Advances in video motion analysis research for mature and emerging application areas

Advances in video motion analysis research for mature and emerging application areas

Adrienne Heinrich

The research described in this thesis has been carried out at Philips Group Innovation - Research, Eindhoven, The Netherlands, as part of the regular research programme.

The cover has been designed by Frank van Heesch & Adrienne Heinrich.

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-3913-0.

Printed on FSC-certified paper by Ipskamp Drukkers, Enschede, The Netherlands.

Advances in video motion analysis research for mature and emerging application areas

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens,
voor een commissie aangewezen door het College voor Promoties,
in het openbaar te verdedigen
op donderdag 10 september 2015 om 16:00 uur

door

Adrienne Heinrich

geboren te Zurich, Zwitserland

Dit proefschrift is goedgekeurd door de promotor en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr.ir. A.C.P.M. Backx
1^{e} promotor:	prof.dr.ir. G. de Haan
co-promotor:	dr.ir. R. Haakma (Philips Group Innovation)
leden:	prof.dr.ir. S. Van Huffel (KU Leuven)
	G.D. Clifford D.Phil. (Georgia Institute of Technology)
	prof.dr.ir. P.H.N. de With
	prof.dr. R.M. Aarts
adviseur:	dr. S. Overeem (Radboudumc)

To my parents, Fe and Manfred,
for their love, honesty and encouragement.

And to my husband, Bram,
for his trust and respect.

And to my sons, Thomas and Nils,
for redefining my life's ambitions.

Contents

Summary

Advances in video motion analysis research for mature and emerging application areas

This thesis aims at enhancing two human activities that together cover almost half of our daily routine, sleeping and watching TV. The research described in this thesis focuses on an area that potentially improves both of these otherwise rather disjunct activities. We address optimization of motion estimation algorithms for TV, and investigate the feasibility of using video motion analysis algorithms to extract valuable information about a person's sleep from video.

Motion estimation (ME) for TV picture rate conversion is a rather mature application area with several decades of on-going research efforts. A number of good ME methods exist and combinations of methods with a large number of parameters are not uncommon. Objective validation of ME methods and the influence of the many parameters that are involved have become more and more important. In this thesis, a methodology has been developed that can be used for the optimization of ME methods. At the same time, the developed methodology takes into account that objective measures cannot fully model the human perception. In a case study with hierarchical 3DRS, one of the state-of-the-art ME algorithms, we explored the extensive parameter space of 13000 motion estimators and provided insights with respect to the importance and the influence of the individual parameters. We found that the motion estimators optimized with the proposed validation scheme are superior to multiple existing techniques as well as standard 3DRS with regard to performance at a low computational complexity. Although the optimization methodology uses performance measures that do not capture the full complexity of human perception, still, a good correspondence with subjectively perceived picture quality is achieved. The conducted perception test confirmed that the components of the proposed methodology are well chosen and yield motion estimators with a good picture rate conversion performance.

Analyzing movements during sleep can provide a wealth of information as body movements can be associated to sleep states and sleep state transitions. Traditional sleep screening is performed in sleep clinics with polysomnography (PSG) studies,

in which a person's sleep is analyzed by a myriad of different on-body sensors (e.g., EEG, EMG, ECG). PSG is considered the gold standard for sleep screening, yet, the PSG measurements are uncomfortable, disturb the natural sleeping behavior and therefore lack reproducibility, and require often time consuming manual analysis. To alleviate these shortcomings, we investigate a new, fully automated, and less invasive monitoring approach. We focus on a video camera-based system for sleep analysis, consisting of a near infrared (NIR) camera and NIR light source. We present methods to extract activity levels, sleep efficiency scores, breathing information, body part movements, infant sleeping pose, and wake-up behavior. Challenging conditions such as a shared bed environment, different camera locations and moving cast shadows are taken into account.

We designed a contactless, off-body video actigraphy system to monitor a sleeping subject's movements. With the aim to analyze competitiveness with wrist actigraphy, we conducted a differentiated comparison between the two actigraphy methods. Video actigraphy contains more comprehensive information and is generally more sensitive than wrist actigraphy. The average PSG to video based sleep efficiency error is comparable to the PSG to wrist actigraphy based error.

In order to discriminate movements of different body parts, we investigated an enhanced K-Means clustering approach for motion vectors. When performing ME on sleep sequences, large environmental variations between recording situations such as viewing angles, blanket types, zoom factors and illumination conditions, can yield different motion vector fields for similar movements. Therefore, our multi-distance clustering algorithm is computing content-dependent weights and is not only based on spatial distances between data points but also on motion vector angle and length.

To realize an easy-to-install system for the end user, we investigated an installation where the camera is conveniently placed on the bedside table of the primary subject who is to be monitored. We designed a method sensitive to small scale movements so that not only activity levels are monitored but also the respiratory waveform can be computed. Our breathing analysis method performed admirably with an overall sensitivity of 87%, precision of 90%, and a breathing rate correspondence of 93%, surpassing the results of state-of-the-art video based breathing algorithms.

To discriminate small movements of a subject from moving cast shadows on a non-planar and dynamic background, our video processing method integrates motion detection, motion estimation and texture analysis, efficiently aggregated in a strong classifier using cascaded AdaBoost. Movement event classification improved threefold with the proposed method in highly varying lighting conditions, compared to state-of-the-art.

The first lifestyle application we investigated illustrates an intelligent baby monitor that warns parents when their baby is turning in its sleep to its belly. By designing a turning movement detector and combining its information with face detection results, we improved the infant's sleeping pose accuracy by 11% compared to a method using

solely face detection.

A personalized wake-up system is envisioned in the second lifestyle application that exposes the sleeping subject to light that is adapting its intensity over time according to the subject's measured activity level. Therefore, we developed a system which can measure the sleeping person's activity and control the light output such that the subject's behavior corresponds to an activity trajectory of a favorable wake-up experience.

Chapter 1

Introduction

The research described in this dissertation leads to the enhancement of two human activities that together cover almost half of our daily routine, sleeping and watching TV. According to the American Time Use Survey 2013 [1], the average time spent watching television in the U.S. amounts to almost three hours a day. Watching TV is our number one leisure activity in terms of an adult's occupied time. The National Sleep Foundation states an average sleep need for an adult of seven to nine hours a day [2]. This thesis describes new methods in video motion analysis that can improve both of these otherwise rather disjunct activities. On the one hand, improvements in video motion estimation can lead to improved TV picture quality. On the other hand, the analysis of motion in videos of sleeping subjects can help in assessing our sleep.

For TV picture quality, we present research on optimizing motion estimation methods for picture rate conversion. In the optimization process we cannot fully rely on an objective performance measure. Up to now, no objective criterion has been accepted in the literature as a standard metric for evaluating motion estimation methods. Multiple performance measures exist but the correlation between the subjectively perceived video quality and the objective scores is poor. Therefore, we researched how to compare different motion estimators even with inferior performance measures. For video analysis of sleep, we investigated the properties and effectiveness of a similar motion estimation technique as for TV for body movement estimation. We explored how the obtained motion fields can serve as a foundation for attaining sleep-relevant information. In this thesis, we present methods to extract activity levels, sleep efficiency scores, breathing information, body part movements, infant sleeping pose, and wake-up behavior. Challenging conditions such as a shared bed environment, different camera locations and moving cast shadows are taken into account.

Hence, in both research areas, TV picture enhancement and sleep analysis from video, capturing movement information in video streams is essential. While addressing optimization of motion estimation algorithms for the mature area (TV), feasibility of video motion analysis algorithms for the emerging research area (Sleep) still needs

to be investigated. In the following sections, these two research areas and their corresponding video processing and analysis challenges are described, and an outline of this thesis is given.

1.1 Picture rate conversion for TV

Rendering good picture quality remains a challenge for video processing algorithms in television sets. Therefore, research is still on-going in areas such as video coding, increasing the perceived sharpness of details, high dynamic range and contrast, wide color gamut rendition, improved picture rate conversion, and 3D rendering [3–7].

Before the 1990s, television sets would repeat incoming frames to accommodate for a higher frame rate. This repetition results in perceived judder and blur. Since the 1990s, the role of motion estimation (ME) for picture rate conversion has become more important. Current TV sets show video signals with several hundreds of frames per second (400 Hz, 600 Hz, or even 800 Hz [8]), whereas the input is often a mere 24 frames per second (fps). This means that per second hundreds of images have to be generated by the television set. Knowing where and what kind of motion takes place in the video stream can help in rendering moving areas sharper (with less motion blur) and allows for a smooth motion portrayal by using the estimated motion information in the picture rate conversion process [9]. For each entity in the image (e.g., a pixel, a block consisting of several pixels, an object) the displacement is defined by a motion vector across which the entity has shifted with regard to the original image(s). By doing this for the entire image, a so-called motion field is returned, i.e., a matrix with entries corresponding to motion vectors of image pixels or blocks. Extracting the motion from video images as closely as possible to the actual motion captured [10], can thus aid applications such as picture rate conversion [11], 2D-to-3D video conversion, and structure from motion. In 2D-to-3D conversion, different cues are used to compute the depth map from a 2D video, of which depth from motion is an essential cue [7]. Similarly, in structure from motion, 3D structures of an object are reconstructed based on the computed motion information of an initial 2D video stream [12]. When the displacement in subsequent images is correctly represented by the motion vectors, they are considered as matching the so-called 'true' motion [10].

Motion estimation (ME) for TV picture rate conversion is a rather mature application area with several decades of on-going research efforts [9, 11, 13, 14]. Nevertheless, because of the increasing spatial resolution (from SD to HD, Full HD and Ultra HD) and picture rates (from 24 fps to more than 800 fps) of video shown on television sets, as well as the increasing size and quality of the television displays, there is continuous pressure to improve the quality of ME algorithms while maintaining acceptable computational complexity. To this end, spatio-temporal prediction methods such as recursive search, e.g., [9, 11, 13], are typically applied in practice (e.g., [15, 16]). A commonly used spatio-temporal prediction method is 3-Dimensional Recursive Search (3DRS) [14]. An output motion vector is selected from a candidate vector set based on the minimization of an energy function (e.g., Sum of Absolute

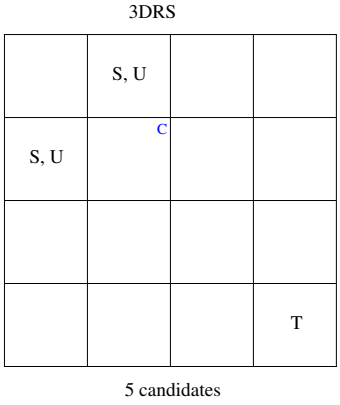


Figure 1.1: 3DRS candidate structure. C denotes the current block for which candidate motion vectors are determined, S a spatial candidate, U a random update vector added to the spatial candidate, and T a temporal candidate.

Differences). The candidate vector set (see Fig. 1.1) consists of prediction vectors from a spatio-temporal neighborhood. Previous estimates are found to be good predictions under the assumption that objects are larger than blocks and that objects have inertia [14]. Additionally, random values are added to the spatial candidates, forming so called update candidates, that help in finding vectors for appearing objects and accommodate for acceleration. Such a motion vector computation is performed sequentially for each location in the image. Generally, spatio-temporal predictors have proven to be a powerful tool in the design of ME algorithms [17–19].

As one of the objectives of this thesis, we investigate the extension of 3DRS ME with the concept of hierarchy [20]. Combinations of 3DRS with concepts borrowed from alternative ME methods have shown to be beneficial in earlier publications, e.g., [21]. Incorporating a hierarchical component to ME methods has appeared to be of advantage [11]. Various realizations of the combined hierarchical 3DRS ME can be thought of (e.g., multi-scale vs. multi-grid ME, different scaling factors and block size

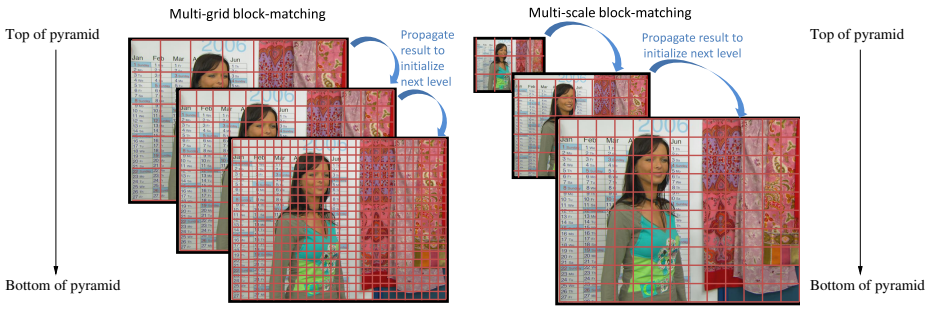


Figure 1.2: Illustration of hierarchical multi-grid and multi-scale approach.

parameters, different candidate structures). For a hierarchical 3DRS ME, an additional candidate vector is added from an 'external' source, i.e., a *hierarchical candidate*. The hierarchical candidate vector can be obtained by both multi-grid (same resolution, multiple block-sizes) and multi-scale (multiple resolution levels) approaches. In order to describe the multi-scale or multi-grid approach, a scale pyramid is used, as shown in Fig. 1.2, where ME is performed on higher scales at the top of the pyramid first and motion vectors are propagated down the pyramid to the lower scales by means of hierarchical candidates.

A large number of ME methods exist and combinations of methods with a large number of parameters are not uncommon. Objective validation of ME methods and the influence of the many parameters that are involved is important [9, 11, 14, 22–25]. Objective measures are proposed in [9, 11, 14, 22–25] to evaluate the ME performance. These methods all strive to best reflect the subjective image quality. Up to now, no objective criterion has been accepted in the literature as a standard metric for evaluating motion estimation methods [10]. In this thesis, we describe a methodology that can be used for the optimization of ME methods with a vast number of parameters. At the same time, the developed methodology takes into account that objective measures cannot fully model the human perception.

1.2 Sleep monitoring with a camera as off-body sensor

Our sleep consists of cycles where in each cycle, we pass through several sleep stages or states, such as light sleep, deep sleep, and REM sleep. Traditional sleep screening and sleep classification are performed in sleep clinics with polysomnography (PSG) studies, in which a person's sleep is analyzed by a myriad of different measuring techniques, such as EEG (electroencephalogram), EMG (electromyogram), ECG (electrocardiogram), EOG (electrooculogram), and respiratory effort belts. PSG is considered the gold standard for sleep screening. Reliability of these measurements suffers from the so-called first night effect when sleeping in a laboratory environment with multiple sensors attached to the head and body instead of the familiar home setting and has negative impact on a person's sleep. This makes an accurate diagnosis more difficult. Additionally, sleep clinics tend to have long waiting lists (weeks to months are not uncommon). The measured signals are usually manually analyzed which is time consuming and cost intensive. Therefore, the current PSG procedure for sleep screening is limited to a one or at most two-night examination.

Although PSG measurements in the sleep clinic are considered the gold standard for sleep screening, more information that is typically obtained from video recordings may aid in making the correct diagnosis. However, differential diagnosis is not always straightforward [26], as is the case for e.g., nocturnal frontal lobe epilepsy and parasomnias, one of the major categories of sleep disorders with abnormal behaviors or experiences [27]. In both disorder types, movement behaviors during sleep are

Figure 1.3: Illustration of a sleep monitoring system using NIR LEDs and a NIR sensitive camera.

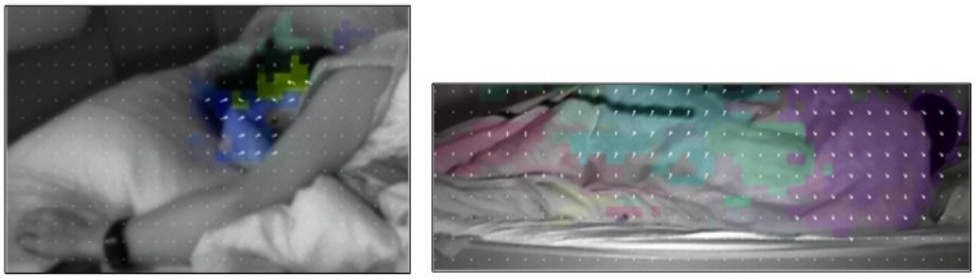


Figure 1.4: Contactless video actigraphy monitoring enables local motion analysis. The motion vectors from the processed video images are color-coded based on the direction and intensity of the motion. Left image: Blue area indicates hand moving to the right, green area indicates head moving to the left. Right image: The different colors indicate different motion of the corresponding body parts.

often dramatic and bizarre [28]. Patient and witness reports may be limited due to the nocturnal occurrences [28, 29]. Therefore, besides the clinical interview, video-polysomnography recordings are suggested to observe the nocturnal behavioral events [30, 31]. In [32], the combination of PSG and manual video analysis has shown to enhance the diagnostic ability for disorders with motor activities. An increase in the positive diagnosis of REM sleep behavior disorder from 80% to 95% of patients has been achieved. The benefit of this manually performed video analysis is similarly recognized in [33], where different movement characteristics are observed in REM sleep (short, jerky movements) than in NREM sleep. The video-polysomnography approach is also found promising regarding sleep-disordered breathing (e.g., obstructive sleep apnea) as there seems to be a relationship between sleep disturbances and body and head position [34, 35].

PSG monitoring is not viable for long-term home monitoring, which in turn may be important to assess problems or disorders which are not clear after a one or two-night assessment. This holds for example for parasomnias where the indicative movement episodes do not necessarily occur on a nightly basis and where complex behaviors and enacted dreams are mostly observed at home [26, 36]. With the goal to overcome the disadvantages of sleeping in a laboratory environment, having to wait for a free bed in a sleep clinic, high cost due to night-shift staff and facility costs, a home PSG system has been proposed in [37]. Although this procedure is better suited for monitoring e.g. a few consecutive nights, and reduces the need and cost for a sleep laboratory and personnel, it is still not a good option for long-term monitoring since on-body sensors are needed and the recorded data has to be manually interpreted by qualified sleep clinicians. In addition, regular consumers are becoming more interested in applications for sleep monitoring or sleep enhancement. It is clear that for this group, fewer and more comfortable sensors are preferred. A monitoring system that comes with comfortable and convenient sensors, with the possibility to perform long-term monitoring and to automatically analyze the recorded data would be appreciated. With the automatic analysis of the recorded data, there is no need anymore for the time consuming manual analysis.

Therefore, quite some research efforts have been carried out in the past decade to monitor sleeping subjects automatically for several nights at home, in the natural sleeping environment, while employing sensors that offer more comfort. These are sensors that can either be installed easily in the bedroom or worn on the body. In the former group, research has been carried out for pressure sensors in the pillow [38], air filled tubes combined with a differential pressure sensor [39] under a 4 cm mattress, piezoelectric sensors, strain-gauge and electret foil sensors [40], microphones [41], and near infrared cameras [42], amongst others. The body-worn sensors include a wrist-worn accelerometer or so-called wrist actigraphy [43], skin conductance measured at the wrist [44], optical heart rate sensors [45] based on photoplethysmography worn on the wrist, combinations of several sensors in one wrist device, e.g., the Empatica E4 wristband [46] measuring optical heart rate, galvanic skin response, skin temperature and 3-axis acceleration. A recent publication from MIT on skin conduction measured

at the wrist offers some information on the possibilities of sleep staging using this technique. Advances on dry electrodes have been used by ZEO to develop a headband with integrated dry electrodes placed at the forehead. The produced sleep staging for REM, light and deep sleep have shown to be acceptable, while inferior in detecting the waking periods. By adding features based on respiration to the already existing actigraphy feature set, sleep state classification could be improved according to [47–49].

Among these sensors, wrist actigraphy has become an essential tool in sleep research and sleep medicine [43]. Body movements are an important behavioral aspect during sleep as shown in [50]. They can be associated to sleep states [51] and sleep state transitions [52]. It was concluded in [53] that frequency and duration of body movements are important characteristics for sleep analysis.

A promising sensor gaining more attention in the recent years is the video camera [42, 54, 55]. Attempts have been made to use the video camera and novel video processing algorithms for movement detection [54], sleep/wake classification [42], sleep/wake classification under varying global lighting conditions [55] and sleep breathing disorder detection [56, 57] leading to this newly emerging application area.

With regard to the sleep research in this thesis, we investigate a video camera-based system for sleep analysis for two main reasons. This remote system is rather unobtrusive without disturbing a person's sleep and we can profit from advanced video processing algorithms developed for more mature video application areas (such as TV). Such a monitoring setup with a near infrared (NIR) camera and NIR light source is illustrated in Fig. 1.3. We analyze body movements of sleeping subjects by adapting ME methods originally developed for the TV application. We developed algorithms based on motion estimation to compute sleep-relevant information. Activity levels and body movements are estimated (see e.g. Fig. 1.4). Motion vector fields resulting from ME methods are further analyzed and processed to comprehend whether they support i) segmenting the person of interest in a shared bed scenario with subject occlusion, ii) assigning motion detection results to true subject movement instead of varying illumination effects, and iii) understanding types of movements such as turning movement when monitoring e.g. the sleeping pose of a baby. Aiding the long-term goals of sleep breathing disorder detection and detailed sleep staging, efforts in small-movement analysis have been undertaken for extracting breathing characteristics.

One of the aspects of this thesis aims at exploring new lifestyle applications by incorporating movement analysis of sleeping subjects. This has been done for an intelligent baby monitor and a personalized wake-up light system where the knowledge of the wake-up behavior over time is used to improve the wake-up experience.

1.3 Problem description and contributions of this thesis

In this section, we formulate the research questions for each of the eight research topics discussed in this thesis. We begin each research topic with the context leading to the corresponding research question (RQ), share evaluation aspects and end with how we investigated the corresponding research question.

Optimization of motion estimators for picture rate conversion

For the mature application area of ME for TV picture enhancement, we are challenged by the extensive parameter space and the lacking means to analyze and validate parameter settings reliably. This is because there are many types of motion estimators (MEs) and, within each type, many parameters, which makes a subjective assessment of alternatives impractical. The parameters can range from a set of values (e.g., block sizes, scaling factors, number of passes) to structures (e.g., motion vector candidate structures) up to sub-methods as part of the ME method (e.g., 2-frame ME vs. 3-frame ME, matching criteria). What complicates the problem even further is that the interaction of the parameters with each other is difficult to measure due to the large amount of possible parameter combinations. In the literature, typically one ME type is selected and proposed, within which some parameters are chosen (typically sub-method parameters) and others are manually optimized (typically value parameters). This is observed in [11] where a multi-grid approach is proposed starting with coarse motion vector blocks (the parameter for the block size is set to 32 $\times$ 32 pixels) followed by a second motion estimation on the same resolution with finer motion vector blocks (the parameter for the finest block size is set to 4 $\times$ 4 pixels). It is documented in [11] that 12 passes are assumed to yield a good result, but this has not resulted from a quantitative analysis. A multi-scale ME is proposed in [58] where the scale factors and corresponding block sizes are empirically determined. Different candidate structures are compared in [22]. Similarly, the effect of different regularization terms is analyzed in order to enforce smooth motion fields [22, 59]. The commonly used manual parameter optimization is unlikely to arrive at an optimum due to the vast size of the optimization space. This brings us to the following research question:

RQ1: *How can a large range of parameter values, parameter types, and interaction of different parameter choices be taken into account in the optimization of motion estimators for picture rate conversion?*

We have not come across a large range parameter optimization in the literature. On a reduced scale, different ME parameters and parameter settings are compared in [22]. The quantitative evaluation of the resulting motion estimators is performed by comparing the ME scores in an M2SE-SI performance space. The M2SE assesses the

prediction accuracy of a motion estimator and the SI the motion field consistency [14, 22]. Other commonly used performance measures comparing different motion estimators or parameters of motion estimators are PSNR [23, 24], TEMC-MSE [9, 11], SSIM [25], and SAMND [9, 11, 25]. Visual comparison of the resulting motion fields with a ground truth motion field is employed in [59]. The ground truth motion vectors are generated with a modified ray tracer. In [60], the angular error and the magnitude of difference with regard to the ground truth motion field are computed for comparing different optical flow motion estimators. By using the ground truth test sequences from the Middlebury University [61], the endpoint error according to the Middlebury benchmark is calculated for the different motion vector fields in [59]. Some ME parameters may alter the computational complexity significantly. Therefore, also the computational complexity is often measured [24, 62].

We have explored the extensive parameter space of hierarchical 3DRS (H3DRS) in an automatic manner and present an analysis of the importance and influence of the various parameters for the application of picture rate conversion. Among the 13000 different motion estimators, the parameter settings have been automatically compared and optimized in order to render superior motion estimators to multiple existing techniques as well as standard 3DRS with regard to performance at a low computational complexity.

Perception-oriented methodology for robust motion estimation design

For the application of picture rate conversion, various objective performance measures have been developed [9, 11, 14, 22–25], such as PSNR [23, 24], M2SE [14, 22] and TEMC-MSE [9, 11] addressing the prediction accuracy, SSIM [25] measuring the similarity between two images, and SI [14, 22] and SAMND [9, 11, 25] measuring the motion field consistency. They have been employed to evaluate the performance of a ME. However, these performance measures do not satisfactorily reflect the perceived subjective image quality. In [63], the objective scores are even found to correlate poorly with the subjectively perceived video quality. Up to now, no objective criterion has been accepted in the literature as a standard metric for evaluating ME methods [10]. Due to the insufficient objective criteria, assessment of a new ME method is typically not only done objectively, but also subjectively on a few test sequences [11, 25, 63]. This leads to the following research question:

RQ2: *Can high quality robust motion estimators be identified while applying metrics with limited validity?*

The evaluation aspects are similar to those of research question RQ1. To evaluate the performance of motion estimators, both quantitative [9, 11, 14, 22–25] and qualitative [11, 25, 63] assessments are usually conducted.

We present an automatic ME design methodology that can deal with performance measures that cannot fully model the human perception. A user study is conducted and indicates that applying this methodology leads to subjectively pleasing upconverted videos despite the inferior objective performance measures. Additionally, the impact and significance of the two chosen performance measures are analyzed and compared.

Body movement analysis during sleep based on video motion estimation

A convenient and quite comfortable solution for monitoring sleep at home for multiple nights is wrist actigraphy [43]. This technology has been used for over 30 years to study sleep/wake patterns [43]. Another interesting modality is the use of a video camera. By analyzing body movements of sleeping subjects recorded with a NIR camera, subjects can sleep in their own familiar bedroom without being disturbed by on-body sensors. Movements of sleeping subjects can be analyzed by performing ME on the recorded video stream (e.g., [54, 64–67]). These computed motion vectors are required to represent movements during sleep, also in low-textured areas such as the blanket. This is one of the key differences between sleep video streams and TV video streams. Texture and edges help in computing truly representative motion with a ME as these areas generally return a better match for the correct motion vector compared to other motion vectors. However, several different motion vectors may seem applicable in homogeneous areas. In these areas, all image blocks look similar and the match criterion or minimization function is thus satisfied for different motion vectors. Different types of textures, strong edges and high contrast areas are typically present in TV picture streams since these make content more appealing to watch. In near infrared monochrome video streams of the bedroom environment, texture, contrast, and dynamic range can be very different from regular TV picture material. Texture differences that may be clearly visible in daylight may be greatly reduced in the monochrome near infrared images (see Fig. 1.5). In order to deal with none or low-textured objects, including less prominent folds in beddings and blankets, the TV ME has to be redesigned. Most common video techniques performing ME for sleep analysis are based on optical flow. Lucas-Kanade [68] and hierarchical Lucas-Kanade using pyramids (e.g., [69]) produce rather noisy vector fields on sleep sequences. Horn-Schunck optical flow ME [70] and hierarchical versions thereof (e.g., [71, 72]) are favored by various researchers [54, 64–67]. Compared to Lucas-Kanade, more consistent motion fields are computed but there are some difficulties in dealing with small movements in very low light conditions. Optical flow methods that produce a dense motion field are computationally quite intense as they perform pixel-based ME. Advances are marked by recent optical flow algorithms (e.g., [73–75]) with a relatively low computational complexity and promising performance on the Middlebury benchmark [61].

Video techniques that compute activity levels of a sleeping subject are either based on straightforward frame differencing [42, 76, 76–78] or computationally more

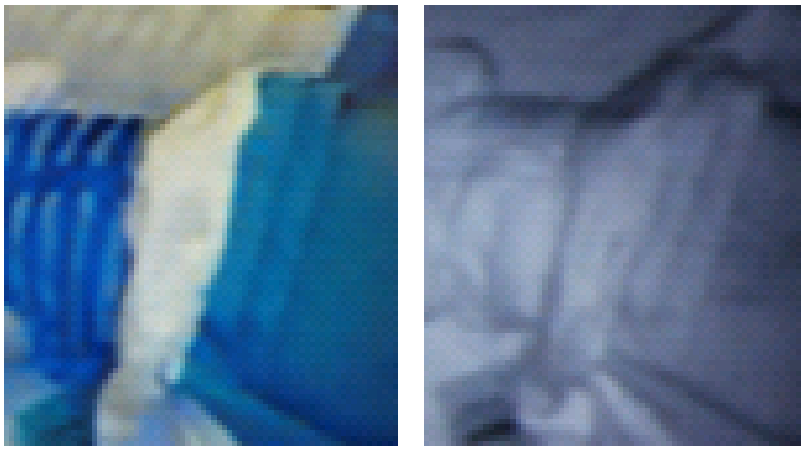


Figure 1.5: Loss of texture when the same scene is captured with an infrared camera (right) compared to a visible light camera (left).

intense optical flow methods [54, 64–67]. Motion vectors derived from optical flow methods are reduced to activity levels, losing the extra directional information. Movement pattern analysis with motion vectors has been explored in [42], but not further applied for sleep analysis. Given the advances of video-based ME techniques and the possibility to record movements originating from the entire body, we address the following research question:

RQ3: *How does video actigraphy compare to wrist actigraphy and is it more sensitive to movements originating from body parts other than the wrist?*

PSG is considered the gold standard for sleep assessment [79] with sleep efficiency being one of the measures derived from the PSG data [80]. Typically, the output of proposed actigraphy methods is compared with PSG [42, 55]. In [42], a video-based sleep monitoring technique is proposed and compared with both PSG and wrist actigraphy. The accuracy of the sleep/wake classifications is computed. Liao and Yang report an 8% performance difference between video and PSG and a 1% performance difference between wrist actigraphy and video actigraphy. We are not aware of a differentiated comparison of video and wrist actigraphy on different movement categories (small, medium, and large) obtained from several nights of sleep. The Bland-Altman analysis is a standard way to measure agreement between two methods applied to the same subjects and also to compare a new technique with the gold standard directly [81].

The candidate structure of the 3DRS motion estimator used for TV applications

has been adapted to better capture the movements of a sleeping person. A video actigraphy method is designed based on video ME and compared with wrist actigraphy in a Bland-Altman analysis. Advantages in full body movement monitoring are observed and similar sleep efficiency values are achieved between video and wrist actigraphy.

Multi-distance motion vector clustering for video-based sleep analysis

When performing ME on sleep sequences, large environmental variations between recording situations such as viewing angles, blanket types, zoom factors and illumination conditions, can yield different motion vector fields for similar movements. For video actigraphy, motion vector direction is a secondary characteristic and is mostly not used to derive an activity level (e.g., [54, 66]). Assuming that image blocks belonging to one body part move with similar speed in similar directions, motion clustering could become interesting in case the clusters coincide with body parts or sub-parts. Mainly body part segmentation and person segmentation could profit from these motion clusters. For some sleep disorders, knowledge on which body part is moving is essential. This is recognized in general by [78] and in particular for the disorder of periodic limb movements by [82–84]. Computation of local motion is important because it allows us to study movement characteristics of different body parts [42]. Additionally, motion vector clusters aid in drawing the boundary between two people lying closely together in the same bed [85]. We use the popular K-Means clustering method [86]. It is simple and straightforward [87], allowing it to run on overnight sleep sequences. The core idea was developed almost half a century ago and is successfully employed nowadays (e.g., [88]), also for motion vector clustering [89]. In K-Means implementations, random seeds are often selected as starting points [87]. Studies conducted by [90] and [91] improve the selection of the seeds based on the spatial distribution of the data but do not take any data-inherent characteristics into account. Weight factors for the different cues may vary depending on the properties (e.g., zoom factor, lighting conditions, viewing angles) of the recorded video sequence. The angle difference between motion vectors can be a much stronger differentiator than the length difference in one viewing angle, whereas the opposite may hold in another viewing angle. In an existing method [92] addressing the challenge of image retrieval, the weight of each descriptor or cue is automatically computed based on a dissimilarity histogram and a capacity graph. This brings us to the following research question:

RQ4: *What are meaningful descriptors for motion vector clustering given the large variety of motion vector field properties produced by the varying recording conditions in sleep monitoring?*

To our best knowledge, there is no review available on the validation of motion vector clustering algorithms for sleep analysis. A motion vector clustering method

is used in the area of meteorology in [93] where it forms one module of the entire system. The performance of the entire system is then evaluated, and the improved results used as a reflection on the clustering algorithm. In [94, 95], motion vector clusters are compared with manual ground truth assignments. Feasibility of an implemented motion vector clustering method is shown with processed images in [89].

We developed a content-dependent clustering method to cluster movements originating from one body part. An enhanced K-Means clustering approach is investigated for motion vectors. Our multi-distance clustering algorithm is not only based on spatial distances between data points but also on differences in motion vector angle and length.

Actigraphy and breathing monitoring of a shared bed from the bedside table

In [42, 55, 96, 97], the camera is mounted high up on the wall or ceiling overlooking the bed. On the one hand, this may yield video data that is easier to process, on the other hand, it does not pose an easy-to-install home system for the end user. Sensor system placement on the bedside table is realized as an alternative in [98]. Movements of sleeping subjects are typically analyzed with the aim to perform sleep/wake classification [43] or to screen for diseases characterized by particular movement patterns [83]. Accurate sleep state classification is more challenging with the comfortable, convenient and easy-to-install sensors offered for the home environment, such as wrist actigraphy [43], radar sensors [99], pressure sensors in the pillow [38], pressure sensors in the bed sheet [100], air filled tubes combined with a differential pressure sensor [39] under a 4 cm mattress, piezoelectric sensors, strain-gauge and electret foil sensors [40]. This is due to the reduced set of physiological signals that can be measured. A more detailed and improved sleep state classification is accomplished by adding respiration features to the set of actigraphy features as shown by [47] and [48]. According to a survey by [101], 62% of the respondents report to sleep with a bed partner. For the camera system in the home setting it is particularly challenging to derive the physiological characteristics of the Primary Subject (PS) in the presence of a bed partner as both subjects are recorded in the image and some of their body parts may lie in close proximity of each other. A side viewing angle makes it even more challenging to distinguish between movements from the bed partner and the PS as they are overlapping in the camera image. Only movement information originating from the PS should contribute to the activity level. The Eulerian Video Magnification approach [102] amplifies breathing motion in the video images for improved visualization of the subtle motion. No breathing parameters are automatically derived from the processed images. Taking these observations into account, we are interested in the following research question:

RQ5: *Is a user-convenient camera placement on the bedside table acceptable for measuring actigraphy and breathing of a sleeping subject in a shared bed?*

Body movement detection in sleep with video analysis methods is compared to manually labeled ground truth in terms of sensitivity and precision in [54], and is compared to wrist actigraphy in terms of accuracy in sleep/wake classification in [42, 55]. To evaluate breathing analysis methods, the computed breathing waveforms are generally not compared with the reference breathing waveform but only the derived breathing rates validated [103, 104]. Other research targeting the breathing disorder obstructive sleep apnea (OSA) measure the performance of their OSA classifier [105, 106]. Manual comparison with a reference signal on single test episodes is also common [56, 65, 107].

We present a system that automatically segments the person of interest in a shared bed with an AdaBoost classifier using among others motion, intensity, focus and Histogram of Oriented Gradients (HOG) features. Actigraphy characteristics are computed and a breathing analysis method is proposed.

Robust and sensitive motion detection for sleep analysis

Sleep experiments are typically conducted in controlled and static illumination conditions [55]. This may work well in a laboratory setting but fails in practice where (sun)light and shadows may move over the sleeping subject. With regard to local dynamic illumination, moving shadows can be caused by e.g., a family member, outside tree branches or moving curtains in the presence of another light source (e.g., street lighting, moonlight, sunlight, indoor lighting). It may become difficult to discern between a shadow propagating over the bed and body movements performed by the sleeping subject. Particularly challenging are body movements performed over small distances and of short durations. These are not uncommon, especially among patients suffering from periodic limb movements in sleep (PLMS) or periodic limb movement disorder (PLMD). They can perform short (down to 0.5 seconds [82]) and small movements [82]. An actigraphy based approach for detecting periodic limb movements is suggested and found reliable and correct in [83] and [84]. Some problems with the usual EMG measurements may even be avoided with actigraphy sensors according to [83].

The commonly used video methods for movement analysis during sleep are ME and frame differencing [42, 54, 64–67, 76–78]. That those systems misdetect a moving shadow as subject movement is recognized by [55] where global artificial light changes were added to the laboratory sleep videos. This approach is however not designed to be insensitive to local illumination changes where subject movements can be distinguished from a moving shadow.

Previous research related to shadow detection imposed various restrictions on the camera [108], the shadow area properties related to the background characteristics

[109, 110], or the background itself [111–113]. For sleep monitoring, only a static camera is assumed, while the assumptions above are relaxed. This is relevant, since the bed is non-planar with beddings lying loosely on and around the subject. Moreover, the background changes dynamically, as subject movements change the folds' location and appearance of the beddings, while fold appearance (texture) varies with moving cast shadows. Strikingly, local intensity changes of 25% between consecutive frames are observed due to cast shadows vs. a 10% change due to subject movement.

