we construct a filter bank of STGFs using different parameters for each filter. Filters operate on each pixel in each frame of a sequence, by considering its spatiotemporal context, i.e., its neighboring pixels, and calculate a response, i.e., the convolution operation. When the context of this pixel is a contour moving with a speed and direction that corresponds to the motion that the filter was tuned to, it gives a maximum response. The further the movement of the contour deviates from the preferred movement of the filter, the further its response decreases. Different from the previous two chapters, we only use Petkov and Subramanian's implementation of Gabor functions to calculate the filter values, since we found in these chapters that the differences between the two implementations were relatively small.

Classifying Human GAIT

Our method convolves a filter bank with a video sequence, yielding G transformed sequences. The body's silhouette is used as a mask to crop and align the responses in each frame. For each of the G sequences the ROI is averaged by the number of frames, resulting in a $W \times H$ response image per filter, where W and H correspond to the width and height of the ROI, respectively.

Feature vectors for classification are the concatenation of the response images per filter. Detecting gait-based gender consists of applying an SVM to a set of labeled training examples.

5.3 EXPERIMENTAL EVALUATION

In this method section, we discuss the dataset, the experimental procedure, the feature construction, the classification and the evaluation.

5.3.1 Dataset

The CASIA Gait Dataset B (Yu et al. 2006) was collected by the Chinese Academy of Science, to create a benchmark set to compare gait recognition methods and to evaluate their performance under varying walking conditions. To this end they recorded participants while walking along a straight path of about 2 to 3 gait cycles in length. Figure 30 illustrates a side view of the path with several snapshots of a participant superimposed. Giving the participants' different step sizes and cadences, each clip contains 2 to 3 gait cycles. A gait cycle starts with two feet next to each other right before the leg starts to move forward to the open position (half a step). Then follows movement from the second leg from the back to start position and ending in the open position again (second leg in front and one full step). The cycle ends when the initial leg is back in the closed position (half step again). Eleven cameras were positioned in a semi-circle directed at the middle of the path. Their viewing angles ranged from 0 degrees (frontal view) up to 180 degrees (rear view) in steps of 18 degrees. Each participant was asked to walk the straight path ten times. In six of these walks, participants wore normal clothes, during two of the walks they carried a bag, and during the remaining two they wore a coat.

The full dataset consists of 124 (31 female and 93 male) participants with ages ranging from 20 to 30 years old who are almost all Asian (with the exception of 1 European participant). The clips are recorded at 25 fps with a frame size of 320×240 pixels. In our experiments we used the data from 62 participants, all 31 female participants and 31 randomly selected male participants. We will be evaluating the difference in performance between static and dynamic Gabor filters using clips obtained in the normal clothing condition for both the 0° (frontal) view as well as the 90° (sideway) view. Besides video clips the CASIA set also contains precalculated silhouette images for most of the corresponding frames. These binary images, obtained by background subtraction and thresholding, show the outline of the participant in white and the rest of the frame in black. These will be helpful for determining the features of the respective methods, as we discuss next.

5.3.2 Implementation Details

To construct the STGF, SGF and GEI images, we rely on the (binary) silhouette images supplied with the dataset. Obviously, they are used to construct the

Figure 30: Example of a gait cycle recorded sideways (90° view). Starting and ending with the legs in the closed position and opening them in between.

Figure 31: Example of a gait cycle frontally recorded (0° view). Starting and ending with the legs in the closed position and opening them in between.

original GEI benchmark images. Furthermore, we use the images as a mask to extract the corresponding regions of STGF and SGF responses. The silhouette images in the dataset were extracted using the method described in the work of Wang, Tan, Ning, and Hu (2003). In short, the method follows a standard background foreground segmentation approach, where the binary residual image after background subtraction is enhanced by dilation, erosion connected component analysis, which are morphological operations that either add or remove pixels based on its surrounding pixels, resulting, in this case, in a person's silhouette. Unfortunately, for some sequences the silhouette images are missing, are not completely connected or show strange artifacts. These cases were left out of our experimental set.

For determining the gait cycles we adopted the method proposed by Han and Bhanu (2004) and modified it slightly. In particular, for each silhouette image of the sideway view we sum all the individual pixels obtaining the participant's size of the silhouette area (Size of Silhouette Area: SSA) per frame. Plotting the SSA as a function of frame (i.e., time) gives a curve with a negative peak when the legs are at the closed position (since the two legs largely overlap resulting in a smaller SSA) and positive plateaus when the legs are open (larger SSA). By finding the negative peaks, we determine the start and end of each gait cycle.

Our method differs from Han and Bhanu (2004)'s method in the way the negative and positive peaks are found. There are many different peak finding algorithms. Han and Bhanu (2004) opted for a maximum entropy spectrum estimation, whereas we used the zero-crossings of the smoothed first-derivative to locate the peaks. The choice of peak finding method does not hinge the results.

Our cycle detection procedure finds 740 individual cycles in the 90° condition for all the remaining sequences of the 62 selected participants. Note that the total number of actual cycles in de video clips ranges from 744 (i.e., 62 participants $\times$ 6 normal condition sequences $\times$ 2 cycles per sequence) to 1116 ($62 \times 6 \times 3$). Figure 30 shows five stages of a gait cycle superimposed on one frame for the sideway view. Since for some frames the silhouette images are missing, we only consider the cycles in the 0° condition that have at least 90% of the corresponding silhouette images available. This results in 613 individual cycles in the frontal view condition. An example of a gait cycle recorded from a 0° view is shown in Figure 31.

Feature Construction

We first describe our construction of the GEI feature, i.e., our alignment and normalization scheme. This will be followed up by the description of our Gabor feature, which takes the silhouette images that the GEI feature is comprised of and uses them as a mask to extract the same human shape in the original frame.

As mentioned in Section 5.1.2 we construct GEI features for sideways and frontal recordings. We use the subject's head to align the binary images in the sideways view, because its size and shape hardly change during the sequence. For each silhouette image we fix the head and allow for a substantial margin to the left and right to make sure we enclose all leg

movements. We then normalize the extracted region to 90×134 pixels for width and height respectively, which corresponds to the average width and height of all sideways silhouette images in the dataset.

Our alignment scheme in the frontal view follows a slightly different approach. We observed that while walking some participants move their upper body from left to right. As a consequence of this movement the head is not always centered in the vertical plane with respect to the center of mass of the body. Instead of using the head to align the silhouettes, we simply take the bounding box enclosing the maximum width and height of the positive area of the binary image (recall that pixels of binary images are either 0 or 1) and resize that region to the mean width and height of the frontal silhouette images (i.e., 41×130 pixels). Figure 29 shows GEI examples for both female and male participants in the sideways as well as in the frontal condition.

Our dynamic Gabor filter feature construction adopts the same approach as we used in the previous chapters. First we construct a filter bank of 56 filters, i.e., filters sensitive to 8 different directions and 7 variations in speed. The first speed preference is set to 0 pixels per frame (PPF), which we consider as the static Gabor filter response. Like in the previous chapters we range the other speeds from 0.5 to 3 PPF, which constitute our dynamic responses. We convolve the filter bank with the original RGB recorded frontal and sideways gait sequences, and average the responses yielding 56 corresponding Gabor images per gait cycle. In order to stay close to the original GEI method where only the silhouette pixels contribute to the residual image, we remove filter responses generated from the static background by setting an empirically determined threshold of 0.5 (Gabor convolutions yield continuous values) to filter out the noise. Filter responses below the threshold are set to 0. We then extract the same region from the Gabor response image as we calculated for the silhouette alignment schemes, and apply the same transformations to normalize the patch to its respective average width and height (i.e., for sideways or frontal view). This procedure allows us to compare our Gabor features to the GEI features fairly. The resulting images (i.e., STGF, SGF and GEI) are the inputs to the classification scheme described next.

5.3.3 Evaluation Procedure

Inspired by Yu et al. (2009), we apply 31−fold cross validation to determine the informativeness of the different features. For each participant we used four recordings of the normal condition with each recording containing 2 to 3 gait cycles. One fold comprises all gait cycles of one male and one female example. In each fold we hold out one male and one female example and train a linear support vector machine with the remaining examples. Performance of the trained SVM is evaluated by testing with the previously unselected male and female participants and is expressed by the correct classification rate, which is determined by the correctness of the predictions of each separate cycle. In other words, we classify a cycle as pertaining to a male or female, instead of classifying one sequence with (possibly) multiple gait cycles. Furthermore, as we did in the previous chapters, we examine the performance of the features on different granularities. Besides taking the

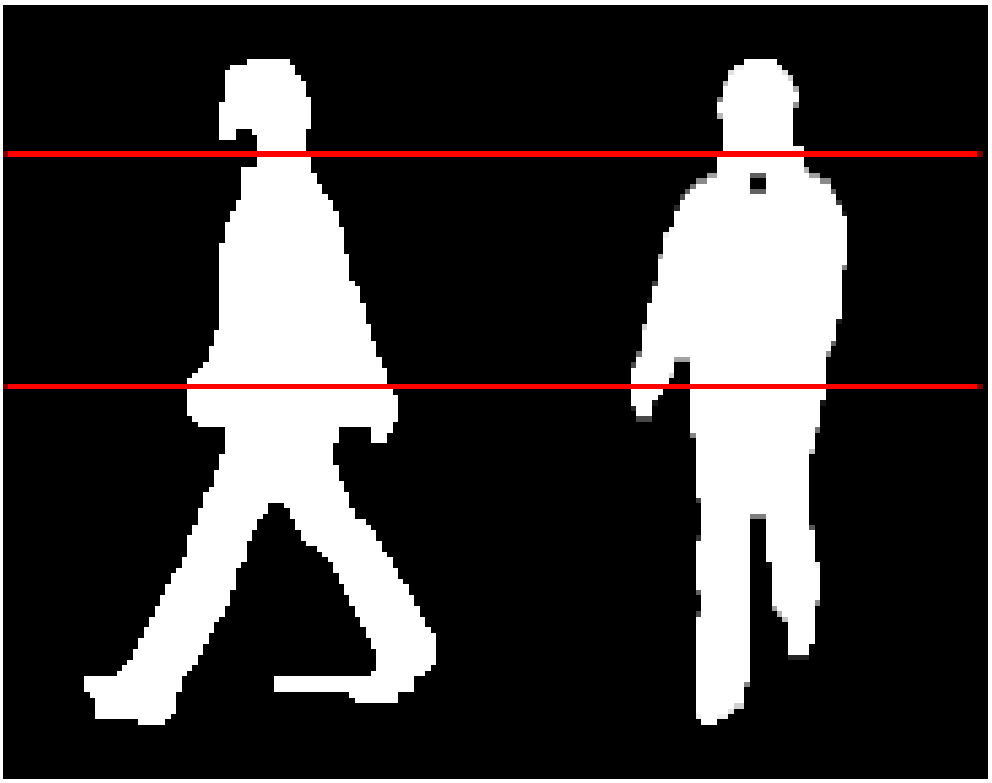


Figure 32: Silhouette images and their corresponding partitioning in head, upper body and lower body.

Table 9: Correct gender classification rates (%±SD) of spatiotemporal GF (STGF), spatial GF (SGF), and GEI, applied to frontal and sideway views of recorded participants.

View	STGF	SGF	GEI
Sideway	**96.5 ± 12.5**	96.0 ± 10.9	95.84 ± 13.0
Frontal	**95.0 ± 13.1**	94.1 ± 13.8	94.6 ± 12.7

whole body into account we also look at the partitions depicted in Figure 32, i.e., head, upper body and lower body. The final level of comparison is at the level of individual speeds of the dynamic Gabor responses. Like we did in the previous chapters we construct individual SVM models that take as input the Gabor responses from filters sensitive to one speed. By looking at the performance of the individual models we can see if certain speeds are more informative to gender classification than others.

5.4 RESULTS

Table 9 summarizes the main result, comparing three different methods (spatiotemporal Gabor filters, spatial Gabor filters, and the GEI baseline), for both frontal and sideways recordings. The first thing to note is that the GEI baseline already scores very high, as expected based on the earlier work discussed above. The results we obtained with this method are highly similar to the earlier reported results on this dataset with the same method (Yu et al.

Table 10: Correct classification rates (%±SD) of spatial GFs (SGF), spatiotemporal GFs (STGF), and GEI applied to sideway en frontal view recorded participants. Performance is evaluated on three levels of granularity, i.e., the head, upper body, and lower body for both sideways as frontal recordings.

View	ROI	STGF	SGF	GEI
Sideway	Head	**97.3 ± 9.7**	94.38 ± 13.9	97.2 ± 9.9
	Upper	**95.2 ± 13.4**	93.37 ± 13.2	92.7 ± 16.0
	Lower	89.1 ± 19.1	**92.5 ± 13.0**	86.7 ± 19.5
Frontal	Head	**94.4 ± 12.3**	91.7 ± 15.2	86.3 ± 20.3
	Upper	**87.5 ± 18.7**	86.8 ± 19.1	82.8 ± 21.3
	Lower	93.2 ± 14.7	**93.7 ± 13.1**	85.3 ± 20.4

2009). The SGF method scores very similar to the GEI method (slightly higher for the sideway recordings and slightly lower for the frontal recordings). Interestingly, the STGF yield the best results in both cases (although the SDs are high), with an increase in correct classification rates of around 1%, even though the GEI and static GF scores are already very high.

Comparing Different Body Parts

Next, we zoom in on the contribution of different body parts. The results of gender classification using specific parts are displayed in Table 10. Again, all three methods generally perform very well and the differences between them are small (and the SDs are high). Importantly, the results of the STGFs are typically better than the static ones, except for the classifications based on the lower part of the body. Here, for both views, the SGFs outperform their STGF counterparts. This is interesting in light of the earlier cited comment from Yu et al. (2009) about lower body gait-based gender classification, which stated that participants had indicated to find static body shape more informative than dynamic information. In the automatic case dynamic information does seem to benefit the performance.

Comparing Different Speeds

In previous chapters, we have seen that usually the best results are obtained when combining different speeds in the Dynamic Gabor Filter approach, and this is what we have reported so far in this chapter. However, it is interesting to compare the performances of different speeds for this task as well. Table 11 summarizes the scores of the filters for the individual speeds plus the combination of all speeds applied to the sideways and frontal recordings for each granularity. Inspection of the table reveals that, for both conditions on the whole frame the combination of all speeds performs best as well as for the upper part granularity in the sideways condition and the for the head granularity in the frontal condition. In the other cases a single speed performs best, although for every case in the table the differences are very small.

Table 11: Correct gender classification rates (%±SD) of STGFs with different speeds (Sp) applied to the head, upper part, lower part and whole frame granularity for both viewing conditions. The view (V) is sideways (S) or frontal (F).

V	Sp	Head	Upper part	Lower part	Frame
S	0.5	97.2 ± 9.8	94.4 ± 14.1	86.1 ± 20.2	95.6 ± 13.0
	1	**97.6 ± 9.1**	95.1 ± 14.0	87.2 ± 20.3	95.6 ± 13.0
	1.5	97.0 ± 10.3	94.4 ± 14.7	88.3 ± 19.5	95.8 ± 12.8
	2	97.0 ± 10.6	94.8 ± 13.9	89.2 ± 18.9	96.3 ± 12.5
	2.5	96.9 ± 10.4	94.7 ± 13.9	90.5 ± 17.3	95.9 ± 12.8
	3	96.6 ± 10.6	94.9 ± 13.2	**90.5 ± 17.3**	95.9 ± 12.8
	all	97.3 ± 9.7	**95.2 ± 13.4**	89.1 ± 19.1	**96.5 ± 12.5**
F	0.5	91.5 ± 14.6	**88.1 ± 18.7**	92.9 ± 14.1	93.8 ± 13.8
	1	92.0 ± 13.1	86.1 ± 20.3	93.1 ± 15.6	93.0 ± 14.5
	1.5	93.1 ± 11.8	85.5 ± 20.9	**93.2 ± 14.2**	92.5 ± 15.6
	2	92.7 ± 11.8	84.2 ± 22.1	91.1 ± 15.6	92.1 ± 17.7
	2.5	92.2 ± 12.9	84.4 ± 22.6	90.6 ± 15.4	92.4 ± 18.6
	3	90.7 ± 15.3	87.7 ± 19.7	90.8 ± 14.6	92.9 ± 16.8
	all	**94.4 ± 12.3**	87.5 ± 18.7	93.2 ± 14.7	**95.0 ± 13.1**

5.5 DISCUSSION

Automatically detecting the gender of walkers based on their gait is a popular task, which has received considerable scholarly attention in recent years. It is generally acknowledged that men and women display a different temporal gait patterns. Therefore, in this chapter, we examined whether adding temporal information from spatiotemporal Gabor filters to the widely used appearance based method GEI, can improve classification results even further. Gait-based gender classification is also an interesting addition to the three social signals that we have studied so far in this thesis (visual speech, smiles and children's facial reactions to arithmetic questions), because human gait is arguably the largest, most prominent moving signal of them all. Moreover it is conveyed at the largest scale we have looked at so far, namely the whole body. For these reasons we were curious to explore the performance of our dynamic filters on this well defined task.

Methodically we compared the GEI method to the static and dynamic Gabor methods described in this chapter, using the publicly available CASIA Gait Dataset (Yu et al. 2006), that has been used in many automatic GAIT detection experiments. In our first experiment we compared the three methods on the frontal and sideways recorded participants in the dataset, where we looked at the whole body and used every filter's response in the classifier to see if adding dynamic information can improve the already impressive GEI accuracy. Second, we zoomed in on the contribution of different body parts to the performance, since we conjectured that the for instance the temporal pattern displayed by the legs when walking differs in speed and amplitude

from the arm and shoulder movement. Finally, we isolated the individual dynamic Gabor responses per speed and evaluated their performance on all body parts.

It is important to note that the results of the GEI baseline method applied to the whole body are already close to ceiling with respect to performance (after all, a classifier can obviously not reach an accuracy above 100% correct). It is well-known that improving scores that are so close to ceiling is hard. When looking at the whole clip, the GEI method yielded scores of 95.8% correct classifications for sideways recordings and 94.6% for frontal recordings. The scores for the spatial Gabor filters were comparable, but those for the spatiotemporal Gabor filters were systematically highest (96.5% for sideways and 95.0% for frontal recordings respectively). Even though this is only a relatively small increase and standard deviations are high, it does suggest that adding the dynamic information does systematically improve performance.

Importantly, this same pattern is revealed when specific regions of interest (head, upper part and lower part) are considered, where the dynamic Gabor filters again systematically outperform the high GEI-baseline, and (with the exception of the lower body) also the static Gabor filters. This suggests that adding dynamic information is indeed beneficial. Yu et al. (2009) discuss earlier gait-based gender classification studies on the CASIA database, showing that these studies have achieved accuracy scores between 85% and 95%, with their own approach (on 31 males and 31 females) yielding the best results (96.0%). This shows that our dynamic Gabor filters produce results that are at least as good as the state-of-the-art.

We also looked at the effects of different speeds, and found – in contrast to previous chapters but similar to Chapter 4 – that combining all speeds does not always yield the best result. It is interesting to observe that the differences in performance between the different speeds themselves are mostly small. This suggests that there is no single speed that is characteristic of gait movements, which might be due to the large scale at which movements occur. After all, during one gait cycle, the feet display more movement than the upper leg, say, which might influence scores for different speeds. Figure 33 displays examples of correct and incorrect classifications of a male, a female and two sequences of the same male but different from the first one.

5.6 CONCLUSION

In this chapter, we explored whether dynamic spatiotemporal Gabor filters (STGFs) are better at classifying full body cues than static, spatial Gabor filters (SGFs). For this, we zoomed in on the task of gender detection based on gait, using a Chinese benchmark gait dataset (CASIA). We compared the performance of our Gabor filters with a state-of-the-art method relying on GEI features, representing human motion during walking in a single image, capturing temporal information as shades of gray. Even though the method relying on the GEI feature already performed very well, we did find that the Gabor filter method yielded higher correct gender classification rates, both for frontal and for sideways recordings, although improvements were

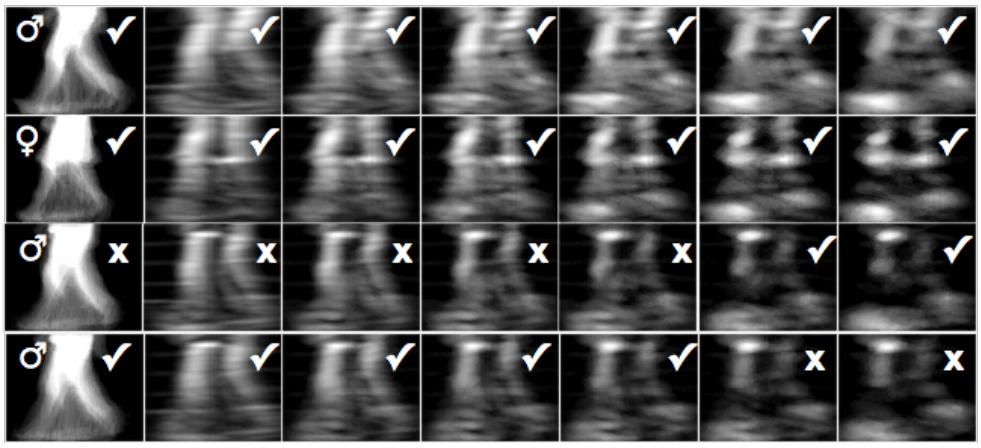


Figure 33: Examples of lower part GEI (first column) and Gabor images with different speeds (from the second column until the last and ranging from 0.5 to 3 PPS) extracted from a male (participant 90), a female (participant 60) and twice the same male (participant 24) from sideways recordings. The "X" and check mark sign indicate a misclassification or a correct classification respectively. Classification was performed on the lower part granularity.

relatively small and standard deviations were high. Crucially, however, we found that STGFs generally outperformed SGFs.

6 GENERAL DISCUSSION AND CONCLUSION

6.1 DISCUSSION

This dissertation investigated to what extent Social Signal Processing (SSP) tasks benefit from taking dynamic information into account, by systematically comparing the contributions of both static and dynamic information to various SSP tasks. More precisely, we used spatial Gabor filters (SGF) and spatiotemporal Gabor filters (STGF) that are able to break down visual signals into structures of visual shape and movement. Despite the importance of Gabor filter methods in SSP, the benefits of STGFs for these tasks received little scholarly attention so far. In fact, to the best of our knowledge, only one previous study has compared dynamic versus static Gabor filters on automatic facial action unit detection (Wu et al. 2010). Their result showed an improvement of dynamic filters over static ones for the classification of facial action units. However, whether a similar benefit can be observed for other social signals remained an open question, and was the central research question of this thesis.

6.1.1 Summary of the Findings

We started our explorations in Chapter 2 with *visual voice activity detection* (VVAD): the task of determining — based on visual information only — whether someone is speaking or not, which can be helpful for applications ranging from speaker identification in multi-party discourse to audio detection in noisy environments. We relied on two different datasets, with different ratios between speech and non-speech: one is the publicly available CUAVE dataset (Patterson et al. 2002) with speakers uttering digits while being filmed both from the front and from the side (relatively much speech), and the other is the LIVER dataset (Joosten et al. 2012), in which participants utter a single word ("liver", hence relatively little speech). We systematically compared dynamic, STGFs (an approach which we dubbed STem-VVAD) with their static, SGF counterparts, relying on the implementation of Petkov and Subramanian (2007), looking at the performance of filters at different speeds and applied to different levels of detail: zooming in on the mouth-region, the face or the whole clip. We found that the best results were obtained by STGFs (of all speeds combined) applied to the mouth region, which revealed a clear improvement over the SGFs applied to the same region. Even though these results are promising, the generalizability over different speakers was not optimal, suggesting that it is important to include speaker-characteristics for visual speech detection.

Visual voice activity detection is, of course, a very basic task (albeit one that was somewhat more difficult than we originally anticipated). Would STGFs also be beneficial for more subtle social signals, that are not inherently

tied to a specific location in the face, such as the mouth? This was addressed in Chapter 3, where we looked at detecting learning difficulties of children based on facial expression analyses. Being able to detect whether or not young learners experience problems with, say, arithmetic assignments is important for, for example, the development of adaptive tutoring applications. For this study we relied on a dataset of children from two age groups, both solving easy and more difficult arithmetic problems, which we collected especially for this purpose. In this chapter, we again compared a dynamic approach, STGFs, with a static variant, SGFs, this time comparing the performance of two implementations, one due to Petkov and Subramanian (2007) and one to Heeger (1987). We also included a more explicit approach in our evaluation, which was based on a model of children's faces using Active Appearance Models (AAMs) (Cootes et al. 2001; Matthews and Baker, 2004; Van der Maaten and Hendriks, 2010). Our results revealed that, for this particular task, the explicit method based on AAMs clearly outperformed all Gabor approaches. More directly relevant for the topic of this thesis, however, we did find that the STGFs (in both implementations) outperformed the SGFs, although the relative improvement was relatively small.