In [111], the previous and current frames are compared by computing the local variance of the intensity ratios. Uniform regions are considered as potential shadow areas. The same effect is observed in [114] using a similar approach. Stauder et al. [114] correctly stressed the assumption of textured objects and a planar background which does not hold for the application at hand. Folds in beddings and blankets yield a non-planar background.

Employing a background model (e.g., [55, 115, 116]) is a very common approach confirmed in [108] and [112] where it is the first choice in the majority of the 50 selected shadow detection methods. A background and/or shadow model cannot be learned for the application discussed in this paper due to the dynamic and non-planar background yielding edges at different locations and at times larger intensity variations under the influence of moving cast shadows compared to subject movement.

Substantial research efforts have been made in determining shadow areas based on edge and texture information [109, 110, 113, 115, 117]. It is assumed that textures in background and shadow areas show high resemblance and/or that a foreground object will have significant interior edges contrary to a shadow region. This assumption is often violated where strong edges due to the blanket folds are manifested despite the cast shadow. Tracking moving objects has proven to be beneficial for shadow detection in [115] and [116].

As both a cast moving shadow and subject movement can be captured by performing ME on the video stream, we raise the following research question:

RQ6: *How can subject movement be distinguished from moving cast shadows for the case of periodic limb movements?*

Concerning the validation of a method addressing the above research question, the following should be taken into account. As gold standard for movement detection for PLMS and PLMD, on-body electromyography (EMG) is utilized [118]. The severity of PLM can be quantified using the so-called periodic limb movement index (PLMI), see [118]. Original PLM patient video data is available in [82]. For assessing the performance of classification algorithms on single samples (e.g., frames), sensitivity, specificity and Matthews Correlation Coefficient [119] are known measures in machine learning [120]. The root mean square (RMS) error is used to measure the difference between an estimated value and the ground truth value and is utilized for motion recognition in sleep in [121].

We propose a camera-based system combining video motion detection, ME and texture analysis with machine learning for sleep analysis. The system is robust to time-varying illumination conditions while using standard camera and infrared illumination hardware. We tested the system for Periodic Limb Movement detection during sleep, while using EMG signals as a reference.

Intelligent baby monitor

In the USA, 4500 infants die annually of sudden infant death syndrome (SIDS) [122]. In order to reduce the risk for SIDS, parents are advised to put their babies to sleep on their back (supine) and not on their stomach (prone). Feedback to the parent on the baby's sleeping pose may be appreciated, particularly after the baby has gained the ability to roll over. Current baby video monitors display the live video stream on the parent unit without performing any automatic analysis on the baby's sleeping position.

Movements of sleeping children are automatically analyzed and compared for different sleep stages in [76]. In [123], reflective markers are attached to the baby's sleeping bag which makes it easier to monitor the movements and pose of a baby. The body pose is recognized in [121] by using an artificial neural network solution and taking into account edge detection, row and column image profile projections. It requires an exhaustive data set to train the neural network and a specific location of the IR camera (i.e., overhead). The authors observe the method's lack of robustness as it cannot deal with a bed quilt, different hairstyles of the sleeping subjects and different clothes. Coarse body part detection of head, torso and upper leg is performed in [97]. It uses mainly models based on edge detections and is therefore invariant to face pose and variations in appearance. Combined with motion information, image areas are identified where body parts are likely to be located. As such, lying on the side can be distinguished from lying in a supine/prone position. However, this method is not capable of distinguishing between supine and prone. Traditional computer vision methods for face detection [124] have been used for infants in [125]. Challenges for an approach based solely on face detection can be the at times unfavorable face pose/angle towards the camera (often non-frontal), low contrast in the infrared images and the rather low quality video images.

With the access to detailed motion vector information from video processing methods discussed in this thesis, we pose the following research question:

RQ7: *How can we detect when a baby is turning from the supine to the prone sleeping pose?*

Regarding the evaluation of binary classification problems as is the case here, accuracy and precision are proposed by [126] as performance measures. Validation of a 2-class categorization method in the area of sleep analysis has been done with sensitivity and precision in [54], and with accuracy in [42, 55].

By designing a turning movement detector based on motion vector information and combining its information with face detection results, we determined the infant's sleeping pose.

Intelligent wake-up light

Analyzing body and breathing movements of sleeping subjects is primarily done for sleep stage classification or monitoring of sleep disorders. Limited research efforts have been spent for incorporating movement analysis in an intelligent wake-up light system. Research has shown [127, 128] that using a so-called artificial dawn wake-up light (e.g., Philips HF3490 [129]) during the wake-up phase (starting 30 minutes before the set wake-up time) results in a better wake-up experience. The light intensity increase is preset and can be too large (resulting in people waking up too early) or too small (resulting in people waking up from deep sleep). Some efforts have been made towards a more personalized wake-up light (e.g., [130]). A correlation exists between the lighter sleep phases and more body movements [43]. This is used in [130] where an intelligent alarm clock wakes the sleeping subject when he/she transitions through the light sleep phase within 30 minutes before the set wake-up time. When the subject does not pass through the light sleep phase in the mentioned time frame, the intelligent alarm clock behaves as any traditional alarm clock. In that case, the subject is still woken up in an abrupt manner by the alarm clock. A similar approach is implemented in the 'iwaku' product [131]. No related work could be found that ensures a wake-up in a light sleep phase at the preferred or set wake-up time. The research question we face in this matter is:

RQ8: *How can we create a good wake-up experience with light?*

Concerning this research question, our objectives were to determine an activity level trajectory with video analysis corresponding to a good wake-up experience and to design a stable wake-up light controller.

Regarding the evaluation of a wake-up light, user studies are typically done (e.g., [127]). In control system design, stability and performance of a controller are commonly investigated. According to [132], a linear time-invariant system is considered stable if all the roots of the transfer function denominator polynomial have negative real parts. Additionally, simulation results can validate the choice of the control loop model and compare it to alternative controller models with less stable elements as is done in [133].

We developed a system which can detect the sleeping person's activity and controls the light output such that the subject's behavior corresponds to an activity trajectory of a favorable wake-up experience.

1.4 Outline

This thesis focuses on video motion analysis methods developed for the mature application area of TV picture enhancement (Chapters 2 and 3) and the emerging application area of sleep analysis from video (Chapters 4-7).

We have explored the extensive parameter space of H3DRS in an automatic manner in Chapter 2 (RQ1). Chapter 3 (RQ2) presents an automatic ME design methodology that can deal with performance measures that cannot fully model the human perception.

A video actigraphy method is designed based on video ME and compared with wrist actigraphy (RQ3) in Chapter 4. Subsequently, a content-dependent clustering method is developed to cluster movements originating from one body part (RQ4). In Chapter 5 (RQ5), the camera system is placed on the bedside table to make product installation as convenient as possible for the user. We present a system that computes actigraphy and breathing characteristics of only the person of interest in a shared bed situation. In Chapter 6 (RQ6), we propose a camera-based system that can robustly discern movements from local time-varying illumination conditions. In Chapter 7, we demonstrate the potential of camera-based movement analysis in sleep related applications outside the common interest of sleep stage classification or monitoring of sleep disorders. The first application targets an intelligent baby monitor that informs parents about changes of their baby's pose in its sleep (RQ7). The second application shows how a sleeping subject's movement pattern can be used to build a personalized wake-up light system (RQ8).

Chapter 8 concludes with the main insights and results obtained in this thesis and offers an outlook to future work.

Bibliography

[1] Bureau of Labor Statistics, "American Time Use Survey Summary", http://www.bls.gov/news.release/atus.nr0.htm, 2013, Accessed 11 April 2014.

[2] National Sleep Foundation, "How Much Sleep Do We Really Need?", http://sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need, 2014, Accessed 11 April 2014.

[3] P. Nguyen, H. Tran, H. Nguyen, X.-N. Nguyen, *et al.*, "Asymmetric diamond search pattern for motion estimation in HEVC", in *Communications and Electronics (ICCE), 2014 IEEE Fifth International Conference on*, July 2014, pp. 434–439.

[4] Y.-S. Moon, Y.-M. Tai, J.H. Cha, and S.-H. Lee, "A simple ghost-free exposure fusion for embedded HDR imaging", in *Consumer Electronics (ICCE), 2012 IEEE International Conference on*, Jan. 2012, pp. 9–10.

[5] Y.-K. Lai and S.-M. Lee, "Wide Color-Gamut Improvement With Skin Protection Using Content-Based Analysis for Display Systems", *Display Technology, Journal of*, **vol. 9**, no. 3 pp. 146–153, March 2013.

[6] M. Cetin and I. Hamzaoglu, "An adaptive true motion estimation algorithm for frame rate conversion of high definition video and its hardware implementations", *Consumer Electronics, IEEE Transactions on*, **vol. 57**, no. 2 pp. 923–931, 2011.

[7] R. Phan and D. Androutsos, "Robust Semi-Automatic Depth Map Generation in Unconstrained Images and Video Sequences for 2D to Stereoscopic 3D Conversion", *Multimedia, IEEE Transactions on*, **vol. 16**, no. 1 pp. 122–136, Jan 2014.

[8] Philips, "Philips 40PFL8605H LED TV", http://www.philips.co.uk/c-p/40PFL8605H_12/8000-series-102-cm-40-inch-full-hd-1080p-digital-tv, 2014, Accessed 12 November, 2014.

[9] J. Wang, D. Wang, and W. Zhang, "Temporal compensated motion estimation with simple block-based prediction", *IEEE Transactions on Broadcasting*, **vol. 49**, no. 3 pp. 241–248, Sep. 2003.

[10] S. Dikbas and Y. Altunbasak, "Novel True-Motion Estimation Algorithm and Its Application to Motion-Compensated Temporal Frame Interpolation", *Image Processing, IEEE Transactions on*, **vol. 22**, no. 8 pp. 2931–2945, 2013.

[11] S.-C. Tai, Y.-R. Chen, Z.-B. Huang, and C.-C. Wang, "A Multi-Pass True Motion Estimation Scheme With Motion Vector Propagation for Frame Rate Up-Conversion Applications", *Display Technology, Journal of*, **vol. 4**, no. 2 pp. 188–197, 2008.

[12] J. Oliensis, "Exact two-image structure from motion", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **vol. 24**, no. 12 pp. 1618–1633, Dec. 2002.

[13] G. de Haan and P.W.A.C. Biezen, "Sub-pixel motion estimation with 3-D recursive search block-matching", *Signal Processing*, **vol. 6**, no. 3 pp. 229–239, June 1994.

[14] G. de Haan, , P.W.A.C. Biezen, H. Huijgen, and O.A. Ojo, "True-motion estimation with 3-D recursive search block matching", *IEEE Trans. Circuits, Syst. Video Techn.*, pp. 368–379, Oct. 1993.

[15] C.N. Cordes and G. de Haan, "Key requirements for high quality picture-rate conversion", *SID Digest of Technical Papers*, **vol. 15**, no. 2 pp. 850–853, June 2009.

[16] E.B. Bellers, "Motion Compensated Frame Rate Conversion for Motion Blur Reduction", *SID Digest of Technical Papers*, **vol. 38**, no. 1 pp. 1454–1457, May 2007.

[17] G.G.C. Lee, M.J. Wang, H.Y. Lin, D.W.C. Su, and B.Y. Lin, "Algorithm/Architecture Co-Design of 3-D Spatio–Temporal Motion Estimation for Video Coding", *IEEE Transactions on Multimedia*, **vol. 9**, no. 3 pp. 455–465, April 2007.

[18] A.M. Tourapis, O.C. Au, and M.L. Liou, "Highly efficient predictive zonal algorithms for fast block-matching motion estimation", *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 12**, no. 10 pp. 934–947, Oct. 2002.

[19] A.M. Tourapis, "Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation", *Proceedings of Visual Communications and Image Processing*, pp. 1069–79, Jan. 2002.

[20] R. Thoma and M. Bierling, "Motion compensating interpolation considering covered and uncovered background", *Signal Processing: Image Communication*, pp. 191–212, Feb. 1989.

[21] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo analysis by hybrid recursive matching for real-time immersive video conferencing", *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 14**, no. 3 pp. 321–334, March 2004.

[22] C. Bartels and G. de Haan, "Smoothness Constraints in Recursive Search Motion Estimation for Picture Rate Conversion", *Circuits and Systems for Video Technology, IEEE Transactions on*, **vol. 20**, no. 10 pp. 1310–1319, 2010.

[23] T. Yamamoto, N. Mishima, T. Ono, and T. Kaneko, "High-accuracy motion estimation with 4-D recursive search block matching", in *Consumer Electronics (GCCE), 2012 IEEE 1st Global Conference on*, 2012, pp. 625–628.

[24] Y. Guo, Z. Gao, L. Chen, and X. Zhang, "Effective early termination using adaptive search order for frame rate up-conversion", in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, 2013, pp. 1416–1419.

[25] R. Han and A. Men, "Frame rate up-conversion for high-definition video applications", *Consumer Electronics, IEEE Transactions on*, **vol. 59**, no. 1 pp. 229–236, 2013.

[26] Paolo Tinuper, Federica Provini, Francesca Bisulli, Luca Vignatelli, *et al.*, "Movement disorders in sleep: guidelines for differentiating epileptic from non-epileptic motor phenomena arising from sleep", *Sleep medicine reviews*, **vol. 11**, no. 4 pp. 255–267, 2007.

[27] G. Matwiyoff, T. Lee-Chiong, *et al.*, "Parasomnias: an overview", *Indian Journal of Medical Research*, **vol. 131**, no. 2 p. 333, 2010.

[28] C. Derry, "Nocturnal frontal lobe epilepsy vs parasomnias", *Current treatment options in neurology*, **vol. 14**, no. 5 pp. 451–463, 2012.

[29] L. Vignatelli, F. Bisulli, A. Zaniboni, I. Naldi, *et al.*, "Interobserver reliability of ICSD–R minimal diagnostic criteria for the parasomnias", *Journal of neurology*, **vol. 252**, no. 6 pp. 712–717, 2005.

[30] C.P. Derry, M. Davey, M. Johns, K. Kron, *et al.*, "Distinguishing sleep disorders from seizures: diagnosing bumps in the night", *Archives of neurology*, **vol. 63**, no. 5 pp. 705–709, 2006.

[31] M. Zucconi and L. Ferini-Strambi, "NREM parasomnias: arousal disorders and differentiation from nocturnal frontal lobe epilepsy", *Clinical Neurophysiology*, **vol. 111** pp. S129–S135, 2000.

[32] Jihui Zhang, Siu Ping Lam, Crover Kwok Wah Ho, Albert Martin Li, *et al.*, "Diagnosis of REM sleep behavior disorder by video-polysomnographic study: is one night enough?", *Sleep*, **vol. 31**, no. 8 p. 1179, 2008.

[33] A. Stefani, D. Gabelia, T. Mitterling, W. Poewe, *et al.*, "A Prospective Video-Polysomnographic Analysis of Movements during Physiological Sleep in 100 Healthy Sleepers.", *Sleep*, 2014.

[34] E.R. van Kesteren, J.P. van Maanen, A.A.J. Hilgevoord, D.M. Laman, and N. de Vries, "Quantitative effects of trunk and head position on the apnea hypopnea index in obstructive sleep apnea", *Sleep*, **vol. 34**, no. 8 p. 1075, 2011.

[35] P. Drakatos, S.E. Higgins, C.A. Kosky, R.T. Muza, and A.J. Williams, "The Value of Video Polysomnography in the Assessment of Intermittent Obstructive Sleep Apnea", *American journal of respiratory and critical care medicine*, **vol. 187**, no. 10 pp. e18–e20, 2013.

[36] B. Mwenge, A. Brion, G. Uguccioni, and I. Arnulf, "Sleepwalking: long-term home video monitoring", *Sleep medicine*, **vol. 14**, no. 11 pp. 1226–1228, 2013.

[37] J. M. Fry, M. A. DiPhillipo, K. Curran, R. Goldberg, and A. S. Baran, "Full polysomnography in the home", *Sleep*, **vol. 21**, no. 6 pp. 635–642, Sep 1998.

[38] T. Harada, A. Sakata, T. Mori, and T. Sato, "Sensor pillow system: monitoring respiration and body movement in sleep", in *Proceedings of 2000 IEEE/RSJ*

International Conference on Intelligent Robots and Systems, Nov. 2000, vol. 1, pp. 351–356.

[39] T. Willemen, B. Haex, J. Vander Sloten, and S. Van Huffel, "Biomechanics based analysis of sleep", in *Proc. of the 5th Dutch conference on Bio-medical engineering*, Jan. 2015, pp. 1–1.

[40] X.L. Aubert and A. Brauers, "Estimation of Vital Signs in Bed from a Single Unobtrusive Mechanical Sensor: Algorithms and Real-life Evaluation", in *30th Annual International Conference of the IEEE EMBS 2008.*, Aug. 2008.

[41] A.K. Ng, T. San Koh, E. Baey, T.H. Lee, *et al.*, "Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea?", *Sleep medicine*, **vol. 9**, no. 8 pp. 894–898, 2008.

[42] W.-H. Liao and C.-M. Yang, "Video-based activity and movement pattern analysis in overnight sleep studies.", in *Int'l Conf. on Pattern Recognition*, 2008, pp. 1–4.

[43] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, *et al.*, "The role of actigraphy in the study of sleep and circadian rhythms", *Sleep*, **vol. 26**, no. 3 pp. 342–392, 2003.

[44] R. Kocielnik, N. Sidorova, F.M. Maggi, M. Ouwerkerk, and J.H.D.M. Westerink, "Smart technologies for long-term stress monitoring at work", in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, June 2013, pp. 53–58.

[45] P. Renevey, J. Sola, P. Theurillat, M. Bertschi, *et al.*, "Validation of a wrist monitor for accurate estimation of RR intervals during sleep", in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, July 2013, pp. 5493–5496.

[46] E. Dolgin, "Technology: Dressed to detect", *Nature*, **vol. 511**, no. 7508 pp. S16–S17, 2014.

[47] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R.M. Aarts, "Respiration amplitude analysis for REM and NREM sleep classification", in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 5017–5020.

[48] W. Karlen, C. Mattiussi, and D. Floreano, "Improving actigraph sleep/wake classification with cardio-respiratory signals", *Conf Proc IEEE Eng Med Biol Soc*, **vol. 2008** pp. 5262–5265, 2008.

[49] K. Kawamoto, K. Hiroyuki, and T. Seiki, "Actigraphic Detection of REM Sleep Based on Respiratory Rate Estimation", *Journal of Medical and Bioengineering*, **vol. 2**, no. 1, 2013.

[50] R. Gardner and W.I. Grossman, "Normal motor patterns in sleep in man", *Advances in Sleep Research*, **vol. 2** pp. 67–107, 1976.

[51] J. Wilde-Frenz and H. Schulz, "Rate and distribution of body movements during sleep in humans", *Percept. Mot. Skills*, **vol. 56** pp. 275–283, 1983.

[52] A. Muzet, P. Naitoh, R.E. Townsend, and L.C. Johnson, "Body movements during sleep as a predictor of state change", *Psychon. Sci.*, **vol. 29** pp. 7–10, 1972.

[53] S. Gori, G. Ficca, Di Nasso, L. I. Murri, and P. Salzarulo, "Body movements during night sleep in healthy elderly subjects and their relationships with sleep stages", *Brain Research Bulletin*, **vol. 63**, no. 5 pp. 393–397, June 2004.

[54] K. Cuppens, L. Lagae, B. Ceulemans, S. Huffel, and B. Vanrumste, "Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy", *Med. Biol. Engineering and Computing*, **vol. 48**, no. 9 pp. 923–931, Sep. 2010.

[55] W.-H. Liao and J.H. Kuo, "Sleep monitoring system in real bedroom environment using texture-based background modeling approaches", *J. Ambient Intelligence and Humanized Computing*, **vol. 4**, no. 1 pp. 57–66, Feb. 2013.

[56] C.W. Wang, A. Ahmed, and A. Hunter, "Vision analysis in detecting abnormal breathing activity in application to diagnosis of obstructive sleep apnoea", in *Conf Proc IEEE Eng Med Biol Soc*, Sep. 2006, pp. 4469–4473.

[57] N. Koolen, O. Decroupet, A. Dereymaeker, K. Jansen, *et al.*, "Automated Respiration Detection from Neonatal Video Data", in *Proc. of the 4th International Conference on Pattern Recognition Applications and Methods. (ICPRAM 2015)*, Jan. 2015, pp. 164–169.

[58] B. Günyel and A.A. Alatan, "Multi-resolution motion estimation for motion compensated frame interpolation", in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, Sept 2010, pp. 2793–2796.

[59] M. Santoro, G. AlRegib, and Y. Altunbasak, "Motion estimation using block overlap minimization", in *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on*, Sept 2012, pp. 186–191.

[60] B. McCane, B. Galvin, and K. Novins, "On the evaluation of optical flow algorithms", in *Fifth International Conference on Control, Automation, Robotics and Vision, Singapore*, Citeseer, 1998.

[61] S. Baker, S. Roth, D. Scharstein, M.J. Black, *et al.*, "A Database and Evaluation Methodology for Optical Flow", *International Journal of Computer Vision*, **vol. 92**, no. 1 pp. 1–31, Mar. 2011.

[62] L. Koskinen, A. Paasio, and K.A.I. Halonen, "Motion estimation computational complexity reduction with CNN shape segmentation", *Circuits and Systems for Video Technology, IEEE Transactions on*, **vol. 15**, no. 6 pp. 771–777, June 2005.

[63] A. D'Angelo, M. Carli, and M. Barni, "Quality evaluation of motion estimation algorithms based on structural distortions", in *Multimedia Signal Processing, 2009. MMSP '09. IEEE International Workshop on*, Oct 2009, pp. 1–6.

[64] K. Nakajima, A. Osa, T. Maekawa, and H. Miike, "Evaluation of Body Motion by Optical Flow Analysis", *Japanese Journal of Applied Physics*, **vol. 36**, no. Part 1, No. 5A pp. 2929–2937, 1997.

[65] K. Nakajima, Y. Matsumoto, and T. Tamura, "Development of real-time image sequence analysis for evaluating posture change and respiratory rate of a subject in bed", *Physiol. Meas.*, **vol. 22**, no. 3 pp. 21–28, 2001.

[66] N. B. Karayiannis, B. Varughese, Guozhi Tao, J. D. Frost, Jr., *et al.*, "Quantifying motion in video recordings of neonatal seizures by regularized optical flow methods", *Trans. Img. Proc.*, **vol. 14**, no. 7 pp. 890–903, Jul. 2005.

[67] S. Kalitzin, G. Petkov, D. Velis, B. Vledder, and F. Lopes da Silva, "Automatic Segmentation of Episodes Containing Epileptic Clonic Seizures in Video Sequences", *Biomedical Engineering, IEEE Transactions on*, **vol. 59**, no. 12 pp. 3379–3385, 2012.

[68] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision", in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, 1981, pp. 674–679.

[69] J.-Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm", *Intel Corporation*, **vol. 5** pp. 1–10, 2001.

[70] B. Horn and B. Schunck, "Determining Optical Flow", *Artificial Intelligence*, **vol. 17**, no. 1-3 pp. 185–203, 1981.

[71] J.L. Barron and M. Khurana, "Determining Optical Flow for Large Motions Using Parametric Models in a Hierarchical Framework", in *In Vision Interface*, 1994, pp. 47–56.

[72] H. Mehrabian, H. Karimi, and A. Samani, "Accurate optical flow field estimation using mechanical properties of soft tissues", in *Proc. SPIE*, 2009, vol. 7262, pp. 72621B–72621B–11.

[73] C. Liu, W.T. Freeman, E.H. Adelson, and Y. Weiss, "Human-assisted motion annotation", in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[74] M. Drulea and S. Nedevschi, "Total variation regularization of local-global optical flow", in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, Oct 2011, pp. 318–323.

[75] D. Sun, S. Roth, and M.J. Black, "Secrets of optical flow estimation and their principles", in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010, pp. 2432–2439.

[76] S. Okada, N. Shiozawa, and M. Makikawa, "Body movement in children with ADHD calculated using video images", in *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*, Jan 2012, pp. 60–61.

[77] J.U. Bak, N. Giakoumidis, G. Kim, H. Dong, and N. Mavridis, "An Intelligent Sensing System for Sleep Motion and Stage Analysis", *Procedia Engineering*, **vol. 41**, no. 0 pp. 1128 – 1134, 2012, international Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012).

[78] C.W. Wang, A. Hunter, and A. Ahmed, "Artificial intelligent vision analysis in obstructive sleep apnoea (OSA)", in *30th Anniversary Conference of the Association for Respiratory Technology and Physiology (ARTP)*, Jan. 2006.

[79] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, for the American Academy of Sleep Medicine. 1st ed. Westchester: IL: American Academy of Sleep Medicine, 2007.

[80] T. Blackwell, S. Redline, S. Ancoli-Israel, J.L. Schneider, *et al.*, "Comparison of sleep parameters from actigraphy and polysomnography in older women: the SOF study", *Sleep*, **vol. 31**, no. 2 pp. 283–291, Feb 2008.

[81] J.M. Bland and D.G. Altman, "Measuring agreement in method comparison studies", *Stat. Methods Med. Res.*, **vol. 8**, no. 2 pp. 135–160, Apr. 1999.

[82] B. Högl, M. Zucconi, and F. Provini, "RLS, PLM, and their differential diagnosis – a video guide.", *Mov Disord*, **vol. 22 Suppl 18** pp. S414–9, 2007.

[83] R.P. Allen and W.A. Hening, "Actigraph Assessment of Periodic Leg Movements and Restless Legs Syndrome", in *Restless Legs Syndrome*, W.B. Saunders, Philadelphia, pp. 142 – 149, 2009.

[84] E. Sforza, M. Johannes, and B. Claudio, "The PAM-RL ambulatory device for detection of periodic leg movements: a validation study.", *Sleep Med*, **vol. 6**, no. 5 pp. 407–13, Sep. 2005.

[85] X. Jin, A Heinrich, C. Shan, and G. de Haan, "Shared-bed person segmentation based on motion estimation", in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, Sept 2012, pp. 137–140.

[86] J.B. MacQueen, "Some Methods for Classification and Analysis of MultiVariate Observations", in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[87] P. Berkhin, "Survey of clustering data mining techniques", *Tech. rep.*, Accrue Software, 2002.

[88] F. Gibou and R. Fedkiw, "A fast hybrid K-means level set algorithm for segmentation", in *4th Annual Hawaii International Conference on Statistics and Mathematics*, 2005, pp. 281–291.

[89] R.V. Babu and K.R. Ramakrishnan, "Background Sprite Generation Using MPEG Motion Vectors.", in *ICVGIP*, Oct. 2002.

[90] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding", in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, 2007, pp. 1027–1035.

[91] A. Likas, N. Vlassis, and J.J. Verbeek, "The global k-means clustering algorithm", *Pattern Recognition*, **vol. 36**, no. 10 pp. 451–461, Feb. 2003.

[92] R. Brunelli and O. Mich, "Histograms analysis for image retrieval", *Pattern Recognition*, **vol. 34**, no. 8 pp. 1625–1637, Aug. 2001.

[93] W.C. Bresky, J.M. Daniels, A.A. Bailey, and S.T. Wanzong, "New methods toward minimizing the slow speed bias associated with atmospheric motion vectors", *Journal of Applied Meteorology and Climatology*, **vol. 51**, no. 12 pp. 2137–2151, 2012.

[94] G.D. Tipaldi and F. Ramos, "Motion Clustering and Estimation with Conditional Random Fields", in *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS'09, 2009, pp. 872–877.

[95] M. Hu, S. Ali, and M. Shah, "Learning motion patterns in crowded scenes using motion flow field.", in *ICPR*, IEEE, Mar. 2008, pp. 1–5.

[96] F.-C. Yang, C.H. Kuo, M.-Y. Tsai, and S.-C. Huang, "Image-based sleep motion recognition using artificial neural networks", in *Machine Learning and Cybernetics, 2003 International Conference on*, 2003, vol. 5, pp. 2775–2780 Vol.5.

[97] C.-W. Wang, A. Hunter, N. Gravill, and S. Matusiewicz, "Real time pose recognition of covered human for diagnosis of sleep apnoea", *Computerized Medical Imaging and Graphics*, **vol. 34**, no. 6 pp. 523 – 533, 2010.

[98] Gear4, "Renew SleepClock", 2014, [Online; accessed 24-May-2014].
URL `http://www.stage.gear4.com/product/_/426/renew-sleepclock/`

[99] P. De Chazal, N. Fox, E. O'Hare, C. Heneghan, *et al.*, "Sleep/wake measurement using a non-contact biomotion sensor", *J Sleep Res*, **vol. 20**, no. 2 pp. 356–366, Jun 2011.

[100] L. Samy, M.-C. Huang, J.J. Liu, W. Xu, and M. Sarrafzadeh, "Unobtrusive Sleep Stage Identification Using a Pressure-Sensitive Bed Sheet", *Sensors Journal, IEEE*, **vol. 14**, no. 7 pp. 2092–2101, July 2014.

[101] National Sleep Foundation, "Sleep in America poll: Adult sleep habits and styles", 2005, [Online; accessed 5-October-2013].
URL `http://www.sleepfoundation.org/article/sleep-america-polls/2005-adult-sleep-habits-and-styles`

[102] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, *et al.*, "Eulerian Video Magnification for Revealing Subtle Changes in the World", *ACM Trans. Graph.*, **vol. 31**, no. 4 pp. 65:1–65:8, July 2012.

[103] M. Bartula, T. Tigges, and J. Muehlsteff, "Camera-based system for contactless monitoring of respiration", in *Engineering in Medicine and Biology Society*

(EMBC), 2013 35th Annual International Conference of the IEEE, 2013, pp. 2672–2675.

[104] F. Al-Khalidi, R. Saatchi, H. Elphick, and D. Burke, "An evaluation of thermal imaging based respiration rate monitoring in children", *American Journal of Engineering and Applied Sciences*, **vol. 4**, no. 4 pp. 586–597, 2011.

[105] E. Gederi and G.D. Clifford, "Fusion of image and signal processing for the detection of obstructive sleep apnea", in *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*, 2012, pp. 890–893.

[106] K.S. Tan, R. Saatchi, H. Elphick, and D. Burke, "Real-time vision based respiration monitoring system", in *Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th International Symposium on*, 2010, pp. 770–774.

[107] M. Frigola, J. Amat, and J. Pagès, "Vision based respiratory monitoring system", in *Proceedings of the 10th Mediterranean Conference on Control and Automation - MED2002 Lisbon, Portugal*, July 2002.

[108] A. Prati, I. Mikic, M.M. Trivedi, and R. Cucchiara, "Detecting Moving Shadows: Algorithms and Evaluation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **vol. 25**, no. 7 pp. 918–923, July 2003.

[109] W. Zhang, X.Z. Fang, X.K. Yang, and Q.M.J. Wu, "Moving Cast Shadows Detection Using Ratio Edge", *Multimedia, IEEE Transactions on*, **vol. 9**, no. 6 pp. 1202–1214, Oct. 2007.

[110] A. Leone and C. Distante, "Shadow detection for moving objects based on texture analysis", *Pattern Recognition*, **vol. 40**, no. 4 pp. 1222–1233, Apr. 2007.

[111] J. Stander, R. Mech, and J. Ostermann, "Detection of moving cast shadows for object segmentation", *IEEE Transactions on Multimedia*, **vol. 1**, no. 1 pp. 65–76, Mar. 1999.

[112] A. Sanin, C. Sanderson, and B.C. Lovell, "Shadow detection: A survey and comparative evaluation of recent methods", *Pattern Recogn.*, **vol. 45**, no. 4 pp. 1684–1695, Apr. 2012.

[113] J.V. Panicker, "Detection of moving cast shadows using edge information", in *Int'l Conf. on Computer and Automation Engineering*, 2010.

[114] E. Kermani and D. Asemani, "A New Illumination-Invariant Method of Moving Object Detection for Video Surveillance Systems", in *Iranian Conf. on Machine Vision and Image Processing*, 2011, pp. 1–5.

[115] L. Unzueta, M. Nieto, A. Cortes, J. Barandiaran, *et al.*, "Adaptive Multicue Background Subtraction for Robust Vehicle Counting and Classification", *Intelligent Transportation Systems, IEEE Transactions on*, **vol. 13**, no. 2 pp. 527–540, June 2012.

[116] Z. Liu, K. Huang, and T. Tan, "Cast Shadow Removal in a Hierarchical Manner Using MRF", *Circuits and Systems for Video Technology, IEEE Transactions on*, **vol. 22**, no. 1 pp. 56–66, 2012.

[117] J. Zhu, K.G.G. Samuel, S.Z. Masood, and M.F. Tappen, "Learning to recognize shadows in monochromatic natural images", in *Computer Vision and Pattern Recognition, 2010 IEEE Conference on*, 2010, pp. 223–230.

[118] M. Zucconi, R. Ferri, R. Allen, P. C. Baier, *et al.*, "The official World Association of Sleep Medicine (WASM) standards for recording and scoring periodic leg movements in sleep (PLMS) and wakeful-ness (PLMW) developed in collaboration with a task force from the International Restless Legs Syndrome Study Group (IRLSSG)", *Sleep Medicine*, **vol. 7**, no. 2 pp. 175–183, Mar. 1972.

[119] B.W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", *Biochim. Biophys. Acta*, **vol. 405**, no. 2 pp. 442–451, Oct. 1975.

[120] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview", *Bioinformatics*, **vol. 16**, no. 5 pp. 412–424, Feb. 2000.

[121] C. Kuo, F. Yang, M. Tsai, and M. Lee, "Artificial neural networks based sleep motion recognition using night vision cameras", *Biomedical engineering: Applications, basis and communications*, **vol. 16**, no. 2 pp. 79–86, Feb. 2004.

[122] S. Parmet, A.E. Burke, and R.M. Golub, "Sudden infant death syndrome", *JAMA*, **vol. 307**, no. 16 p. 1766, 2012.

[123] A. Jeung, H. Mostafavi, M.L. Riaziat, *et al.*, "Method and system for monitoring breathing activity of a subject", *Patent US 7403638 B2*, July 2008.

[124] P. Viola and M.J Jones, "Robust real-time face detection", *International journal of computer vision*, **vol. 57**, no. 2 pp. 137–154, 2004.

[125] M.N. Mansor, M.N. Rejab, S.H.-F.S.A. Jamil, A.H.-F.S.A. Jamil, *et al.*, "Fast infant pain detection method", in *Computer and Communication Engineering (ICCCE), 2012 International Conference on*, July 2012, pp. 918–921.

[126] S. S. Coughlin, B. Trock, M.H. Criqui, L.W. Pickle, *et al.*, "The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test", *Journal of Clinical Epidemiology*, **vol. 45**, no. 1 pp. 1 – 7, 1992.

[127] M. Van De Werken, M.C. Gimenez, B. De Vries, D.G. Beersma, *et al.*, "Effects of artificial dawn on sleep inertia, skin temperature, and the awakening cortisol response", *J Sleep Res*, **vol. 19**, no. 3 pp. 425–435, Sep. 2010.

[128] M. C. Gimenez, M. Hessels, M. van de Werken, B. de Vries, *et al.*, "Effects of artificial dawn on subjective ratings of sleep inertia and dim light melatonin onset", *Chronobiol. Int.*, **vol. 27**, no. 6 pp. 1219–1241, Jul 2010.

[129] Philips, "Philips HF3490 Wake-up Light", http://www.philips.co.uk/c-p/HF3490_01, 2014, Accessed 20 March, 2014.

[130] R. Liao, "2-D/3-D registration of C-ARM CT volumes with fluoroscopic images by spines for patient movement correction during electrophysiology", in *Proc. of the Int'l Conf. on Biomedical Imaging*, 2010, pp. 1213–1216.

[131] iwaku, "iwaku wake-up light", http://www.iwaku.com/product/item13, 2014, Accessed 20 March, 2014.

[132] G.F. Franklin, D.J. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 4th ed., 2001, ISBN 0130323934.

[133] C. Setz, A. Heinrich, Ph. Rostalski, G. Papafotiou, and M. Morari, "Application of model predictive control to a cascade of river power plants", in *Proceedings of the IFAC World Congress*, July 2008.

Chapter 2

Optimization of hierarchical 3DRS motion estimators for picture rate conversion

Abstract

There is a continuous pressure to improve the quality of motion-compensated picture rate conversion methods while maintaining acceptable computational complexity. Since the concept of hierarchy can be advantageously applied to many motion estimation methods, we have extended and improved the current state-of-the-art motion estimation method in this field, 3-Dimensional Recursive Search (3DRS), with this concept. We have explored the extensive parameter space and present an analysis of the importance and influence of the various parameters for the application of picture rate conversion. Since well-performing motion estimation methods for picture rate conversion show a trade-off between prediction accuracy and spatial motion field consistency, determining the optimal trade-off is an important part of the analysis. We found that the proposed motion estimators are superior to multiple existing techniques as well as standard 3DRS with regard to performance at a low computational complexity.

This chapter is published as: A. Heinrich, C. Bartels, R.J. van der Vleuten, C.N. Cordes, G. de Haan; Optimization of hierarchical 3DRS motion estimators for picture rate conversion, *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 2, pp. 262-274, Mar. 2011.

2.1 Introduction

Motion estimation (ME) is an essential part of the picture rate conversion methods that are applied to eliminate film judder in high-end televisions [134]. Because of the increasing spatial resolution (from SD to HD and Full HD) and picture rates (from 24 fps to more than 200 fps) of video shown on those televisions, as well as the increasing size and quality of the television displays, there is a continuous pressure to improve the quality of the ME algorithm while maintaining acceptable computational complexity. To this end, spatio-temporal prediction methods such as recursive search, e.g. [135–137], are typically applied in practice (e.g. [138, 139]). Generally, spatio-temporal predictors have proven to be a powerful tool in the design of motion estimation algorithms [140–142].