The results of Chapter 3 suggest that when a social signal is very subtle and not associated with a specific facial area, detecting this signal is hard (although, as noted, STGFs still did somewhat better than SGFs). In Chapter 4 we therefore looked at a signal that is both more subtle than visual voice activity detection, but more localized than learning problem assessment: the classification of smiles as either genuine or not. Being able to distinguish smiles that are caused by genuine happiness (Duchenne smile) from merely social (non-Duchenne) smiles is helpful for automatic emotion recognition systems, but has practical application as well, including for example, the ongoing development of digital photo camera's that automatically decide when a portrait picture would be optimally taken. It has been argued that the speed with which a smile appears on the face (with Duchenne smiles being slower) is a potentially important cue (Krumhuber et al. 2009; Schmidt et al. 2006). Therefore it offers an excellent opportunity to study the added value of dynamic information in STGFs for smile classification (again using both the aforementioned implementations), which we did based on the UvA-NEMO Smile database of spontaneous and posed smiles (Dibeklioğlu et al. 2015). Since head movements might have a profound effect on smile classification, we compared results for both 'raw' (unprocessed) faces and automatically 'fixed' ones. We found, once again, a benefit of dynamic filters; both STGF implementations clearly outperformed the corresponding SGFs, on every granularity and for both aligned and unaligned faces. Interestingly and somewhat unexpectedly, fixing the faces resulted in a drop in performance.

Finally, in Chapter 5 we moved beyond facial signals, and studied the impact of Gabor filters on full body movements. In particular, we looked at an arguably basic, full-body task: gender classification based on a person's gait, which has potential practical applications, for example, for shops which want to automatically track the number of male and female shoppers inside. It has been shown that general bodily movement characteristics are helpful for this (Kozlowski and Cutting, 1977). In the final experimental study of this thesis we studied how Gabor filters perform on this task, comparing

STGFs and SGFs. We used the CASIA Gait Dataset B (Yu et al. 2006), which is often used for comparing gait recognition methods, and also included a state-of-the-art GEI-based system for the sake of comparison. We applied these systems both to frontal and sideways clips of people walking, at different levels of detail: the head as well as the upper and the lower body. We found that the results for the GEI-method were already rather good, but we did observe that the Gabor filter methods lead to even better gender classification rates, for both frontal and sideways recordings. Most importantly, we found, once again, that STGFs generally outperformed SGFs.

6.1.2 Discussion

Because the results of all individual studies were already discussed in the separate chapters, here we will concentrate on the main discussion points of this thesis.

The Added Value of Dynamic Information

The central question in this thesis was: does adding dynamic information to Gabor filters benefit the automatic analysis of human social signals? So, does it? Does adding dynamic information to a Gabor filter improve classification results? In all chapters we found that dynamic, STGFs outperform their static counterpart SGFs, suggesting that movement information indeed is beneficial. Importantly, however, the benefits of STGFs over SGFs differed between the various signals, ranging from a rather modest improvement, in the case of learning difficulty assessments (Chapter 3) to substantial, for instance in the case of visual voice activity detection (Chapter 2).

In general, it seems that especially for clear movements, in a specific location, STGFs performed substantially better than SGFs, but less so for movements that are not clearly located in one or more specific places (most notably in the case of children solving arithmetic problems, Chapter 3). This makes sense: the filters are applied at local regions, so if movements can not be pinpointed to a particular location dynamic Gabor filters are less likely to pick them up. Because of the global nature of the head movement, many different filters will be activated, thereby causing the relevant signal to disappear in the general noise. Interestingly, in Chapter 3 we saw that Gabor filters (both static and dynamic ones) were outperformed by the AAM method, the success of which can be attributed to the rigid movement of the head, instead of local facial muscle movement.

The Impact of Different Speeds

Throughout this thesis, we have systematically compared the performance of STGFs at different speeds. In many cases, the best performance was obtained by combining information from all speeds (with smiles, Chapter 4, as a notable exception), but it is interesting to also ask which individual speed performs best for which task.

It is likely that different social signals are associated with different "signature" speeds. For example, intuitively, movements of the mouth may differ in

size and speed from movements of, let us say, the eyelids, and this becomes even more evident when comparing facial gestures with full-body gestures (as in Chapter 5).

Still, in general, we found that differences between speeds were often subtle. For most of our experiments, filters tuned to low speeds (i.e., 0.5 pixels per frame) gave the best individual results. The most notable exception was found in Chapter 5 for gait based gender classification. When we zoomed in on the legs in the sideway condition, we found that the higher speeds lead to better results than the lower speeds. The movement of the legs in this condition probably represented the largest displacement in pixels per frame displayed of all video sequences in this dissertation. In general, we did find that combining different speeds usually lead to the best performance, which is presumably due to the fact that different speeds code for partially different information.

Comparing Implementations

We experimented with two different implementations of dynamic Gabor filters, one due to Petkov and Subramanian (Petkov and Subramanian, 2007) and one due to Heeger (Heeger, 1987). While the first has studied STGFs in the context of modeling cells in the primary visual cortex, the second applied them in the context of estimating image velocity. Due to the biological plausibility of the first implementation, it differs in choice of parameters and in their relations from the second one, although both based on the same mathematical principles of Gabor filters. Moreover, Petkov and Subramanian's implementation is able to construct both velocity tuned filters and frequency tuned filters, whereas the Heeger implementation only allows for the construction of frequency tuned filters. In general, where both approaches were applied to the same data (Chapter 3 and Chapter 4) we found little differences in performance, suggesting that the specific implementation (and whether it is biologically inspired or not) did not so much matter, at least for these tasks.

Future Research: (Deep) Learning of Feature Representations

In this thesis, we studied the benefits of STGFs for various SSP tasks, ranging from subtle to more obvious signals, and we found that these benefits differed somewhat for different signals. To be able to better predict for which kind of task adding dynamic information is beneficial, it would be helpful to look at a broader range of tasks. This applies especially to subtle signals, such as for instance stress detection (Koldijk, 2016), which is clearly important for a range of applications such as stress-related absenteeism reduction programs or simulators that train emergency response personnel and monitor their stress levels, even though there is no single cue that is clearly associated with stress. At the other extreme, it would be interesting to look at full body expressions in more detail. In this thesis we looked at gait-based gender detection, but it would be interesting to also look at bodily signals that are more clearly socially relevant, such as full body emotional expressions (Gelder, 2006) and manual co-speech gestures (McNeill, 1996).

More technically, it would be interesting to get a better understanding of what the contributions are of individual features for classification. In our experiments we employed a method that aggregates Gabor features over areas of interest. It might be worthwhile to examine different feature constructions (e.g., only using the maximum filter response) and explore the optimal setting for different scenarios, an approach sometimes referred to as a bag-of-features approach.

A general limitation of the approach described in this thesis is that the filters in the STGF filterbank have been manually constructed. As a result, it is conceivable that we may have missed certain combinations of values in the frequency domain that would have given better results than the ones we manually selected. An alternative would be to automatically learn optimal filters based on input data, which has the advantage that the statistical representations of the domain are learned. It would be interesting to see which performance increase (if any) this would yield.

One way to do this would be to make use of the recent developments in the field of deep learning. For example, one could explore the use of deep learning for generating shape and appearance features (Jaiswal and Valstar, 2016), which has yielded promising results relevant to this thesis (Egede, Valstar, and Martinez, 2017). An interesting use case would be to apply these models to learn the optimal combination and configuration of filters.

While the use of deep learning techniques for social signal processing is currently increasingly popular, they have also two disadvantages which we briefly mention here. One is that deep learning is difficult to understand, in the sense that deep learning models still mostly operate as a black box. Additionally, for good performance they require very large sets of training examples, which are not always readily available for the signals studied in this thesis.

In this thesis we wanted to investigate in a systematic manner what the added value of temporal information is to Gabor filters in the context of social signal processing. The goal was not *per se* to achieve state-of-the-art results, and indeed various solutions exist which perform better than our approaches (typically involving deep learning techniques). In fact, it will be very interesting to see what effects these techniques for learning features and filters will have on performance, and we hope to address this in future work.

6.2 CONCLUSION

In this thesis, we set out to examine whether adding temporal information to spatial Gabor filters leads to better predictive performances of automatic systems in the context of social signal processing. Based on the experiments performed in this thesis, we conclude that this is indeed often the case.

This is especially true when salient movements are explicitly present in specific facial or bodily areas. In those cases, adding temporal information generally lead to better predictive results. Comparable or even better results might be obtainable with methods zooming in on a narrow region, for example a facial landmark, or by explicitly tracking fiducial points, but this comes at the price of adding manual annotations or using point-light displays

while the Gabor filter method does not require such additional efforts. When movements were very subtle or not specifically associated with specific facial parts (most notable in our study of learning difficulty assessment) we found the Gabor filter method to perform less well, although here too the dynamic, STGF method outperformed the static, SGF method.

Additionally, based on our studies we can conclude that combining information of different filters, tuned to various speeds, generally leads to better performance than using filters with one specific speed preference. In general, the social signals we studied did not seem to be associated with a specific "signature" speed. We found little or no performance differences between the two different implementations of STGFs we used.

Taking everything together, we can conclude that when one wants to use Gabor filters for the automatic analysis of social signals, it is better to use spatiotemporal Gabor filters rather than the more common spatial filters.

REFERENCES

Adolphs, R. (2002a). "Neural systems for recognizing emotion." *Current opinion in neurobiology* 12.2, pp. 169–177 (cit. on p. 2).

— (2002b). "Recognizing emotion from facial expressions: Psychological and neurological mechanisms." *Behavioral and cognitive neuroscience reviews* 1.1, pp. 21–62 (cit. on p. 40).

Ambady, N. and Rosenthal, R. (1992). "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." *Psychological bulletin* 111.2, p. 256 (cit. on p. 47).

Amelsvoort, M. van, Joosten, B., Krahmer, E., and Postma, E. (2013). "Using non-verbal cues to (automatically) assess children's performance difficulties with arithmetic problems." *Computers in Human Behavior* 29.3, pp. 654–664 (cit. on pp. 41, 43, 47, 48).

Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., and Solomon, P. E. (2009). "The painful face – Pain expression recognition using active appearance models." *Image and Vision Computing* 27.12, pp. 1788–1796 (cit. on p. 42).

Aubrey, A., Rivet, B., Hicks, Y., Girin, L., Chambers, J., and Jutten, C. (2007). "Two novel visual voice activity detectors based on appearance models and retinal filltering." In: *Proceedings of the 15th European Signal Processing Conference, EUSIPCO-2007*, pp. 2409–2413 (cit. on p. 20).

Aviezer, H., Trope, Y., and Todorov, A. (2012). "Body cues, not facial expressions, discriminate between intense positive and negative emotions." *Science* (cit. on p. 77).

Banda, N. and Robinson, P. (2011). "Multimodal Affect Recognition in Intelligent Tutoring Systems." In: *Affective Computing and Intelligent Interaction*. Ed. by S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin. Vol. 6975. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 200–207 (cit. on p. 41).

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., et al. (2006). "Top-down facilitation of visual recognition." *Proceedings of the National Academy of Sciences of the United States of America* 103.2, pp. 449–454 (cit. on p. 5).

Beritelli, F., Casale, S., and Cavallaero, A. (1998). "A robust voice activity detector for wireless communications using soft computing." *Selected Areas in Communications, IEEE Journal on* 16.9, pp. 1818–1829 (cit. on p. 19).

Bettadapura, V. (2012). "Face Expression Recognition and Analysis: The State of the Art." [Online]: https://arxiv.org/abs/1203.6722 (cit. on p. 60).

Bhanu, B. and Han, J. (2002). "Individual recognition by kinematic-based gait analysis." In: *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. IEEE, pp. 343–346 (cit. on p. 79).

Bosch, N., Chen, Y., and D'Mello, S. (2014). "It's Written on Your Face: Detecting Affective States from Facial Expressions while Learning Computer

Programming." In: *Intelligent Tutoring Systems*. Ed. by S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia. Vol. 8474. Lecture Notes in Computer Science. Springer International Publishing, pp. 39–44 (cit. on p. 41).

Bouguet, J.-Y. (2000). "Pyramidal Implementation of the Lucas Kanade Feature Tracker." *Intel Corporation, Microprocessor Research Labs* (cit. on p. 21).

Bouma, H., Burghouts, G., Hollander, R. den, Van Der Zee, S., Baan, J., Hove, J.-M. ten, Diepen, S. van, Haak, P. van den, and Rest, J. van (2016). "Measuring cues for stand-off deception detection based on full-body nonverbal features." In: *SPIE Security+ Defence*. International Society for Optics and Photonics, pp. 99950M–99950M (cit. on p. 77).

Chang, J.-H., Kim, N. S., and Mitra, S. K. (2006). "Voice activity detection based on multiple statistical models." *Signal Processing, IEEE Transactions on* 54.6, pp. 1965–1976 (cit. on p. 19).

Chew, S. W., Lucey, P., Lucey, S., Saragih, J., Cohn, J. F., Matthews, I., and Sridharan, S. (2012). "In the Pursuit of Effective Affective Computing: The Relationship Between Features and Registration." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 42.4, pp. 1006–1016 (cit. on p. 62).

Chu, W.-S., De la Torre, F., and Cohn, J. F. (2013). "Selective Transfer Machine for Personalized Facial Action Unit Detection." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 42).

Cohn, J. F. (2010). "Advances in Behavioral Science Using Automated Facial Image Analysis and Synthesis." *IEEE Signal Processing Magazine* 27.6, p. 128 (cit. on p. 4).

Cohn, J. F. (2007). "Foundations of human computing: Facial expression and emotion." In: *Artifical Intelligence for Human Computing*. Springer, pp. 1–16 (cit. on p. 42).

Cohn, J. F. and Schmidt, K. L. (2004). "The timing of facial motion in posed and spontaneous smiles." *International Journal of Wavelets, Multiresolution and Information Processing* 2.2, pp. 121–132 (cit. on pp. 60, 61, 75).

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). "Active appearance models." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.6, pp. 681–685 (cit. on pp. 4, 17, 20, 43, 44, 92, 112).

Coulson, M. (2004a). "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence." *Journal of nonverbal behavior* 28.2, pp. 117–139 (cit. on pp. 3, 4).

— (2004b). "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence." *Journal of nonverbal behavior* 28.2, pp. 117–139 (cit. on p. 77).

Craig, S. D., D'Mello, S., Witherspoon, A., and Graesser, A. (2008). "Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive–affective states during learning." *Cognition and Emotion* 22.5, pp. 777–788 (cit. on p. 41).

Craig, S., Graesser, A., Sullins, J., and Gholson, B. (2004). "Affect and learning: an exploratory look into the role of affect in learning with AutoTutor." *Journal of Educational Media* 29.3, pp. 241–250 (cit. on p. 41).

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. Perennial Modern Classics. Harper & Row (cit. on p. 39).

Daugman, J. G. (1985). "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters." *Journal of the Optical Society of America A* 2.7, pp. 1160–1169 (cit. on pp. 9, 11, 22).

De la Torre, F., Chu, W.-S., Xiong, X., Vicente, F., Ding, X., and Cohn, J. (2015). "IntraFace." In: *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, pp. 1–8 (cit. on p. 4).

Dellwo, V., Leemann, A., and Kolly, M.-J. (2012). "Speaker idiosyncratic rhythmic features in the speech signal." In: *Proceedings of Interspeech, Portland (USA)* (cit. on p. 23).

Derpanis, K. G. (2007). "Gabor filters." Lecture note. URL: http://www.cse.yorku.ca/~kosta/CompVis_Notes/gabor_filters.pdf (cit. on p. 7).

Dibeklioğlu, H., Salah, A., and Gevers, T. (2015). "Recognition of Genuine Smiles." *IEEE Transactions on Multimedia* 17.3, pp. 279–294 (cit. on pp. 17, 42, 60–62, 69, 74, 75, 92, 112).

Dibeklioğlu, H., Salah, A. A., and Gevers, T. (2012). "Are you really smiling at me? Spontaneous versus posed enjoyment smiles." In: pp. 525–538 (cit. on pp. 61, 62, 64, 65, 68).

Dibeklioğlu, H., Valenti, R., Salah, A. A., and Gevers, T. (2010). "Eyes do not lie: Spontaneous versus posed smiles." In: *Proceedings of the 18th ACM international conference on Multimedia*. ACM, pp. 703–706 (cit. on p. 75).

D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., Kaliouby, R. el, Picard, R. W., et al. (2008). "AutoTutor detects and responds to learners affective and cognitive states." In: *Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems* (cit. on pp. 41, 42).

Dragon, T., Arroyo, I., Woolf, B., Burleson, W., Kaliouby, R. el, and Eydgahi, H. (2008). "Viewing Student Affect and Learning through Classroom Observation and Physical Sensors." In: *Intelligent Tutoring Systems*. Ed. by B. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie. Vol. 5091. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 29–39 (cit. on pp. 41, 42).

Du, S., Tao, Y., and Martinez, A. M. (2014). "Compound facial expressions of emotion." *Proceedings of the National Academy of Sciences* 111.15, E1454–E1462 (cit. on p. 60).

Egede, J., Valstar, M., and Martinez, B. (2017). "Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation." [Online]: https://arxiv.org/abs/1701.04540 (cit. on p. 95).

Ekman, P. (1973). *Darwin and facial expression: A century of research in review.* Academic Press (cit. on p. 40).

— (1992a). "An argument for basic emotions." *Cognition & Emotion* 6.3-4, pp. 169–200 (cit. on p. 40).

— (1992b). "Facial expressions of emotion: New findings, new questions." *Psychological science* 3.1, pp. 34–38 (cit. on p. 59).

Ekman, P., Davidson, R. J., and Friesen, W. V. (1990). "The Duchenne smile: Emotional expression and brain physiology: II." *Journal of personality and social psychology* 58.2, p. 342 (cit. on pp. 4, 60).

Ekman, P. and Friesen, W. V. (1969). "The repertoire of nonverbal behavior: Categories, origins, usage, and coding." *Semiotica* 1.1, pp. 49–98 (cit. on p. 2).

— (1975). *Unmasking the face: A guide to recognizing emotions from facial clues.* Prentice-Hall (cit. on pp. 2, 40).

Ekman, P. and Friesen, W. V. (1976). "Measuring facial movement." *Environmental psychology and nonverbal behavior* 1.1, pp. 56–75 (cit. on pp. 17, 59).

Ekman, P., Friesen, W. V., and Hager, J. (1978). "The Facial Action Coding System (FACS): A technique for the measurement of facial action. Palo Alto." *Palo Alto: Consulting Psychologists Press* (cit. on p. 4).

El Kaliouby, R. and Robinson, P. (2005). "Real-time inference of complex mental states from facial expressions and head gestures." In: *Real-time vision for human-computer interaction*. Springer, pp. 181–200 (cit. on pp. 41, 42).

Evangelidis, G. D. and Psarakis, E. Z. (2008). "Parametric image alignment using enhanced correlation coefficient maximization." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.10, pp. 1858–1865 (cit. on p. 66).

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). "LIBLINEAR: A library for large linear classification." *The Journal of Machine Learning Research* 9, pp. 1871–1874 (cit. on pp. 27, 68).

Fasel, B. and Luettin, J. (2003). "Automatic facial expression analysis: a survey." *Pattern Recognition* 36.1, pp. 259–275 (cit. on p. 60).

Field, D. J. (1987). "Relations between the statistics of natural images and the response properties of cortical cells." *Journal of the Optical Society of America A* 4.12, pp. 2379–2394 (cit. on p. 22).

Fischer, S., Šroubek, F., Perrinet, L., Redondo, R., and Cristóbal, G. (2007). "Self-Invertible 2D Log-Gabor Wavelets." *International Journal of Computer Vision* 75.2, pp. 231–246 (cit. on p. 7).

Freund, Y. and Schapire, R. E. (1995). "A desicion-theoretic generalization of on-line learning and an application to boosting." In: *Computational learning theory*. Springer, pp. 23–37 (cit. on p. 20).

Frijda, N. H. (1986). *The Emotions*. Cambridge University Press (cit. on p. 2).

Furui, S. (1997). "Recent Advances in Speaker Recognition." In: *AVBPA '97: Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication*. Springer-Verlag (cit. on p. 19).

Gabor, D. (1946). "Theory of communication. Part 1: The analysis of information." *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering* 93.26, pp. 429–441 (cit. on pp. 7, 9).

Gatica-Perez, D. (2009). "Automatic nonverbal analysis of social interaction in small groups: A review." *Image and Vision Computing* 27.12, pp. 1775–1787 (cit. on p. 41).

Gelder, B. de (2006). "Towards the neurobiology of emotional body language." *Nature Reviews Neuroscience* 7.3, pp. 242–249 (cit. on pp. 77, 94).

Ghosh, P. K., Tsiartas, A., and Narayanan, S. (2011). "Robust voice activity detection using long-term signal variability." *Audio, Speech, and Language Processing, IEEE Transactions on* 19.3, pp. 600–613 (cit. on p. 19).

Girard, J. M., Cohn, J. F., and Torre, F. D. la (2014). "Estimating smile intensity: A better way." *Pattern Recognition Letters*, pp. -. ISSN: 0167-8655. DOI: http://dx.doi.org/10.1016/j.patrec.2014.10.004. URL: http://www.sciencedirect.com/science/article/pii/S0167865514003080 (cit. on p. 74).

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press (cit. on p. 57).

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., and Lester, J. C. (2013). "Automatically recognizing facial indicators of frustration: a learning-centric analysis." In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, pp. 159–165 (cit. on pp. 39, 41).

Gray, J. A. (1982). *The neuropsychology of anxiety*. Oxford: Oxford university press (cit. on p. 2).

Gregory, R. (1970). *The Intelligent Eye*. London: Weidenfeld and Nicolson (cit. on pp. 5, 6).

Grigorescu, S. E., Petkov, N., and Kruizinga, P. (2002). "Comparison of texture features based on Gabor filters." *IEEE Transactions on Image processing* 11.10, pp. 1160–1167 (cit. on p. 7).

Gross, M., Crane, E., and Fredrickson, B. (2007). "Effect of felt and recognized emotions on gait kinematics." In: *American Society for Biomechanics Annual Conference* (cit. on pp. 3, 4).

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). "Multi-PIE." *Image and Vision Computing* 28.5. Best of Automatic Face and Gesture Recognition 2008, pp. 807–813. ISSN: 0262-8856. DOI: http://dx.doi.org/10.1016/j.imavis.2009.08.002. URL: http://www.sciencedirect.com/science/article/pii/S0262885609001711 (cit. on p. 50).

Han, J. and Bhanu, B. (2004). "Statistical feature fusion for gait-based human recognition." In: *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. IEEE, II–842–II–847 Vol.2 (cit. on p. 84).

Han, J. and Bhanu, B. (2006). "Individual Recognition Using Gait Energy Image." *IEEE Trans. Pattern Anal. Mach. Intell. ()* 28.2, pp. 316–322 (cit. on pp. 17, 79).

Hart, T. (2008). "Interiority and Education." *Journal of Transformative Education* 6.4, pp. 235–250 (cit. on p. 41).

Hateren, H. van and Ruderman, D. (1998). "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265.1412, pp. 2315–2320 (cit. on p. 22).

He, Q. and Debrunner, C. (2000). "Individual Recognition from Periodic Activity Using Hidden Markov Models." *Workshop on Human Motion*, pp. 47–52 (cit. on p. 79).

Heeger, D. J. (1987). "Model for the extraction of image flow." *Journal of the Optical Society of America A* 4.8, pp. 1455–1471 (cit. on pp. 9, 10, 12, 13, 16, 24, 51, 56, 57, 63, 74, 92, 94, 112).

Hoque, M. E., McDuff, D., and Picard, R. W. (2012). "Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles." *T. Affective Computing ()* 3.3, pp. 323–334 (cit. on p. 61).

Howell, J. L. and Shepperd, J. A. (2013). "Reducing health-information avoidance through contemplation." *Psychological science*, p. 0956797613478616 (cit. on p. 41).

Hu, M., Wang, Y., Zhang, Z., and Wang, Y. (2010). "Combining Spatial and Temporal Information for Gait Based Gender Classification." *ICPR*, pp. 3679–3682 (cit. on p. 78).

Itti, L. and Koch, C. (2001). "Computational modelling of visual attention." *Nature reviews. Neuroscience* 2.3, p. 194 (cit. on p. 5).

Izard, C. E. (1977). *Human emotions*. New York: Plenum Press (cit. on p. 2).