Combinations of 3-Dimensional Recursive Search (3DRS) [143] with concepts borrowed from alternative ME methods have shown to be beneficial in earlier publications, e.g. [144]. In this paper, we extend the 3DRS technique with the concept of hierarchy [145] which has been advantageously applied to many ME methods. This work shows the effect of parameter optimization on the quality of hierarchical motion estimation, and most importantly, that the commonly used manual parameter optimization is unlikely to arrive at an optimum due to the vast size of the optimization space. This work thus clearly indicates that sufficient attention has to be placed on the optimization of a hierarchical motion estimator, as otherwise the result is likely to be suboptimal. Therefore, we explore the extensive parameter space and provide insights with respect to the importance and the influence of the individual parameters. Since well-performing ME methods for picture rate conversion show a trade-off between prediction accuracy and spatial motion field consistency, the optimal trade-off is analyzed.

In Section 2.2, we introduce the concept of hierarchy for ME and describe its integration with 3DRS. The designed motion estimators and their parameters are formally defined in Section 2.3. In Section 2.4, we explore the parameter space in order to find the optimal range of parameter settings. A further in-depth analysis of several parameters, the performance and complexity analysis of the proposed motion estimators and existing techniques is then given in Section 2.5 and Section 2.6, respectively. Section 2.7 summarizes our conclusions.

2.2 Hierarchical 3DRS motion estimation

Hierarchical ME is introduced in Section 2.2.1, followed by a discussion of its integration into 3DRS in Section 2.2.2.

2.2.1 Multi-scale and multi-grid hierarchical motion estimation

The ME methods discussed in this paper are based on the principle of block-matching [146]. According to this principle, the image is divided into blocks and for each block a reference image is searched for the best-matching block (according to a pre-defined cost function). In case of motion in the sequence, the best-matching block will be located at a different spatial position in the reference image than in the current image. The vector over which the matching block has been shifted compared to the block in the current image is called the motion vector. The complexity of finding the best-matching block obviously depends directly on the number of different vectors (with corresponding cost function evaluations) that are examined. Hierarchical ME methods can take large search ranges efficiently into account [147] which reduces the risk of being trapped in suboptimal local minima.

In this paper, we investigate a hierarchical ME approach using resolution down-scaling, which we call multi-scale block-matching ME. Using down-scaling, the coarser motion vectors are obtained from block-matching at a lower spatial resolution and can be successively refined at higher resolutions. We will combine the multi-scale approach with a hierarchical ME method known as multi-grid block-matching [146]. In this method, a coarse motion vector is first found using a large block size and this vector is successively refined for the smaller blocks into which the larger blocks are decomposed (using a quad-tree decomposition). By combining the multi-scale and multi-grid approaches, we are flexible in investigating the effects of using different block sizes and scale factors.

2.2.2 Hierarchical 3DRS block-matching

3DRS [135] selects the output motion vector $\vec{d}$ from a candidate vector set C, that is based on prediction vectors from a spatio-temporal neighborhood. This process comprises two steps:

1. For each (block/pixel) location $\vec{x}$ in frame number n, construct a candidate set C, e.g.

$$C = \left\{ \begin{array}{c} \vec{d}(\vec{x} + k \cdot \vec{u}_x - \vec{u}_y, n), \\ \vec{d}(\vec{x} - \vec{u}_x, n),\ \vec{d}(\vec{x}, n-1),\ \vec{d}(\vec{x} + \vec{u}_x, n-1), \\ \vec{d}(\vec{x} + l \cdot \vec{u}_x + \vec{u}_y, n-1), \\ \vec{d}(\vec{x} - \vec{u}_x, n) + \vec{\eta},\ \vec{d}(\vec{x} - \vec{u}_y, n) + \vec{\eta} \end{array} \right\}, \qquad (2.1)$$

$$k = -1, 0, 1, \quad l = -1, 0, 1$$

   where $\vec{u}_x$, $\vec{u}_y$ are unit vectors on the block/pixel grid, and $\vec{\eta}$ is a random value. Usually this random value is drawn from a fixed update set [135].

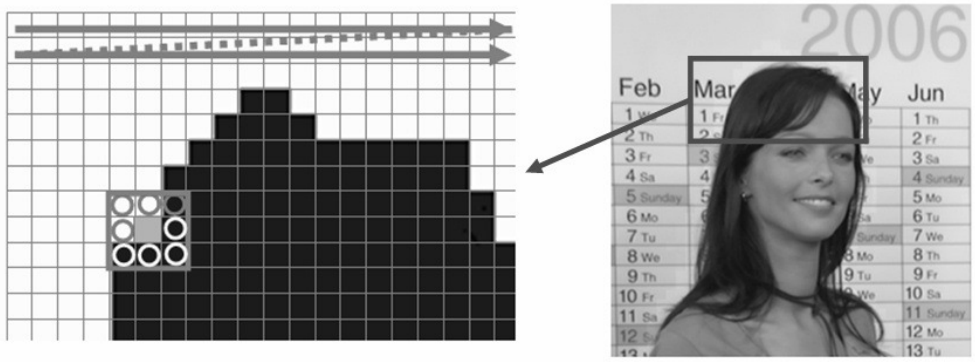


Figure 2.1: Configuration of the spatial and temporal candidates for the scanning direction indicated by the gray arrows in the block grid. The light gray block is the current block. The spatial candidates are indicated with the gray circles, the temporal candidates with the white circles (see Eq. (2.1)).

2. The estimated value for $\vec{d}(\vec{x}, n)$ then is

$$\vec{d}(\vec{x}, n) = \arg \min_{\vec{d}_c \in C} \left(E_{\mathrm{m}}(\vec{x}, \vec{d}_c, n) + E_{\mathrm{p}}(\vec{d}_c^{\,*}) \right) \tag{2.2}$$

where E_{m} is a common match term for which we use the Sum of Absolute Differences (SAD), while E_{p} is a block size dependent penalty term to bias the preference among the different types of candidates $\vec{d}_c$ which is denoted by $\vec{d}_c^{\,*}$. Regarding the three candidate types, we distinguish between the spatial, temporal and update predictors which will be elaborated on in the following. We refer to the sum of the match term and the penalty term as the energy function.

Important is that steps 1 & 2 are performed *sequentially* for each location. Hence, the newly estimated value is assigned to the location before moving to the next location. Therefore this new value becomes part of the candidate set of the next location, directly influencing the estimate for that next location.

The underlying idea of 3DRS is that "objects are larger than blocks" and therefore already estimated neighboring vectors are good predictions for the current value to be estimated. These neighboring values are called *spatial* candidates ($\vec{d}(\cdot, n)$ in Eq. (2.1)). Unfortunately, not all neighboring values have already been estimated. However, previous estimates both in time and iteration are also good predictions, assuming "objects have inertia". These predictions from previous estimates are called *temporal* candidates ($\vec{d}(\cdot, n-1)$ in Eq. (2.1)), but are somewhat less reliable than spatial candidates, because of the motion of the objects and the change in motion. The reliability of the different types of predictors is taken into account by the penalty in Eq (2.2).

The *scanning direction* determines the order in which block-based motion estimation is performed. Fig. 2.1 shows the configuration of spatial and temporal candidates

when processing from top-left to bottom-right. This is the scanning order assumed in Eq. (2.1). Processing solely in this order means that good motion vector estimates can only propagate in one direction. If a good estimate is found at the bottom of the image it can take some time before it is propagated to the top.

To improve the propagation of good estimates it is beneficial to *vary the scanning direction*. In practice two mechanisms are used for this. In the first option after a scan from top to bottom, the next scan is run from bottom to top. This is alternated continuously. The second option is called *meandering* meaning that after scanning a line from left to right, the next line is scanned from right to left. If both mechanisms are used, good estimates can propagate in four different directions, which ensures a quick spreading of good estimates all over the image. For an even faster propagation, two scans (both top to bottom and bottom to top) per image can be performed in a meandering manner. The example candidate set from Eq. (2.1) assumes scanning from top to bottom and from left to right, i.e., along the unit vectors defining the axes of the image. If the scanning direction is changed the unit vectors $\vec{u}_x$ and $\vec{u}_y$ in Eq. (2.1) should be changed in unit vectors defining the current scanning direction, $\vec{s}_x$ and $\vec{s}_y$.

Another important aspect concerns the *update* candidates ($\vec{d}(\cdot, n)+\vec{\eta}$ in Eq. (2.1)). Both spatial and temporal candidates contain values that already have been estimated. However, new values need to be introduced as well to find vectors for appearing objects and to accommodate for acceleration. This is achieved by adding small random values to spatial candidates. These random values can be drawn from a random distribution, e.g., a Gaussian distribution ($\mathcal{N}$), but typically they are drawn from a fixed update set (U_S) [135].

$$\eta_{x,y} \sim \mathcal{N}(0, \sigma) \quad \text{where} \quad \vec{\eta} = \begin{pmatrix} \eta_x \\ \eta_y \end{pmatrix}, \quad \text{or} \quad \vec{\eta} \in U_S \tag{2.3}$$

These vectors can be small since objects have inertia and in order to promote smoothness, $\sigma \leq 2$. Hence, the motion of objects will only change gradually. To find a motion vector that differs significantly from previously estimated vectors, it takes several consecutive updates. This process is called *convergence*. Updated vectors are considered the least reliable predictors and are therefore assigned the highest penalty.

Typically the penalty for spatial candidates is fixed to zero. For 16-bit image data, we empirically determined that the penalty E_p of 128 per pixel in a block for temporal candidates and 512 per pixel in a block for update candidates produces good results.

The evaluation of the energy function in Eq. (2.2) is the most expensive part. In the case of 3DRS the energy function only needs to be evaluated for a few candidates, regardless of the range of the motion vectors. The size of the candidate set can be tuned to achieve good quality with a minimum number of candidates. The candidate set from Eq. (2.1) contains 11 candidates. However because of the smoothness of the motion field, often neighboring candidate locations result in the same prediction. Therefore the number of candidates can be sub-sampled without significant loss in quality (see the 3DRS candidate structure in Fig. 2.3). In this paper, we add an

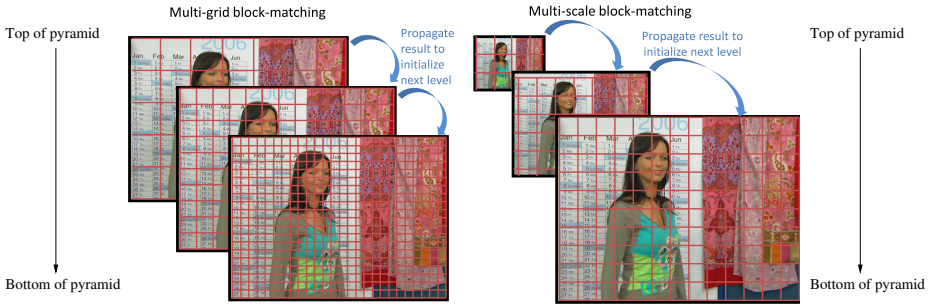


Figure 2.2: Illustration of multi-grid and multi-scale approach.

additional candidate vector from an 'external' source, i.e., a *hierarchical candidate*. The hierarchical candidate vector can be obtained by both multi-grid (same resolution, multiple block-sizes) and multi-scale (multiple resolution levels) approaches. The 3DRS block matching method with the additional hierarchical candidate is performed on each hierarchical level (except on the coarsest level, where the hierarchical candidate is not available and thus the non-hierarchical 3DRS is performed). On each hierarchical level, the temporal candidate vectors are propagated from the previously computed vectors on the highest-resolution image and down-scaled with regard to all the scales used in the current hierarchical ME scan. This shows better performance than when possibly unconverged temporal candidates from the same scale are used.

The introduction of a hierarchical candidate vector increases the number of candidate vector evaluations that are performed, compared to non-hierarchical 3DRS. In order to profit from the hierarchical candidate without a complexity increase, we could e.g., skip the ME on the full-resolution image and use motion vectors that are up-scaled from a lower-resolution estimation or modify the motion vector candidate structures (see Section 2.3.2).

2.3 Hierarchical motion estimation definition and parameters

First, the hierarchical motion estimators, as well as their parameters, are defined in Section 2.3.1. Next, the chosen parameter values are discussed in Section 2.3.2 Candidate Structures, Section 2.3.3 Scans, Section 2.3.4 Scale Parameter Sets, and Section 2.3.5 Block Sizes. Finally, an example motion estimator configuration is given in Section 2.3.6.

2.3.1 Definitions

In order to describe the multi-scale or multi-grid approach, a scale pyramid is used, as shown in Fig. 2.2, where ME is performed on higher scales at the top of the

fine	Lowest/Finest scale on which ME is performed	Scalar
coarse	Highest/Coarsest scale on which ME is performed	Scalar
$\mathbf{sf_w}$, $\mathbf{sf_h}$	Scaling factor width and height for resizing scales	Array
$\mathbf{blk_w}$, $\mathbf{blk_h}$	Block width and height of each scale	Array
scan	Amount of ME scans performed per scale	Array

Table 2.1: Parameters for the hierarchical motion estimator design.

pyramid first and motion vectors are propagated down the pyramid to the lower scales by means of hierarchical candidates. The parameters involved in the design of hierarchical motion estimators are explained in the following and an overview is given in Table 2.1.

The relevant parameters for the scale structure are the **fine** scale, the **coarse** scale and the scaling factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$. The **fine** scale and the **coarse** scale denote the levels of the pyramid (see Fig. 2.2) where ME is performed, e.g., **fine** $= 1$, **coarse** $= 2$. **fine** is the finer scale (for multi-scale ME) or the one with a finer block grid than **coarse** in the case that the coarse and fine scale have the same size (multi-grid ME). If the full resolution is included as a scale on which ME is performed, **fine** $= 0$ is chosen (otherwise **fine** $= 1$). The scale factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$ determine the size of the scales. The scaling factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$ for width and height are arrays which indicate how much one scale is down-scaled in comparison to the next lower scale in the pyramid. The first component, i.e., $\mathbf{sf_w}[0]$, denotes the scaling factor between the full resolution image and the following higher level of the pyramid; the second component, i.e., $\mathbf{sf_w}[1]$, denotes the scaling factor between the first down-scaled image and the next higher scale in the pyramid etc. In the case of a multi-scale motion estimator (right image in Fig. 2.2), the image dimensions become smaller as we ascend in the pyramid. However, when a multi-grid (left image in Fig. 2.2) motion estimator is designed, two scales have the same dimension, thus the corresponding scaling factor component equals 1. As the spatial resolution of two vector fields from two different scales may not be equal, this may require scaling of the vector field as well, which is implemented as nearest neighbor scaling.

The block width and block height dimensions $\mathbf{blk_w}$ and $\mathbf{blk_h}$ are arrays where the elements $\mathbf{blk_w}[i], i = 0, \ldots,$ **coarse**, indicate the block sizes for each scale i in the pyramid. The equivalent block width dimensions on the full resolution image can be computed as in

$$\mathbf{blk_w}'[0] = \mathbf{blk_w}[\mathbf{coarse}] \cdot \prod_{i=0}^{\mathbf{coarse}-1} \mathbf{sf_w}[i], \text{for } \mathbf{coarse} > 0. \tag{2.4}$$

The equivalent formula is applied to the block height dimensions. The number of ME scans performed on each scale is defined by the parameter **scan** which is also an array. In this experiment, all the elements were chosen to take identical values, i.e., either all equal to 1 or all equal to 2.

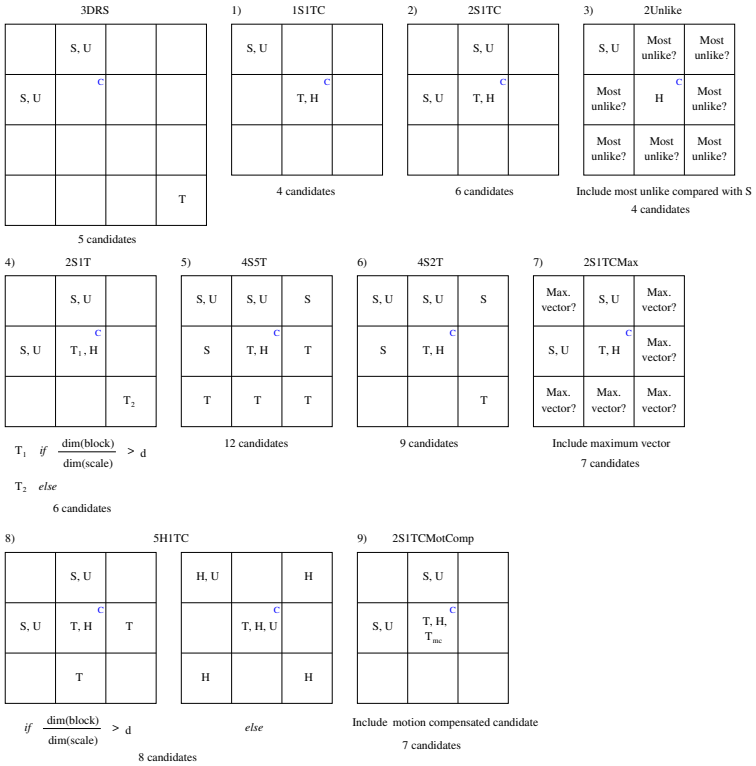


Figure 2.3: The usual 3DRS candidate structure, as well as nine different subsamplings of the spatio-temporal neighborhood of a block (candidate structures 1,...,9) are shown. C denotes the current block for which candidate motion vectors are determined, S a spatial candidate, U a random update vector added to the spatial candidate, T a temporal candidate, and H the hierarchical candidate resulting from the ME scan on a coarser grid or on a coarser scale. For candidate structures 4 and 8, $d = 1/60$.

The random update vectors in both, positive and negative, horizontal and vertical direction are chosen with quarter-pixel accuracy. The lengths of the update vectors are discretized to 0.25, 0.5, 1 and 2. The length of the update vectors is not changed throughout the scales and thus not down-scaled proportionally to the scaling factor, which should favor a fast convergence speed. In order to find the zero motion of stationary image parts such as subtitles and logos faster, the zero vector is included as an additional motion vector candidate with a high penalty, set equal to the update penalty.

2.3.2 Candidate structures

Different numbers of candidates and different approaches are applied to determine the motion vector candidates, as shown in Fig. 2.3. In contrast to the usual 3DRS candidate structure (also shown in Fig. 2.3), the temporal candidate is closer to the current block for all the hierarchical approaches because, for coarse scales, the tem-

Scale structure A	1 scale, fine 1
Scale structure B	2 scales, fine 0
Scale structure C	1 scale, fine 1, multi-grid
Scale structure D	2 scales, fine 1
Scale structure E	1 scale, fine 0, multi-grid

Table 2.2: Investigated scale structures.

poral candidate may come to lie outside the object in which the current block is located.

Candidate structures 1 and 2 are intended to determine the performance of simple candidate structures that resemble the usual 3DRS structure. Candidate structures 5 and 6, on the other hand, are intended to determine the performance of very complex candidate structures with many candidates.

Candidate structures 3, 4, 7, 8 and 9 are quite complex in their design. The goal of candidate structures 3 and 7 is to speed up the convergence. Candidate structure 3 includes a candidate which least resembles the spatial candidate S, with respect to its angle, whereas candidate structure 7 adds the longest vector which is computed by comparing the sum $|v_x| + |v_y|$ of the absolute value of the vector components v_x, v_y. Candidate structures 4 and 8 choose different types of candidates (8) or a different location of the candidates (4) depending on the ratio between the block size and the scale dimension dim(block)/dim(scale). Candidate structure 9 includes a motion-compensated candidate by projecting the motion vectors found in the previous scan to the new block locations in the current image.

2.3.3 Scans

In our experiments, the motion estimation scans are performed in a meandering manner either once or twice for each scale of a designed motion estimator.

2.3.4 Scale parameter sets

Different scale structures were selected by varying the values of **fine**, **coarse**, $\mathbf{sf_w}$ and $\mathbf{sf_h}$. We chose simple structures involving at most two scales on which ME is performed. The benefit of multi-scale and multi-grid ME in comparison with 3DRS ME on a down-scaled version of the input image were investigated. The multi-scale estimators are B and D in Table 2.2 (2 scales), where B performs the last ME step on the full resolution and D on a down-scaled version. The multi-grid estimators are C and E in Table 2.2 (1 scale), where C performs ME only on a down-scaled image and E on the full resolution. Finally, the 3DRS ME on a down-scaled version of the input is described by scale structure A in Table 2.2.

The investigated parameter settings of the scale structures shown in Table 2.2 are given in Table 2.3.

fine	**coarse**	$\mathbf{sf_w}$	$\mathbf{sf_h}$	Scale structure
1	1	2	2	A
1	1	4	4	A
1	1	8	8	A
1	1	2	4	A
0	1	2	2	B
0	1	4	4	B
1	2	2,1	2,1	C
1	2	4,1	4,1	C
1	2	8,1	8,1	C
1	2	2,1	4,1	C
1	2	2,2	2,2	D
1	2	2,4	2,4	D
0	1	1	1	E

Table 2.3: Parameter settings regarding scale.

2.3.5 Block sizes

For each row in Table 2.3, the block width and block height are selected from the set of possible block sizes $\{2, 4, 8, 16, 32, 64\}$. Non-square blocks are included as well, however only when the block width is twice as large as the block height. When the last ME scan is performed on the full resolution image (**fine** = 0), the block width and height for scale 0 are chosen to be either 2, 4 or 8.

2.3.6 Example

An example of a hierarchical motion estimator configuration is given in Table 2.4. The candidate and scale structures and block sizes have been explained in Section 2.3.2, Section 2.3.3, Section 2.3.4, and Section 2.3.5, respectively.

2.4 Quantitative analysis

In this section, the performance of the different motion estimator (ME) parameter settings will be evaluated for the application of picture rate conversion (see [148], Chapter 4). Therefore, we selected ten Full HD test sequences (see Fig. 2.4) with a

Candidate structure	Scale structure	$\mathbf{sf_w}$	$\mathbf{sf_h}$	$\mathbf{blk_w}$	$\mathbf{blk_h}$	**scan**
2S1TC	C	4,1	4,1	8, 2, 8	8, 2, 8	2

Table 2.4: Multi-grid motion estimator configuration example.

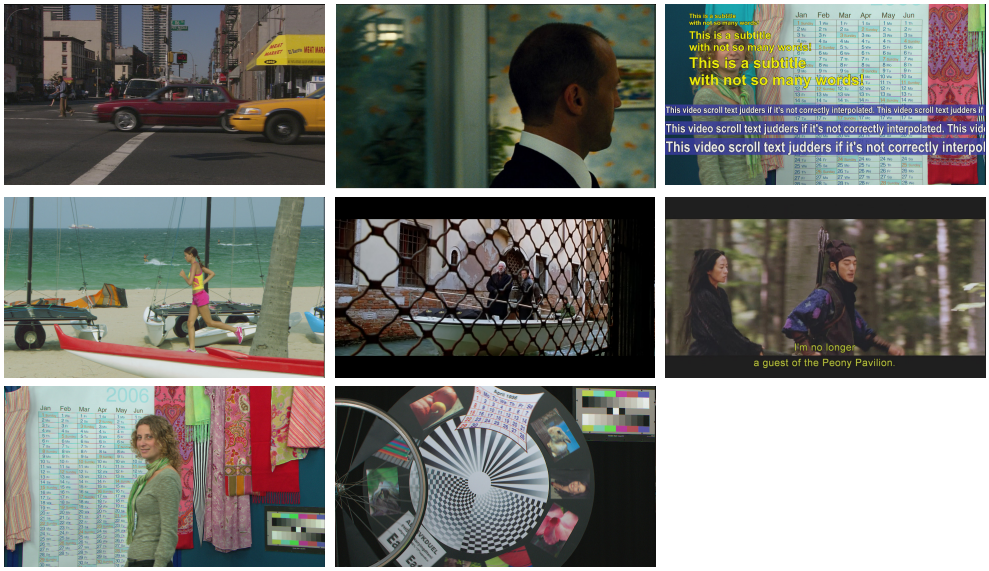


Figure 2.4: Test sequences used for the quantitative evaluation. Note that the bottom row is reused for still images and de-interlacing with a typical de-interlacer, e.g. [149].

duration of 3 frames that address common challenges in ME, such as several layers with different motion, repetitive structures, small objects, subtitles and ticker tapes, de-interlaced images with typical de-interlacers of average quality (e.g. [149]), large motion, and occlusion areas. To ensure a satisfactory performance with less challenging test material, we also included fairly straightforward sequences for ME as well as a repeated still image. We expect a well-performing ME to have a good average performance for all challenges. For individual challenges, we acknowledge that other ME parameter settings may render a better result, however, the objective in the ME design for picture rate conversion remains a good overall performance. Therefore, the average performance over all test sequences is compared. In order to analyze the behavior of the motion estimators with respect to convergence speed and steady state performance, two motion vector initializations are chosen as described in Section 2.4.1. The objective measures used to evaluate the motion estimator performance are introduced in Section 2.4.2, followed by the evaluation itself in Section 2.4.3.

2.4.1 Motion vector initialization

In order to examine the different motion estimators with respect to the convergence speed of the motion vectors and regarding their performance in the steady state, when the motion field is already converged, two different initializations are chosen. Firstly, for evaluating the convergence speed, a zero vector initialization is used. Such unconverged states occur frequently, not only in scene changes but, more importantly,

when a tracked object reappears from behind an occluding area or when accelerations and irregular motions are involved (e.g., up and downwards moving head of walking person). Secondly, an initialization with converged motion vectors is performed. To save computation time for the analysis, the motion vectors used for the second initialization are computed with a fixed multi-grid ME of which the parameter settings are given in Table 2.4.

2.4.2 Performance measures

Two fundamental characteristics are recognized as the basis of ME design: The brightness constancy assumption when the true motion is found and the smoothness constraints to enforce consistent motion fields within a moving object. The trade-off between smoothness terms and brightness constancy in the form of luminance comparisons has already become apparent in the early optical flow advances [150]. The metrics developed for high-performance ME methods for retiming show comparable features and the known trade-off between prediction accuracy and spatial motion field consistency. It is recognized in [136] and [143], that accurate predictions at a highest possible consistency are necessary for a satisfactory viewing experience. Relevant metrics addressing the temporal continuity and spatial consistency of the motion vectors are documented in [136] and [143]. The prediction accuracy and temporal continuity are quantitatively assessed with the 'M2SE' [143],

$$\mathrm{M2SE}(n) = \frac{1}{n_h \cdot n_w} \sum_{\vec{x} \in W} (F_o(\vec{x}, n) - F_i(\vec{x}, n))^2, \tag{2.5}$$

and the spatial inconsistency measure 'SI' is based on [143],

$$\mathrm{SI}(n) = \sum_{\vec{x_b} \in W_b} \sum_{\substack{k=-1 \\ l=-1}}^{1} \left(\frac{|\Delta_x(\vec{x_b}, k, l, n)| + |\Delta_y(\vec{x_b}, k, l, n)|}{8 * N_b} \right), \tag{2.6}$$

where n_h and n_w are the image height and width in pixels, respectively, W is the set of all the pixels in the entire image, $F_o(\vec{x}, n)$ the luminance of the original image at position $\vec{x}$ and at the temporal position n. F_i is the motion compensated average of frames $n-1$ and $n+1$ by applying the vectors estimated for frame n, $\vec{x_b}$ the position of the block b among the set of all the blocks W_b in the entire image, N_b the number of blocks in an image and

$$\Delta_x(\vec{x_b}, k, l, n) = d_x(\vec{x_b}, n) - d_x(\vec{x_b} + \begin{pmatrix} k \\ l \end{pmatrix}, n), \tag{2.7}$$

$$\Delta_y(\vec{x_b}, k, l, n) = d_y(\vec{x_b}, n) - d_y(\vec{x_b} + \begin{pmatrix} k \\ l \end{pmatrix}, n), \tag{2.8}$$

where d_x and d_y are the computed motion vectors. Different block sizes in the SI metric, e.g., 8×8 pixel blocks vs. 1×1 pixel blocks, return different results due to the metric bias towards larger motion vector blocks, thus appropriate block dimensions should be chosen for the set of MEs one would like to compare (8×8 pixel blocks in this paper).

The PSNR depends on the number of bits NB used for representing the video data and is calculated from the M2SE ($\mathrm{PSNR}(n) = 10 \cdot \log_{10}((2^{\mathrm{NB}} - 1)^2 / \mathrm{M2SE}(n))$). It measures how well the interpolation result corresponds to true motion using temporally extrapolated motion vectors, whereas the SI indicates the spatial smoothness of the computed motion field. The motion field and interpolated images are evaluated after performing ME on the second image of the input sequence since the pixels from a previous image are included in the M2SE computation. Note that, for these measures, ME is performed at the original image position and not at the interpolated position. Thus, a motion vector is assigned to each occurring element in the original image rendering it unlikely to miss small objects which may be the case when ME is performed on the interpolated position.

All the motion vectors are computed without applying any post processing such as block erosion [151] in order to facilitate an easier analysis of the results. This is assumed correct because the SI and PSNR measures are expected to indicate the same tendency and ranking of the MEs with and without applying block erosion.

2.4.3 Performance evaluation

Since both a high PSNR as well as a consistent motion field are characteristics of a good ME, a *PSNR - Consistency* plot as shown in Fig. 2.5 is introduced as a means to capture the achieved PSNR performance in relation to the consistency of the motion field. The inverse mean of the PSNR and the mean inconsistency values (SI) are plotted in the following sections by computing the average performance of all parameter setting combinations with regard to the different test sequences. The optimal ME with a high PSNR and a low inconsistency is located in the bottom left corner. A ME which surpasses all the others in one regard (either consistency or PSNR) is called a Pareto-optimal or an 'optimal' motion estimator. The set of optimal MEs lies on the 'optimal trade-off curve' as described by [152].

The PSNR-Consistency graphs in Fig. 2.5 and Fig. 2.6 depict the metric results of 13320 hierarchical MEs (6660 MEs in the steady state and 6660 MEs in the unconverged state) which are created based on all possible parameter combinations (i.e., varying candidate structures, scale structures, block sizes and scans) described in Section 2.3. For each ME, the average performance with respect to the different test sequences was computed. A wide spread of the MEs in the unconverged state is visible in Fig. 2.5. The best MEs lie close to the optimal trade-off curve. Therefore, the optimal contour lines of the hierarchical MEs are depicted in Fig. 2.6 where a compromise between PSNR and consistency performance is attained. It is visible that an improvement in both motion field consistency and PSNR is achievable for a

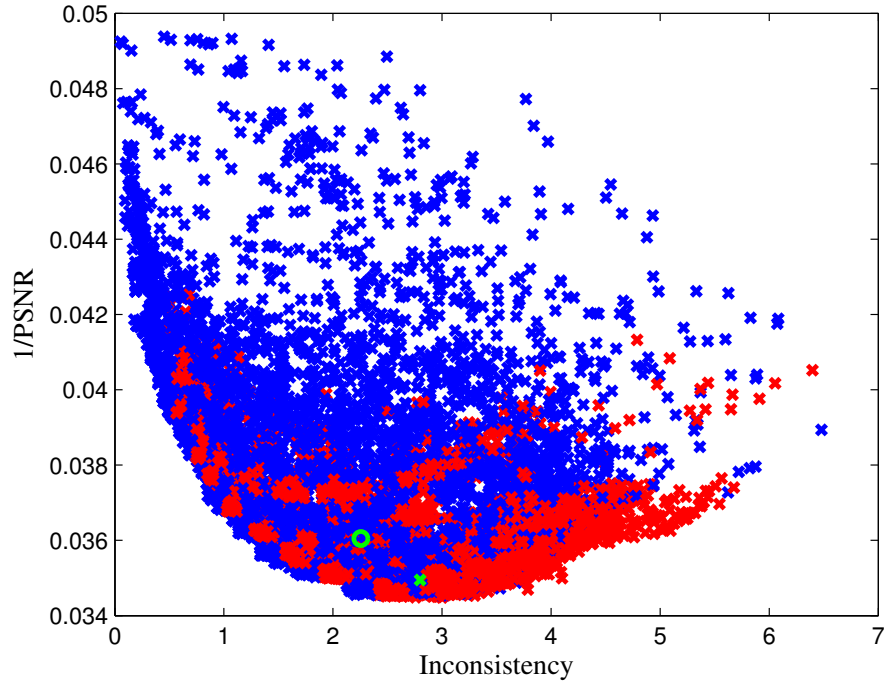


Figure 2.5: PSNR-Consistency trade-off graph: Steady state (converged): hierarchical MEs (red), 3DRS ME (green x); Unconverged state: hierarchical MEs (blue), 3DRS ME (green o).

hierarchical ME with regard to the traditional 3DRS ME.

The contour lines in Fig. 2.6 indicate often a better consistency performance in the unconverged than in the steady state. Note that the unconverged state denotes merely that the motion vector initialization was chosen to be zero but does not exclude the fact that a converged motion field may result already in the second image. This was the case for the MEs on the contour line. Furthermore, the lower consistency in the steady state might be related to the fact that a converged default motion field is used in the initialization which is not computed with the tested ME but with the one given in Section 2.4.1.

Since it is not evident which part of the optimal contour line a good visual quality corresponds to, seven optimal MEs were chosen for an initial analysis. Their characteristics are summarized in Table 2.5.

Along the curve of Fig. 2.6, from low to high PSNR up to Opt. ME 6, the following motion field improvements are observed. The related picture rate conversion benefits are confirmed when playing back the interpolated sequence.

- Improved spatial consistency of the motion field

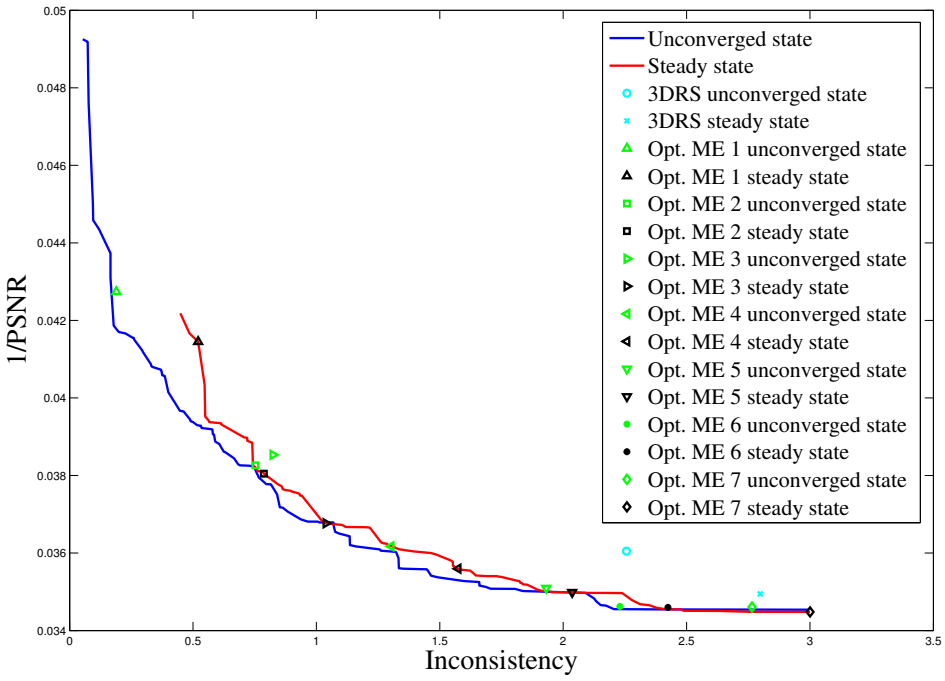
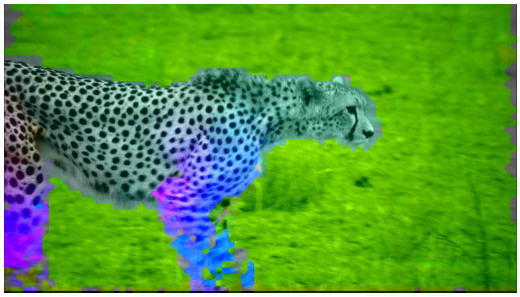


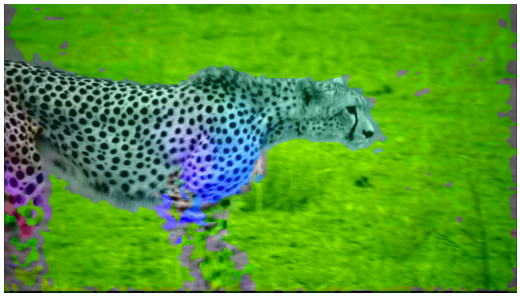
Figure 2.6: Contour lines of optimal hierarchical MEs in steady state (red) and unconverged state (blue); 3DRS and 7 optimal hierarchical MEs are indicated as well.

	Candidate structure	Scale structure	sf_w	sf_h	blk_w	blk_h	scan
Opt. ME 1	2Unlike	D	2, 2	2, 2	8, 64, 32	8, 64, 32	2
Opt. ME 2	5H1TC	D	2, 4	2, 4	8, 64, 16	8, 32, 8	2
Opt. ME 3	2S1TCMotComp	C	2,1	2,1	8, 32, 32	8, 16, 32	2
Opt. ME 4	2S1TC	C	2,1	2,1	8, 16, 64	8, 8, 32	2
Opt. ME 5	4S2T	D	2, 2	2, 2	8, 8, 8	8, 8, 4	2
Opt. ME 6	2S1TC	D	2, 2	2, 2	8, 8, 16	8, 4, 8	2
Opt. ME 7	4S5T	B	4	4	8, 4	8, 4	2

Table 2.5: Selected MEs on optimal contour line. All of them incorporate the zero vector as an additional candidate.