Jain, A. K. and Farrokhnia, F. (1990). "Unsupervised texture segmentation using Gabor filters." In: *Systems, Man and Cybernetics, 1990. Conference Proceedings., IEEE International Conference on*. IEEE, pp. 14–19 (cit. on p. 22).

Jain, A. K. and Farrokhnia, F. (1991). "Unsupervised texture segmentation using Gabor filters." *Pattern Recognition* 24.12, pp. 1167–1186 (cit. on p. 7).

Jaiswal, S. and Valstar, M. (2016). "Deep learning the dynamic appearance and shape of facial action units." In: *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, pp. 1–8 (cit. on p. 95).

Johnston, L., Miles, L., and Macrae, C. N. (2010). "Why are you smiling at me? Social functions of enjoyment and non-enjoyment smiles." *British Journal of Social Psychology* 49.1, pp. 107–127 (cit. on p. 59).

Jones, J. P. and Palmer, L. A. (1987). "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex." *Journal of Neurophysiology* 58.6, pp. 1233–1258 (cit. on p. 22).

Joosten, B., Amelsvoort, M. van, Krahmer, E., and Postma, E. (2011a). "Thin slices of head movements during problem solving reveal level of difficulty." In: *International Conference on Audio-Visual Speech Processing (AVSP 2011)*. Ed. by G. Salvi, J. Beskow, O. Engwall, and S. Al Moubayed. Stockholm, pp. 85–88.

Joosten, B., Amelsvoort, M. van, Postma, E., and Krahmer, E. (2011b). *Thin slices of head movements during problem solving reveal level of difficulty*. Poster presented at the 23rd Benelux Conference on Artificial Intelligence. Ghent, November 2011.

Joosten, B., Postma, E., and Krahmer, E. (2014). *Visual Lip Reading Using Spatiotemporal Gabor Filters*. Poster presented at the 41th annual meeting of the Australasian Experimental Psychology Society. Brisbane, Australia, April 2014.

— (2015). "Voice activity detection based on facial movement." *Journal on Multimodal User Interfaces* 9.3, pp. 183–193 (cit. on pp. 70, 71).

Joosten, B., Postma, E., Krahmer, E., Swerts, M., and Kim, J. (2011c). *Automated Measurement of Spontaneous Surprise*. Talk presented at the 23rd Benelux Conference on Artificial Intelligence. Ghent, November 2011.

— (2012). "Automated Measurement of Spontaneous Surprise." In: *Proceedings of Measuring Behavior*. Ed. by A. J. Spink, F. Grieco, O. E. Krips, L. W. S. Loijens, L. Noldus, and P. H. Zimmerman. Utrecht, pp. 385–389 (cit. on pp. 16, 22, 25, 91, 111).

Kapoor, A., Burleson, W., and Picard, R. W. (2007). "Automatic prediction of frustration." *International Journal of Human-Computer Studies* 65.8, pp. 724–736 (cit. on p. 41).

Kapoor, A. and Picard, R. W. (2005). "Multimodal affect recognition in learning environments." In: *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, pp. 677–682 (cit. on pp. 41, 42).

Kass, M., Witkin, A., and Terzopoulos, D. (1988). "Snakes: Active contour models." *International Journal of Computer Vision* 1.4, pp. 321–331 (cit. on p. 20).

Kinnunen, T. and Li, H. (2010). "An overview of text-independent speaker recognition: From features to supervectors." *Speech Communication* 52.1, pp. 12–40 (cit. on p. 19).

Kleinschmidt, M. and Gelbart, D. (2002). "Improving word accuracy with Gabor feature extraction." In: *International Conference on Spoken Language Processing, Denver, CO*, pp. 25–28 (cit. on p. 22).

Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Cengage Learning (cit. on pp. 2, 40).

Koldijk, S. (2016). "Context-Aware Support for Stress Self-Management: From Theory to Practice." Radboud University (cit. on p. 94).

Kort, B., Reilly, R., and Picard, R. W. (2001). "An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion." In: *Advanced Learning Technologies, IEEE International Conference on*. IEEE Computer Society, pp. 0043–0043 (cit. on p. 39).

Kozlowski, L. T. and Cutting, J. E. (1977). "Recognizing the sex of a walker from a dynamic point-light display." *Perception & Psychophysics* 21.6, pp. 575–580 (cit. on pp. 17, 78, 92, 112).

Krahmer, E. and Swerts, M. (2005). "Audiovisual prosody and feeling of knowing." *Journal of Memory and Language* 53.1, pp. 81–94 (cit. on pp. 20, 23).

Kraut, R. E. and Johnston, R. E. (1979). "Social and emotional messages of smiling: An ethological approach." *Journal of personality and social psychology* 37.9, p. 1539 (cit. on p. 59).

Krumhuber, E., Manstead, A., and Cosker, D. (2009). "Effects of dynamic attributes of smiles in human and synthetic faces: A simulated job interview setting." *Journal of Nonverbal* (cit. on pp. 17, 59, 92, 112).

Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L., and Kappas, A. (2007). "Facial dynamics as indicators of trustworthiness and cooperative behavior." *Emotion* 7.4, pp. 730–735 (cit. on p. 60).

Kusakunniran, W., Wu, Q., Li, H., and Zhang, J. (2009). "Multiple views gait recognition using View Transformation Model based on optimized Gait Energy Image." In: *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, pp. 1058–1064 (cit. on p. 80).

Lee, H., Kwon, T., and Cho, D.-H. (2005). "An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system." *IEEE Communications Letters* 9.8, pp. 691–693 (cit. on p. 19).

Lehman, B., Matthews, M., D'Mello, S., and Person, N. (2008). "What are you feeling? Investigating student affective states during expert human tutoring sessions." In: *Intelligent Tutoring Systems*. Springer, pp. 50–59 (cit. on pp. 39, 40).

Li, J. and Allinson, N. M. (2008). "A comprehensive review of current local features for computer vision." *Neurocomputing* 71.10, pp. 1771–1787 (cit. on p. 7).

Liao, S., Zhu, X., Lei, Z., Zhang, L., and Li, S. (2007). "Learning multi-scale block local binary patterns for face recognition." *Advances in biometrics*, pp. 828–837 (cit. on p. 26).

Little, J. and Boyd, J. (1998). "Recognizing people by their gait: the shape of motion." *Videre: Journal of Computer Vision Research* (cit. on p. 79).

Littlewort, G. C., Bartlett, M. S., Salamanca, L. P., and Reilly, J. (2011a). "Automated measurement of children's facial expressions during problem solving tasks." In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, pp. 30–35 (cit. on p. 41).

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., and Bartlett, M. (2011b). "The computer expression recognition toolbox (CERT)." In: *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 298–305 (cit. on pp. 4, 22, 41–44).

Liu, C., Ham, J., Postma, E., Midden, C., Joosten, B., and Goudbeek, M. (2012). "How to Make a Robot Smile? Perception of Emotional Expressions from Digitally-Extracted Facial Landmark Configurations." In: *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings*. Ed. by S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, and M.-A. Williams. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 26–34.

— (2013). "Representing Affective Facial Expressions for Robots and Embodied Conversational Agents by Facial Landmarks." *International Journal of Social Robotics* 5.4, pp. 619–626.

Liu, H. and Wu, P. (2012). "Comparison of methods for smile deceit detection by training AU6 and AU12 simultaneously." In: *2012 19th IEEE International Conference on Image Processing*, pp. 1805–1808 (cit. on p. 60).

Liu, Q., Wang, W., and Jackson, P. (2011). "A visual voice activity detection method with adaboosting." *Sensor Signal Processing for Defence (SSPD 2011)*, pp. 1–5 (cit. on p. 20).

Long, F., Wu, T., Movellan, J. R., and Bartlett, M. S. (2012). "Learning spatiotemporal features by using independent component analysis with application to facial expression recognition." *Neurocomputing* (cit. on p. 57).

Lowe, D. G. (1999). "Object recognition from local scale-invariant features." In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 2, 1150–1157 vol.2 (cit. on p. 7).

Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). "Coding facial expressions with gabor wavelets." In: *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205 (cit. on pp. 22, 60).

MacLennan, B. (1991). *Gabor representations of spatiotemporal visual images*. Tech. rep. CS-91-144. University of Tennessee. Computer Science Department (cit. on pp. 7–10).

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc. ISBN: 0716715678 (cit. on p. 7).

Masters, J. C., Barden, R. C., and Ford, M. E. (1979). "Affective states, expressive behavior, and learning in children." *Journal of Personality and Social Psychology* 37.3, p. 380 (cit. on p. 39).

Matthews, I. and Baker, S. (2004). "Active Appearance Models Revisited." *International Journal of Computer Vision* 60.2, pp. 135–164 (cit. on pp. 4, 17, 43, 44, 92, 112).

McGurk, H. and MacDonald, J. (1976). "Hearing lips and seeing voices." *Nature* 264, pp. 746–748 (cit. on p. 19).

McNeill, M. (1996). *Hand and Mind: What Gestures Reveal about Thought)*. University Of Chicago Press, IL (cit. on pp. 3, 94).

Meyer, D. K. and Turner, J. C. (2006). "Re-conceptualizing emotion and motivation to learn in classroom contexts." *Educational Psychology Review* 18.4, pp. 377–390 (cit. on p. 39).

Movellan, J. (2005). *Tutorial on gabor filters*. Tech. rep. UCSD MPLab (cit. on p. 7).

Nakano, M., Mitsukura, Y., Fukumi, M., and Akamatsu, N. (2002). "True smile recognition system using neural networks." In: *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on*. Vol. 2, 650–654 vol.2 (cit. on p. 60).

Navarathna, R., Dean, D., Sridharan, S., Fookes, C., and Lucey, P. (2011). "Visual voice activity detection using frontal versus profile views." In: *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*. IEEE, pp. 134–139 (cit. on p. 21).

Ng, C., Tay, Y., and Goi, B.-M. (2012). "Recognizing Human Gender in Computer Vision: A Survey." In: *PRICAI 2012: Trends in Artificial Intelligence*. Ed. by P. Anthony, M. Ishizuka, and D. Lukose. Vol. 7458. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 335–346 (cit. on p. 78).

Niedenthal, P. M. and Mermillod, M. (2010). "The Simulation of Smiles (SIMS) model: Embodied simulation and the meaning of facial expression." *Behavioral and brain* (cit. on pp. 17, 59, 60).

Niyogi, S. and Adelson, E. (1994). "Analyzing and recognizing walking figures in XYT." In: *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pp. 469–474 (cit. on p. 79).

Ortony, A. and Turner, T. J. (1990). "What's basic about basic emotions?" *Psychological review* 97.3, p. 315 (cit. on p. 2).

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT press (cit. on p. 5).

Pantic, M. and Rothkrantz, L. J. M. (2000). "Automatic analysis of facial expressions: the state of the art." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.12, pp. 1424–1445 (cit. on p. 60).

Pantic, M. and Bartlett, M. S. (2007). *Machine analysis of facial expressions*. I-Tech Education and Publishing (cit. on p. 4).

Patterson, E. K., Gurbuz, S., Tufekci, Z., and Gowdy, J. N. (2002). "Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus." *EURASIP Journal on Applied Signal Processing* 2002, pp. 1189–1201 (cit. on pp. 16, 21–23, 25, 91, 111).

Petkov, N. and Subramanian, E. (2007). "Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal Gabor filters with surround inhibition." *Biological Cybernetics* 97.5-6, pp. 423–439 (cit. on pp. 9, 13–17, 22, 24, 26, 50, 56, 57, 63, 66, 74, 81, 91, 92, 94, 111, 112).

Pfister, T., Li, X., Zhao, G., and Pietikäinen, M. (2011). "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework." In: *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, pp. 868–875 (cit. on p. 75).

Picard, R. W. (1997). "Affective Computing" (cit. on p. 1).

Pollick, F. E., Kay, J. W., Heim, K., and Stringer, R. (2005). "Gender Recognition From Point-Light Walkers." *Journal of Experimental Psychology: Human Perception and Performance* 31.6, pp. 1247–1265 (cit. on p. 78).

Pollick, F. E., Paterson, H. M., Bruderlin, A., and Sanford, A. J. (2001). "Perceiving affect from arm movement." *Cognition* 82.2, B51–B61 (cit. on pp. 3, 4).

Porter, P. (1954). "Another puzzle-picture." *American Journal of Psychology* 67, pp. 550–551 (cit. on p. 6).

Potamianos, G., Neti, C., Luettin, J., and Matthews, I. (2012). "Audiovisual automatic speech recognition." In: *Audiovisual Speech Processing*. Ed. by G. Bailly, P. Perrier, and E. Vatikiotis-Bateson. Cambridge University Press, pp. 193–247 (cit. on p. 20).

Radlak, K., Radlak, N., and Smolka, B. (2018). "Static Posed Versus Genuine Smile Recognition." In: *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017*. Ed. by M. Kurzynski, M. Wozniak, and R. Burduk. Springer International Publishing, pp. 423–432 (cit. on pp. 60, 61).

Ramírez, J., Segura, J. C., Benítez, C., Torre, Á. de la, and Rubio, A. (2004). "Efficient voice activity detection algorithms using long-term speech information." *Speech Communication* 42.3-4, pp. 271–287 (cit. on p. 19).

Reisenzein, R., Bordgen, S., Holtbernd, T., and Matz, D. (2006). "Evidence for strong dissociation between emotion and facial displays: The case of surprise." *Journal of Personality and Social Psychology* 91.2, p. 295 (cit. on p. 42).

Reynolds, D. (2002). "An overview of automatic speaker recognition." In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4072–4075 (cit. on p. 19).

Rijsbergen, C. van (1979). *Information Retrieval. 1979*. Butterworth (cit. on p. 27).

Rothblum, E. D., Solomon, L. J., and Murakami, J. (1986). "Affective, cognitive, and behavioral differences between high and low procrastinators." *Journal of Counseling Psychology* 33.4, pp. 387–394. ISSN: 0022-0167 (cit. on p. 40).

Sariyanidi, E., Gunes, H., and Cavallaro, A. (2015). "Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37.6, pp. 1113–1133 (cit. on p. 60).

Schmidt, K. L., Ambadar, Z., Cohn, J. F., and Reed, L. I. (2006). "Movement differences between deliberate and spontaneous facial expressions: Zygo-

maticus major action in smiling." *Journal of Nonverbal* (cit. on pp. 17, 59, 92, 112).

Scott, D., Jung, C., Bins, J., Said, A., and Kalker, A. (2009). "Video Based VAD Using Adaptive Color Information." In: *Proceedings of 11th IEEE Internation Symposium on Multimedia (ISM '09)*, pp. 80–87 (cit. on p. 21).

Sénéchal, T., Turcot, J., and El Kaliouby, R. (2013). "Smile or smirk? Automatic detection of spontaneous asymmetric smiles to understand viewer experience." In: *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, pp. 1–8 (cit. on p. 61).

Shutler, J. D., Nixon, M. S., and Harris, C. J. (2000). "Statistical gait description via temporal moments." In: *Image Analysis and Interpretation, 2000. Proceedings. 4th IEEE Southwest Symposium*. IEEE, pp. 291–295 (cit. on p. 79).

Sidney, D., Craig, S., Gholson, B., Franklin, S., Picard, R. W., and Graesser, A. (2005). "Integrating affect sensors in an intelligent tutoring system." In: *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces*, pp. 7–13 (cit. on p. 41).

Siritanawan, P., Kotani, K., and Chen, F. (2014). "Independent Subspace of Dynamic Gabor Features for Facial Expression Classification." *Multimedia (ISM), 2014 IEEE International Symposium on*, pp. 47–54 (cit. on p. 57).

Sodoyer, D., Rivet, B., Girin, L., Savariaux, C., Schwartz, J.-L., and Jutten, C. (2009). "A study of lip movements during spontaneous dialog and its application to voice activity detection." *The Journal of the Acoustical Society of America* 125, p. 1184 (cit. on pp. 20, 21).

Sodoyer, D., Rivet, B., Girin, L., Schwartz, J.-L., and Jutten, C. (2006). "An analysis of visual speech information applied to voice activity detection." In: *Proceedings of 2006 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 1184–1196 (cit. on p. 19).

Sohn, J., Kim, N. S., and Sung, W. (1999). "A statistical model-based voice activity detection." *Signal Processing Letters, IEEE* 6.1, pp. 1–3 (cit. on p. 19).

Stekelenburg, J. J. and Vroomen, J. (2012). "Electrophysiological evidence for a multisensory speech-specific mode of perception." *Neuropsychologia* 50.7, pp. 1425–1431 (cit. on p. 19).

Suppes, P. (1966). "The Uses of Computers in Education." *Scientific American* 215.3, pp. 206–220 (cit. on p. 40).

Surakka, V. and Hietanen, J. K. (1998). "Facial and emotional reactions to Duchenne and non-Duchenne smiles." *International Journal of Psychophysiology* 29.1, pp. 23–33 (cit. on p. 60).

Szeliski, R. (2010). *Computer Vision* (cit. on p. 7).

Tao, D., Li, X., Wu, X., and Maybank, S. J. (2007). "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29.10, pp. 1700–1715 (cit. on p. 80).

Tian, Y.-L., Kanade, T., and Cohn, J. F. (2005). "Facial expression analysis." In: *Handbook of face recognition*. Springer, pp. 247–275 (cit. on p. 4).

Tiawongsombat, P., Jeong, M.-H., Yun, J.-S., You, B.-J., and Oh, S.-R. (2012). "Robust visual speakingness detection using bi-level HMM." *Pattern Recognition* 45.2, pp. 783–793 (cit. on p. 21).

Trutoiu, L. C., Hodgins, J. K., and Cohn, J. F. (2013). "The temporal connection between smiles and blinks." In: *2013 10th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2013)*. IEEE, pp. 1–6 (cit. on p. 61).

Valstar, M. F., Gunes, H., and Pantic, M. (2007). "How to distinguish posed from spontaneous smiles using geometric features." In: *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*. ACM Request Permissions (cit. on p. 61).

Valstar, M. F., Pantic, M., Ambadar, Z., and Cohn, J. F. (2006). "Spontaneous vs. posed facial behavior: automatic analysis of brow actions." In: *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*. ACM (cit. on p. 60).

Van den Stock, J., Righart, R., and De Gelder, B. (2007). "Body expressions influence recognition of emotions in the face and voice." *Emotion* 7.3, p. 487 (cit. on pp. 3, 4, 77).

Van der Maaten, L. and Hendriks, E. (2010). "Capturing Appearance Variation in Active Appearance Models." In: *Computer Vision and Pattern Recognition Workshops*. San Francisco, pp. 34–41 (cit. on pp. 17, 44, 50, 92, 112).

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). "Social signal processing: Survey of an emerging domain." *Image and Vision Computing* 27.12, pp. 1743–1759 (cit. on pp. 1, 3, 4, 111).

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schroeder, M. (2012). "Bridging the gap between social animal and unsocial machine: A survey of social signal processing." *IEEE Transactions on Affective Computing* 3.1, pp. 69–87 (cit. on p. 1).

Viola, P. and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features." In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Vol. 1, I-511-I-518 (cit. on pp. 4, 50).

Visser, M., Krahmer, E., and Swerts, M. (2014). "Contextual effects on surprise expressions: A developmental study." *Journal of Nonverbal Behavior* 38.4, pp. 523–547 (cit. on p. 42).

Von Helmholtz, H. (1924). *Treatise on physiological optics*. Vol 3. NY (cit. on p. 7).

Wallbott, H. G. (1998). "Bodily expression of emotion." *European journal of social psychology* 28.6, pp. 879–896 (cit. on p. 77).

Wang, L., Tan, T., Ning, H., and Hu, W. (2003). "Silhouette analysis-based gait recognition for human identification." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 25.12, pp. 1505–1518 (cit. on p. 84).

Wassenhove, V. van, Grant, K. W., and Poeppel, D. (2005). "Visual speech speeds up the neural processing of auditory speech." *Proceedings of the National Academy of Sciences of the United States of America* 102.4, pp. 1181–1186 (cit. on p. 19).

Wells, G. L. and Petty, R. E. (1980). "The effects of overt head movements on persuasion: Compatibility and incompatibility of responses." *Basic and Applied Social Psychology* 1.3, pp. 219–230 (cit. on p. 56).

Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., and Movellan, J. (2009). "Toward Practical Smile Detection." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31.11, pp. 2106–2111 (cit. on p. 62).

Whitehill, J., Bartlett, M., and Movellan, J. (2008). "Automatic facial expression recognition for intelligent tutoring systems." In: *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, pp. 1–6 (cit. on p. 41).

Wilting, J., Krahmer, E., and Swerts, M. (2006). "Real vs. acted emotional speech." In: *INTERSPEECH*. Vol. 2006, 9th (cit. on p. 3).

Wu, P., Liu, H., and Zhang, X. (2014). "Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors." In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pp. 1240–1244 (cit. on p. 61).

Wu, P.-p., Liu, H., Zhang, X.-w., and Gao, Y. (2017). "Spontaneous versus posed smile recognition via region-specific texture descriptor and geometric facial dynamics." *Frontiers of Information Technology & Electronic Engineering* 18.7, pp. 955–967 (cit. on p. 75).

Wu, T., Bartlett, M., and Movellan, J. R. (2010). "Facial expression recognition using Gabor motion energy filters." In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 42–47 (cit. on pp. 14, 22, 24, 36, 57, 91).

Xiong, X. and De la Torre, F. (2013). "Supervised Descent Method and its Applications to Face Alignment." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 67).

Yoo, J.-H., Nixon, M. S., and Harris, C. J. (2002). "Extracting Human Gait Signatures by Body Segment Properties." *SSIAI*, pp. 35–39 (cit. on p. 79).

Yu, S., Tan, D., and Tan, T. (2006). "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition." In: *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. IEEE, pp. 441–444 (cit. on pp. 17, 78, 82, 88, 93, 112).

Yu, S., Tan, T., Huang, K., Jia, K., and Wu, X. (2009). "A Study on Gait-Based Gender Classification." *IEEE Transactions on Image Processing ()* 18.8, pp. 1905–1910 (cit. on pp. 78, 81, 85–87, 89).

Zheng, S., Zhang, J., Huang, K., He, R., and Tan, T. (2011). "Robust view transformation model for gait recognition." In: *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, pp. 2073–2076 (cit. on p. 78).

Zorn, E. and Lokesh, R. (2010). *Motion filtering: frequency domain approach*. Tech. rep. ECE-2010-01. Department of Electrical and Computer Engineering, Boston University. URL: http://www.bu.edu/vip/files/pubs/reports/EZLR10-04buece.pdf (cit. on p. 51).

SUMMARY

Human communication is often comprised of more than just the spoken words. We can use a wide array of so-called social signals, short-spanned temporal sequences of non-verbal cues, that tells the receiver something about our attitude, mental state or other personal characteristics. Computers naturally lack these skills, however in some situations, human computer interaction would greatly benefit if they were able to interpret these signals automatically. Efforts towards equipping computers with “social skills" are channeled in the emerging field called Social Signal Processing (SSP) (Vinciarelli et al. 2009).

This dissertation took a systematic approach to examine if adding dynamic information to various Social Signal Processing tasks leads to an increase in performance, by evaluating the performance of each task using both static and dynamic information. In order to generate both types of information we used spatial Gabor filters (SGF) and spatiotemporal Gabor filters (STGF) to decompose visual signals in terms of shape and movement. Even though Gabor filters are widely used in SSP, the contribution of STGFs for these tasks has been little studied as of yet. In four experiments, exploring different social signals we sat out to investigate whether STGFs improve detection performance compared to SGFs.

The first signal we explored was visual speech, in the context of visual voice activity detection (VVAD) which is the task of determining whether a person is speaking or not, using only visual cues and not the auditory signal. VVAD has many applications ranging from speaker identification in multi-party discourse to audio detection in noisy environments. We relied on two different datasets, with different ratios between speech and non-speech: one is the publicly available CUAVE dataset (Patterson et al. 2002) with speakers uttering digits while being filmed both from the front and from the side (relatively much speech), and the other is the LIVER dataset (Joosten et al. 2012), in which participants utter a single word (“liver”, hence relatively little speech). We systematically compared dynamic, STGFs (an approach which we dubbed STem-VVAD) with their static, SGF counterparts, relying on the implementation of Petkov and Subramanian (2007), looking at the performance of filters at different speeds and applied to different levels of detail: zooming in on the mouth-region, the face or the whole clip. We found that the best results were obtained by STGFs (of all speeds combined) applied to the mouth region, which revealed a clear improvement over the SGFs applied to the same region. Even though these results are promising, the generalizability over different speakers was not optimal, suggesting that it is important to include speaker-characteristics for visual speech detection.