(a) Steady state motion field, Opt. ME 5



(b) Steady state motion field, 3DRS

(c) Unconverged motion field, Opt. ME 5

(d) Unconverged motion field, 3DRS

Figure 2.7: Motion field visualized with color overlay of 3DRS and Opt. ME 5 in example sequence chosen for subjective assessment.

- Better alignment of motion vectors with the edge of moving objects, resulting in reduced artifacts in occlusion regions
- Improved temporal consistency of the motion field, resulting in reduced flickering
- Higher convergence speed (for the steep part of the contour line)

When the metric results indicate a high spatial consistency of the motion field, there are large temporal motion field inconsistencies and artifacts due to erroneous consistent motion vectors across the motion edges. Opt. ME 1-3 show indeed spatially consistent vector fields (e.g., large zero vector areas) but their quality is unacceptable due to the produced local judder when played back. Along the contour line, an improvement of the motion vectors regarding the object alignment (clearly better with Opt. ME 4) is visible which causes less occlusion artifacts. The motion field is temporally still quite inconsistent which produces flickering and visibly varying artifacts at motion boundaries. Overall, Opt. ME 6 shows the best visual quality. The hierarchical MEs Opt. ME 5 and Opt. ME 6 obtain a similar PSNR value as 3DRS but a higher consistency measure for both the unconverged and steady states. When comparing 3DRS with Opt. ME 5 (Fig. 2.7(a) and Fig. 2.7(b)) in the steady state, a smoother motion field (encoded in color) is clearly visible with Opt. ME 5 in the background and in the legs of the leopard. For the unconverged state, the motion fields obtained by performing ME between the second and the third image are shown in Fig. 2.7(c) and Fig. 2.7(d). The consistency increase and faster convergence of Opt. ME 5 in comparison with 3DRS is apparent as well. This corresponds well with the expected added value of the hierarchical candidate with respect to larger search ranges and the rare selection of false local minima. The performance degrades when going beyond Opt. ME 6. The motion field of Opt. ME 7 is perceived as noticeably more inconsistent which leads to disturbing flickering artifacts. The relevance of the spatial inconsistency metric is confirmed since a high PSNR at the cost of consistency is not preferred.

We expect the performance gain achieved with the addition of hierarchical layers to stagnate when more than a couple of layers are applied. In order to get a visual impression of the qualitative contribution when more than 2 scales are used, the visually best performing ME, Opt. ME 6, was extended to a ME with 5 scales. An informal subjective evaluation showed hardly any visible differences in terms of the listed motion field improvements mentioned earlier. This is in correspondence with the quantitative improvements of 0% in PSNR and 7% in SI, indicating that using more than 2 scales has only minor performance benefits.

2.5 Detailed parameter analysis

In order to carry out a more representative analysis and to allow for slight imperfections in the metrics, also the MEs within a particular distance from the optimal

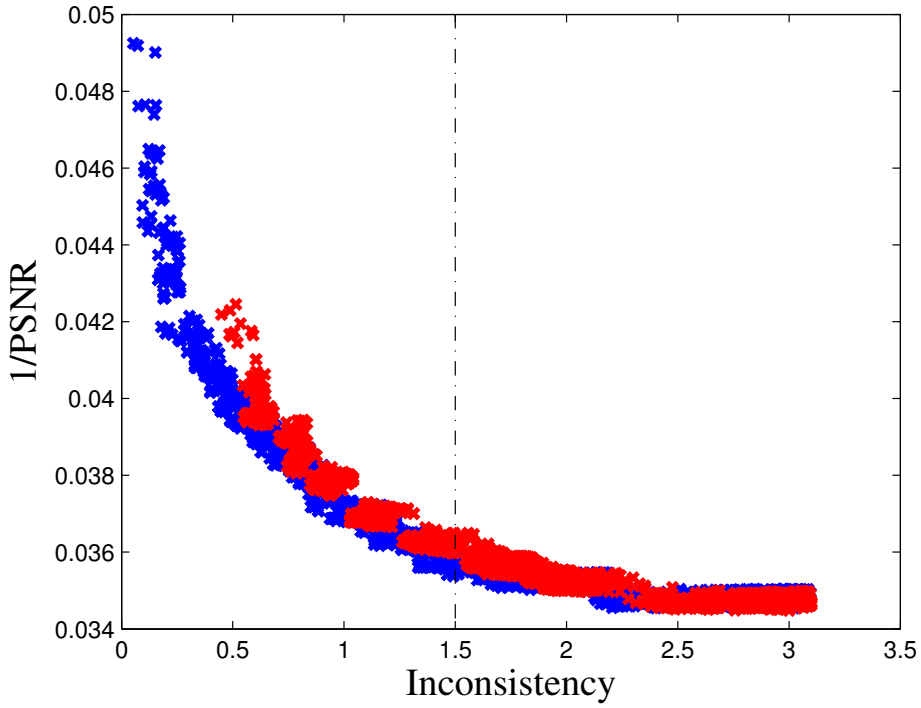


Figure 2.8: Range of optimal hierarchical MEs in steady state (red) and unconverged state (blue).

contour lines are investigated. For the 16-bit HD data, the considered range was chosen to be $\delta(1/\text{PSNR}) = 0.0005$ and $\delta(\text{SI}) = 0.1$. The MEs within this range are shown in Fig. 2.8. The contour lines are further divided into two segments as shown with the dashed line. When comparing common settings among the MEs, it is chosen to take into account all the MEs on the right hand side of the dashed line where the PSNR hardly decreases and room for a rather large improvement in terms of consistency is given.

Section 2.5.1 discusses the candidate structures, Section 2.5.2 the amount of scans, Section 2.5.3 investigates the scale parameter sets, and Section 2.5.4 the block sizes. The resulting optimal hierarchical MEs are summarized in Section 2.5.5.

2.5.1 Candidate structures

An overview of the performance of the optimal MEs regarding the candidate structures is given in Fig. 2.9. The contour lines of all the optimal MEs for each candidate structure are given in Fig. 2.9. The description of the different candidate structures can be found in Fig. 2.3. For comparison, the performance of 3DRS is illustrated as well. It is clearly visible that the two candidate structures with the least (4) can-

didates (1S1TC and 2Unlike) perform worst. This indicates the necessary number of prediction vectors for a satisfactory performance. The importance of spatial predictors on a large resolution with small block sizes is apparent in the suboptimal performance of the candidate structure 5H1TC. The optimal MEs of the other candidate structures in Fig. 2.9 achieve a more or less similar metric result. Especially for the unconverged state, a significant increase in consistency (around 1) and PSNR (around 1.2 dB) compared with 3DRS is found. In the steady state there is mainly room for a consistency increase.

ME candidate structures that yield more often an optimal ME are preferred as they are assumed less sensitive to varying settings than other candidate structures and thus more robust. The goal is to find optimum settings for both the unconverged and the steady state. For a practical implementation in real-time applications, however, it can be useful to discriminate between the unconverged and the steady state and choose the best candidate structure for each state. With respect to the range of optimal MEs, the distribution of the candidate structures is given in Fig. 2.10 (the steady state case is comparable to the unconverged state). The numbers on the x-axis correspond to the candidate structure numbers in Fig. 2.3 and the y-axis to the ME count. For both states, the distribution in the interesting segment indicates that a good performance can be achieved with the candidate structures 2, 4, 5, 6 and 9. Hence, it may be interesting to use the most straightforward candidate structure, 2, as it contains the least number of candidates and does not require a complex implementation (for candidate structure 9 which only involves one more candidate a higher computational complexity is expected due to the motion compensated candidate). These results suggest that a minimum number of prediction vectors is needed for a satisfactory performance and that most of the necessary information is contained in this minimum candidate set.

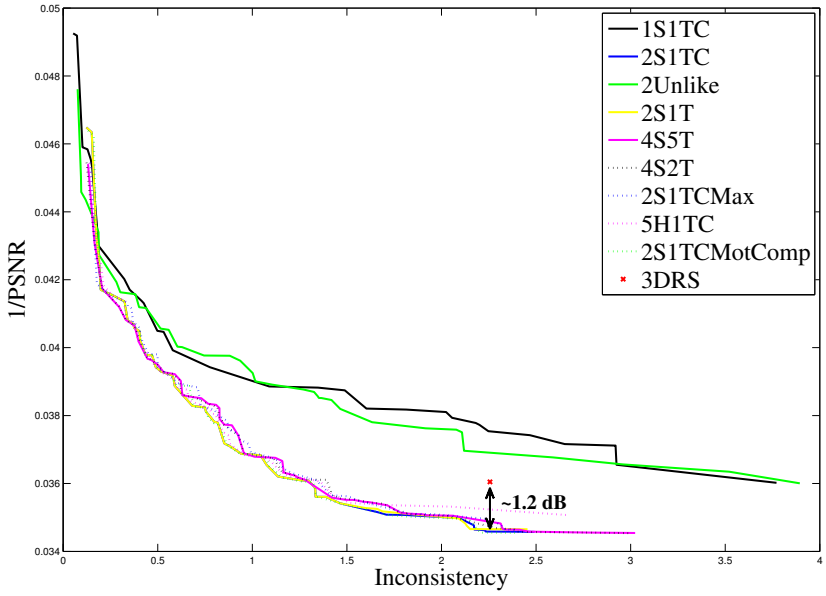
2.5.2 Scans

Performing two ME scans per scale generally renders a better overall performance than only one scan per scale (occurrence rate of 67% vs. 33% respectively) since good motion vectors found close to object edges can be refined and propagated to other parts within the object. Two scans with an occurrence rate of 81% in the unconverged state are found to be particularly useful for a fast convergence.

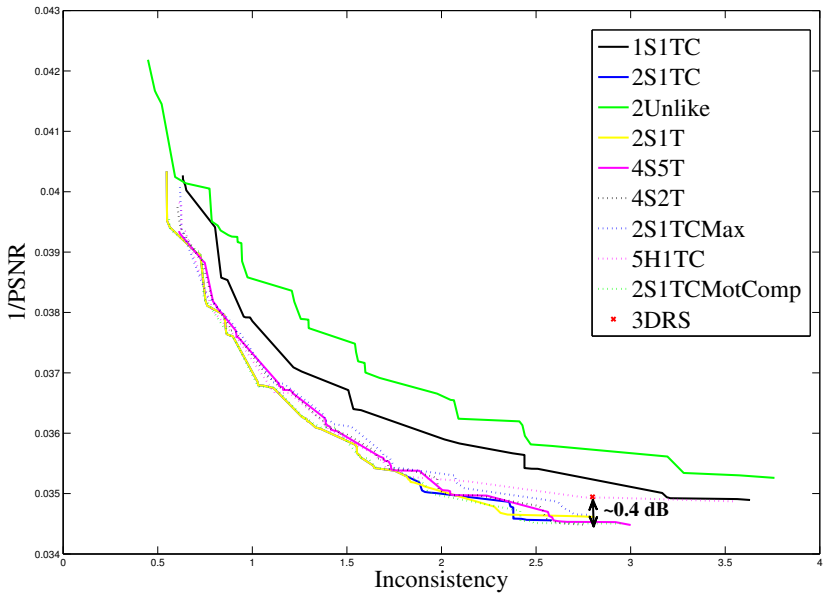
Note that the total number of scans performed when computing the resulting motion field of one image is dependent on the number of scans per scale and the number of scales or block grids used. For e.g., the multi-grid approach with 1 scan, the total number of scans is equivalent to the case with the scale structure A (indicated in Table 2.2) and 2 scans.

2.5.3 Scale parameter sets

In this section, the scale structures and scaling factors are discussed. Particularly the unconverged state shows a clear discrepancy between the scale structures. The contour



(a) Unconverged state



(b) Steady state

Figure 2.9: Contour lines of optimal MEs in unconverged (a) and steady (b) state for different candidate structures.

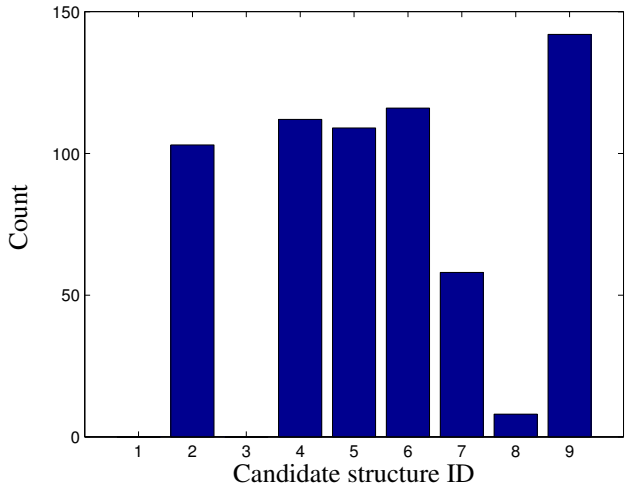


Figure 2.10: Distribution of the candidate structures among the range of optimal MEs in the unconverged state.

lines in Fig. 2.11 depict the optimal quantitative performances of the MEs for each scale structure in the unconverged state. Particularly the scale structures '2 scales, fine 0' and '1 scale, fine 0, multi-grid' show a significant decrease in consistency and/or PSNR suggesting suboptimal high-frequency content (e.g., noise) in the full resolution image. With the selected test sequences addressing natural content, removing the higher frequencies by downscaling the input image thus does not show any visible drawbacks. In the steady state there is no noticeable difference among the other three scale structures. However, when the motion vector is not yet converged the multi-scale/multi-grid approach ('1 scale, fine 1, multi-grid', '2 scales, fine 1') seems beneficial for both PSNR and consistency which confirms the hypothesis of the added value of an hierarchical candidate.

In the analysis of the distribution of the five scale structures (analogously to the candidate structures in Fig. 2.10) we found that the scale structures C and D are the most represented groups (84%). Scale structure D occurs around 35% more often than C, thus using two different scales seems to be of advantage.

When analyzing the distributions of the scaling factors the following is observed. The finest scale on which ME is performed is dominated by the scaling factor 2. The full resolution with scaling factor 1 is rarely chosen. Apparently, the highest frequencies (such as noise) in the image do not contribute to a more accurate ME. In the multi-scale approach, the coarse scale which is added for fast convergence and consistency shows, as expected, higher scaling factors (4 and 8 occur approximately equally often).

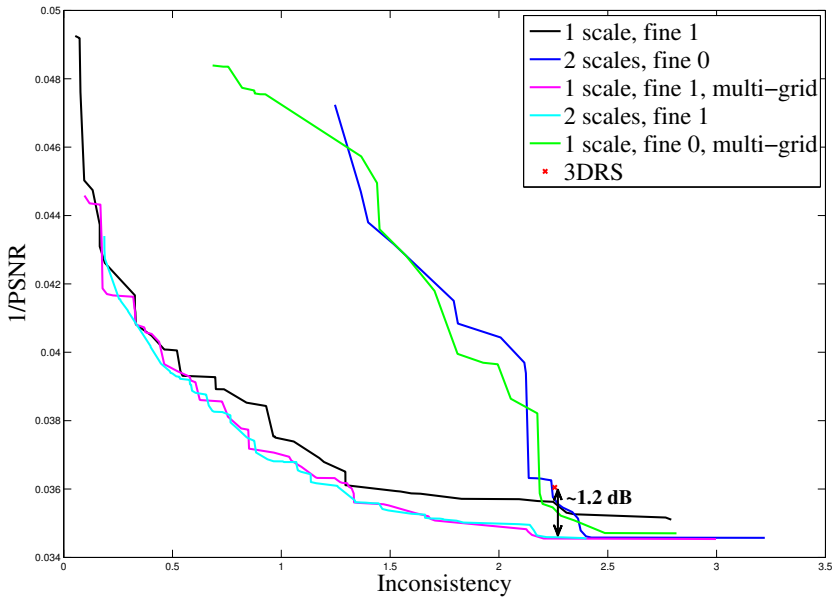


Figure 2.11: Contour lines of optimal MEs in unconverged state for different scale structures.

2.5.4 Block sizes

We assume that larger blocks and/or coarser scales would improve the convergence speed and large object area smoothness, and smaller blocks on the fine scale would serve as a refinement of the motion field obtained on the coarse scale. For the fine scale in the context of HD sequences, block dimensions in the neighborhood of 8×8 blocks on the full resolution would be plausible since experience on SD content has shown that 8×8 blocks are a good trade-off between PSNR and SI [153].

The distribution of the block dimensions for multi-scale MEs is illustrated in Fig. 2.12. The data reveals that the block sizes of MEs using 1 scale are similarly distributed as the ones of the fine level of the multi-scale MEs (in order to avoid repetition and limit the figures, only the graphs corresponding to the multi-scale case are shown). The dominant width and height dimensions in the well performing segment range from [8,32] and [4,16] respectively. The block sizes of the multi-grid motion estimators which are included in the 1-scale case are more concentrated than the ones of the multi-scale MEs (see the large spread of the coarse level block sizes) indicating that more similar block dimensions are selected when the same scale is re-used. On average, the block width and height of the selected MEs on the coarse scale range from [32, 128] and [16,64], respectively. Based on these observations, we conclude that multi-scale MEs make use of the varying frequency content and are more robust when different block sizes are used.

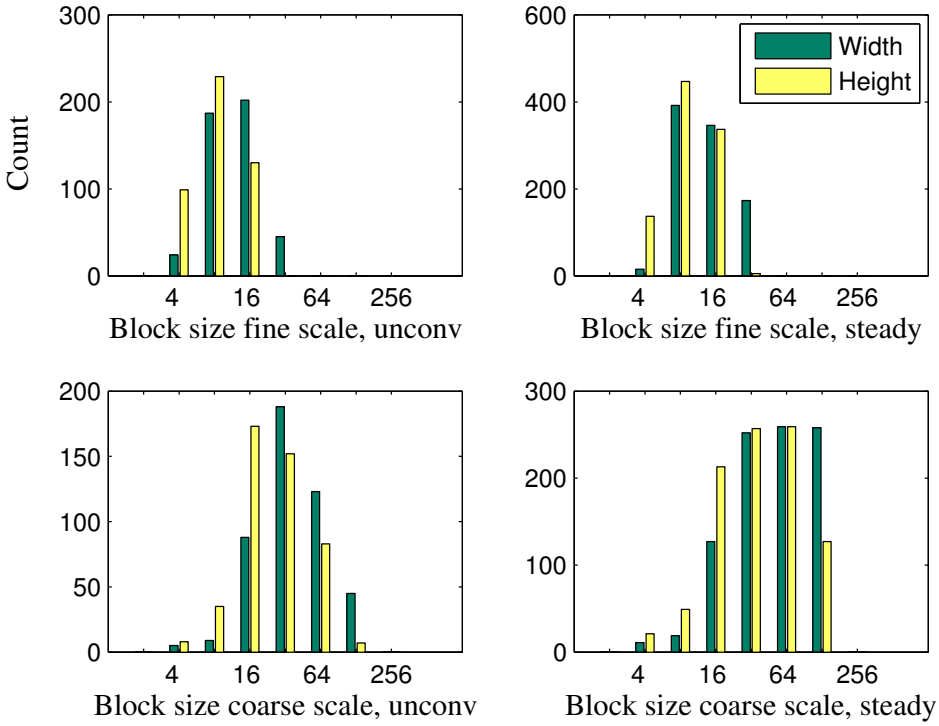


Figure 2.12: Range of optimal MEs using 2 scales. Distribution of equivalent block sizes for full resolution image.

2.5.5 Optimal hierarchical motion estimators

Based on the parameter analysis in the previous sections, we propose to employ the multi-scale MEs with candidate structure 2S1TC, scale structure D and 2 scans. An overview of the proposed parameter settings of this ME type is given in Table 2.6 where block settings, performance and complexity are rendered. The fourth row of Table 2.6 shows the mean performance for the 62 most robust MEs. The range of block width and height settings were analyzed in more detail. Therefore, their distributions were considered as probability distributions of settings for well performing MEs and their expectation value a good approximation of a robust well-performing ME given in the seventh row. The selected block sizes indicate that larger blocks are suited for HD content. When applying one of the two scale factor settings of scale structure D given in Table 2.3, the resulting ME happens to coincide with Opt. ME 6 in Figure 2.6 which was visually perceived as the most pleasing ME among the seven MEs on the contour line.

	Block width full res.	Block height full res.	mean PSNR	mean SI	# Block comparisons n_{BC}
3DRS unconv. + steady	8	8	28.17	2.53	388800
3DRS unconv.	8	8	27.74	2.26	388800
3DRS steady	8	8	28.60	2.80	388800
Range of MEs	[8,32], [32,128]	[4,16], [16,64]	28.46	2.18	[60242,963900]
Low-complexity MEs	32, 128	16, 64	27.98	1.45	60242
High-complexity MEs	8, 32	4, 16	28.81	3.01	963900
Proposed MEs	16, 64	8, 32	28.91	2.37	240975
Proposed MEs unconv.	16, 64	8, 32	28.90	2.27	240975
Proposed MEs steady	16, 64	8, 32	28.91	2.46	240975
HRNM [154] unconv.	8	8	28.02	1.87	777600
HRNM [154] steady	8	8	29.32	1.48	777600
FS [155] unconv.	16	16	25.78	15.00	1056370680
FS [155] steady	16	16	25.78	15.00	1056370680
TSS [156] unconv.	16	16	22.80	3.90	201000
TSS [156] steady	16	16	22.80	3.90	201000
OTS [157] unconv.	16	16	23.79	6.64	152271
OTS [157] steady	16	16	23.79	6.64	152271
DS [158] unconv.	16	16	23.95	7.24	292462
DS [158] steady	16	16	23.95	7.24	292462
HEXBS [159] unconv.	16	16	23.90	7.28	214186
HEXBS [159] steady	16	16	23.90	7.28	214186
MVFAST [160] unconv.	16	16	28.12	4.44	131207
MVFAST [160] steady	16	16	28.15	4.43	130824
EPZS [142] unconv.	16	16	27.84	3.62	133821
EPZS [142] steady	16	16	28.69	3.77	84137
MRST [147] unconv.	16,16,16,16	16,16,16,16	28.26	5.26	9199420
MRST [147] steady	16,16,16,16	16,16,16,16	28.60	5.16	9182275
MPMVP [137] unconv.	32,16, 8, 4	32,16, 8, 4	27.41	3.99	3125265
MPMVP [137] steady	32,16, 8, 4	32,16, 8, 4	28.13	3.80	3120024

Table 2.6: Performance and complexity analysis of proposed MEs, 3DRS and various techniques documented in literature. Block width and block height indicate the equivalent block sizes for the full resolution where the selected settings for the **fine** and **course** scales can be a range ([..]) of values.

2.6 Results

We measured the quantitative performance and computational complexity of various MEs. The computational complexity is expressed in the number of block comparisons n_{BC}. For the proposed MEs, this can also be derived from

$$n_{\mathrm{BC}} = \sum_{i \in W_{\mathrm{scales}}} \frac{n_h \cdot n_w \cdot \mathrm{scan}(i) \cdot n_{\mathrm{cand}}}{\mathrm{blk}_{\mathrm{w,fullRes}}(i) \cdot \mathrm{blk}_{\mathrm{h,fullRes}}(i)}, \tag{2.9}$$

where W_{scales} is the set of all the used scales, $\mathrm{scan}(i)$ the number of scans on scale i, n_{cand} the number of motion vector candidates ($n_{\mathrm{cand}} = 7$ for the hierarchical MEs due to the addition of the zero vector candidate), $\mathrm{blk}_{\mathrm{w,fullRes}}$ and $\mathrm{blk}_{\mathrm{h,fullRes}}$ the width and height of the equivalent block sizes for the full resolution scale.

Table 2.6 gives an overview of the SI and M2SE-PSNR values and the number of block comparisons of the different recursive-search MEs that are proposed, as well as the benchmark results from several methods described in literature. These include full-search (FS) and reduced-search pattern based methods, i.e., three-step-search (TSS) [156], one-at-a-time search (OTS) [157], diamond search (DS) [158] and hexagon-based-search (HEXBS) [159], as well as algorithms based on spatio-temporal predictors, i.e., the predictive (zonal) search methods MVFAST [160] and EPZS [142], the recursive search methods 3DRS [135] and HRNM [154], and combined hierarchical-predictive methods, i.e., the MRST-method proposed in [147] and MPMVP from [137]. In the steady state, we simulate the convergence mode for these methods by iterating the corresponding MEs ten times on the first image. Note that the M2SE-PSNR metric favors 'true' motion, i.e., MEs with a better vector field consistency can outperform a full-search method. Furthermore, all methods from literature were adapted and tested with smaller block dimensions (e.g., 8×8), however, no improvement in PSNR and SI was observed.

In comparison with standard 3DRS with two scans, the proposed hierarchical MEs achieve a complexity reduction of 38% while outperforming 3DRS on average by 0.7 dB. This holds particularly for the unconverged state with an improvement of more than 1 dB and 7% in consistency. Even the sophisticated HRNM ME [154], with a significantly higher complexity due to 3-picture estimates, is surpassed in the unconverged state (PSNR difference of 0.9 dB). However, in the steady state, HRNM shows a clearly better performance than any of the hierarchical MEs. From these observations we conclude that for the unconverged state, a combination of the hierarchical approach and HRNM may be beneficial for both computational complexity and performance.

The benchmark further shows that the non-predictive (reduced-)search methods FS, TSS, OTS, DS, and HEXBS are generally unsuitable for picture rate conversion. As these methods purely optimize for minimal 'residue' in the match criterion, they produce highly inconsistent vector fields (with PSNR values smaller than 24 and/or SI values larger than 4). The predictive search methods generally perform better, as they (implicitly) enforce vector field consistency, with the methods EPZS and MRST

achieving the best steady-state PSNR performance (slightly below the proposed MEs). Among these, when taking the computational complexity into account, EPZS is identified as the ME achieving the best compromise between performance and complexity. Yet, its spatial inconsistency is more than 50% larger than the SI values of the proposed MEs, and this has a large impact on the perceived picture quality. We conclude from these results that the proposed hierarchical MEs are superior to multiple existing techniques as well as standard 3DRS with regard to combined PSNR/SI performance at a low computational complexity.

2.7 Conclusion

Hierarchical ME promises fast convergence of motion vectors, a high motion field consistency and a small M2SE error. In this paper, we introduced the concept of hierarchical ME to 3D-Recursive Search (3DRS), and we performed a design-space exploration of the extensive parameter space to provide insights into the importance and influence of individual parameters. In particular, a quantitative analysis was performed by determining the PSNR and Spatial Inconsistency (SI) of 13320 hierarchical MEs to show the trade-off between spatial consistency and match quality.

In general, we found that applying the hierarchical approach to 3DRS does not require complex candidate structures in order to perform well. In fact, straightforward candidate structures having relatively few candidates already offer a good overall performance, i.e., one that is close to the optimal trade-off curve. Furthermore, we identified that multi-scale MEs are amongst the best performing hierarchical MEs, closely followed by multi-grid MEs on down-scaled images, with these being hindered by a lower robustness with respect to varying block sizes.

Based on the design-space exploration, a ME configuration is proposed that offers an improvement of more than 1 dB over 3DRS in the unconverged state, and of 0.7 dB on average. At the same time, the computational complexity is reduced by 38%. When benchmarking the proposed MEs to various other techniques, the results show a superior combination of PSNR/SI performance while offering a low computational complexity.

We also showed that, in comparison to a sophisticated ME approach using 3-picture estimates (HRNM), the proposed hierarchical MEs offer better results in terms of both image quality and complexity in the unconverged state. Therefore, as future work, a combination of the hierarchical approach and HRNM may be investigated to identify whether the combination offers further improvements in performance and/or computational complexity.

Bibliography

[134] G. de Haan and J. Kettenis, "System-on-Silicon for high quality display format conversion and video enhancement", in *Proc. ISCE*, Erfurt, Germany, Sep. 2002, pp. E1–E6.

[135] G. de Haan and P.W.A.C. Biezen, "Sub-pixel motion estimation with 3-D recursive search block-matching", *Signal Processing*, **vol. 6**, no. 3 pp. 229–239, June 1994.

[136] J. Wang, D. Wang, and W. Zhang, "Temporal compensated motion estimation with simple block-based prediction", *IEEE Transactions on Broadcasting*, **vol. 49**, no. 3 pp. 241–248, Sep. 2003.

[137] S.-C. Tai, Y.-R. Chen, Z.-B. Huang, and C.-C. Wang, "A Multi-Pass True Motion Estimation Scheme With Motion Vector Propagation for Frame Rate Up-Conversion Applications", *Display Technology, Journal of*, **vol. 4**, no. 2 pp. 188–197, June 2008.

[138] C.N. Cordes and G. de Haan, "Key requirements for high quality picture-rate conversion", *SID Digest of Technical Papers*, **vol. 15**, no. 2 pp. 850–853, June 2009.

[139] E.B. Bellers, "Motion Compensated Frame Rate Conversion for Motion Blur Reduction", *SID Digest of Technical Papers*, **vol. 38**, no. 1 pp. 1454–1457, May 2007.

[140] G.G.C. Lee, M.J. Wang, H.Y. Lin, D.W.C. Su, and B.Y. Lin, "Algorithm/Architecture Co-Design of 3-D Spatio–Temporal Motion Estimation for Video Coding", *IEEE Transactions on Multimedia*, **vol. 9**, no. 3 pp. 455–465, April 2007.

[141] A.M. Tourapis, O. C. Au, and M. L. Liou, "Highly efficient predictive zonal algorithms for fast block-matching motion estimation", *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 12**, no. 10 pp. 934–947, Oct. 2002.

[142] A.M. Tourapis, "Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation", *Proceedings of Visual Communications and Image Processing*, pp. 1069–79, Jan. 2002.

[143] G. de Haan, , P.W.A.C. Biezen, H. Huijgen, and O.A. Ojo, "True-motion estimation with 3-D recursive search block matching", *IEEE Trans. Circuits, Syst. Video Techn.*, pp. 368–379, Oct. 1993.

[144] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo analysis by hybrid recursive matching for real-time immersive video conferencing", *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 14**, no. 3 pp. 321–334, March 2004.

[145] R. Thoma and M. Bierling, "Motion compensating interpolation considering covered and uncovered background", *Signal Processing: Image Communication*, pp. 191–212, Feb. 1989.

[146] F. Dufaux and F. Moscheni, "Motion Estimation Techniques for Digital TV: A Review and a New Contribution", *Proc. IEEE*, pp. 858–876, June 1995.

[147] J. Chalidabhongse and C.C.J. Kuo, "Fast motion vector estimation using multiresolution-spatio-temporal correlations", *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 7**, no. 3 pp. 477–488, June 1997.

[148] G. de Haan, *Video Processing for multimedia systems*, University Press Eindhoven, 2000, ISBN 90-9014015-8.

[149] G. de Haan and E.B. Bellers, "De-interlacing of video data", *IEEE Transactions on Consumer Electronics*, **vol. 43**, no. 3 pp. 819–825, Aug. 1997.

[150] B. Horn and B. Schunck, "Determining Optical Flow", *Artificial Intelligence*, **vol. 17**, no. 1-3 pp. 185–203, 1981.

[151] G. de Haan and H. Huijgen, "Motion Estimator for TV Picture Enhancement", *Signal Processing of HDTV, III, H. Yasuda and L. Chiariglione, eds. Elseviers Science Publishers B.V.*, 1992.

[152] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, ISBN 052183378.

[153] G. de Haan, "Motion Estimation and Compensation, an integrated approach to consumer display field-rate conversion", Ph.D. thesis, Delft University of Technology, 1992.

[154] E. Bellers, J.W. van Gurp, J.G.W.M. Janssen, J.R. Braspenning, and R. Wittebrood, "Solving occlusion in Frame-Rate up-Conversion", in *Digest of the ICCE*, Jan. 2007, pp. 1–2.

[155] J.R. Jain and A.K. Jain, "Displacement Measurement and Its Application in Interframe Image Coding", *IEEE Trans. Commun.*, pp. 1799–1808, Dec. 1981.

[156] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro, "Motion-Compensated Interframe Coding for Video Conferencing", in *Proc. Nat. Telecom. Conf.*, Nov./Dec. 1981, pp. G 5.3.1–G 5.3.5.

[157] R. Srinivasan and K. Rao, "Predictive Coding Based on Efficient Motion Estimation", *Communications, IEEE Transactions on*, **vol. 33**, no. 8 pp. 888 – 896, Aug. 1985.

[158] S. Zhu and K.K. Ma, "A new diamond search algorithm for fast block-matching motionestimation", *IEEE Transactions on Image Processing*, **vol. 9**, no. 2 pp. 287–290, Feb. 2000.

[159] C. Zhu, X. Lin, and L.-P. Chau, "Hexagon-based search pattern for fast block motion estimation", *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 12**, no. 5 pp. 349–355, May 2002.

[160] P.I. Hosur and K.K. Ma, "Motion vector field adaptive fast motion estimation", in *Second International Conference on Information, Communications and Signal Processing (ICICS)*, Dec. 1999.

Chapter 3

Perception-oriented methodology for robust motion estimation design

Abstract

Optimizing a motion estimator (ME) for picture rate conversion is challenging. This is because there are many types of MEs and, within each type, many parameters, which makes subjective assessment of all the alternatives impractical. To solve this problem, we propose an automatic design methodology that provides 'well-performing MEs' from the multitude of options. Moreover, we prove that applying this methodology results in subjectively pleasing quality of the upconverted video, even while our objective performance metrics are necessarily suboptimal. This proof involved a user rating of 93 MEs in 3 video sequences. The 93 MEs were systematically selected from a total of 7000 ME alternatives. The proposed methodology may provide an inspiration for similar tough multi-dimensional optimization tasks with unreliable metrics.

This chapter is published as: A. Heinrich, R.J. van der Vleuten, G. de Haan; Perception-oriented methodology for robust motion estimation design, *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp., June 2014.

3.1 Introduction

Motion estimation (ME) is an essential part of picture rate conversion methods that are applied to eliminate film judder, reduce flicker and eliminate blur in high-end televisions [161]. Because of the increasing spatial resolution (from SD to HD, Full HD and Ultra HD) and picture rates (from 24 fps to more than 200 fps) of video shown on those televisions, as well as the increasing size and quality of the television displays, there is continuous pressure to improve the quality of ME algorithms while maintaining acceptable computational complexity.

Optimizing a motion estimator (ME) for picture rate conversion is challenging. This is because there are many types of MEs and, within each type, many parameters, which makes *subjective assessment* of alternatives impractical.

For the application of picture rate conversion, various *objective metrics* have been developed and employed to evaluate the performance of a ME, e.g [161–167]. Unfortunately, these performance measures represent a *necessarily suboptimal* approach to reflect the perceived *subjective image quality*.

In this paper we propose a robust ME design methodology that, while applying such suboptimal metrics, can still identify good MEs automatically and identify the MEs with a consistently good performance for a multitude of challenges. Moreover, we present a user study to support this perception-oriented ME-design methodology and its assumptions. Users rated 93 MEs in 3 video sequences. The 93 MEs were systematically selected from a total of 7000 ME alternatives. The current paper can be seen as an extension of earlier work described in [168] and [169]. The proposed methodology may provide an inspiration for similar tough multi-dimensional optimization tasks with unreliable metrics.

In Section 3.2, we present the proposed robust ME design methodology followed by video data analysis results on ten test sequences supporting the robustness claims made in [168] and [169]. Section 3.3 describes the user perception test to confirm and improve the proposed design methodology. The results are presented in Section 3.4 and discussed in Section 3.5 where the perceived quality of MEs is incorporated in an improved ME design methodology. Conclusions are drawn in Section 3.6.

3.2 Proposed motion estimation design methodology and robustness analysis

This section presents the proposed motion estimation methodology and related robustness experiments.

3.2.1 Proposed motion estimation design methodology

In order to automatically identify parameter settings of robust MEs for upconverting video sequences, we present a methodology that can successfully deal with performance measures that are suboptimal in the sense that they do not fully reflect the

perceived video quality. A three-step approach is suggested where, first, the variety of conditions under which the MEs should perform well is defined and appropriate test data is selected. Second, a contour line or trade-off curve illustrates the achieved compromise between the motion vector prediction accuracy and consistency (see Fig. 3.3). Third, an attractive segment is identified containing all MEs within a defined distance from an attractive section of the contour line. Histogram analysis provides the distribution of MEs within the attractive segment to identify the parameter settings of the MEs that are least sensitive to varying settings and thus most robust.

Test data selection

A ME should perform well under a majority of considered conditions for the picture rate conversion application. These conditions are included in the test data which should address ME challenges such as repetitive structures, small objects, subtitles and ticker tapes, several layers with different motion, de-interlaced images with typical de-interlacers of average quality (e.g., [170]), large motion, and occlusion areas. To ensure a satisfactory performance with less challenging test material, also fairly straightforward sequences for ME should be included, as well as repeated still images. Out of a pre-selection of 20 sequences, the ten test data sets were selected according to these criteria. Some of the 10 remaining test sequences were used for verification purposes of the results. A snapshot of each Full-HD test sequence is shown in Fig. 3.1. The two sequences shown in the bottom row are reused for the repeated still image sequence and the de-interlaced sequence. The test set should thus address the main challenges posed by the application and include standard test sequences. Additionally, each algorithm (in our case, each motion estimator type) may introduce new problem cases that are less prominent with other algorithms. For these new problem cases, additional test data is included. With this in mind, a good representation of all types of motion should be accomplished.