Visual voice activity detection is, of course, a very basic task (albeit one that was somewhat more difficult than we originally anticipated). Would STGFs also be beneficial for more subtle social signals, that are not inherently tied to a specific location in the face, such as the mouth? This was addressed in our second study, where we looked at detecting learning difficulties of children

based on facial expression analyses. Being able to detect whether or not young learners experience problems with, say, arithmetic assignments is important for, for example, the development of adaptive tutoring applications. For this study we relied on a dataset of children from two age groups, both solving easy and more difficult arithmetic problems, which we collected especially for this purpose. In this chapter, we again compared a dynamic approach, STGFs, with a static variant, SGFs, this time comparing the performance of two implementations, one due to Petkov and Subramanian (2007) and one to Heeger (1987). We also included a more explicit approach in our evaluation, which was based on a model of children's faces using Active Appearance Models (AAMs) (Cootes et al. 2001; Matthews and Baker, 2004; Van der Maaten and Hendriks, 2010). Our results revealed that, for this particular task, the explicit method based on AAMs clearly outperformed all Gabor approaches. More directly relevant for the topic of this thesis, however, we did find that the STGFs (in both implementations) outperformed the SGFs, although the relative improvement was relatively small.

The results of our second study suggest that when a social signal is very subtle and not associated with a specific facial area, detecting this signal is hard (although, as noted, STGFs still did somewhat better than SGFs). In our third we therefore looked at a signal that is both more subtle than visual voice activity detection, but more localized than learning problem assessment: the classification of smiles as either genuine or not. Being able to distinguish smiles that are caused by genuine happiness (Duchenne smile) from merely social (non-Duchenne) smiles is helpful for automatic emotion recognition systems, but has practical application as well, including for example, the ongoing development of digital photo camera's that automatically decide when a portrait picture would be optimally taken. It has been argued that the speed with which a smile appears on the face (with Duchenne smiles being slower) is a potentially important cue (Krumhuber et al. 2009; Schmidt et al. 2006). Therefore it offers an excellent opportunity to study the added value of dynamic information in STGFs for smile classification (again using both the aforementioned implementations), which we did based on the UvA-NEMO Smile database of spontaneous and posed smiles (Dibeklioğlu et al. 2015). Since head movements might have a profound effect on smile classification, we compared results for both 'raw' (unprocessed) faces and automatically 'fixed' ones. We found, once again, a benefit of dynamic filters; both STGF implementations clearly outperformed the corresponding SGFs, on every granularity and for both aligned and unaligned faces. Interestingly and somewhat unexpectedly, fixing the faces resulted in a drop in performance.

Finally, in our fourth we moved beyond facial signals, and studied the impact of Gabor filters on full body movements. In particular, we looked at an arguably basic, full-body task: gender classification based on a person's gait, which has potential practical applications, for example, for shops which want to automatically track the number of male and female shoppers inside. It has been shown that general bodily movement characteristics are helpful for this (Kozlowski and Cutting, 1977). In the final experimental study of this thesis we studied how Gabor filters perform on this task, comparing STGFs and SGFs. We used the CASIA Gait Dataset B (Yu et al. 2006), which is often used for comparing gait recognition methods, and also included a state-of-the-art

GEI-based system for the sake of comparison. We applied these systems both to frontal and sideways clips of people walking, at different levels of detail: the head as well as the upper and the lower body. We found that the results for the GEI-method were already rather good, but we did observe that the Gabor filter methods lead to even better gender classification rates, for both frontal and sideways recordings. Most importantly, we found, once again, that STGFs generally outperformed SGFs.

Based on the results of this thesis we can conclude that adding temporal information to spatial Gabor filters often improves the predictive quality of automated systems for social signal detection, especially in the cases where the informative visual cues are explicitly present in the facial or bodily areas.

DANKWOORD (ACKNOWLEDGMENTS)

Zo, en dan als laatste nog mijn dankwoord, het allerlaatste stukje tekst dat ik nog schrijf voor deze dissertatie. Lange tijd heb ik uitgekeken naar het schrijven van dit stuk, omdat dat zou betekenen dat mijn werk was goedgekeurd en dat ik binnenkort op zou mogen voor mijn verdediging. Ik zou op mijn gemak een rustig avondje uitkiezen, een goed glas wijn inschenken een fijne plek opzoeken en in een rits iedereen bedanken die mij geholpen en gesteund hebben om dit proefschrift af te ronden, die mij dierbaar zijn en altijd hebben geloofd/gehoopt dat het op een dag toch echt klaar zou zijn en die ik gewoon leuk vind om in mijn boek te noemen. Zo is het dus *niet* gegaan. Nu zou ik kunnen zeggen dat dat komt omdat ik het moeilijk vind dat mijn tijd als promovendus nu echt tot een einde komt en dat ik het schrijven van deze tekst daarom zo lang het uitgesteld, maar dat kan ik niet met een droog gezicht volhouden, al laat het afsluiten van deze mijlpaal me natuurlijk niet helemaal onberoerd. Ik kijk namelijk met ontzettend veel plezier terug op mijn tijd in Tilburg, waar ik enorme vrijheid had om mijn onderzoek zo uit te voeren als ik zelf wilde. Daarnaast denk ik ook graag terug aan alle zaken die horen bij het leven als Ph.D. student zoals het bedenken en uitvoeren van experimenten, conferenties bezoeken, studenten begeleiden en zo af en toe lesgeven, mijn tijd voor onderzoek in Australië en de omgang met alle collega's. Maar ja, dat schrijven... Dat ging niet altijd even soepel, zeker niet toen ik het afronden van mijn proefschrift moest combineren met een uitdagende baan. Waarschijnlijk daarom zit ik nu in Namibië tijdens een lange autorit pas dit stukje te schrijven. Een fijne plek uitzoeken is me tenminste wel gelukt. Nou hier komt ie dan.

Eric en Emiel, ik wil beginnen met jullie te bedanken voor de kans die jullie me gegeven hebben om dit promotietraject te mogen doen. Bedankt ook voor alle inspiratie, discussies, inzichten, hulp, mental coaching, koffietjes in Den Bosch, muziektips en jullie eindeloos geduld met mij (dat ik vast meer dan eens op de proef heb gesteld). Fijn dat jullie deur altijd voor me open stond. Hierdoor hervond ik elke keer nieuwe energie om weer verder te gaan en waardoor dit boek nu ook daadwerkelijk af is. Ik hoop dat we in de toekomst ook nog eens zullen samenwerken of gewoon af en toe eens zullen bijkletsen.

Mijn collega's in Tilburg, bedankt voor de fijne sfeer die ik altijd heb ervaren als ik naar mijn werk kwam. Bedankt voor de samenwerking, de hulpvaardigheid, de gezellig praatjes en de potlucks die we zo nu en dan hadden. Mandy, mijn lieve maatje in Australië, zonder jou was ik daar waarschijnlijk nooit naar toegegaan. Ik ben je voor altijd dankbaar dat je dat toen zo goed hebt doorgezet en dat ik daardoor nu tot de aanhang van jouw kookclub behoor. Bedankt ook voor het onvoorwaardelijk opbeuren wanneer ik weer eens dacht dat het allemaal niet zo lekker ging met mijn proefschrift, jouw vertrouwen in mij was altijd een duwtje in de rug. Mijn nieuwe collega's bij TNO, bedankt voor jullie grenzeloze enthousiasme dat jullie uitstralen in jullie werk, dat is voor mij zeer inspirerend. Judith, bedankt voor het

meedenken bij het combineren van het afronden van mijn proefschrift en werk bij TNO. Dankzij jouw flexibiliteit heb ik grote stappen kunnen zetten.

Mayo my Goose, je noemde jezelf de afgelopen periode weleens voor de grap de werkpolitie, maar zonder jouw schop onder mijn kont zou ik nu nog steeds zeggen dat het bijna klaar is. Nu is het dan echt zo ver en is het eindelijk Mayo en Bart fun time. Deze reis is daarvan al een goed begin. Ik verheug me op al onze komende avonturen waar die dan ook moge zijn. Bedankt lieverd dat je mij er zo goed doorheen hebt gesleept, best wingman ever!

Lieve papa en mama, jullie hebben mij tijdens mijn promotietijd altijd gesteund met een grenzeloos vertrouwen op een goede afloop, de heerlijkste borden zelfgemaakt eten en de lekkerste wijntjes uit de kelder als ik weer eens bij jullie thuis kwam na een lange dag schrijven. Bedankt voor jullie onvoorwaardelijke liefde en steun. Ik hoop dat we nog vaak mooie reizen met zijn allen zullen maken.

Lieke, lief zusje, jou wil ik bedanken omdat je “altijd je heerlijke zelf bent gebleven." Ik hoop dat we nog vaak samen om bizarre grappen zullen lachen en bedankt dat je altijd achter me staat.

Mats, bedankt voor jouw creativiteit, zonder jouw had ik niet zo gave cover op dit boek gehad.

De boykes (Arjen, Bas, Jef, Lex, Marvin, Menno, Rajen, Robin, Roel, Tristan), bedankt voor de boykes- weekenden/sinterklaasavonden, “puurtjes pakken", en alle avonden afleiding.

PUBLICATION LIST

JOURNAL PAPERS

Joosten, B., Postma, E., and Krahmer, E. (2015). "Voice activity detection based on facial movement." *Journal on Multimodal User Interfaces* 9.3, pp. 183–193 (cit. on pp. 70, 71).

Amelsvoort, M. van, Joosten, B., Krahmer, E., and Postma, E. (2013). "Using non-verbal cues to (automatically) assess children's performance difficulties with arithmetic problems." *Computers in Human Behavior* 29.3, pp. 654–664 (cit. on pp. 41, 43, 47, 48).

Liu, C., Ham, J., Postma, E., Midden, C., Joosten, B., and Goudbeek, M. (2013). "Representing Affective Facial Expressions for Robots and Embodied Conversational Agents by Facial Landmarks." *International Journal of Social Robotics* 5.4, pp. 619–626.

PAPERS IN CONFERENCE PROCEEDINGS (PEER REVIEWED)

Joosten, B., Postma, E., Krahmer, E., Swerts, M., and Kim, J. (2012). "Automated Measurement of Spontaneous Surprise." In: *Proceedings of Measuring Behavior*. Ed. by A. J. Spink, F. Grieco, O. E. Krips, L. W. S. Loijens, L. Noldus, and P. H. Zimmerman. Utrecht, pp. 385–389 (cit. on pp. 16, 22, 25, 91, 111).

Liu, C., Ham, J., Postma, E., Midden, C., Joosten, B., and Goudbeek, M. (2012). "How to Make a Robot Smile? Perception of Emotional Expressions from Digitally-Extracted Facial Landmark Configurations." In: *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings*. Ed. by S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, and M.-A. Williams. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 26–34.

Joosten, B., Amelsvoort, M. van, Krahmer, E., and Postma, E. (2011a). "Thin slices of head movements during problem solving reveal level of difficulty." In: *International Conference on Audio-Visual Speech Processing (AVSP 2011)*. Ed. by G. Salvi, J. Beskow, O. Engwall, and S. Al Moubayed. Stockholm, pp. 85–88.

ABSTRACTS OF CONFERENCE PRESENTATIONS (PEER REVIEWED)

Joosten, B., Postma, E., and Krahmer, E. (2014). *Visual Lip Reading Using Spatiotemporal Gabor Filters*. Poster presented at the 41th annual meeting of the Australasian Experimental Psychology Society. Brisbane, Australia, April 2014.

Joosten, B., Amelsvoort, M. van, Postma, E., and Krahmer, E. (2011b). *Thin slices of head movements during problem solving reveal level of difficulty*. Poster presented at the 23rd Benelux Conference on Artificial Intelligence. Ghent, November 2011.

Joosten, B., Postma, E., Krahmer, E., Swerts, M., and Kim, J. (2011c). *Automated Measurement of Spontaneous Surprise*. Talk presented at the 23rd Benelux Conference on Artificial Intelligence. Ghent, November 2011.