We expect a well-performing ME to have a good average performance for all challenges. For individual challenges, we acknowledge that other ME parameter settings may render a better result, however, the objective in the ME design for retiming video sequences remains a good overall performance. Therefore, the average performance over all test sequences is compared.

Performance measures

The chosen performance measures which the trade-off curve is dependent on are based on fundamental characteristics that are recognized as the basis of ME design: The brightness constancy assumption when the true motion is found and the smoothness constraints to enforce consistent motion fields within a moving object. The trade-off

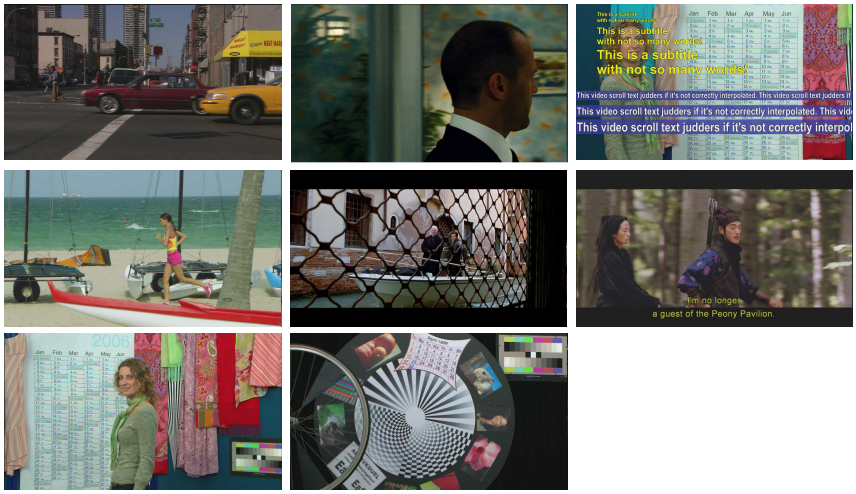


Figure 3.1: Snapshots of test sequences used for the quantitative evaluation.

between smoothness terms and brightness constancy in the form of luminance comparisons has already become apparent in the early optical flow implementations [171].

Similarly, [161, 162, 165, 166] and [167] recognize that accurate predictions at a highest possible consistency are necessary for a satisfactory viewing experience. Relevant metrics addressing the prediction accuracy, temporal continuity and spatial consistency of the motion vectors are documented in [161, 162, 165, 166] and [167]. The prediction accuracy and temporal continuity are quantitatively assessed with the 'M2SE' [162, 167],

$$\mathrm{M2SE}(n) = \frac{1}{n_h \cdot n_w} \sum_{\vec{x} \in W} (F_o(\vec{x}, n) - F_i(\vec{x}, n))^2, \tag{3.1}$$

and the spatial inconsistency measure 'SI' based on [167],

$$\mathrm{SI}(n) = \sum_{\vec{x_b} \in W_b} \sum_{\substack{k=-1 \\ l=-1}}^{1} \left(\frac{|\Delta_x(\vec{x_b}, k, l, n)| + |\Delta_y(\vec{x_b}, k, l, n)|}{8 * N_b} \right), \tag{3.2}$$

where n_h and n_w are the image height and width in pixels, respectively, W is the set of all the pixels in the entire image, $F_o(\vec{x}, n)$ the luminance of the original image at position $\vec{x}$ and at the temporal position n. F_i is the motion compensated average of frames $n-1$ and $n+1$ by applying the vectors estimated for frame n, $\vec{x_b}$ the position of the block b among the set of all the blocks W_b in the entire image, N_b the number of blocks in an image and

$$\Delta_x(\vec{x_b}, k, l, n) = d_x(\vec{x_b}, n) - d_x(\vec{x_b} + \begin{pmatrix} k \\ l \end{pmatrix}, n), \tag{3.3}$$

$$\Delta_y(\vec{x_b}, k, l, n) = d_y(\vec{x_b}, n) - d_y(\vec{x_b} + \binom{k}{l}, n), \tag{3.4}$$

where d_x and d_y are the computed motion vectors.

The PSNR measure is calculated from the M2SE: $\mathrm{PSNR}(n) = 10 \cdot \log_{10}((2^{\mathrm{NB}} - 1)^2/\mathrm{M2SE}(n))$, where NB is the number of bits used for representing the video data.

The *PSNR - Consistency* plot as e.g. shown in Fig. 3.2 with 6660 hierarchical Recursive Search (RS) MEs [168] is introduced displaying the statistics for each ME computed over all test sequences. The *PSNR - Consistency* plot captures the achieved PSNR performance in relation to the consistency of the motion field. The inverse mean of the PSNR and the mean inconsistency values (SI) are plotted by computing the average performance of all parameter setting combinations with regard to the different test sequences. The optimal ME with a high PSNR and a low inconsistency is located in the bottom left corner. We call a ME which is not surpassed by any other ME in both regards (consistency and PSNR) an 'optimal' ME. This set of optimal MEs lies on the 'contour line' (blue line in Fig. 3.3) or 'trade-off curve' as described in [172].

Identification of robust ME settings

A subset of MEs considered 'well-performing' MEs are found close to the contour line. The selection criteria for this group of MEs are the following:

1. Well performing MEs should be located close to a so-called attractive section of the contour line,

2. the attractive section is bounded by a minimal PSNR and a maximal Inconsistency (dashed lines in Fig. 3.3),

3. the area spanned by the attractive section and the maximum ΔSI and ΔPSNR distance contains well performing MEs. This area is called attractive segment (red arrows in Fig. 3.3).

Among the MEs with a satisfactory performance (e.g., blue MEs in Fig. 3.4), the distribution of the parameter settings is analyzed and their values compared in a histogram parameter analysis. The most robust ME settings are identified by high counts in the corresponding parameter histogram. A ME parameter setting that yields more often an optimal ME is preferred as it is assumed less sensitive to varying settings of other parameters. Among the high counts the setting closest to the expectation value is selected.

The attractive segment and the corresponding attractive contour line section have been defined based on the authors' observation of different ME performances on ten Full HD test sequences.

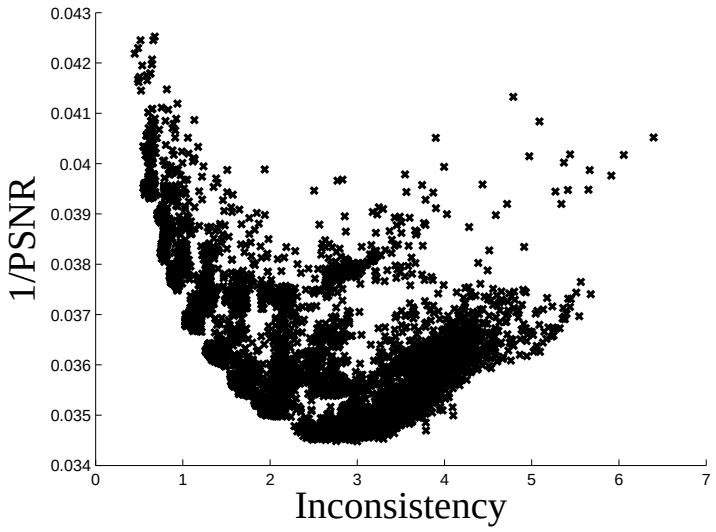


Figure 3.2: PSNR-Consistency trade-off plot of 6660 RS MEs.

3.2.2 Robustness analysis of proposed methodology

A design space exploration of thousands of MEs (resulting from combinations of 8 different parameters, each parameter with 2 to 8 different possible values) has been carried out on ten test sequences (see snapshots in Fig. 3.1). The proposed methodology has been applied to two different ME types: Hierarchical Recursive Search block matching (RS) and Phase Plane Correlation (PPC). Both are relevant motion estimators for the application of picture rate conversion and are commercially available in products. Spatio-temporal prediction methods such as RS, e.g., [161–164, 173–176], are applied in practice (e.g., [177, 178]), and so are alternatives based on PPC [179, 180].

Recursive search motion estimation

We investigated a hierarchical ME approach using resolution down-scaling, which we call multi-scale block-matching ME. Using down-scaling, the coarser motion vectors are obtained from block-matching at a lower spatial resolution and can be successively refined at higher resolutions. We will combine the multi-scale approach with a hierarchical ME method known as multi-grid block-matching [181]. In this method, a coarse motion vector is first found using a large block size and this vector is successively refined for the smaller blocks into which the larger blocks are decomposed (using a quad-tree decomposition). By combining the multi-scale and multi-grid approaches, we aim at reducing the computational complexity and are flexible in investigating the

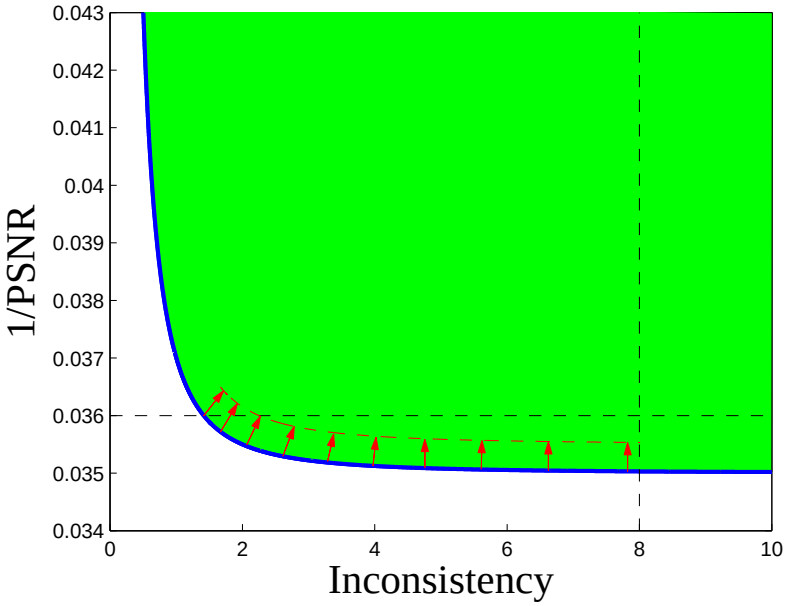


Figure 3.3: The PSNR-Consistency trade-off graph of the ME design space where the green shaded area indicates the area of possible MEs which are bounded by the contour line (blue). The black dashed lines indicate the minimal PSNR and maximal Inconsistency for the attractive contour line section. The range of well performing MEs in the attractive segment are highlighted by the red arrows.

effects of using different block sizes and scale factors.

The multi-scale and multi-grid approach are illustrated by the scale pyramid shown in Fig. 3.5, where ME is performed on higher scales at the top of the pyramid first and motion vectors are propagated down the pyramid to the lower scales by means of hierarchical candidates.

The block-matching method we apply is the RS ME of [167]. In contrast to the usual RS candidate structure of [167], the temporal candidate is closer to the current block for all the hierarchical approaches because, for coarse scales, the temporal candidate may come to lie outside the object in which the current block is located. An overview of the investigated candidate structures is given in Fig. 3.6.

The parameters involved in the design of hierarchical motion estimators are explained in the following and an overview is given in Table 3.1.

The relevant parameters for the scale structure are the **fine** scale, the **coarse** scale and the scaling factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$. The **fine** scale and the **coarse** scale denote the levels of the pyramid (see Fig. 3.5) where ME is performed, e.g., **fine** $= 1$, **coarse** $= 2$. **fine** is the finer scale (for multi-scale ME) or the one with a finer block grid than **coarse** in the case that the coarse and fine scale have the same size (multi-grid ME). If the full resolution is included as a scale on which ME is performed, **fine** $= 0$ is chosen (otherwise **fine** $= 1$). The scale factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$ determine the size of the scales. The scaling factors $\mathbf{sf_w}$ and $\mathbf{sf_h}$ for width and height indicate how much one

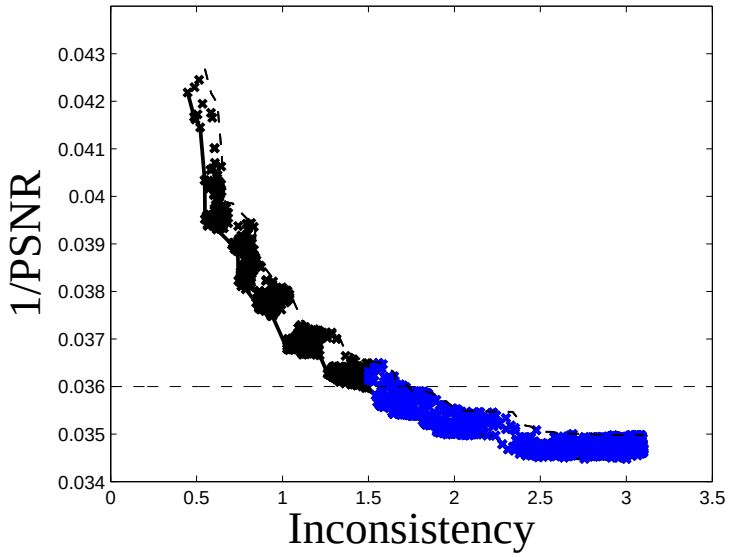


Figure 3.4: RS MEs (black) within a limited distance from the contour line; 1745 highlighted MEs (blue) in the attractive segment. Dashed line indicates minimal PSNR for well performing ME.

scale is down-scaled in comparison to the next lower scale in the pyramid. In the case of a multi-scale motion estimator (right image in Fig. 3.5), the image dimensions become smaller as we ascend in the pyramid. However, when a multi-grid (left image in Fig. 3.5) motion estimator is designed, two scales have the same dimension, thus the corresponding scaling factor component equals 1. As the spatial resolution of two vector fields from two different scales may not be equal, this may require scaling of

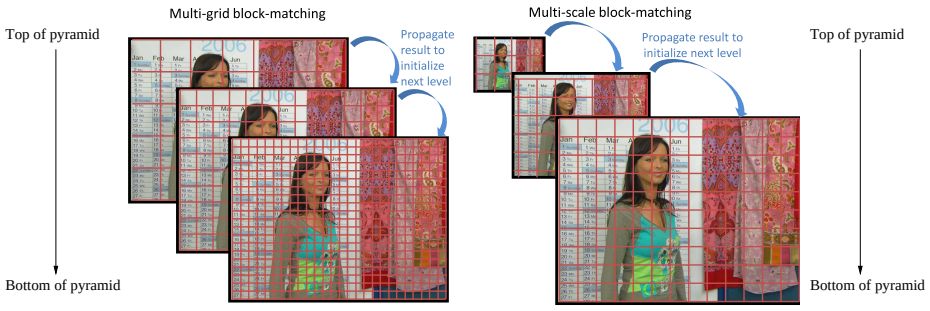


Figure 3.5: Illustration of multi-grid (left) and multi-scale (right) motion estimation approach. In both cases, ME is performed on higher scales at the top of the pyramid first and motion vectors are propagated down the pyramid to the lower scales by means of hierarchical candidates.

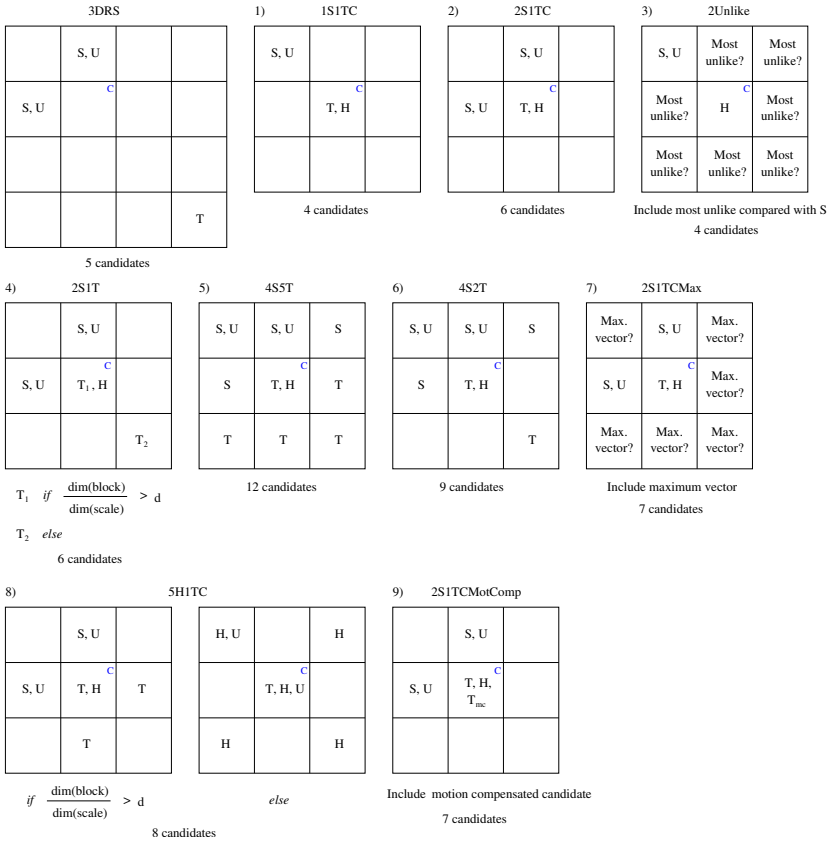


Figure 3.6: The usual 3DRS candidate structure, as well as nine different subsamplings of the spatio-temporal neighborhood of a block (candidate structures 1,...,9) are shown. C denotes the current block for which candidate motion vectors are determined, S a spatial candidate, U a random update vector added to the spatial candidate, T a temporal candidate, and H the hierarchical candidate resulting from the ME scan on a coarser grid or on a coarser scale. For candidate structures 4 and 8, $d = 1/60$.

the vector field as well, which is implemented as nearest neighbor scaling.

The block width and block height dimensions $\mathbf{blk_w}$ and $\mathbf{blk_h}$ are arrays where the elements $\mathbf{blk_w}[i], i = 0, \ldots, \mathbf{coarse}$, indicate the block sizes for each scale i in the pyramid.

The PSNR-Consistency plot and the contour line of the optimal RS MEs are given in Fig. 3.2 and Fig. 3.7. Note that the SI measure is computed based on 8×8 pixel blocks. The MEs within the attractive segment are shown in Fig. 3.4. Based on the parameter histogram analysis, which is elaborately discussed in [168], 62 multi-scale MEs have been identified out of the 1745 MEs in the attractive segment (see Fig. 3.4). An overview of the proposed parameter settings of this ME type is given in Table 3.2 where block settings and performance are rendered. The first row of Table 3.4 shows the mean performance for the 62 robust MEs. From the expectation value of the

fine	Lowest/Finest scale on which ME is performed
coarse	Highest/Coarsest scale on which ME is performed
$\mathbf{sf_w}$, $\mathbf{sf_h}$	Scaling factor width and height for resizing scales
$\mathbf{blk_w}$, $\mathbf{blk_h}$	Block width and height of each scale
cand.struc.	Selected candidate structure
scan	Amount of ME scans performed per scale

Table 3.1: Parameters for the hierarchical motion estimator design.

block dimension distributions of the 62 MEs we determined the proposed ME settings given in the second row. The resulting ME happens to coincide with Opt. ME 6 in Fig. 3.7 which was visually perceived as the most pleasing ME among the seven MEs on the contour line.

	Block width full res.	**Block height full res.**	**mean PSNR**	**mean SI**
62 Robust RS MEs	[8,32], [32,128]	[4,16], [16,64]	28.46	2.18
Proposed RS ME	16, 64	8, 32	28.91	2.46

Table 3.2: Block settings and performance in PSNR and SI of 62 robust RS MEs and the proposed RS ME. Block width and block height indicate the equivalent block sizes for the full resolution where the selected settings for the **fine** and **course** scales can be a range ([..]) of values.

Phase Plane Correlation Motion Estimation

PPC was developed in the '80s [179] and is employed in state-of-the-art products [180]. Instead of obtaining motion vector candidates from a spatio-temporal neighborhood as in RS, PPC retrieves the motion vectors by performing phase correlation in the Fourier domain between spatially corresponding blocks from consecutive images. A correlation plane of displacement peaks is returned of which a subset is used as motion vector candidates in a consequent block matching operation on smaller blocks. Among the most dominant peaks in the obtained displacement field, the peak yielding the minimal match error between the motion compensated and the original smaller blocks is selected.

We implemented PPC-based MEs based on [179] (which may not reflect current product implementations) with the parameter variations as given in Table 3.3. In total, 1800 ME parameter combinations are investigated. Initially, a two-dimensional Fourier transform is performed on the larger blocks with dimensions m_lxm_l. The n_p most dominant displacement peaks are considered motion candidates for the smaller blocks with dimensions m_sxm_s. Another parameter is the block step size s_b based on which the pixel locations of the next m_lxm_l block are selected for the next FFT

m_l	$\{32, 64, 128\}$
m_s	$\{1, 2, 4, 8, 16\}$
n_p	$\{1, 2, 3, ..., 20\}$
s_b	$\{16, ..., m_l\}$
a_s	$\{0, 1\}$

Table 3.3: Parameter settings for PPC MEs.

operation. The values for s_b are set in the range between the largest m_s setting (16) and the current m_l dimension. The displacement of the larger blocks can be determined with pixel or sub-pixel accuracy. The binary variable a_s indicates a sub-pixel accuracy of 0.25 pixels when $a_s = 1$ and pixel accuracy when $a_s = 0$.

The contour line (with the 8×8 block-based SI measure) of the PPC MEs is given in Fig. 3.7.

The histogram analysis is conducted for all parameters in Table 3.3. Among the 54 MEs within the attractive segment, we found that the block dimensions of both the larger and smaller blocks converge to $m_l = 128$ and $m_s = 16$. This is expected since large m_l dimensions are necessary to capture larger movements. Taking 16×16 blocks to perform block matching on the candidate peaks is already proven in the RS study to be a suitable value when we are dealing with HD sequences. The robust number of candidate peaks n_p is determined to be $n_p = 13$. Largely overlapping blocks are favored with a block step size tending to $s_b = 32$.

In the authors' perception, the computed robust ME settings (referred to as the PPC proposed ME and RS Opt. ME 6 in Fig. 3.7) corresponded with better upconverted quality than other ME settings (e.g., other optimal MEs in Fig. 3.7 or MEs in the attractive segment). The results have also shown that a comparison between two MEs from different ME types (i.e., from RS and from PPC) is possible when MEs are sufficiently far apart in the PSNR-Consistency trade-off graph as is the case between PPC and RS MEs (see Fig. 3.7).

The proposed methodology should find robust MEs even with suboptimal performance measures. The SI metric is evidently suboptimal in the sense that the SI output is dependent on the selected block dimension, thus the 8×8 block-based SI measure is not comparable with the 1×1 pixel-based SI measure. For the case of PPC, we have added a pixel-based SI evaluation (see Fig. 3.8), where a motion vector is assigned to each pixel in the image instead of validating the SI performance on only 8×8 pixel motion vector blocks. The distances ΔSI and ΔPSNR to the attractive section were chosen such that approximately the same number of MEs ended up in the attractive segment. When comparing the 8×8 block based SI with the pixel-based SI, we found that different MEs are returned in the attractive segment. 22% of the MEs in the attractive segment of the 8×8 block-based SI measure are not present in the attractive segment of the pixel-based approach. Nevertheless, the histogram analysis

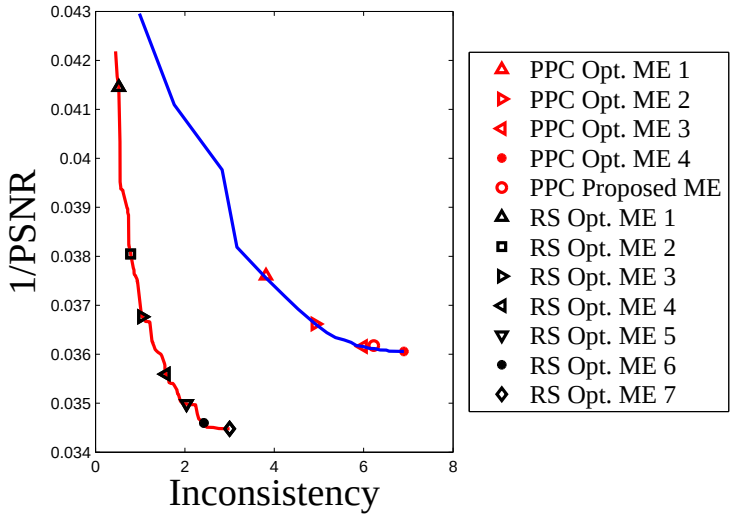


Figure 3.7: Contour line of RS (red line) and PPC (blue line) MEs; several optimal MEs are highlighted.

reveals that the same robust ME is computed in the case of the pixel-based SI. In Fig. 3.8, it is apparent that the computed robust ME is located halfway between Opt. ME 3 and Opt. ME 4, whereas in Fig. 3.7, the same ME is located closer to Opt. ME 3, which underlines the incongruent output of the two SI metrics.

The performance of the computed robust ME is analyzed and compared to other MEs within the attractive segment to determine the robustness of the chosen pa-

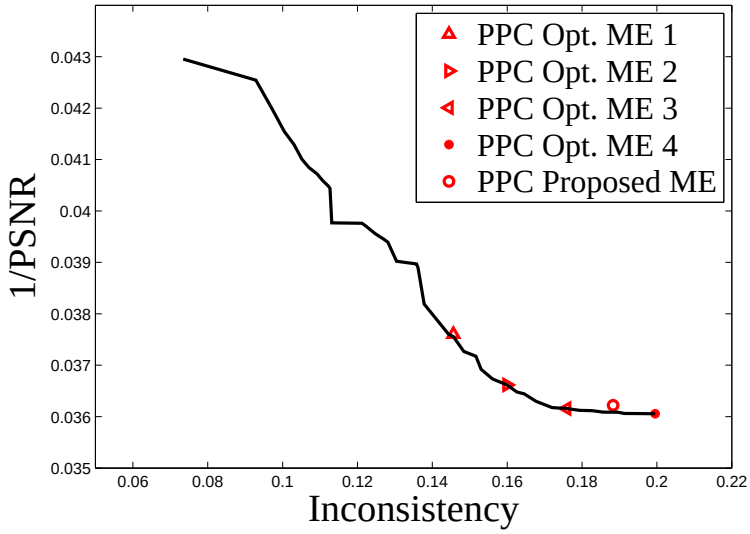


Figure 3.8: Contour line of PPC MEs derived from pixel-based SI metric with highlighted optimal MEs.

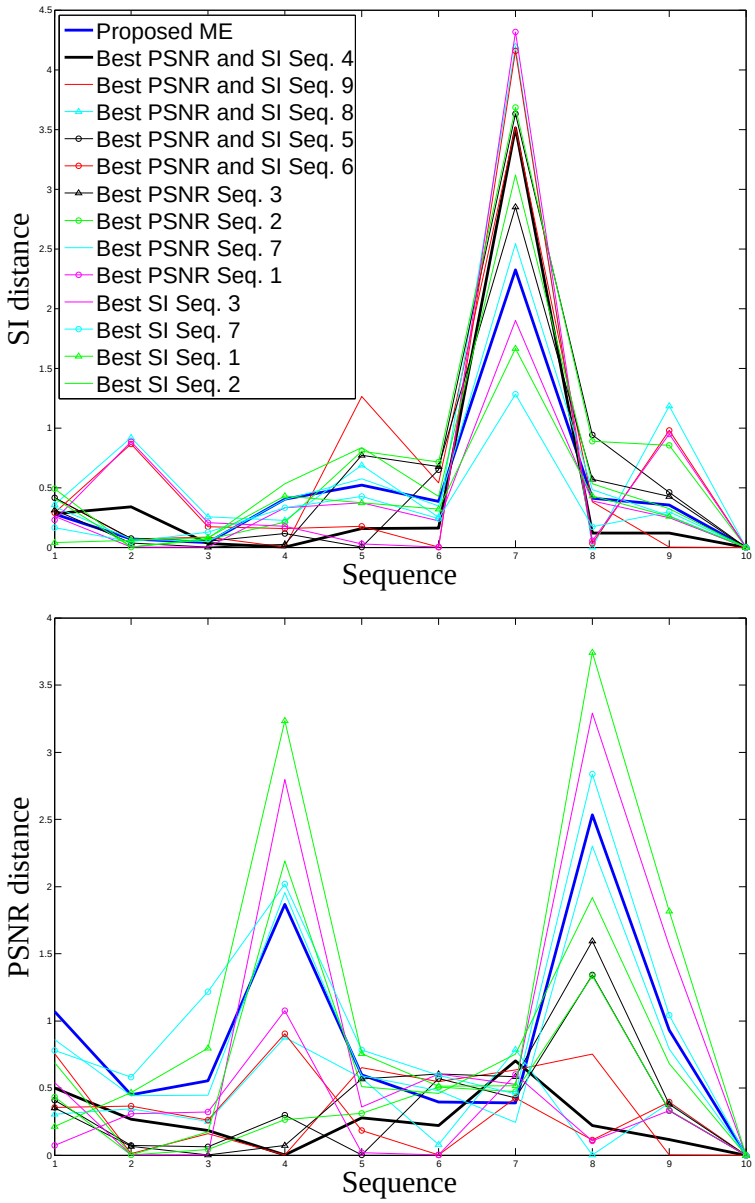


Figure 3.9: Top: SI distances to the contour line for the best performing RS MEs within the attractive segment in PSNR and SI per sequence. Bottom: PSNR distances to the contour line for the best performing RS MEs within the attractive segment in PSNR and SI per sequence. The computed robust ME is highlighted with a thicker blue line.

	Block width full res.	Block height full res.	mean PSNR	mean SI
Robust RS ME	16, 64	8, 32	28.91	2.46
3DRS [173]	8	8	28.60	2.80
HRNM [182]	8	8	29.32	1.48
FS [183]	16	16	25.78	15.00
3SS [184]	16	16	22.80	3.90
OTS [185]	16	16	23.79	6.64
DS [186]	16	16	23.95	7.24
HEXBS [187]	16	16	23.90	7.28
MVFAST [188]	16	16	28.15	4.43
TCSBP [161]	16	16	28.31	4.07
EPZS [189]	16	16	28.69	3.77
MRST [190]	16,16,16,16	16,16,16,16	28.60	5.16
MPMVP [176]	32,16, 8, 4	32,16, 8, 4	28.13	3.80

Table 3.4: Performance comparison of the computed robust RS ME and various techniques documented in literature. Block width and block height indicate the equivalent block sizes for the full resolution **fine** and **course** scales.

rameter settings. A robust ME is expected not to perform badly on any of the test sequences. Therefore, the PSNR and SI distance to the contour line are displayed in Fig. 3.9, where the computed robust RS ME is plotted against the best MEs (in either PSNR or SI) for each sequence. Only one ME (highlighted in black in Fig. 3.9) reveals on average smaller distances in both PSNR and SI. However, its SI distance from the contour line for the 'WalkingMan' test sequence (sequence 7 in top image of Fig. 3.9) shown in the middle snapshot of the first row in Fig. 3.1 is clearly larger than the SI distance of the computed robust ME. Hence, the ME computed with the proposed methodology is not surpassed in robustness by any of the best MEs per test sequence.

The computed robust RS ME is a multi-scale recursive search ME with candidate structure 2S1TC (see [168] for details), employing two scales where the first scale is a downscaled version of the full resolution image by a factor of 2 and the second scale a downscaled version by a factor of 4, where 2 estimation scans are performed on each scale. The block size settings and PSNR/SI performance are given in the first row of Table 3.4.

To further confirm that the proposed methodology returns well-performing MEs which can compete with other techniques, a benchmark is provided in Table 3.4. An overview is given of the SI and M2SE-PSNR values of the investigated recursive-search MEs as well as the benchmark results from several methods described in literature implemented by us. These include full-search (FS) and reduced-search pattern based methods, i.e. three-step-search (TSS) [184], one-at-a-time search (OTS) [185], diamond search (DS) [186] and hexagon-based-search (HEXBS) [187], as well as algo-

rithms based on spatio-temporal predictors, i.e. the predictive (zonal) search methods MVFAST [188] and EPZS [189], the RS methods 3DRS [173], HRNM [182] and TCSBP [161], and combined hierarchical-predictive methods, i.e. the MRST-method proposed in [190] and MPMVP from [176]. Note that the M2SE-PSNR metric favors 'true' motion, i.e. MEs with a better vector field consistency can outperform a full-search method. Furthermore, all methods from literature were adapted and tested with smaller block dimensions (e.g., 8×8), however, no improvement in PSNR and SI was observed.

The benchmark shows that the computed hierarchical RS ME is outperformed solely by the sophisticated HRNM ME which employs 3-picture estimates. We suggest from these results that the proposed methodology does yield superior performing MEs among the thousands of ME parameter combinations.

The proposed methodology has been tested with a large set of parameters (6660 RS MEs) and a smaller set of parameters (1800 PPC MEs). In both cases, the methodology returned robust MEs. As long as there are sufficient permutations of parameter settings, the methodology should be able to compute reoccurring 'well-performing' settings yielding robust motion estimators.

3.3 Perception test for assessing ME quality

A subset of MEs is considered as 'well performing' when they satisfy particular selection criteria. The properties of these MEs are further analyzed to determine the settings of a robust ME. We have conducted a user study on a limited set of test sequences to gain insight into the validity of the assumptions made in Section 3.2. In particular, the definition of the chosen attractive segment and the quality influence of PSNR and Inconsistency performance measures are investigated.

3.3.1 Video sequence selection

Snapshots of the three video sequences used in the user study are shown in Fig. 3.10. Sequence A shows a person walking in front of a calendar and other objects with high contrast and many details. Departing cars are accelerating over an intersection in sequence B. In sequence C, a motor boat passes behind an iron grid fence. This sequence contains different motions and occlusion. The sequences posed different ME challenges resulting in MEs on different PSNR and Inconsistency quality scales (see ME clusters in Fig. 3.11). The video sequences were converted from 24 fps (sequences A and C) or from 30 fps (sequence B) to 60 fps using the MEs selected in Section 3.3.2. The video clips were 1.5 s - 2 s long and presented to the viewers in an uninterrupted loop.

Figure 3.10: Snapshots of the selected video sequences.

3.3.2 ME selection

The objective of the user study is to answer the following research questions which help in improving and verifying the ME design methodology.

1. Is a contour line analysis of MEs sufficient such that the attractive segment can be limited to the attractive section of the contour line?

2. Is the attractive segment appropriately chosen in Section 3.2 such that a ME inside the attractive segment scores significantly better than a ME outside the attractive segment?

3. Do the PSNR and SI measures show similar importance for assessing the ME quality?

Therefore, three different partitions or areas were selected corresponding to the attractive section of the contour line, the original attractive segment and an attractive segment biased towards high PSNR values.

The selection of the specific MEs within each partition was done in a systematic way. A grid was used for selecting motion estimators with approximately the same distances in 1/PSNR and Inconsistency. The grid is shown in Fig. 3.11 as black dots, where the colors indicate the sequence (A (red), B (blue), and C (green)). This grid consisted of equally separated points which figured as guiding target positions. Points were chosen systematically according to their PSNR and SI values (not according to their parameter settings which can vary largely in a non-consistent way), from dense points closer to the optimum to less dense further away from the optimum. When a grid point was selected, the closest motion estimator was chosen. From the optimum (in terms of 1/PSNR and SI), the first ME was selected for each sequence. The next five MEs then were selected with a distance of two grid points between each other. The MEs farthest away from the optimum were selected with a distance of four to five grid points from each other. An overview of the 93 selected motion estimators in the 1/PSNR-Inconsistency plot is given in Fig. 3.12. The black rectangle indicates the area of interest shown in Fig. 3.13 with the resulting ME selection. Due to performance variations of the same ME for different sequences, different MEs may be

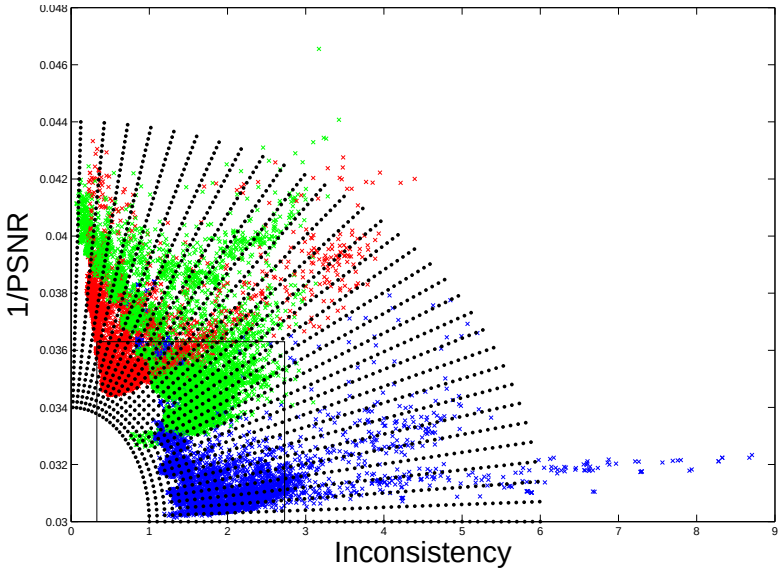


Figure 3.11: All MEs: sequences are indicated by the colors red (A), blue (B), and green (C). A grid is laid over to systematically select motion estimators.

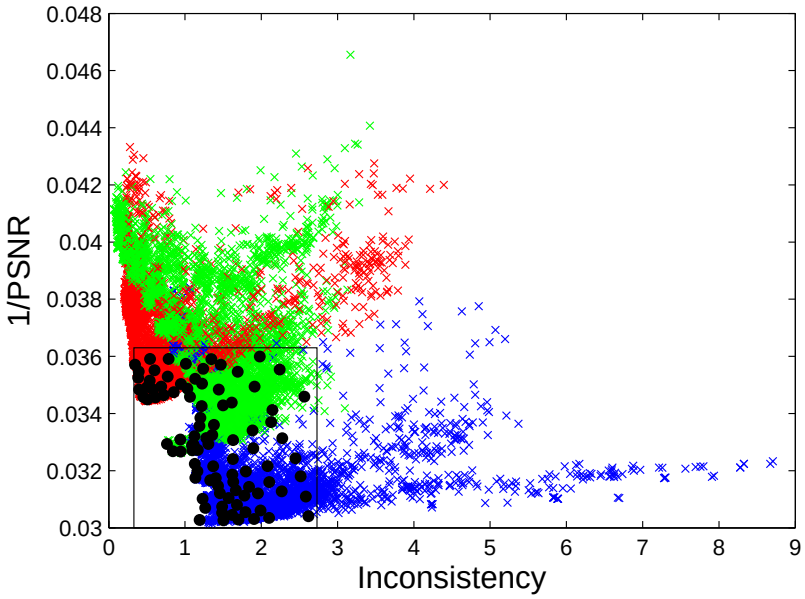


Figure 3.12: Selected MEs as black dots in the total 1/PSNR-Inconsistency plot.

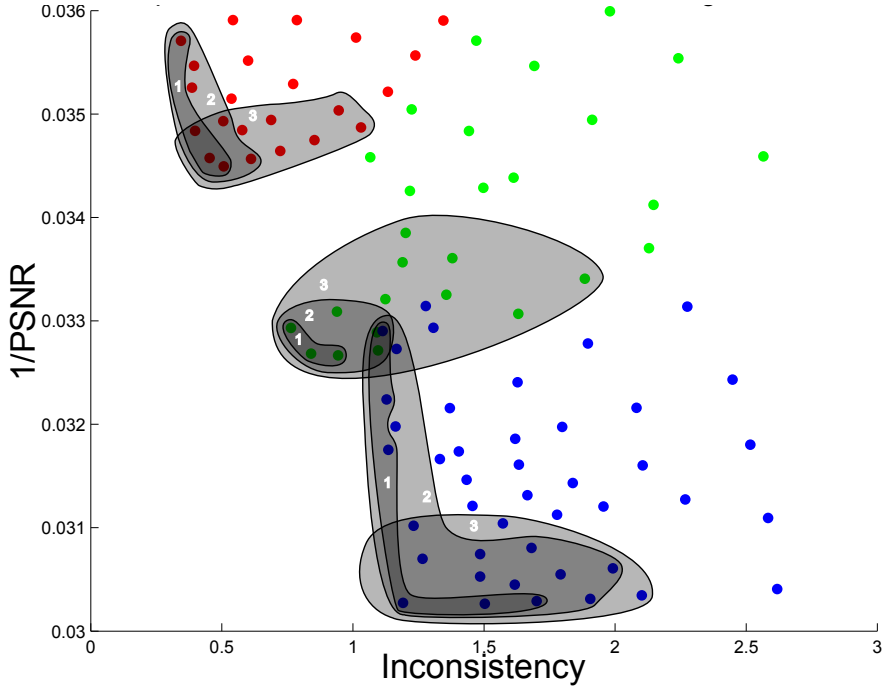


Figure 3.13: Selected MEs from sequences A (red), B (blue), and C (green). Three partitions are indicated (grey) and numbered 1, 2, 3.

selected for sequence A than for sequence B. In total, 93 ME-sequence combinations were selected.

Three partitions were selected for each sequence to draw MEs from (see Fig. 3.13). They differ in position and size in the 1/PSNR-Inconsistency plot. The first partition is limited to the contour line, the second partition describes the attractive segment proposed in [169], and the third partition is biased to high PSNR scores, largely disregarding the SI score. This bias is chosen to assess the influence of the PSNR and SI measures on the perceived video quality. The bias towards PSNR is present in a number of publications evaluating MEs where no inconsistency measure is taken into account (e.g., [163, 191, 192]). This third partition encompasses all MEs within a distance of ΔPSNR from the ME with the highest PSNR score. The range of ΔPSNR varied between $0.35 < \Delta\text{PSNR} < 0.55$, selecting a similar number of MEs for each sequence. MEs from the three partitions were compared with MEs outside the partitions in Fig. 3.13.

3.3.3 User study setup

The experiment was set up in an enclosed testing room without any windows. The room was dimly lit with two identical floor lamps, each consisting of two halogen

Figure 3.14: User study setup: participant rating the quality. The second screen was positioned outside the field of view of the participant and its brightness level was set to minimum.

lamps. With each lamp, the spot was directed to the wall and the main lamp was directed diffusely to the white ceiling, giving a domestic impression (see Fig. 3.14). In the back of the testing room, the main screen displaying the video sequences was located. In front of it a table was positioned on which a second screen (the instruction and score entry screen), a computer keyboard and a computer mouse were present. Participants were seated on a chair behind the table. The second screen was positioned outside the field of view of the participants and its brightness level was set to minimum to reduce its influence as much as possible. The reason for introducing a second screen is to give participants the possibility to adjust the score while reviewing the test sequences. The distance between the center of the main screen and the forehead of the participant was approximately twice the diagonal length of the screen minus 10 percent, in our case, with a 46 inch screen, 82.8 inch. This ratio between screen diagonal and viewing distance was based on the SMPTE standard 196M-2003 [193].

A video streaming system was used for both presentation and response recording. The main screen used for video presentation was a 46 inch Sony LCD TV screen (Sony KDL-46HX920) with a LED back light and had a 16:9 aspect ratio. The minimum luminance was 0 cd/m^2 (due to local dimming), the maximum luminance was 600

cd/m^2. The instruction screen used was a Philips LCD monitor screen (Philips Brilliance 240B) with a 24 inch diameter and a 16:10 aspect ratio. The brightness was set to the screen's minimum. The other settings were left at factory defaults.

3.3.4 Participants and procedure

In total, 24 participants joined the perception test, among them 12 male and 12 female subjects, ranging from 21 to 51 years old. None of the participants had any professional experience in video processing.

The participants were seated in front of two displays on which the instructions and video sequences were presented. Written instructions were given on the instruction display and example trials were presented on the main TV display. In the first phase of the experiment, subjects got familiar with the stimuli. Every video sequence was presented at two performance levels (i.e., a high quality video sequence and a low quality video sequence) that indicated the range of performances. In the second phase, the training phase, six samples were presented to let the participants get used to the rating slider. Participants were asked to rate the video sequences on a quality scale from 0 to 10 points (10 denotes highest quality). While the video sequence was presented, participants judged the quality of the video sequence by positioning the slider with the computer mouse on the instruction screen.

In the third phase, the test phase, all conditions were presented in a randomized order, and ratings had to be given. After pressing the confirmation button below the rating slider, the next sequence was presented on the main display. On average, the experiment took 25 minutes, approximately 8 minutes per sequence.

3.4 Results

In total, 93 ME-sequence combinations (stimuli) were rated on quality by 24 participants. We conducted a one-way ANOVA with the 93 stimuli as independent variable and the user scores as dependent variable. An ANOVA per sequence and per partition was carried out. The motion estimators and their mean quality score are plotted in the top rows of Fig. 3.15, Fig. 3.16, and Fig. 3.17, where the color of the dots indicate the mean quality score. The mean score did not exceed the level of 8, thus the range of the color bar is limited to 8 in the corresponding figures.

When partition 1 (contour line MEs, Fig. 3.13) was compared to the other MEs in its specific sequence, no significant difference between the partition and the rest of the MEs was found ($p_A = .131$, $p_B = .205$, $p_C = .407$). Significance is judged when the p-value $< .05$. p_i denotes the p-value for sequence i where $i \in \{\text{A, B, C}\}$. For partition 2 (attractive segment as defined in [169]), a significantly higher mean score on quality was obtained than for the other MEs in the sequence: $p_A < .05$, $p_B < .001$, $p_C < .05$. The difference between partition 3 (high PSNR influence) and the rest of the quality scores appeared to be significant too: $p_A < .001$, $p_B < .001$, $p_C < .001$.

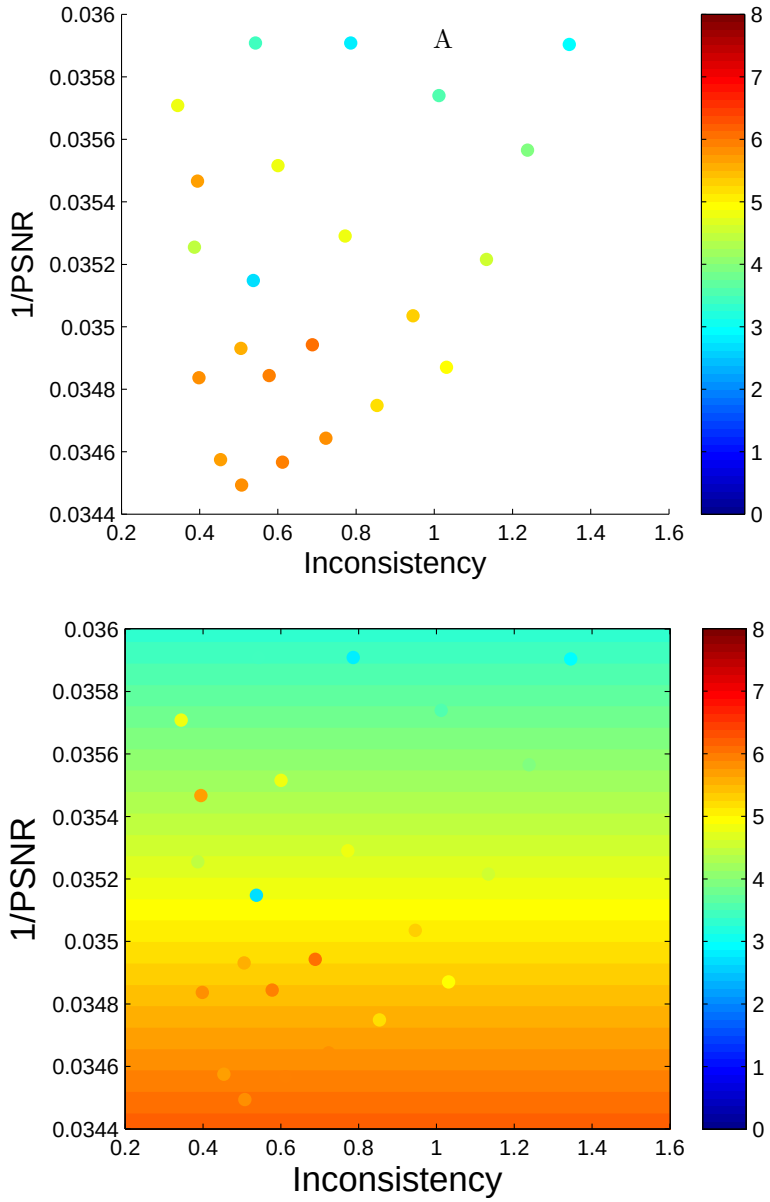


Figure 3.15: Top: Mean quality scores for sequence A indicated in color in the 1/PSNR-Inconsistency plot. Bottom: Quality Means and estimation by the model for sequence A.

For the quality measures, regression analysis was performed for the predictors Inconsistency and 1/PSNR. For the three video sequences, specific regression models were calculated. Models are represented in the scatter plots in the bottom rows of

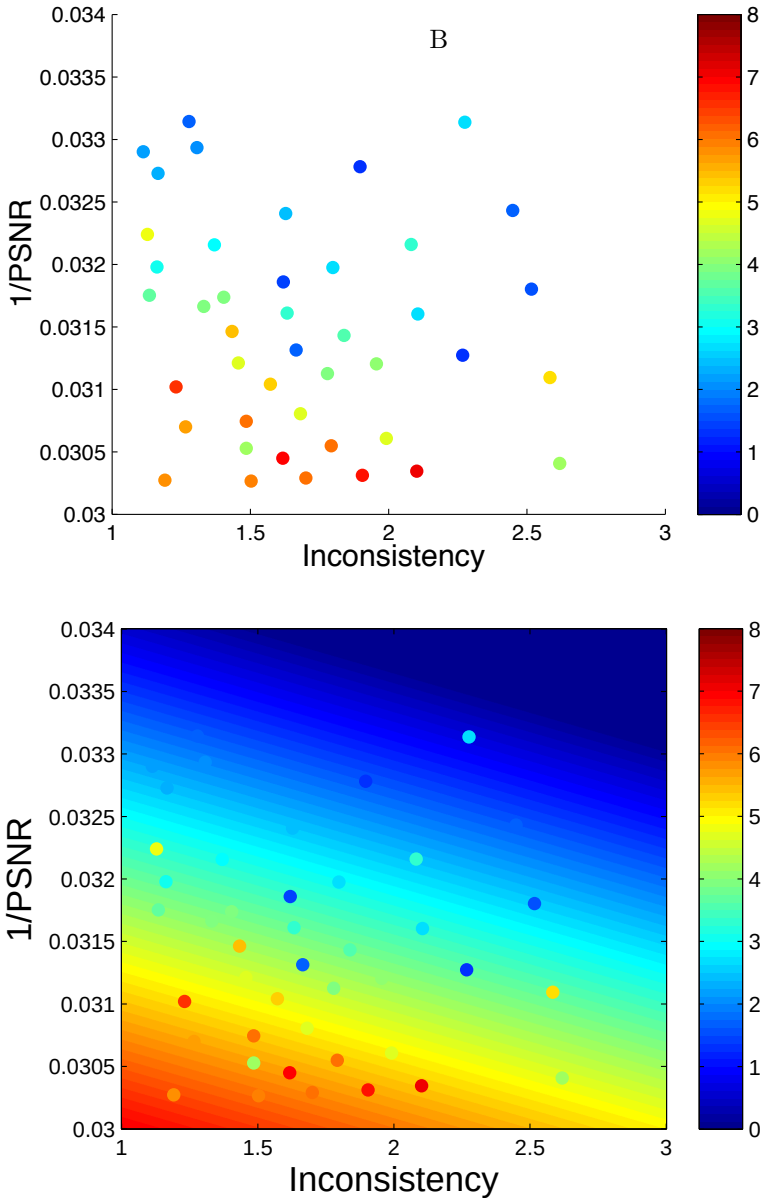


Figure 3.16: Top: Mean quality scores for sequence B indicated in color in the 1/PSNR-Inconsistency plot. Bottom: Quality Means and estimation by the model for sequence B.

Fig. 3.15, Fig. 3.16 and Fig. 3.17, where the background color indicates the expected value by the model. The colored dots represent the mean quality value of the motion estimators obtained by the user study, plotted against Inconsistency and 1/PSNR.

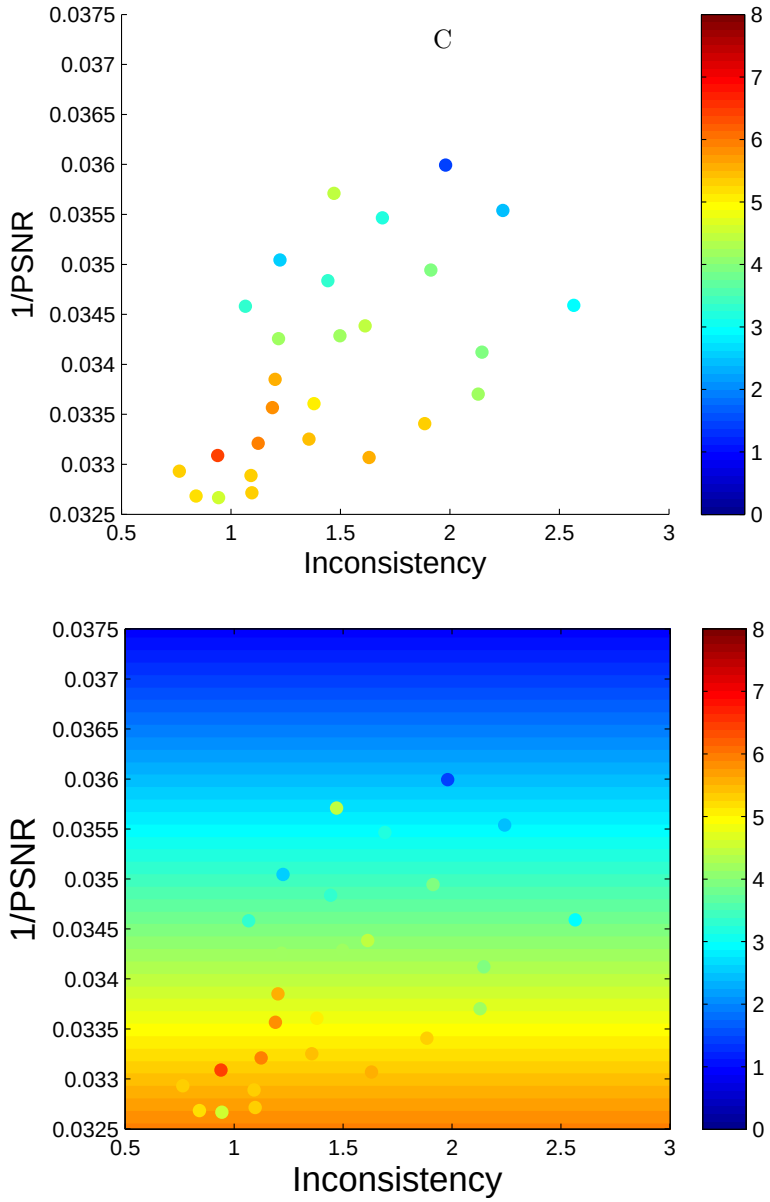


Figure 3.17: Top: Mean quality scores for sequence C indicated in color in the 1/PSNR-Inconsistency plot. Bottom: Quality Means and estimation by the model for sequence C.

For sequence A (bottom image in Fig. 3.15), variance in quality can only be explained by the 1/PSNR predictor. The percentage explained variance is 58.9% (R^2=.589). This model appears to be significant: $p_A < .001$. Taking into account

that Inconsistency would not improve the model fit, the resulting equation for the quality Q is as shown in (3.5).

$$Q_{\text{Seq.A}} = 69.138 - 1829.001 \cdot 1/\text{PSNR} \tag{3.5}$$

For sequence B (presented in bottom image in Fig. 3.16), a model based solely on 1/PSNR explains 60.2% of the variance significantly ($p_B < .001$). An improvement of 6% is found with the model with both predictors Inconsistency and 1/PSNR where 66.1% of the variance is significantly explained ($p_B < .001$). In (3.6), the quality formula is derived.

$$Q_{\text{Seq.B}} = 56.694 - 1.018 \cdot \text{SI} - 1621.128 \cdot 1/\text{PSNR} \tag{3.6}$$

Similarly to sequence A, variance within the sequence C group of MEs (see bottom image in Fig. 3.17) can only be explained by 1/PSNR ($p_C < .001$). The percentage explained variance is 68.3%. The equation of the quality model is shown in (3.7). Inconsistency as a predictor would not improve the model fit.

$$Q_{\text{Seq.C}} = 38.660 - 894.966 \cdot 1/\text{PSNR} \tag{3.7}$$

3.5 Discussion

The results of the ME perception test give us insight into the relevance of the contour line, the attractive segment as defined in Section 3.2 and the influence of the performance measures PSNR and SI.

3.5.1 Contour line vs. attractive segment(s)

To be able to select a group of MEs within each sequence to be the best performing, three partitions have been compared. For all three video sequences, the data analysis of the user scores yielded no significant difference between the MEs on the contour line (with a PSNR of at least 27.8 and the SI limit set to 8) and the other MEs. This confirms the initial hypothesis and observation of the authors that the performance measures are suboptimal and do not necessarily yield the perceptually best MEs on the contour line.

However, MEs within the attractive segment as defined in Section 3.2 were evaluated significantly higher on quality, supporting the choice of the attractive segment. Also the MEs in the partition heavily influenced by high PSNR scores with a large range of SI values was rated significantly higher. This suggests that another attractive segment with a PSNR bias exists returning a robust, well performing ME. To some extent this is recognized in the proposed methodology. A larger variation in

the SI performance measure is allowed than in the PSNR metric (see SI and PSNR limits marked with dashed lines in Fig. 3.3). Accordingly, when incorporating the third partition as an attractive segment in the proposed methodology and applying the methodology on the ten test sequences illustrated in Fig. 3.1, the ME with the same robust settings is computed as with the original attractive segment.

3.5.2 Regression analysis

Regression models for the quality as a function of 1/PSNR and Inconsistency SI were estimated. These models explained 59%-68% of the variance in quality. More than half of the variance in quality is explained by the PSNR measure (in sequences A, B, C) and by the SI measure (in sequence B). For MEs with lower PSNR (sequences A and C, red and green MEs in Fig. 3.11), the PSNR measure plays the only role in assessing the ME quality. Users may not see the difference in inconsistency when the PSNR is too bad. For sequence B with the highest PSNR MEs, a slight influence of the SI measure became visible (6% improvement compared to PSNR only). The authors have analyzed MEs for test sequences outside the set in Fig. 3.10 and can observe a marginal improvement with the PSNR-dependent importance of SI. For large SI values (e.g., with some PPC ME settings the SI value goes beyond 13 for PSNR values close to the best PPC PSNR values), a clear degradation in performance is observed. MEs with large SI values should be discarded and therefore, the vertical cut-off line in Fig. 3.3 limiting the attractive segment should be well chosen for the motion estimator types at hand. Assumptions that the SI measure would be more meaningful even in the smaller SI value ranges have not manifested themselves. Other performance measures should be investigated in future work to increase the percentage of explained variance in quality.

3.6 Conclusions

A computer-aided design methodology is proposed that can deal with suboptimal performance measures. A three-step approach is employed where, first, the variety of conditions under which the motion estimators should perform well is defined and appropriate test data is selected. Second, a contour line or trade-off curve illustrates the achieved compromise between the motion vector prediction accuracy and consistency. Third, an attractive segment of well performing MEs is identified containing all motion estimators within a defined distance from an attractive section of the contour line.

In order to validate and improve the methodology, we have conducted a perception test to come to a perception-oriented motion estimation design methodology which corresponds well with the perceived video quality. In the user study, TV viewers rated 93 different motion estimators in 3 video sequences. User ratings indicate that well performing motion estimators should not be limited to the contour line. The proposed attractive segment has been confirmed.

High quality ratings are also given to a partition dominated by high PSNR scores while maintaining a large variation in consistency. The higher impact of the PSNR measure compared to the inconsistency measure is supported by the conducted regression analysis. The inconsistency measure has influence on the perceived video quality, only for the sequence with motion estimators at high PSNR scores, and yields even there an improvement of only 6%. A clear degradation in performance is observed for MEs with large SI values. These should be discarded and therefore, the vertical cut-off line limiting the attractive segment should be well chosen. Other performance measures should be investigated in future work to increase the percentage of explained variance in quality.

The proposed methodology may provide an inspiration for similar tough multi-dimensional optimization tasks with suboptimal metrics.

Bibliography

[161] J. Wang, D. Wang, and W. Zhang, "Temporal compensated motion estimation with simple block-based prediction", *IEEE Transactions on Broadcasting*, **vol. 49**, no. 3 pp. 241–248, Sep. 2003.

[162] C. Bartels and G. de Haan, "Smoothness Constraints in Recursive Search Motion Estimation for Picture Rate Conversion", *Circuits and Systems for Video Technology, IEEE Transactions on*, **vol. 20**, no. 10 pp. 1310–1319, 2010.

[163] T. Yamamoto, N. Mishima, T. Ono, and T. Kaneko, "High-accuracy motion estimation with 4-D recursive search block matching", in *Consumer Electronics (GCCE), 2012 IEEE 1st Global Conference on*, 2012, pp. 625–628.

[164] Y. Guo, Z. Gao, L. Chen, and X. Zhang, "Effective early termination using adaptive search order for frame rate up-conversion", in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, 2013, pp. 1416–1419.

[165] R. Han and A. Men, "Frame rate up-conversion for high-definition video applications", *Consumer Electronics, IEEE Transactions on*, **vol. 59**, no. 1 pp. 229–236, 2013.

[166] S.-C. Tai, Y.-R. Chen, Z.-B. Huang, and C.-C. Wang, "A Multi-Pass True Motion Estimation Scheme With Motion Vector Propagation for Frame Rate Up-Conversion Applications", *Display Technology, Journal of*, **vol. 4**, no. 2 pp. 188–197, 2008.

[167] G. de Haan, , P.W.A.C. Biezen, H. Huijgen, and O.A. Ojo, "True-motion estimation with 3-D recursive search block matching", *IEEE Trans. Circuits, Syst. Video Techn.*, pp. 368–379, Oct. 1993.

[168] A. Heinrich, C. Bartels, R.J. Van der Vleuten, C.N. Cordes, and G. de Haan, "Optimization of Hierarchical 3DRS Motion Estimators for Picture Rate Conversion", *IEEE Journal of Selected Topics in Signal Processing*, **vol. 5**, no. 2 pp. 262–274, Apr. 2011.

[169] A. Heinrich *et al.*, "Robust Motion Estimation Design Methodology", in *Proceedings of the 2010 Conference on Visual Media Production (CVMP)*, Nov. 2010, pp. 49–57.

[170] G. de Haan and E.B. Bellers, "De-interlacing of video data", *IEEE Transactions on Consumer Electronics*, **vol. 43**, no. 3 pp. 819–825, Aug. 1997.

[171] B. Horn and B. Schunck, "Determining Optical Flow", *Artificial Intelligence*, **vol. 17**, no. 1-3 pp. 185–203, 1981.

[172] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, ISBN 052183378.

[173] G. de Haan and P.W.A.C. Biezen, "Sub-pixel motion estimation with 3-D recursive search block-matching", *Signal Processing*, **vol. 6**, no. 3 pp. 229–239, June 1994.

[174] M. Cetin and I. Hamzaoglu, "An adaptive true motion estimation algorithm for frame rate conversion of high definition video and its hardware implementations", *Consumer Electronics, IEEE Transactions on*, **vol. 57**, no. 2 pp. 923–931, 2011.

[175] S. Dikbas and Y. Altunbasak, "Novel True-Motion Estimation Algorithm and Its Application to Motion-Compensated Temporal Frame Interpolation", *Image Processing, IEEE Transactions on*, **vol. 22**, no. 8 pp. 2931–2945, 2013.

[176] S.-C. Tai *et al.*, "A Multi-Pass True Motion Estimation Scheme With Motion Vector Propagation for Frame Rate Up-Conversion Applications", *Display Technology, Journal of*, **vol. 4**, no. 2 pp. 188–197, June 2008, ISSN 1551-319X.

[177] C.N. Cordes and G. de Haan, "Key requirements for high quality picture-rate conversion", *SID Digest of Technical Papers*, **vol. 15**, no. 2 pp. 850–853, June 2009.

[178] E.B. Bellers, "Motion Compensated Frame Rate Conversion for Motion Blur Reduction", *SID Digest of Technical Papers*, **vol. 38**, no. 1 pp. 1454–1457, May 2007.

[179] G.A. Thomas, "Television Motion Measurement for DATV and other applications", *Tech. Rep. PH-283*, BBC Research Department, 1987.

[180] Snell's Alchemist Ph.C-HD motion-compensated standards converter, see http://www.snellgroup.com/news-and-events/press-releases/625/snell-announces-1080p-support-for-alchemist-ph.c-hd.

[181] F. Dufaux and F. Moscheni, "Motion Estimation Techniques for Digital TV: A Review and a New Contribution", *Proc. IEEE*, pp. 858–876, June 1995.

[182] E. Bellers, J.W. van Gurp, J.G.W.M. Janssen, J.R. Braspenning, and R. Wittebrood, "Solving occlusion in Frame-Rate up-Conversion", in *Digest of the ICCE*, Jan. 2007, pp. 1–2.

[183] J.R. Jain and A.K. Jain, "Displacement Measurement and Its Application in Interframe Image Coding", *IEEE Trans. Commun.*, pp. 1799–1808, Dec. 1981.

[184] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro, "Motion-Compensated Interframe Coding for Video Conferencing", in *Proc. Nat. Telecom. Conf.*, Nov./Dec. 1981, pp. G 5.3.1–G 5.3.5.

[185] R. Srinivasan and K. Rao, "Predictive Coding Based on Efficient Motion Estimation", *Communications, IEEE Transactions on*, **vol. 33**, no. 8 pp. 888 – 896, Aug. 1985.

[186] S. Zhu and K.K. Ma, "A new diamond search algorithm for fast block-matching motionestimation", *IEEE Transactions on Image Processing*, **vol. 9**, no. 2 pp. 287–290, Feb. 2000.

[187] C. Zhu, X. Lin, and L.-P. Chau, "Hexagon-based search pattern for fast block motion estimation", *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 12**, no. 5 pp. 349–355, May 2002.

[188] P.I. Hosur and K.K. Ma, "Motion vector field adaptive fast motion estimation", in *Second International Conference on Information, Communications and Signal Processing (ICICS)*, Dec. 1999.

[189] A.M. Tourapis, "Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation", *Proceedings of Visual Communications and Image Processing*, pp. 1069–79, Jan. 2002.

[190] J. Chalidabhongse and C.C.J. Kuo, "Fast motion vector estimation using multiresolution-spatio-temporal correlations", *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 7**, no. 3 pp. 477–488, June 1997.

[191] T. Tsai and H. Lin, "Hybrid Frame Rate Up Conversion Method Based on Motion Vector Mapping", *Circuits and Systems for Video Technology, IEEE Transactions on*, **vol. 23**, no. 11 pp. 1901–1910, Nov. 2013.

[192] H. Liu, R. Xiong, D. Zhao, S. Ma, and W. Gao, "Multiple Hypotheses Bayesian Frame Rate Up-Conversion by Adaptive Fusion of Motion-Compensated Interpolations", *Circuits and Systems for Video Technology, IEEE Transactions on*, **vol. 22**, no. 8 pp. 1188–1198, 2012.

[193] Society of Motion Picture & Television Engineers, SMPTE 196M-2003, Motion-Picture Film - Indoor Theater and Review Room Projection - Screen Luminance and Viewing Conditions, Jan. 2003.

Chapter 4

Video based movement analysis during sleep

4.1 Body movement analysis during sleep based on video motion estimation

Abstract

To assess sleep in the home situation, wrist actigraphy is often used. However, it requires an on-body sensor which may disturb sleep and primarily collects data on the movement of one wrist only. Video actigraphy, by estimating motion from captured infrared images, overcomes these issues. In this work, we compare activity levels from wrist and full body video actigraphy in a home setting. Video actigraphy correctly found 19% more small and medium movements that were missed by the wrist sensor. We further show that similar values of sleep efficiency (SE, %) are obtained from simultaneous recordings of video and wrist actigraphy, and we compared both to reference SE values provided by a full polysomnography (PSG). The proposed video actigraphy proved convenient and easy to use in real home situations. It successfully found movements originating from under the blanket, and turned out to be robust to various sleeping positions, different illumination conditions, viewing angles, beds and blankets. Our results suggest that on-body actigraph sensors can be successfully replaced with the proposed video-based solution for the application of unobtrusive sleep monitoring and analysis.

This chapter is published as:

1. A. Heinrich, X. Aubert, G. de Haan; Body movement analysis during sleep based on video motion estimation, *IEEE International Conference on e-Health Networking, Applications & Services (Healthcom)*, pp. 539-543, Oct. 2013.

2. A. Heinrich, X. Zhao, G. de Haan; Multi-distance motion vector clustering algorithm for video-based sleep analysis, *IEEE International Conference on e-Health Networking, Applications & Services (Healthcom)*, pp. 223-227, Oct. 2013.

4.1.1 Introduction

Body movements are an important behavioral aspect of sleep as shown in [194] and [195]. They can be associated to sleep states [196] and be connected to the sleep states' transitions [197]. It has been shown that frequency and duration of body movements are important characteristics for sleep analysis [198]. In order to measure the amount and intensity of nocturnal activity in an automatic manner, so-called actigraph monitors can be worn around one wrist or ankle during sleep episodes. During the last decades, actigraphy (activity-based monitoring) has become an essential tool in sleep research and sleep medicine. Research confirms that it is a cost-effective method for assessing specific sleep disorders and circadian rhythms compared to traditional polysomnography based analysis [199]. Typically, actigraphy is obtained from a wrist device equipped with a small electronic accelerometer and is worn by a person on the non-dominant arm. It has been largely proven that there is good correlation between the output data of actigraphy and polysomnography (PSG) in sleep analysis, e.g., estimation of wake/sleep time, estimation of coarse sleep stages [200, 201]. However, when wearing an actigraph, the sleep quality may be affected due to the discomfort of applying contact sensors on the body.

Other contact-based sensors in bed or on person are considered in sleep research, e.g., pressure sensors in the pillow in [202]. Other mechanical sensor examples of interest for sleep monitoring are given by piezoelectric sensors, strain-gauge and electret foil sensors as mentioned in [203].

It is challenging for sensors that primarily measure movement from a specific body part to render accurate and objective movement information of the sleeping subject, e.g., the actigraph around the wrist will less clearly record other body part movements. Aside from sleep comfort and accurate measurements, installation convenience and short waiting times for the output the next morning are of high importance to the user. Attaching sensors to the body, fixing them to the bed or having to purchase specific bed items to set up the system are all obstacles. Therefore, we propose a real-time feasible system with low computational complexity that performs an off-body motion analysis in sleep where the sleeping subjects can enjoy the natural sleeping environment without being attached to any sensors and which can provide a higher sensitivity and more information than the currently widely used wrist actigraph for movement monitoring in sleep research.

Section 4.1.2 discusses related work and the proposed video-based activity estimation system. Also the experimental setup and data collection are described. The evaluation of the proposed video actigraphy system is presented in Section 4.1.3. Conclusions are drawn in Section 4.1.4.

4.1.2 Methods - Video-based activity estimation system

We developed an off-body near infrared (NIR) sleep monitoring system with a NIR camera, a NIR light source being invisible to the human eyes and a video analysis algorithm to extract the movements of the sleeping person. It can handle many

viewing angles and NIR lighting settings, which makes installation in the bedroom easy. The NIR sleep monitoring system can perform the analysis in real-time and with volatile memory, so that privacy issues are limited.

Related work

NIR video processing is becoming a relevant area as shown by the number of publications over the last years, e.g., [204]. More work is carried out in this area for the application of sleep analysis [205–211]. In [205], a new edge detector is proposed which can deal with the lower dynamic range and contrast in IR images which is claimed to be useful for upper body detection [206] and the diagnosis of obstructive sleep apnea. In [207], the body position and body direction is recognized by using an artificial neural network solution. It requires an exhaustive data set to train the neural network and a specific location of the IR camera (overhead, not possible from the side). Due to the importance of the training, the latter approach is limited and does not work with a bed quilt, different hairstyles of the sleeping subject and different clothes. These three characteristics are all likely to change over time.

In order to analyze general human motion (not targeted on sleeping persons), [208] suggests a version of MHI (motion history image) where a successive layering of image silhouettes is analyzed. The gradient computation of the motion history image returns the motion of the segments in the image. [209, 212] follow a similar approach as [208], but targeted at the application of sleep. The challenges of this approach are the correct assignment of motion direction and magnitude to image areas within the body silhouette due to the variety of spatial and temporal motion regions. Additionally, the gradient of complex motions cannot be found since neighboring segments do not necessarily need to belong to the same movement when different body parts are moving at the same time.

Besides performing MHI, [209, 212] also report video-based activity monitoring. To this end, frame differencing is applied in order to estimate the activity level. A similar approach is described in [210] where the resulting activity count is based on a body part dependent weighting scheme. Hereby motion detection is performed, not motion estimation, thus no information on the direction and the amount of movement is obtained. Local motion analysis with motion vector information is beneficial for several movement based applications for sleep such as advanced video actigraphy, bed-sharing, sleep disorder analysis, e.g., periodic limb movement disorder (PLMD), and body part segmentation.

Several motion estimation methods exist returning motion vectors for image patches/pixels. Spatio-temporal prediction methods have proven to be powerful in the design of motion estimation algorithms [213–218]. They render consistent motion fields in textured and non-textured areas at a real-time feasible computational complexity as is shown in [219, 220]. A motion estimation method is documented in [211] and [221], where the optical flow method is applied to determine motion and even particularly cloth motion in [221] and is also applied in the area of sleep analysis

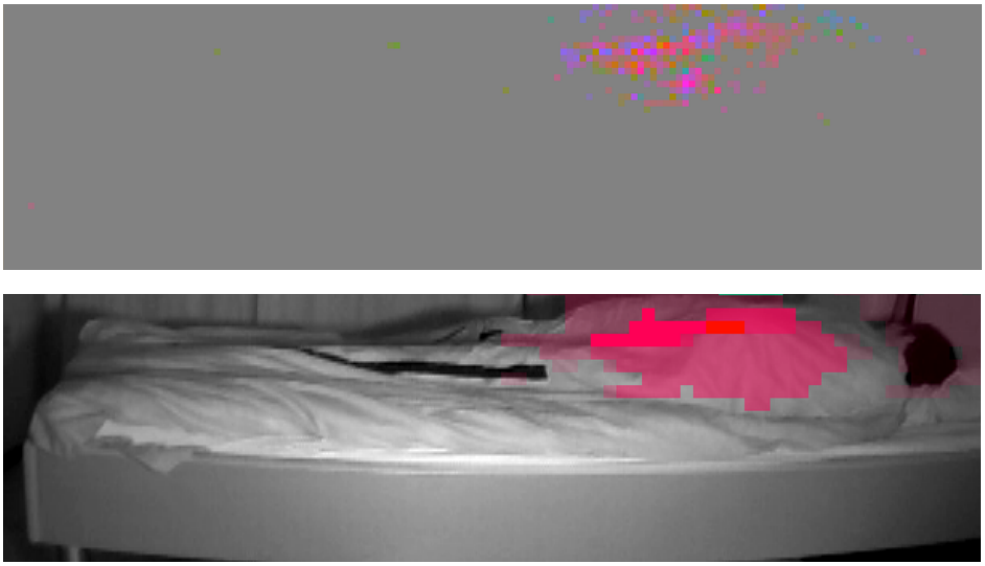


Figure 4.1: Resulting UV color-coded motion vector field with optical flow (top) and recursive search (bottom) motion estimation. The motion vectors from the processed video images are color-coded based on the direction and intensity of the motion.