SIKS DISSERTATIONS

2011 01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models

02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems

04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference

05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage

07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction

08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues

09 Tim de Jong (OU), Contextualised Mobile Media for Learning

10 Bart Bogaert (UvT), Cloud Content Contention

11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective

12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining

13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling

14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets

15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval

16 Maarten Schadd (UM), Selective Search in Games of Different Complexity

17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness

18 Mark Ponsen (UM), Strategic Decision-Making in complex games

19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles

20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems

22 Junte Zhang (UVA), System Evaluation of Archival Description and Access
23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media
24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior
25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics
26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots
27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns
28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure
29 Faisal Kamiran (TUE), Discrimination-aware Classification
30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions
31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality
32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science
33 Tom van der Weide (UU), Arguing to Motivate Decisions
34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations
35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training
36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach
37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference
38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization
39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games
40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development
41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control
42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution
43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge
44 Boris Reuderink (UT), Robust Brain-Computer Interfaces
45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work
47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression
48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent
49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

---

2012 01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda
02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models
03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories
04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications
05 Marijn Plomp (UU), Maturing Interorganisational Information Systems
06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks
07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions
08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories
09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices
19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution

20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascall (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques

45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

---

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation

17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System
26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning
27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance
28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience
29 Iwan de Kok (UT), Listening Heads
30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support
31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications
32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development
33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere
34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search
35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction
36 Than Lam Hoang (TUe), Pattern Mining in Data Streams
37 Dirk Börner (OUN), Ambient Learning Displays
38 Eelco den Heijer (VU), Autonomous Evolutionary Art
39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems
40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games
41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning
42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning

43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

---

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data

02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method

03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions

04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation

05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability

06 Damian Tamburri (VU), Supporting Networked Software Development

07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior

08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints

09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language

10 Ivan Salvador Razo Zapata (VU), Service Value Networks

11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support

12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control

13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

14 Yangyang Shi (TUD), Language Models With Meta-information

15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare

16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria

17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability

18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations

19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support

20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link

21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments

22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33 Tesfa Tegegne (RUN), Service Discovery in eHealth
34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection

26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure

27 Sándor Héman (CWI), Updating compressed colomn stores

28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO

29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains

30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning

31 Yakup Koç (TUD), On the robustness of Power Grids

32 Jerome Gard (UL), Corporate Venture Management in SMEs

33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources

34 Victor de Graaf (UT), Gesocial Recommender Systems

35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

---

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VU), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UVA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UVA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TUE), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UVA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VU), Logics for causal inference under uncertainty

23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VU), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (UvT), From IT-Business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT

30 Wilma Latuny (UvT), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TUE), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julia S. Mollee (VU), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TUE), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future; Exploring the added value of computational models for increasing the use of renewable energy in the residential sector

14 Bart Joosten (UvT) Detecting Social Signals with Spatiotemporal Gabor Filters

TICC PH.D. SERIES

1. Pashiera Barkhuysen. *Audiovisual Prosody in Interaction.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.

2. Ben Torben-Nielsen. *Dendritic Morphology: Function Shapes Structure.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.

3. Hans Stol. *A Framework for Evidence-based Policy Making Using IT.* Promotor: H.J. van den Herik. Tilburg, 21 January 2009.

4. Jeroen Geertzen. *Dialogue Act Recognition and Prediction.* Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.

5. Sander Canisius. *Structured Prediction for Natural Language Processing.* Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.

6. Fritz Reul. *New Architectures in Computer Chess.* Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.

7. Laurens van der Maaten. *Feature Extraction from Visual Data.* Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).

8. Stephan Raaijmakers. *Multinomial Language Learning.* Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.

9. Igor Berezhnoy. *Digital Analysis of Paintings.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.

10. Toine Bogers. *Recommender Systems for Social Bookmarking.* Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.

11. Sander Bakkes. *Rapid Adaptation of Video Game AI.* Promotor: H.J. van den Herik. Co-promotor: P. Spronck. Tilburg, 3 March 2010.

12. Maria Mos. *Complex Lexical Items.* Promotor: A.P.J. van den Bosch. Co-promotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).

13. Marieke van Erp. *Accessing Natural History. Discoveries in data cleaning, structuring, and retrieval.* Promotor: A.P.J. van den Bosch. Co-promotor: P.K. Lendvai. Tilburg, 30 June 2010.

14. Edwin Commandeur. *Implicit Causality and Implicit Consequentiality in Language Comprehension.* Promotores: L.G.M. Noordman, W. Vonk. Co-promotor: R. Cozijn. Tilburg, 30 June 2010.

15. Bart Bogaert. *Cloud Content Contention.* Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.

16. Xiaoyu Mao. *Airport under Control.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotores: N. Roos, A. Salden. Tilburg, 25 May 2011.

17. Olga Petukhova. *Multidimensional Dialogue Modelling.* Promotor: H. Bunt. Tilburg, 1 September 2011.

18. Lisette Mol. *Language in the Hands.* Promotores: E.J. Krahmer, A.A. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (cum laude).

19. Herman Stehouwer. *Statistical Language Models for Alternative Sequence Selection.* Promotores: A.P.J. van den Bosch, H.J. van den Herik. Co-promotor: M.M. van Zaanen. Tilburg, 7 December 2011.

20. Terry Kakeeto-Aelen. *Relationship Marketing for SMEs in Uganda.* Promotores: J. Chr. van Dalen, H.J. van den Herik. Co-promotor: B.A. Van de Walle. Tilburg, 1 February 2012.

21. Suleman Shahid. *Fun & Face: Exploring non-verbal expressions of emotion during playful interactions.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.

22. Thijs Vis. *Intelligence, Politie en Veiligheidsdienst: Verenigbare Grootheden?* Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).

23. Nancy Pascall. *Engendering Technology Empowering Women.* Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.

24. Agus Gunawan. *Information Access for SMEs in Indonesia.* Promotor: H.J. van den Herik. Co-promotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.

25. Giel van Lankveld. *Quantifying Individual Player Differences.* Promotores: H.J. van den Herik, A.R. Arntz. Co-promotor: P. Spronck. Tilburg, 27 February 2013.

26. Sander Wubben. *Text-to-text Generation Using Monolingual Machine Translation.* Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.

27. Jeroen Janssens. *Outlier Selection and One-Class Classification.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.

28. Martijn Balsters. *Expression and Perception of Emotions: The Case of Depression, Sadness and Fear.* Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.

29. Lisanne van Weelden. *Metaphor in Good Shape.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 28 June 2013.

30. Ruud Koolen. *"Need I say More? On Overspecification in Definite Reference."* Promotores: E.J. Krahmer and M.G.J. Swerts. Tilburg, 20 September 2013.

31. J. Douglas Mastin. *Exploring Infant Engagement. Language Socialization and Vocabulary. Development: A Study of Rural and Urban Communities in Mozambique.* Promotor: A.A. Maes. Co-promotor: P.A. Vogt. Tilburg, 11 October 2013.

32. Philip C. Jackson. Jr. *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language.* Promotores: H.C. Bunt and W.P.M. Daelemans. Tilburg, 22 April 2014.

33. Jorrig Vogels. *Referential choices in language production: The Role of Accessibility.* Promotores: A.A. Maes and E.J. Krahmer. Tilburg, 23 April 2014.

34. Peter de Kock. *Anticipating Criminal Behaviour.* Promotores: H.J. van den Herik and J.C. Scholtes. Co-promotor: P. Spronck. Tilburg, 10 September 2014.

35. Constantijn Kaland. *Prosodic marking of semantic contrasts: do speakers adapt to addressees?* Promotores: M.G.J. Swerts and E.J. Krahmer. Tilburg, 1 October 2014.

36. Jasmina Marić. *Web Communities, Immigration and Social Capital.* Promotor: H.J. van den Herik. Co-promotores: R. Cozijn and M. Spotti. Tilburg, 18 November 2014.

37. Pauline Meesters. *Intelligent Blauw.* Promotores: H.J. van den Herik and T.A. de Roos. Tilburg, 1 December 2014.

38. Mandy Visser. *Better use your head. How people learn to signal emotions in social contexts.* Promotores: M.G.J. Swerts and E.J. Krahmer. Tilburg, 10 June 2015.

39. Sterling Hutchinson. *How symbolic and embodied representations work in concert.* Promotores: M.M. Louwerse and E.O. Postma. Tilburg, 30 June 2015.

40. Marieke Hoetjes. *Talking hands. Reference in speech, gesture and sign.* Promotores: E.J. Krahmer and M.G.J. Swerts. Tilburg, 7 October 2015.

41. Elisabeth Lubinga. *Stop HIV. Start talking? The effects of rhetorical figures in health messages on conversations among South African adolescents.* Promotores: A.A. Maes and C.J.M. Jansen. Tilburg, 16 October 2015.

42. Janet Bagorogoza. *Knowledge Management and High Performance. The Uganda Financial Institutions Models for HPO.* Promotores: H.J. van den Herik and B. van der Walle, Tilburg, 24 November 2015.

43. Hans Westerbeek. *Visual realism: Exploring effects on memory, language production, comprehension, and preference.* Promotores: A.A. Maes and M.G.J. Swerts. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 10 February 2016.

44. Matje van de Camp. *A link to the Past: Constructing Historical Social Networks from Unstructured Data.* Promotores: A.P.J. van den Bosch and E.O. Postma. Tilburg, 2 Maart 2016.

45. Annemarie Quispel. *Data for all: How professionals and non-professionals in design use and evaluate information visualizations.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 15 June 2016.

46. Rick Tillman. *Language Matters: The Influence of Language and Language Use on Cognition.* Promotores: M.M. Louwerse and E.O. Postma. Tilburg, 30 June 2016.

47. Ruud Mattheij. *The Eyes Have It.* Promoteres: E.O. Postma, H. J. Van den Herik and P.H.M. Spronck. Tilburg, 5 October 2016.

48. Marten Pijl. *Tracking of human motion over time.* Promotores: E. H. L. Aarts and M. M. Louwerse Co-promotor: J. H. M. Korst. Tilburg, 14 December 2016.

49. Yevgen Matusevych. *Learning constructions from bilingual exposure: Computational studies of argument structure acquisition.* Promotor: A.M. Backus. Co-promotor: A.Alishahi. Tilburg, 19 December 2016.

50. Karin van Nispen. *What can people with aphasia communicate with their hands? A study of representation techniques in pantomime and co-speech gesture.* Promotor: E.J. Krahmer. Co-promotor: M. van de Sandt-Koenderman. Tilburg, 19 December 2016.

51. Adriana Baltaretu. *Speaking of landmarks. How visual information inuences reference in spatial domains.* Promotores: A.A. Maes and E.J. Krahmer. Tilburg, 22 December 2016.

52. Mohamed Abbadi. *Casanova 2, a domain specific language for general game development.* Promotores: A.A. Maes, P.H.M. Spronck and A. Cortesi. Co-promotor: G. Maggiore. Tilburg, 10 March 2017.

53. Shoshannah Tekofsky. *You Are Who You Play You Are. Modelling Player Traits from Video Game Behavior.* Promotores: E.O. Postma and P.H.M. Spronck. Tilburg, 19 June 2017.

54. Adel Alhuraibi. *From IT-business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT.* Promotores: H.J. van den Herik and Prof. dr. B.A. Van de Walle. Co-promotor: Dr. S. Ankolekar. Tilburg, 26 September 2017.

55. Wilma Latuny. *The Power of Facial Expressions.* Promotores: E.O. Postma and H.J. van den Herik. Tilburg, 29 September 2017.

56. Sylvia Huwaë. *Different Cultures, Different Selves? Suppression of Emotions and Reactions to Transgressions across Cultures.* Promotores: E.J. Krahmer and J. Schaafsma. Tilburg, 11 October, 2017.

57. Mariana Serras Pereira. *A Multimodal Approach to Children's Deceptive Behavior.* Promotor: M. Swerts. Co-promotor: S. Shahid Tilburg, 10 January, 2018.

58. Emmelyn Croes. *Meeting Face-to-Face Online: The Effects of Video-Mediated Communication on Relationship Formation.* Promotores: E.J. Krahmer and M. Antheunis. Co-promotor A.P. Schouten. Tilburg, 28 March 2018.

59. Lieke van Maastricht. *Second Language Prosody: Intonation and Rhythm in Production and Perception.* Promotores: E.J. Krahmer and M. Swerts. Tilburg, 9 May 2018.

60. Nanne van Noord. *Learning visual representations of style.* Promotores: E.O. Postma and M. Louwerse. Tilburg, 16 May 2018.

61. Ingrid Masson Carro. *Handmade: On the Cognitive Origins of Gestural Representations.* Promotor: E.J. Krahmer. Co-promotor M.B. Goudbeek. Tilburg, 25 June 2018.

62. Bart Joosten. *Detecting Social Signals with Spatiotemporal Gabor Filters.* Promotores: E.O. Postma and E.J. Krahmer. Tilburg, 29 June 2018.