[222]. Optical flow is based on gradient computation and therefore works well in textured areas, but it has difficulty in non-textured areas as reported in [221], which is often the case for the sleep application where a blanket can have very low frequency patterns. In our experiments, the resulting optical flow vector field tends to be rather noisy and inconsistent (see Fig. 4.1 for a comparison of vector fields obtained with optical flow (top) and recursive search (bottom)). This is a drawback when applications are considered where a reliable estimation of the local motion direction and magnitude are important.

Proposed video actigraphy method

Spatio-temporal prediction methods such as recursive search form the basis of our approach, e.g., [213–215]. In order to extract the motion information of a sleeping person, recursive search motion estimation (RS) is performed on two consecutive video images. This technique returns motion vectors per block and thus allows for a local analysis of coarse body motion. For the purpose of one full-body actigraphy score, motion detection methods would be sufficient where the number of detected motion pixels or blocks in the image indicate the activity level. With the aim towards a flexible system which can easily be extended with local motion analysis, we designed a motion estimation method returning close to true motion fields of the sleeping subject's movements. The video-based actigraphy measure a is derived from the computed motion vectors $\vec{d}$ per block b (b is an element among the set of all blocks

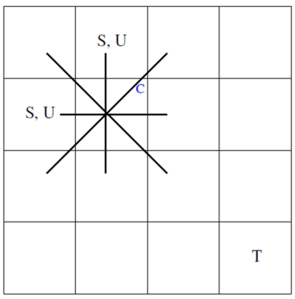


Figure 4.2: Proposed candidate structure where C indicates the current block, S the spatial candidates, U the update candidates, T the temporal candidate and the four lines the means of the motion vectors in the neighboring blocks.

in B) in frame n where the sum of the resulting absolute motion vector components (L_1-norm) ($|\vec{d}(x)|$,$|\vec{d}(y)|$) is used as an estimation of the activity level

$$a(n) = \sum_{b \in B} |d_b(x, n)| + |d_b(y, n)|. \tag{4.1}$$

For the application at hand, the motion vector candidate sets proposed in [214] and [215] are adapted by changing the spatio-temporal candidate locations and by adding mean motion vectors to the candidate set. Limb movements under the blanket and small movements in the blanket may occur and spread over a larger area where the motion within this area would only vary slightly. The addition of the mean candidates showed an improved motion magnitude correlation with the 'ground truth' from 130 annotated short video clips of sleeping subjects. The output motion vector $\vec{d}$ at each image block location $\vec{c}$ is selected from a candidate set C that is based on prediction vectors: firstly, from a spatio-temporal neighborhood for propagation of good estimates and secondly, based on the addition of small random values to the spatio-temporal candidate vectors. This enables faster convergence by finding vectors for newly appearing objects in the video and accommodation for acceleration. It is important that the motion estimator is provided with very similar motion vector candidates to the neighborhood motion for a faster convergence and for capturing local motion changes. Therefore, four mean candidates were added to the candidate set. These four candidates are the means of the motion vectors surrounding the current image block (typically 8×8 pixels) in the four directions starting from the horizontal direction with an angular spacing of 45°. The candidate structure is shown in Fig. 4.2 where c indicates the current block, S the spatial candidates, U the update candidates, T the temporal candidate and the four lines represent the means of the $2k + 1$ motion vectors in the neighboring blocks. The output motion vector $\vec{d}$ can be

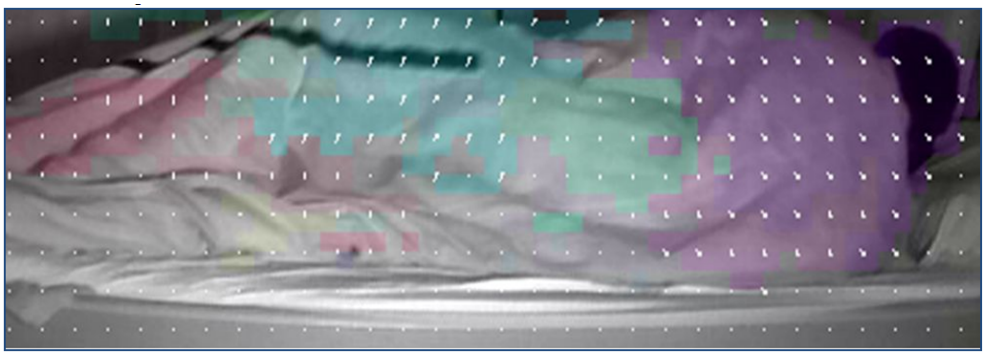


Figure 4.3: Resulting UV color-coded motion vector field when applying the proposed video-based activity estimation method. The motion vectors from the processed video images are color-coded based on the direction and intensity of the motion. The different colors indicate different motion of the corresponding body parts.

computed for each block location $\vec{c}$ in frame number n as in

$$C = \left\{ \begin{array}{c} \vec{d}(\vec{c} - \binom{u}{0}, n), \\ \vec{d}(\vec{c} - \binom{0}{u}, n), \\ \vec{d}(\vec{c} + 2\binom{u}{u}, n-1), \\ \frac{1}{2k+1}\sum_{l=-k}^{k} \vec{d}(\vec{c} + l\binom{u}{0}, n+m), \\ \frac{1}{2k+1}\sum_{l=-k}^{k} \vec{d}(\vec{c} + l\binom{0}{u}, n+m), \\ \frac{1}{2k+1}\sum_{l=-k}^{k} \vec{d}(\vec{c} + l\binom{u}{u}, n+m), \\ \frac{1}{2k+1}\sum_{l=-k}^{k} \vec{d}(\vec{c} + l\binom{u}{-u}, n+m), \\ \vec{d}(\vec{c} - \binom{u}{0}, n) + \vec{\eta}, \\ \vec{d}(\vec{c} - \binom{0}{u}, n) + \vec{\eta}, \end{array} \right\}, \tag{4.2}$$

where u signifies units on the block grid, $2k+1$ the number of blocks we would like to consider for the mean candidates,

$$m = \left\{ \begin{array}{ll} 0, & \text{if } l < 0 \\ -1, & \text{if } l \geq 0 \end{array} \right. \tag{4.3}$$

and where $\vec{\eta}$ is a random value. This random value is drawn from a fixed update set in accordance with [223] for accelerated convergence.

An example of the obtained motion vector field is given in Fig. 4.3 when applying the proposed video-based activity estimation method to a real-life sleep situation.

Test data

Two independent data sets are collected in two experiments. In the first experiment (home setting), six independent volunteers, not part of the research team, 2 female

and 4 males, aged between 24 and 55, not complaining of any sleep disturbances (measured with the Pittsburgh Sleep Quality Index (PSQI) questionnaire [224]), were monitored at home overnight (recording length approximately 8 hours), using both wrist actigraphy (WA) and the video actigraphy (VA) system. For measuring wrist accelerations, two Philips DirectLife Activity Monitors with a (for wrist actigraphy) high sampling frequency of 1 Hz were worn on each wrist. The video actigraphy system was set up by the participants in their bedroom in line-of-sight with the bed. Their whole night's sleep has been recorded with a frame rate of 10 to 15 fps, stored on a laptop and further compared offline with the wrist activity measurements. Note, that viewing angle and NIR light intensity varied between test subjects as the participants were the ones to choose the camera location, adjust the NIR light intensity for the recording and start the recording, based on a limited set of instructions. The recording of one female participant was used as training data for adapting the motion estimation algorithm.

In order to measure the ability of the proposed video actigraphy method for estimating the sleep efficiency, we collected a second video data set (sleep clinic) of four subjects spending one night in a sleep laboratory. The four subjects were independent volunteers and not part of the research team. Each subject wore an actigraphy device (Actiwatch-Spectrum from Philips Minimitter, Bend, OR) on the non-dominant wrist during the whole night while undergoing a full PSG [225] (Alice 5, Respironics) and being filmed by NIR camera. The actigraphy measurements were collected with an epoch length of 60 seconds. A time marker was provided by the subjects by means of a vigorous hand-shaking to indicate the beginning and end of their intended rest period. These data are a subset of those described and analyzed in more detail in [226].

Quantitative evaluation

Experiment 1 (home setting) evaluates the activity count sequences of video and wrist actigraphy with a ground truth observation ('no motion' / 'motion' annotation) obtained by visual inspection of the video recordings. This first data set containing five video sequences has been processed with the developed video actigraphy method (see Fig. 4.4 for a visual comparison between video and wrist actigraphy). 188 motion events in the video recordings have first been visually identified as small (S), medium (M), large (L) and no (0) motion. These 188 events are all movement events of the 5 subjects that have been observed in the video recordings. They contain movements of different duration, movements over different distances performed by single and multiple body parts, whole body movements, movements performed under/above the blanket, movements partially occluded by other body parts, and movements performed in different sleeping positions. The actigraphy counts from the proposed video system and from the reference wrist actigraphy system were subsequently manually classified into small (S), medium (M), large (L) and no (0) motion. This has been carried out by making use of activity amplitude thresholds and movement durations (related to the

area covered by the VA and WA detected motions). The movements were classified and the two measurement methods compared in a Bland-Altman inspired plot [227].

Experiment 2 (sleep clinic) examines the suitability of using video actigraphy for sleep monitoring by comparing its sleep efficiency score with the PSG and wrist actigraphy reference values. The sleep efficiency parameter (defined as the ratio of total sleep time over total time in bed, expressed in %) has been estimated for four subjects spending one night in a sleep laboratory. The time markers indicating beginning and end of intended rest were automatically detected to give the rest-interval bounds.

The video motion analysis data have been processed in three steps. First, the video data collected at a rate of about 15 fps have been converted to minute sequence by averaging over consecutive sub-intervals of 60 seconds. Second, the obtained values have been clipped above the 90th percentile value from the whole night recording (to resemble the Actiwatch activity counts of a night), and down-scaled to the typical Actiwatch range of activity-counts. Third, the sequence of video-derived activity-counts has been extracted within the time interval provided by the hand-shaking markers. For each subject, this provided a pair of activity-count sequences defined over the same time-interval and spanning the same range of values, as shown in Fig. 4.4 where the common range of activity-counts has been set to [0 1000] arbitrary units.

The resulting activity counts (either from video or wrist actigraphy) have been processed by the same given classifier (see [200]) tagging each minute epoch in terms of sleep or wake state. The sleep efficiency (SE) parameter has been estimated from the sleep/wake tag sequences spanning the time interval spent in bed. Both the classification algorithm and the sleep efficiency computation are based on an algorithm derived from Respironics Actiware [228], using a medium wake threshold value of 40. This classification algorithm is based on the calculation of linear combinations of activity counts over a time-window of five minutes, centered on each epoch.

4.1.3 Results

In the first experiment, 188 movements were classified into small (S), medium (M), large (L) and no (0) motion and the two measurement methods video and wrist actigraphy compared in a Bland-Altman inspired plot [227] presented in Fig. 4.5. The difference of the two results is displayed on the vertical axis versus their arithmetic mean on the horizontal axis. The size of the circles has been scaled proportionally to the number of occurrences belonging to the corresponding motion category. The number of occurrences appear inside the respective circles for the four most frequent cases with, respectively, 33 (17.5%), 23 (12.2%), 53 (28.2%) and 38 (20.2%), representing about 80% of all detected movements. The level of correspondence between the two methods may be evaluated from the number of near-zero differences, which are the circles close to the horizontal zero level in Fig. 4.5.

The motion data obtained by the VA system correspond in 61% of the 188 cases to the WA signals for small, medium, and large motions, which is derived from summing

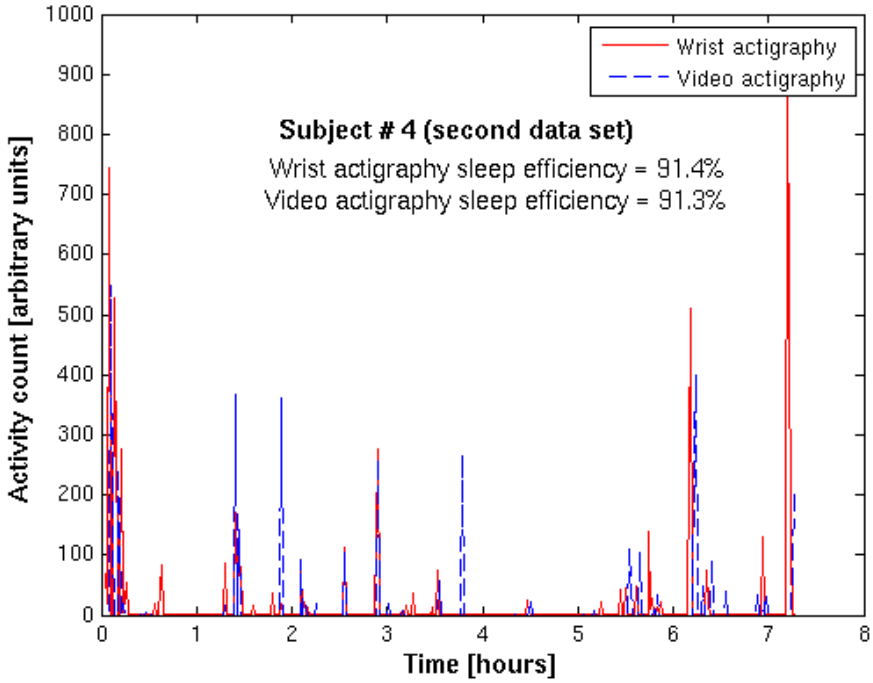


Figure 4.4: Activity counts during a subject's sleep with video actigraphy (blue) and wrist actigraphy (red).

up the contributions of the three circles on the zero-difference level ($12.2\% + 28.2\% + 20.2\% = 60.6\%$). Ground truth comparisons revealed that the reference wrist-sensor failed to detect motion in 20.2% of all body part movements. These small (17.0%) and medium (3.2%) motions not detected by wrist actigraphy, e.g., small leg movements, represent an opportunity area for the video system. Only one event (0.5%) was falsely detected by the video actigraphy system due to a light beam passing the bed. An appropriate selection of the emitted NIR light frequency and a corresponding narrow-band filter on the camera sensor would make the system more robust to local lighting changes.

Furthermore, movements from arm, leg, head and torso, or caused by tossing and turning, were detected even though the subjects were sleeping under a blanket in various positions. The VA system could handle the different illumination conditions, viewing angles, beds, and blankets well.

The participants were asked to set up the system and perform a recording themselves. Since the users chose the lighting settings, the acquired video images differed in their overall brightness. Nevertheless, the proposed video actigraphy algorithm could process the videos with the same motion estimator settings. The large amount of useful video data returned by inexperienced users having received limited instructions

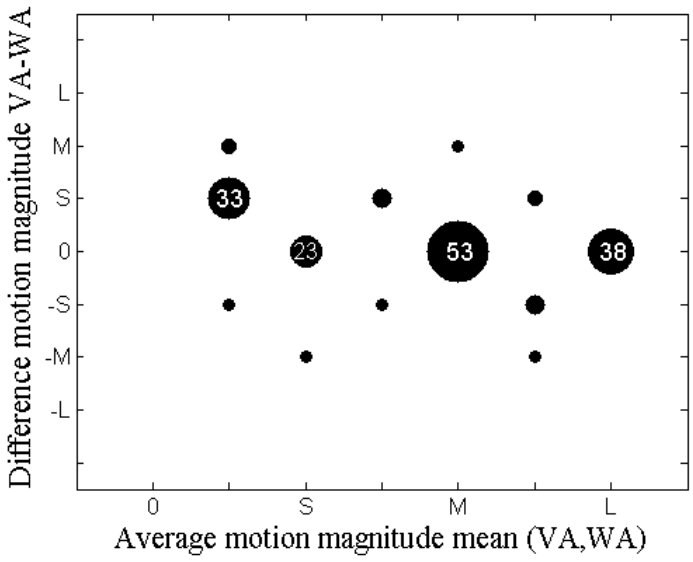


Figure 4.5: Activity count validation for small (S), medium (M), large (L) and no (0) motion. Differences of VA and WA (y-axis) vs. their arithmetic mean (x-axis). The size of the circles indicates the number of occurrences.

gives an indication that the proposed system is convenient and easy to use.

From Fig. 4.4, it may be seen that both activity-count sequences obtained, respectively, from the wrist device and the video motion analysis, are quite similar although a slight drift along the time axis may be observed in the second part of the night. This is due to the non-uniform sampling of the video recording. It may also be observed in Fig. 4.4 that the video-motion analysis produces a few additional activity peaks of medium amplitude, compared to the wrist-device measurements, while the latter shows more small amplitude peaks.

The obtained SE values are shown in Table 4.1, together with the PSG reference values. The SE values derived from Video Actigraphy (VA) are similar to the wrist actigraphy (WA). They are in fair agreement with the PSG values with an average error of 5.8% vs. 4.5% for VA and WA, respectively. It must be pointed out that the applied procedure did not exploit the full potential of video actigraphy, both in terms of time resolution and in terms of dynamic range. The goal of the experiments described in this section is just to present a first demonstration that video actigraphy is able to provide at least comparable SE results with respect to on-body accelerometer-based actigraphy.

4.1.4 Conclusion

In this work, we have shown that our proposed contactless, off-body video actigraphy system could successfully replace on-body actigraphs to monitor a sleeping person's

Subject	SE [%] PSG	SE [%] VA	SE [%] WA
1	80.2%	88.2%	89.0%
2	89.9%	78.0%	85.9%
3	91.3%	92.5%	88.0%
4	93.3%	91.3%	91.4%
Average error	0%	5.8%	4.5%

Table 4.1: Sleep Efficiency (SE) estimations from PSG, Video (VA) and Wrist Actigraphy (WA) on data collected in the second experiment.

movements. Different types of motion from different body parts are well estimated and correspond exactly to the reference wrist actigraphy results in more than 60% of the cases. The remaining movements where some disagreement appears between the two VA and WA methods, may actually represent opportunity areas for the video system. Indeed, we have shown that the video actigraphy method contains more comprehensive information and is generally more sensitive than wrist actigraphy. This is especially true for the 20% of body motions correctly detected by the video actigraphy system and missed by the reference wrist sensor. Our current setup showed one false positive due to a short episode of light change in the room. Future work can investigate the appropriate selection of the emitted NIR light frequency and a corresponding narrow-band filter on the camera sensor which would make the system more robust to local lighting changes.

Concerning the estimation of sleep efficiency from activity patterns, the results obtained from video-based motion analysis can be regarded as quite positive, given that the sleep-wake classification algorithm has been optimized for wrist actigraphy. The average video-based sleep efficiency error amounts to 5.8% vs. 4.5% with wrist actigraphy. These similar results give a first indication that video actigraphy can compete with wrist actigrahpy. Besides, the method used for deriving activity-counts from video-motion is also susceptible of improvements, both in terms of amplitude dynamic and time-resolution, below the usual minute epochs. The system is convenient and easy to use in real home situations. For the limited set of 9 test subjects we found that it is robust to movements originating from under the blanket, to various sleeping positions, different illumination conditions, viewing angles, beds and blankets.

Future work will explore the system's opportunities for improved sleep-wake classification and go beyond the possibilities of wrist actigraphy, such as motion analysis of specific body parts over time, which is relevant for e.g., periodic limb movement disorder. With further developement, the system may become a cheap and easy to use solution for personalized sleep evaluation and early screening of sleep disorders in the home environment.

4.2 Multi-distance motion vector clustering algorithm for video-based sleep analysis

Abstract

Overall health and daily functioning deteriorate with poor sleep. Sleep monitoring can help identify causes of sleep problems. As an advantage over traditional wrist actigraphy used in home sleep monitoring solutions, video contains more comprehensive movement information. Particularly, different body movements can be distinguished which is beneficial for a more detailed sleep analysis. We developed an efficient K-Means clustering method with a multi-distance seeding technique to find the dominant cluster candidates. An integrated multi-distance dissimilarity measure was used for the subsequent clustering. We present an automatic content-dependent weight tuning method for the dissimilarity measure to balance between different distance descriptors. This discriminative algorithm partitions similar body movements in the same cluster. We were able to produce several dissimilarity measures producing clusters that agreed 67% with manual clustering of motion vectors by one expert. Similar clustering characteristics were preferred by both the five expert annotators and the suggested clustering algorithm. This gives us confidence that the proposed optimization method can be used in the future.

4.2.1 Introduction

Overall health and daily functioning deteriorate with poor sleep. To improve one's sleep, sleep monitoring can help identifying causes of poor sleep. Actigraphy is a method of monitoring human rest/activity cycles, and it has been used to study sleep patterns for over 20 years [199]. In order to acquire the data for sleep analysis, the patient needs to wear a so-called Actiwatch (e.g., [228]) on the wrist of his/her non-dominant arm before he/she goes to bed. While research results have shown that wrist actigraphy is a cost-effective method for assessing specific sleep disorders and circadian rhythms [199], it also impacts the sleeper's comfort as it is an on-body sensor. Moreover, it is sometimes lacking specificity regarding sleep-wake classification [229]. As an alternative, an infrared video based sleep analysis approach (video actigraphy) can offer a more comfortable solution while at the same time providing insight into the local motion analysis of a sleeping subject. The more comprehensive local motion information obtained with video can show the strength of body movements, the direction of the movements and how the different body movements are correlated with each other. This last element is particularly interesting since it enables monitoring of two sleepers in a bed-sharing context with one sensor, sleep disorder analysis, e.g., periodic limb movement disorder (PLMD), and body part segmentation.

In order to obtain information on body part motion from the motion vector field, clustering movements corresponding to the same body part is beneficial. Similarly, the relevance of such local motion information of a sleeping subject is recognized in

[230], however no automatic interpretation of the obtained motion vectors is done (e.g., which vectors should be treated as spanning over one dominant motion area or which vectors describe the movement of the same body part). In order to address this overall goal to study movement characteristics of different body parts, we estimate movements in the entire body and transform them into representative motion vectors by clustering similar motion vectors. This work will focus on the second step of this process, i.e., clustering the motion vectors such that the results are discriminative and agree with expert clustering.

The proposed clustering method is outlined in Section 4.2.2. Section 4.2.3 and Section 4.2.4 describe the test data and the evaluation methods, respectively. Section 4.2.5 presents results and discussion, and Section 4.2.6 concludes this work.

4.2.2 Proposed clustering method

In order to perform motion vector clustering, K-Means [231] clustering is used. The core idea was developed almost half a century ago and is successfully employed nowadays (e.g., in [232]) because it is one of the simplest unsupervised machine learning algorithms to solve the clustering problem [233]. [233] even claimed it to be the most popular clustering method employed in scientific and industrial applications as it is simple and straightforward. Alternative clustering methods could be employed as well, however, K-Means' implementation simplicity and computational efficiency allows it to run on large data sets, e.g., on over-night video sequences.

In K-Means implementations, random seeds are often selected as starting points [233]. The K-Means performance thus depends on these randomly selected seeds, resulting in a non-deterministic outcome and, at times, accordingly in non-representative clusters. Some research has been invested into finding a good set of seeds, e.g., [234, 235]. Both studies improve the selection of the seeds based on the spatial distribution of the data but do not take any data-inherent characteristics into account. For the application at hand where we are dealing with motion vectors, we apply a seeding technique in our proposed clustering method which adds angle and length information to the spatial distance cue to determine an initial set of clusters and their starting points. Clustering motion vectors is also the objective in [236], yet, K-Means^{++} proposed in [234] is employed which we believe can be improved on for motion vector data. Since cues other than the spatial proximity are considered, we call our approach multi-distance K-Means clustering. It resembles the 'clustering with obstructed distance' approach documented in [237] where two data points are assigned to different clusters due to a separating obstacle (gray line in Fig. 4.6) although their Euclidean distance is small. Such an obstacle could be represented by e.g., a large difference in motion vector direction. The obstacle or obstructed distance is what we describe with a multi-distance seeding technique and a multi-distance dissimilarity measure in the subsequent cluster assignments. The idea is illustrated in Fig. 4.6.

The flow graph of the whole clustering process is shown in Fig. 4.7. In the following sections, we present an efficient seeding technique (Pre-Clustering) to find an appro-

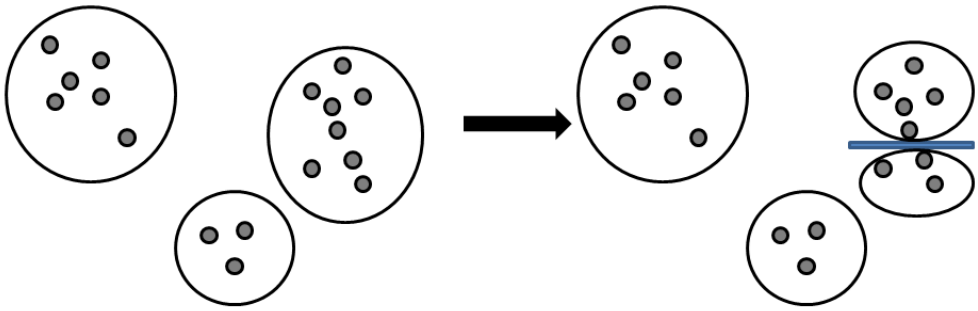


Figure 4.6: Traditional K-Means cluster assignment (left) and proposed multi-distance clustering approach where data properties can lead to assigning two data points to two different clusters despite a small Euclidean distance (right).

priate set of starting points and number of clusters. Unlike the traditional K-Means algorithm, in which only one parameter is selected for measuring the data's similarity to the centroids, in the proposed Adapted K-Means Algorithm, motion vectors' angle and length information as well as their spatial distance are taken into account as discriminating characteristics. These three measure descriptors will be weighted automatically and optimally according to the proposed indexing effectiveness metric (Optimized Dissimilarity Measure) to get the most representative clustering results.

Multi-distance seeding technique

Motion vector clustering is the assignment of dividing a set of motion vectors into subsets (called clusters) so that the vectors in the same cluster are similar to each other, i.e., neighboring vectors with similar direction and similar length. This section

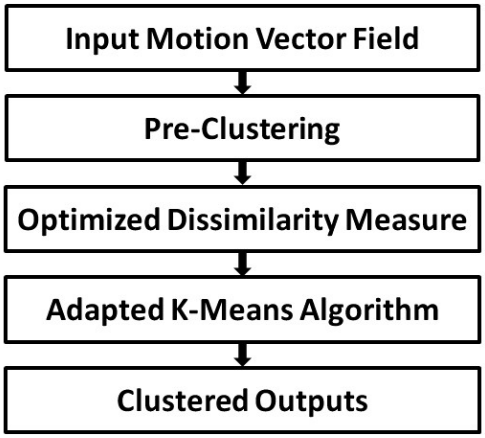


Figure 4.7: Clustering flow of the proposed content-dependent weight tuning method.

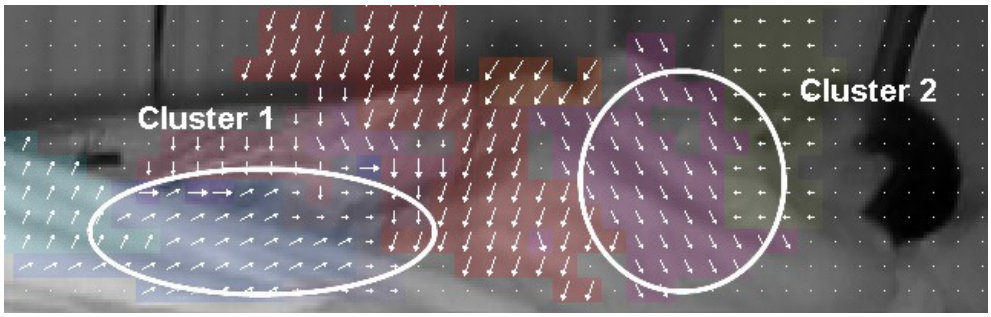


Figure 4.8: Example input frame for clustering motion vectors. Two clusters are indicated with different colors and arrows.

features a pre-clustering process to find the starting points of the dominant movements with the highest potential to form final clusters.

Pre-cluster calculation A snapshot illustrating the clustering task is shown in Fig. 4.8. The motion vector information is obtained by performing a variant of Recursive Search Motion Estimation (ME) [213, 214] which has been proven to provide good results for the application of sleep analysis in [238]. Our ME assigns one motion vector to each 8×8 pixel block in the frame which is illustrated with an arrow and color coding. Based on subjective inspection, the blue motion vectors around the legs (cluster 1) and purple vectors around the chest (cluster 2) can potentially form final motion vector clusters, because they cover a relatively large area and contain a lot of similar motion vectors.

In order to initially identify different motion vector clusters based on the motion vectors and their connectivity, an improved *Row-by-Row Labeling* algorithm [239], which is classical for connected components labeling, is proposed. This allows us to place an obstacle between two motion vectors if their angle or length disagree although these two motion vectors are direct neighbors. The algorithm makes two passes over the image: one pass to record equivalences and assign temporary labels and the second to replace each temporary label by the label of its equivalence class, which is implemented with a *Union-Find* algorithm [239]. The traditional binary image connectivity check is carried out by checking the blocks of the current component and its neighbor, i.e., if both blocks have the same label it means they are connected. However, for motion vector pre-clustering, the angle and length of the motion vector are also considered in this multi-distance approach. See Fig. 4.9 for an example. On the left the binary output is shown after applying the traditional *Row-by-Row Labeling Algorithm* and *union-find* algorithm. The image on the right displays the modified *Row-by-Row Labeling Algorithm* and *Union-Find* algorithm by taking into account the motion vectors' angle and length. Note that for the right example in Fig. 4.9, one block consists of 8×8 pixels. The data point patterns of the two images are the same, but since the motion vectors of the blue and red parts are different, they are

separated into two clusters. Therefore, we altered the algorithm to single out clusters of motion vectors with similar motion directions and magnitudes.

Starting point selection We have empirically observed in recorded videos of sleeping people that we can expect fewer than 10 motion clusters in the typical sleep movements, e.g., turning, stretching legs, moving arms and head, etc. Therefore, we select ten clusters with the highest potential to represent dominant body part movements. Here, the idea of *energy level* is introduced. A pre-cluster's energy level is the product of its mean vector length of all vectors in the cluster and its area (number of motion vectors). According to different application scenarios and sensitivity requirements, the top (in our case ten) pre-clusters or the pre-clusters of which the energy level is above a threshold are selected. The centroids of the chosen pre-clusters are taken as the starting points for the K-Means algorithm. With this method, both intense motions (e.g., rapidly moving hands) and slow motions (e.g., turning) are detected. Fig. 4.10 is the pre-clustered result of Fig. 4.8. The left part shows all the pre-clusters obtained, and the right part shows the top 7 clusters of which the energy levels are higher than the pre-defined empirically determined threshold of 60.

Objective multi-distance dissimilarity measure selection and content-dependent weight optimization

In the seeding process, the starting points for the K-Means algorithm have been determined. The next step is to find a good dissimilarity measure to be used in the adapted K-Means algorithm to cluster not only according to spatial distance but also according to distances in motion vector length and angle. The dissimilarity measure consists of the appropriate descriptors and their corresponding coefficients. We present here a technique to compute an automatic, content-dependent dissimilarity measure with the goal to discriminate well among different motion vectors.

The following descriptors are used for the dissimilarity measure:

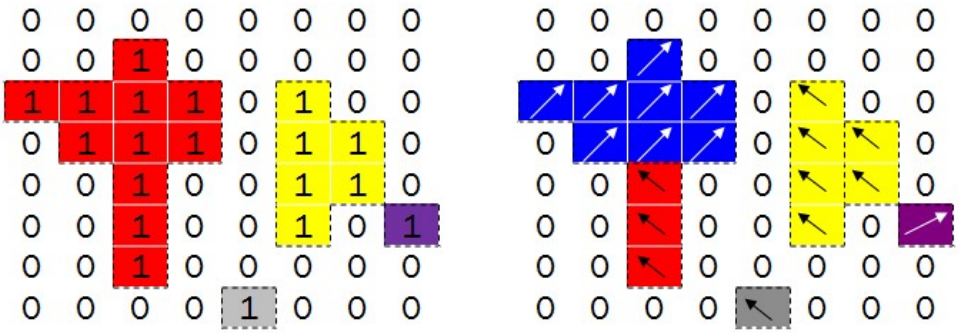


Figure 4.9: Binary image and motion vector samples after applying the traditional (left) and improved (right) row-by-row labeling and union-find algorithm.

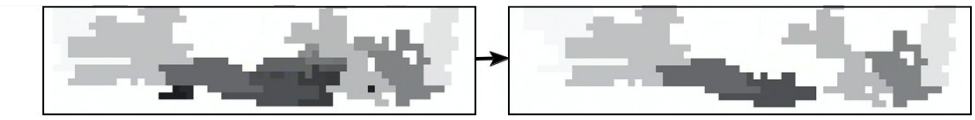


Figure 4.10: All pre-clustered results (left) of Fig. 4.8 and top 7 pre-clustered results according to the energy level computation (right).

- Motion vector length difference
- Motion vector angle difference
- Spatial distance (the distance between the coordinates of two blocks)
  - L_1-norm distance (Manhattan distance) or
  - L^2-norm distance (Euclidean distance) or
  - Square root of L^2-norm distance

Descriptor	**Measure 1**	**Measure 2**
Length difference	$\|\vec{A}\|$-$\|\vec{B}\|$	$\|\vec{A}\|$-$\|\vec{B}\|$
Angle difference	$\|1-\cos(\angle\vec{A}-\angle\vec{B})\|$	$\|1-\cos(\angle\vec{A}-\angle\vec{B})\|$
Spatial dist. betw. motion vectors	L^2-norm	L_1-norm
Descriptor	**Measure 3**	**Measure 4**
Length difference	$\|\vec{A}\|$-$\|\vec{B}\|$	$\|\vec{A}_x-\vec{B}_x\|$, $\|\vec{A}_y-\vec{B}_y\|$
Angle difference	$\|1-\cos(\angle\vec{A}-\angle\vec{B})\|$	
Spatial dist. betw. motion vectors	$\sqrt{L^2-\text{norm}}$	L_1-norm
Descriptor	**Measure 5**	**Measure 6**
Length difference	$\left\|\vec{A}-\vec{B}\right\|$	$\left\|\vec{A}-\vec{B}\right\|$
Angle difference		$\|1-\cos(\angle\vec{A}-\angle\vec{B})\|$
Spatial dist. betw. motion vectors	L^2-norm	L^2-norm

Table 4.2: Summary of six dissimilarity measures. $\vec{A}$ and $\vec{B}$ represent two motion vectors.

Table 4.2 shows an overview of the examined combinations of these descriptors. They can be combined to Measures 1, 2, 3. Furthermore, we added three additional measures in our investigation to analyze the trade-off between efficiency and importance of the angle difference between motion vectors (Measure 4), angle difference integrated with length difference in one descriptor (Measure 5), angle difference added separately to Measure 5 (Measure 6). So in total, six dissimilarity measures are proposed and evaluated. For angle difference, $|1-\cos(\angle\vec{A}-\angle\vec{B})|$ is used so that the difference is monotonically increasing.

Since there are multiple descriptors compared to the traditional K-Means algorithm, a weighting factor tuning method is desired. A similar tuning method is proposed for image retrieval in [240] where the discriminative power of each descriptor is automatically computed based on a dissimilarity histogram. We have adapted this method to our application with motion vectors as follows. The equivalent histogram capacity graph C of a dissimilarity measure descriptor D is the distribution of the dissimilarity between all possible motion vector pairs. It is used to quantify the indexing effectiveness E of motion vector descriptors and identify the best dissimilarity measure for a clustering method. The formula for calculating E is given by

$$E = \int_0^{100} x \cdot C(x) \mathrm{d}x,$$

where x is the normalized dissimilarity between 0 and 100. The weighting factor W_i of each descriptor D_i within a dissimilarity measure is proportionally assigned according to the indexing effectiveness of the corresponding dissimilarity measure descriptor, which objectively shows the importance of the descriptor. Formally,

$$W_i = \frac{E_i}{\sum_{j=1}^{N} E_j},$$

where i, j denote the i-th and j-th descriptor, N the number of descriptors in one dissimilarity measure. The final dissimilarity measure D_{tot} is a weighted sum of the descriptors.

$$D_{tot} = \sum_{i=1}^{N} W_i \cdot D_i.$$

4.2.3 Test data

The proposed method has been designed based on eight representative short movement sequences from four different test subjects recorded overnight with a frame rate of 10 fps. The test subjects were considered healthy sleepers not complaining of any sleep disturbances (measured with the Pittsburgh Sleep Quality Index (PSQI) questionnaire [224]). The movement clips contained different types of motion with few (approximately 5) and multiple (> 5) motion areas of different dimensions and under different viewing angles. For the evaluation of the six dissimilarity measures, nine other 10-second test sequences were extracted from the four overnight recordings, which were taken to compare the clustering results with the subjective impression of clustering. One to two sequences were selected out of each overnight recording. Each sequence consists of 100 frames with a resolution of 640 $\times$ 480 pixels.

4.2.4 Evaluation methods

The evaluation is performed according to two aspects, namely the clustering ability among different dissimilarity measures and the quantitative agreement with the visual ground truth cluster annotation.

Clustering ability A

We aim for a clustering algorithm that accumulates similar motion vectors in one cluster. When motion vector pairs not belonging to the same cluster are counted, a high occurrence rate is expected for large dissimilarities and a low occurrence rate for small dissimilarities (as these occur in the same cluster). Therefore, we propose to measure the clustering performance based on the histogram dissimilarity graph C_d. A linear penalty function $p(x) = x - 100$ (for high occurrences in low dissimilarities) and linear reward function $r(x) = x$ (for high occurrences in high dissimilarities) are applied when computing the clustering ability A,

$$A = \int_0^{100} C_d(x) \cdot p(x) \cdot r(x) \mathrm{d}x.$$

Only motion vector pairs from different clusters are taken into account in C_d (computed with one optimized dissimilarity measure).

Quantitative agreement with expert cluster annotation

In order to assess how well the indexing effectiveness clustering method reflects the subjective assignment of motion vectors to a sleeping subject's moving body part, an evaluation was carried out on a subset of the nine test sequences. In total, 50 frames with different movement information were selected out of the eight movement test sequences. For each frame, we annotated several rectangular image areas that should belong to one cluster based on subjective visual assessment. Two examples are shown in Fig. 4.12. A pixel-level comparison between the processed clusters computed with the six dissimilarity measures and the pre-defined rectangles yields the matching scores (percentage of matched pixels).

Measuring reliability of the single expert annotation and general expert clustering behavior is computed by comparing manual annotations from five different sleep or patient behavior experts on ten images.

4.2.5 Results and discussion

Clustering ability A

An example of dissimilarity graph C_d of a test sequence is shown in Fig. 4.11. The first maximum of Measure 3 occurs at larger dissimilarities than the first maxima of the other measures. Measure 5 and 6 have a high occurrence peak for small dissimilarities.

From the results listed in Table 4.3, we can conclude that, firstly, Measure 3 has the best performance. Furthermore, Measure 1 and 2 have similar results which implies that the L_1 and L^2 norms have similar influence on clustering results for this data set. Measure 3 uses $\sqrt{L^2-\text{norm}}$ which makes it score higher. Measure 4 with a cheap distance computation (L_1-norm) and without any angle calculation scores better than Measure 5 and 6 which combine angle and length difference in one descriptor.

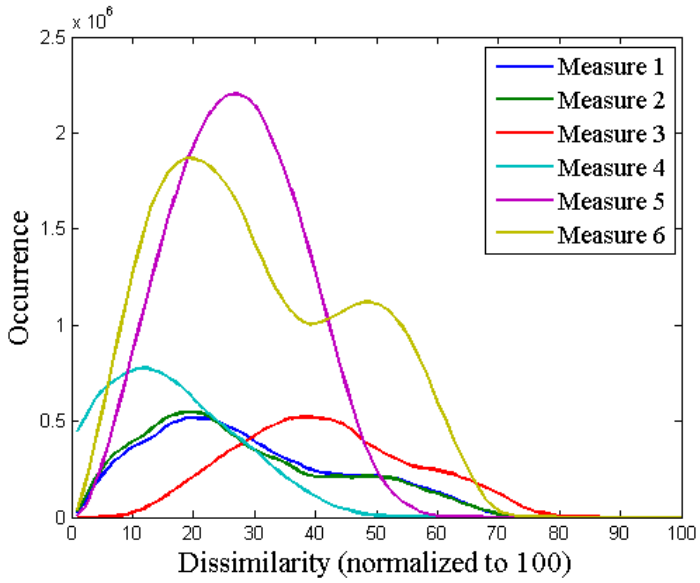


Figure 4.11: Dissimilarity graph C_d of a test sequence.

Quantitative agreement with expert cluster annotation

The matching scores with expert annotation are given in Table 4.4.

The correspondence with the annotated frames agrees highly with the results obtained with the clustering ability A. The best matching scores are achieved by the first three measures which differ only in the descriptor for the spatial distance. From their identical matching scores, we can conclude that the variation in the spatial distance computation (L^2-norm, L_1-norm, $\sqrt{L^2-\text{norm}}$) does not play a discriminative role for the cluster assignment. Dissimilarity Measure 4 neglects the angle discriminator while achieving a higher computational efficiency and obtains a slightly lower matching score than the first three measures. That the explicit integration of the angle difference yields better results is also confirmed by the improved matching score with Measure 6 compared to Measure 5. Angle and length difference combined in one descriptor as is done in Measure 5 yields the lowest matching score and is thus not proposed for motion vector clustering methods when movements of sleeping subjects are concerned.

Measure	1	2	3	4	5	6
Mean A	0.75	0.66	0.89	0.41	0.13	0.25
StD A	0.17	0.20	0.31	0.20	0.31	0.19
Rank	2	3	1	4	6	5

Table 4.3: Performance (normalized) and rank of the six measures on clustering ability A.

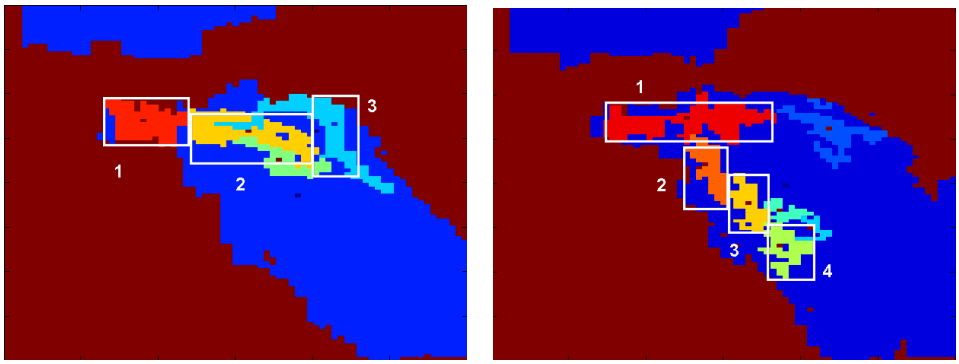


Figure 4.12: Cluster annotation with rectangles in two example frames. The different colors within one rectangle denote different motion clusters computed using a dissimilarity measure.

Overall, the optimized automatic clustering method achieves an agreement of 67% with the annotated motion vector cluster formation. When comparing manual annotations from five different experts, a mutual agreement of only 65% is achieved, suggesting a careful look at the quantitative results regarding the subjective annotation. All expert annotators consistently drew large clusters comprising also motion vectors in different directions (e.g., upper leg and lower leg moving in different directions). The clusters were often linked to body parts suggesting a combination of body part segmentation with motion cluster analysis for future sleep analysis work. The clusters were never intertwined and clearly separated.

An example clustering output is shown in Fig. 4.13 where it is illustrated that the results of Measure 5 can be quite different from Measure 3. In the clustering results given in subfigure C, different colors denote different clusters and the cluster centroids are indicated by light green squares. Measure 3 shows clear and recognizable clusters. However, the clusters of Measure 5 are less recognizable because the clusters are intertwined. This may be due to the usage of the $\left\|\vec{A}-\vec{B}\right\|$ descriptor, which does not separate the motion vectors' length and angle differences in two descriptors. In agreement with the clustering ability, the manual cluster annotation of one expert and the communalities among the five experts, Measure 5 and 6 with multiply intertwined clusters score low, supporting the proposed clustering algorithm.

Measure	1	2	3	4	5	6
Matching score	67%	67%	67%	64%	62%	63%

Table 4.4: Average matching score in percentage of matched pixels for the six dissimilarity measures.

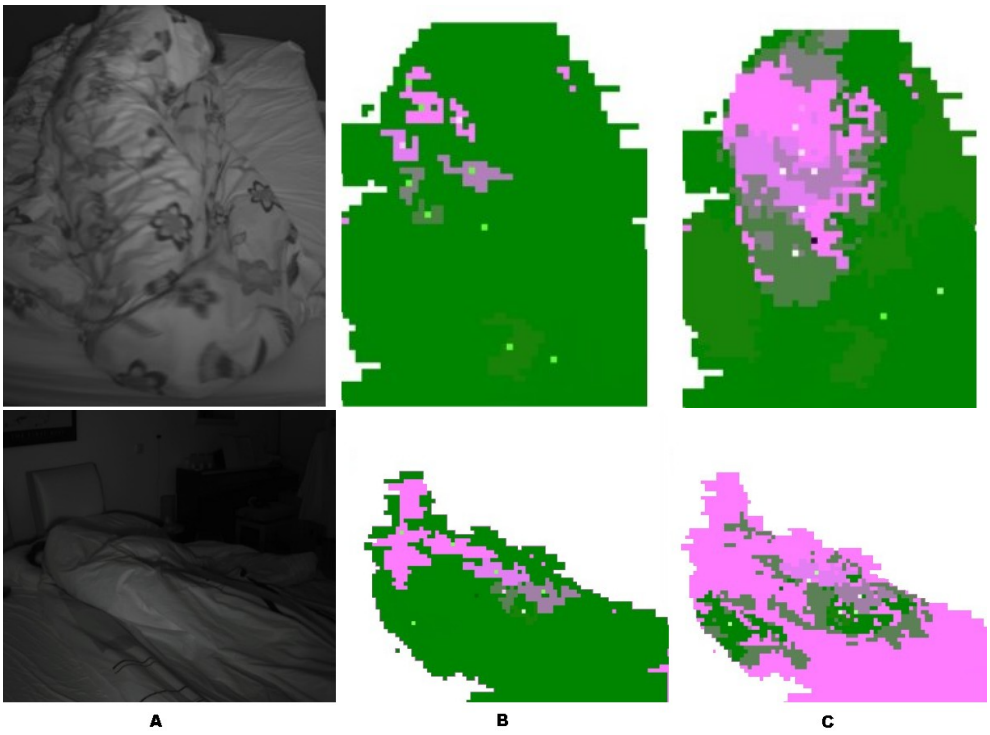


Figure 4.13: Clustering results for two sequences: original frame (A), Measure 3 (B) vs. Measure 5 (C). Different colors denote different clusters, cluster centroids are indicated by light green squares.

4.2.6 Conclusion

In this work, a multi-distance motion vector clustering algorithm based on an enhanced K-Means algorithm is proposed where not only spatial distances between data points but also motion vector angle and length descriptors are used in the dissimilarity measures. A multi-distance seeding method computes the K-Means starting points representing the best cluster candidates. An automatic content-dependent weight tuning method is applied to optimally balance between different dissimilarity measure descriptors. Six dissimilarity measures are investigated with an adapted histogram analysis and compared with manually annotated ground truth clusters. We were able to produce several dissimilarity measures with promising first clustering results where motion vectors of one body part movement are assigned to the same cluster.

Both evaluations, clustering ability A and agreement with manually annotated ground truth, agree with each other, supporting the proposed motion vector clustering algorithm using a content-dependent weight optimization for different distance descriptors. The common clustering behaviors between the high scoring measures and the five expert annotators give us confidence that the proposed optimization method

can be used in future for similar applications such as body part segmentation of a sleeping person.

In both, the clustering ability ranking and the matching scores regarding the manual cluster annotation, the Dissimilarity Measure 1, 2 and 3 perform best. The clustering ability validation suggests an improved clustering performance when using $\sqrt{L^2-\text{norm}}$.

Furthermore, both evaluations agree that an explicit descriptor for the angle difference is beneficial for the cluster assignment. Additional descriptors worth examining in future are body part labels and a temporal descriptor where the cluster history is taken into account for the current cluster computation.

Bibliography

[194] R. Gardner and W.I. Grossman, "Normal motor patterns in sleep in man", *Advances in Sleep Research*, **vol. 2** pp. 67–107, 1976.

[195] A. Muzet, "Dynamics of body movements in normal sleep", *Sleep*, pp. 232–234, 1988.

[196] J. Wilde-Frenz and H. Schulz, "Rate and distribution of body movements during sleep in humans", *Percept. Mot. Skills*, **vol. 56** pp. 275–283, 1983.

[197] A. Muzet, P. Naitoh, R.E. Townsend, and L.C. Johnson, "Body movements during sleep as a predictor of state change", *Psychon. Sci.*, **vol. 29** pp. 7–10, 1972.

[198] S. Gori, G. Ficca, Di Nasso, L. I. Murri, and P. Salzarulo, "Body movements during night sleep in healthy elderly subjects and their relationships with sleep stages", *Brain Research Bulletin*, **vol. 63**, no. 5 pp. 393–397, June 2004.

[199] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, *et al.*, "The role of actigraphy in the study of sleep and circadian rhythms", *Sleep*, **vol. 26**, no. 3 pp. 342–392, 2003.

[200] R.J. Cole *et al.*, "Automatic sleep/wake identification from wrist activity", *Sleep*, **vol. 15**, no. 5 pp. 461–469, 1992.

[201] C.P. Pollak *et al.*, "How Accurately Does Wrist Actigraphy Identify the States of Sleep and Wakefulness?", *Sleep*, **vol. 24**, no. 8 pp. 957–965, 2001.

[202] T. Harada, A. Sakata, T. Mori, and T. Sato, "Sensor pillow system: monitoring respiration and body movement in sleep", in *Proceedings of 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov. 2000, vol. 1, pp. 351–356.

[203] X.L. Aubert and A. Brauers, "Estimation of Vital Signs in Bed from a Single Unobtrusive Mechanical Sensor: Algorithms and Real-life Evaluation", in *30th Annual International Conference of the IEEE EMBS 2008.*, Aug. 2008.

[204] W. Liao and D. Li, "Homomorphic Processing Techniques For Near-Infrared Images", in *Proceedings of 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Dec. 2000.

[205] C.W. Wang, "Real time sobel square edge detector for night vision analysis", *LNCS*, **vol. 4141** pp. 404–413, 2006.

[206] C.W. Wang, A. Ahmed, and A. Hunter, "Locating the Upper Body of Covered Humans in application to Diagnosis of Obstructive Sleep Apnea", in *Proceedings of the World Congress on Engineering, Vol I*, July 2007.

[207] C. Kuo, F. Yang, M. Tsai, and M. Lee, "Artificial neural networks based sleep motion recognition using night vision cameras", *Biomedical engineering: Applications, basis and communications*, **vol. 16**, no. 2 pp. 79–86, Feb. 2004.

[208] G.R. Bradski and J.W. Davis, "Motion segmentation and pose recognition with motion history gradients", *Machine vision and applications, Special issue IEEE WACV*, **vol. 13**, no. 3 pp. 174–184, 2002.

[209] W. Liao and C. Yang, "Video-based Activity and Movement Pattern Analysis in Overnight Sleep Studies", in *19th International Conference on Pattern Recognition*, Dec. 2008.

[210] C.W. Wang, A. Ahmed, and A. Hunter, "Artificial intelligent vision analysis in obstructive sleep apnoea (OSA)", in *30th Anniversary Conference of the Association for Respiratory Technology and Physiology (ARTP)*, Jan. 2006.

[211] K. Nakajima, Y. Matsumoto, and T. Tamura, "Development of real-time image sequence analysis for evaluating posture change and respiratory rate of a subject in bed", *Physiol. Meas.*, **vol. 22**, no. 3 pp. 21–28, 2001.

[212] W.-H. Liao and J.H. Kuo, "Sleep monitoring system in real bedroom environment using texture-based background modeling approaches", *J. Ambient Intelligence and Humanized Computing*, **vol. 4**, no. 1 pp. 57–66, Feb. 2013.

[213] G. de Haan *et al.*, "True-motion estimation with 3-D recursive search block matching", *IEEE Trans. Circuits, Syst. Video Techn.*, pp. 368–379, Oct. 1993.

[214] J. Wang, D. Wang, and W. Zhang, "Temporal compensated motion estimation with simple block-based prediction", *IEEE Transactions on Broadcasting*, **vol. 49**, no. 3 pp. 241–248, Sep. 2003.

[215] S.-C. Tai, Y.-R. Chen, Z.-B. Huang, and C.-C. Wang, "A Multi-Pass True Motion Estimation Scheme With Motion Vector Propagation for Frame Rate Up-Conversion Applications", *Display Technology, Journal of*, **vol. 4**, no. 2 pp. 188–197, June 2008.

[216] G.G.C. Lee, M.J. Wang, H.Y. Lin, D.W.C. Su, and B.Y. Lin, "Algorithm/Architecture Co-Design of 3-D Spatio–Temporal Motion Estimation for Video Coding", *IEEE Transactions on Multimedia*, **vol. 9**, no. 3 pp. 455–465, April 2007.

[217] A.M. Tourapis, O.C. Au, and M.L. Liou.

[218] A.M. Tourapis, "Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation", *Proceedings of Visual Communications and Image Processing*, pp. 1069–79, Jan. 2002.

[219] C.N. Cordes and G. de Haan, "Key requirements for high quality picture-rate conversion", *SID Digest of Technical Papers*, **vol. 15**, no. 2 pp. 850–853, June 2009.

[220] E.B. Bellers, "Motion Compensated Frame Rate Conversion for Motion Blur Reduction", *SID Digest of Technical Papers*, **vol. 38**, no. 1 pp. 1454–1457, May 2007.

[221] V. Scholz and M.A. Magnor, "Cloth Motion from Optical Flow", in *VMV'04*, 2004, pp. 117–124.

[222] K. Cuppens, L. Lagae, B. Ceulemans, S. Van Huffel, and B. Vanrumste, "Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy.", *Med. Biol. Engineering and Computing*, **vol. 48**, no. 9 pp. 923–931, Sep. 2010.

[223] G. de Haan and P.W.A.C. Biezen, "Sub-pixel motion estimation with 3-D recursive search block-matching", *Signal Processing*, **vol. 6**, no. 3 pp. 229–239, June 1994.

[224] D.J. Buysse, C.F. Reynolds III, T.H. Monk, S.R. Berman, and D.J. Kupfer, "The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research", *Psychiatry Research*, **vol. 28**, no. 2 pp. 193–213, May 1989.

[225] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, for the American Academy of Sleep Medicine. 1st ed. Westchester: IL: American Academy of Sleep Medicine, 2007.

[226] S. Devot, R. Dratwa, and E. Naujokat, "Sleep/Wake Detection Based on Cardiorespiratory Signals and Actigraphy", in *Proc. IEEE conference of EMBC*, Sep. 2010.

[227] J.M. Bland and D.G. Altman, "Measuring agreement in method comparison studies", *Stat. Methods Med. Res.*, **vol. 8**, no. 2 pp. 135–160, Apr. 1999.

[228] "Actiwatch for Sleep Evaluation, Philips/Respironics at http://www.learnactiware.com", .

[229] L. de Souza, A.A. Benedito-Silva, M.L.N. Pires, D. Poyares, *et al.*, "Further validation of actigraphy for sleep studies", *Sleep*, **vol. 26**, no. 1 pp. 81–85, Feb. 2003.

[230] W.-H. Liao and C.-M. Yang, "Video-based activity and movement pattern analysis in overnight sleep studies", in *19th International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[231] J. B. MacQueen, "Some Methods for Classification and Analysis of MultiVariate Observations", in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[232] F. Gibou and R. Fedkiw, "A fast hybrid K-means level set algorithm for segmentation", in *4th Annual Hawaii International Conference on Statistics and Mathematics*, 2005, pp. 281–291.

[233] P. Berkhin, "Survey of clustering data mining techniques", *Tech. rep.*, Accrue Software, 2002.

[234] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding", in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, 2007, pp. 1027–1035.

[235] A. Likas, N. Vlassis, and J.J. Verbeek, "The global k-means clustering algorithm", *Pattern Recognition*, **vol. 36**, no. 10 pp. 451–461, Feb. 2003.

[236] L. Grèzes-Besset, J. Schaerer, P. Clarysse, and D. Sarrut, "Apparent motion clustering on Cone-Beam fluoroscopic images for thorax tracking", in *International Conference on the Use of Computers in Radiation Therapy (ICCR)*, June 2010.

[237] A.K.H. Tung, J.Hou, and J. Han, "Spatial clustering in the presence of obstacles", in *Proc. 2001 Intl. Conf. On Data Engineering (ICDE)*, 2001, pp. 359–367.

[238] A. Heinrich and H. van Vugt, "A new video actigraphy method for non-contact analysis of body movement during sleep", *Journal of Sleep Research*, **vol. 19** p. S283, Sep. 2010.

[239] L. Shapiro and G. Stockman, *Computer vision, binary image analysis*, Prentice Hall, 2002, ISBN "0130307963".

[240] R. Brunelli and O. Mich, "Histograms analysis for image retrieval", *Pattern Recognition*, **vol. 34**, no. 8 pp. 1625–1637, Aug. 2001.

Chapter 5

Video based actigraphy and breathing monitoring of a shared bed from the bedside table

Abstract

Good sleep is an important factor for a high quality of life. Presence of sleep disorders requires patients to undergo polysomnography examinations at a sleep clinic, which involves attaching multiple head and body sensors. This results in an uncomfortable and unnatural sleep setting for the patients, complicating an accurate diagnosis. Currently, this diagnosis is done manually, making the entire process time consuming and cumbersome. We propose a camera based system capable of analyzing the subjects in their own sleeping environment. The system segments the primary subject from the background and any other bed occupants, then computes the actigraphy and the breathing characteristics of the subject. Segmentation is performed with an AdaBoost classifier using among others motion, intensity, focus and Histogram of Oriented Gradients (HOG) features. A Sum of Absolute Differences (SAD) operation on the pixels within the segmented area of the primary subject returns the actigraphy signal. The breathing characteristics are extracted based on small motion analysis of consecutive and reference frames. The proposed system has been evaluated on 4 healthy adults for actigraphy and on 5 healthy adults for breathing analysis using a Texas Instrument Chronos wrist watch and inductive respiratory belts as references respectively. Evaluation was performed using the metrics accuracy, precision, sensitivity and breathing rate correspondence. The proposed system has an average accuracy of 88% at a precision of 79% for segmentation of the primary subject. It can detect movements up to an accuracy of 85% while outperforming wrist actigraphy at 75% accuracy. State-of-the-art video based breathing algorithms were surpassed with an overall sensitivity

This chapter is published as: A. Heinrich, F. van Heesch, B. Puvvula, M. Rocque; Video based actigraphy and breathing monitoring of shared beds from the bedside table, *Journal of Ambient Intelligence and Humanized Computing*, vol. 6, no. 1, pp. 107-120, Feb. 2015.

of 87%, precision of 90%, and a breathing rate correspondence of 93%.

5.1 Introduction

The prevalence of sleep problems and costs induced by sleep disorders are high. A polysomnography (PSG) examination in a sleep clinic is considered the gold standard for sleep screening and includes attaching multiple sensors to the head (e.g., EEG, EMG, EOG) and the body (e.g., respiratory effort belts) in order to measure various physiological signals. Patients can suffer from the so-called first night effect from sleeping in a laboratory environment instead of their familiar home setting making an accurate diagnosis more difficult. Additionally, sleep clinics tend to have long waiting lists (weeks to months are not uncommon). The measured signals are usually manually analyzed which is time consuming and cost intensive. With the goal to overcome these disadvantages, home PSG is employed since the 1980s and since then further improved on [241]. As such, this procedure is better suited for monitoring e.g. a couple of consecutive nights, but not for long-term monitoring as on-body sensors and qualified sleep clinicians are still needed.

For consumer applications, fewer and more comfortable sensors are favorable. These less obtrusive sensors include amongst others wrist actigraphy [242], radar sensors [243], pressure sensors in the pillow [244], pressure sensors in the bed sheet [245], air filled tubes combined with a differential pressure sensor [246] under a 4 cm mattress, piezoelectric sensors, strain-gauge and electret foil sensors [247]. Most common are so-called wrist actigraphy devices that are worn around the wrist measuring the subject's activity for deriving sleep/wake episode estimates [242, 248]. The importance of body movements as a behavioral aspect during sleep [249], their association to sleep states [250, 251] and sleep state transitions [252] have been found. To this end, the movements of sleeping subjects are typically analyzed with the aim to perform sleep/wake classification [242] or to screen for diseases characterized by particular movement patterns [253]. A more detailed and improved sleep state classification is accomplished by adding a set of features based on respiration as shown by [254] and [255].

Recent commercial sleep monitoring products are based on less obtrusive sensors such as sensors under the bedsheet [256], under the mattress [257], radar [258, 259] and accelerometer sensors [258]. We propose a camera-based system for computing activity levels and the respiratory waveform where subjects can sleep in their own bedroom without being disturbed by on-body sensors. Sleep laboratory costs are reduced by automatically analyzing the recorded data. The optical sensor allows us to measure both body movements and breathing of a sleeping subject, thereby surpassing the limitations of wrist actigraphy. One of the challenges for the camera system in the home setting is to be able to derive the physiological characteristics of the Primary Subject (PS) in the presence of a bed partner. According to a survey by [260], 62% of the respondents report to sleep with a bed partner. Therefore, we propose a method for differentiating the PS from the bed partner in the image before calculating the actigraphy signal and breathing waveform. To realize an easy-to-use system for the end user, the solution we present in this paper does not require to

mount a camera high up on the wall or ceiling overlooking the bed, contrary to other proposed systems [261–264]. Instead, the camera can be more conveniently placed on the bedside table of the PS who is to be monitored. It can be a standalone device or integrated in existing bedside table products, such as a wake-up light. The viewing angle makes it more challenging to distinguish between movements from the bed partner and the PS as they appear in close proximity of each other and are overlapping in the camera image. A near-infrared camera is used in order to visualize recordings in a dark environment. This introduces the additional challenge to deal with none or low-textured objects. Texture differences that may be clearly visible in daylight may be greatly reduced in the infrared image.

The goal of this research is to develop a sleep monitoring application that can monitor the movements and breathing waveforms of the PS in the presence of a bed partner. Therefore, a segmentation algorithm initially discriminates between the area of the PS and the bed partner, so that, subsequently, the actigraphy and breathing signals of the PS can be computed. We will present the design of the experiments in Section 5.2, the proposed segmentation, actigraphy and breathing methods in Section 5.3, the evaluation methods for each experiment in Section 5.4 and the results and discussion in Section 5.5. Conclusions are drawn in Section 5.6.

5.2 Design of experiment

In this section, the design of the experiments is described for the analysis of activity and breathing during sleep. The sleep analysis application envisions a product in which the illumination source and the camera are combined. Such a product could be easily placed on the bedside table. The camera and the light source are thus in close proximity of each other and near the head of the PS. To avoid the influence of varying heights due to different bedside tables a tripod is used as shown in Fig. 5.1. The height of the camera causes the bed partner to be occluded by the PS for the largest part of his/her body and for most of the time. The viewing angle of the camera has been set such that the PS is completely visible from head to foot. For the video recording, an IDS μEye CMOS camera (USB UI-1220SE-M) with a Fujinon FE185C086HA-1 fish-eye lens has been used. For illumination in the dark, an infrared light source with maximum intensity at wavelength 825 nm is used. The resolution of the image is 752×480 and the frame rate has been set to 10 fps (frames per second).

The wrist actigraphy has been registered with a Texas Instruments Chronos watch and is used as state-of-the-art comparison for the segmentation and actigraphy algorithms presented in this paper.

The recorded data sets are split into two categories, natural and synthetic. The two natural data sets have been acquired in a natural sleeping environment at a subject's home (DATA Home1 and DATA Home2). The movements of the subjects in these data sets are not restricted and reflect the real sleep situation. The two synthetic data sets (DATA Synt1 and DATA Synt2) have been acquired both at a sleep laboratory (Sleep Area of Philips Research Eindhoven) and at home where body movements

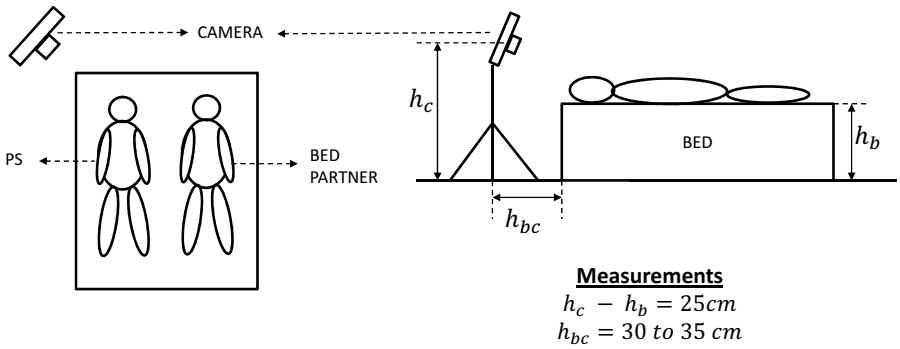


Figure 5.1: Schematic representation of the experimental setup for the actigraphy measurements.

were performed according to a movement protocol. Small movements (single arm, single leg and head movements) and large movements (turning around, moving both arms and legs simultaneously) were performed by both, PS and bed partner, both simultaneously and sequentially.

In a separate experiment for breathing analysis, five subjects (2 females, 3 males) were equipped with the inductive respiratory chest and abdominal belts [265] which are used as reference signals. The test subjects were monitored in four different sleeping poses at the Sleep Area of Philips Research Eindhoven. Snapshots of a test sequence with the four different body positions are given in Fig. 5.2. The subjects were selected based on their differences in Body Mass Index (BMI) (varying from 17.63 to 29.07 with a mean BMI of 22.67). A sweep over several breathing frequencies and depths was carried out in each of the four body positions. The participants followed a metronome in order to approach the breathing rate of the protocol. In each body position, the subjects started with their natural or instinctive breathing rhythm for 2 min, followed by 1 min of 23 breaths per minute (bpm), 1 min of 19 bpm, 1 min of 15 shallow bpm, 1 min of 15 deep bpm, 2 min of 12 bpm, 2 min of 10 bpm and concluding with 3 deep sighs. This data set has been used for the evaluation of the designed breathing analysis algorithm. For the algorithm design, a different data set has been used collected in the framework of another study. Few parameter values were tuned with 2 short breathing periods from the evaluation data set.

Both experiments have been approved by the internal ethics committee of Philips Research.

5.3 Proposed method

In this section, we will describe the proposed method to analyze breathing and activity during sleep using the setup described in the previous section. The method consists of three steps: the PS segementation, the PS actigraphy measurement and

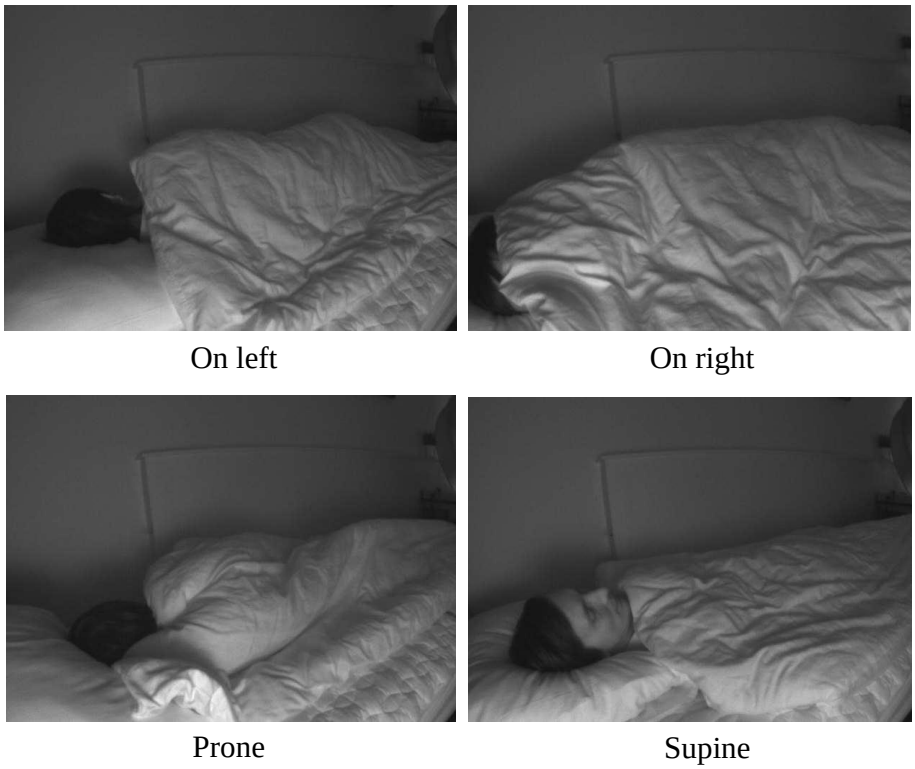


Figure 5.2: Snapshots of a breathing test sequence with four different body positions recorded from the bedside table.

the PS breathing algorithm. They are described in Sections 5.3.1, 5.3.2 and 5.3.3, respectively. PS breathing monitoring is only performed when no subject movement is detected in the PS actigraphy signal.

5.3.1 PS segmentation

In order to automatically segment the PS from its background for each picture in the video, a segmentation algorithm has been implemented using AdaBoost [266]. AdaBoost determines the weights of many weak classification features, such that they form a Least Mean Square (LMS) optimal strong classifier. The following weak classification features have been designed that attempt to segment the PS:

Intensity Given the setup of the camera and the IR light source (Fig. 5.1), the distance to the camera is likely to correlate with the signal intensity. Given that the PS is the closest object in view, it makes sense to use signal intensity as a feature. As such, the camera intensity is normalized using histogram normalization and pixels above a threshold intensity (i.e., 0.5, empirically determined)

are classified as PS.

Focus For recordings in a dark environment, cameras with a large aperture are preferred. Large apertures typically imply a shallow depth of field and, hence, focus can be used to segment a certain depth range. By setting the camera focus such that the PS is in focus, a focus estimator can be used as a feature. The focus estimate, as described by [267], has been used, yielding a focus estimate for each pixel location. This focus estimate result is then binarized, such that focussed edges are found. These edges are then connected, using edge-linking [268], to form regions. The detected focus regions are considered to be part of the PS.

Motion Motion can be used as a classification cue to segment the PS from the background, as the background typically has different motion characteristics compared to the PS. To determine the local motion for each frame, we have used the motion estimation algorithm as described in [269] and [270]. The method estimates the true motion on a (8×8) block basis. The method is known to propagate motion vectors in areas with a homogeneous intensity. This is undesirable for our application and has been circumvented by adding a preference towards the zero-vector (no motion). This has been implemented by adding the zero-vector to the motion vector candidate list. To convert motion estimates to PS classifications, the motion vectors have been spatially clustered based on motion vector size and direction. Depending on the location of the cluster they are considered part of the PS, bed partner or background. Blocks with zero-motion vectors are considered to be part of the background. All pixels within a block are assigned the block label.

HOG The histogram of local and normalized oriented gradients (HOG) feature [271] is a commonly used feature in object classification. The HOG feature is considered, because we expect the PS to have different shape characteristics than the background. HOG histograms are created for the subsequent AdaBoost classification.

Location The setup of the camera with respect to the bed is set such that the PS is in the center of the camera's view. To exploit this, the pixels distances to the center (vertically and horizontally) are used as features. Binary features are created by thresholding multiple distances.

Edges Although edge information is indirectly encoded in both the HOG and focus features, edge information has been added as an additional feature. Multiple edge detectors [268] have been implemented using a fixed kernel size of 3×3 pixels for Laplacaian, Sobel, Prewitt, Laplacian of Gaussian (LOG), and Canny operators and 2×2 for the Robert operator. Edge intensity thresholds are used to create binary decisions.

In order to combine the features described above in an LMS-optimal way, cascaded AdaBoost [272] has been used. In our case, the weights of the features have been

determined using 240 manually annotated frames: 80 frames each from DATA Home1 and DATA Home2, and 40 frames each from DATA Synt1 and DATA Synt2, as ground truth. The 240 frames were selected such that they contain most of the typical poses of a person in a bed.

5.3.2 PS actigraphy

A video based actigraph generation algorithm based on frame differencing, by computing the SAD, and motion estimation is presented by [264] and [273]. The video actigraph signals generated by these methods have been compared with wrist actigraph signals generated by an accelerometer. The video actigraphy method has been found to be sensitive to movements originating from under the blanket, to movements originating from body parts other than the wrist, and robust to various sleeping positions and different illumination conditions. However, in both methods, the camera has been mounted high up on a wall or ceiling. In our study, the camera is mounted on the bedside table and a bed partner is in the camera view. By segmenting the PS first we are robust to bed partner movements. The video actigraph is further computed by performing a SAD operation on the pixels within the segmented area of the PS.

5.3.3 PS breathing

For breathing and sleep monitoring, the breathing rate and global changes in the breathing rate, the regularity/irregularity of the breathing episodes and the tidal volume are important characteristics. Therefore, the breathing waveform is an asset as various features representing above characteristics can be derived from it.

After segmenting the PS according to the method described in Section 5.3.1, the breathing signal is extracted. In this work, the video sequences were obtained from the breathing data set in Section 5.2 with single sleeping subjects. This setting can be easily extended to the shared-bed scenario as the segmentation and actigraphy algorithms employ the same camera placement, light settings and yield high segmentation accuracies as is shown in Section 5.5.

Simple approaches for computing a breathing signal from a video recorded with a near-infrared camera are based on subtracting consecutive frames from each other and mapping the sum of pixels in the difference image to the breathing signal (e.g., [274]). Besides being very sensitive to noise and lighting changes, this method cannot differentiate between inhale and exhale phases and sophisticated post-processing is needed to retrieve the breathing waveform.

In [275], an image sequence is compared with a breath motion template. This template is formed by accumulating differences between consecutive frames and a background model generated during a 'breathing only' period.

A number of other methods for breathing monitoring [276–279] use optical flow approaches based on [280]. For monitoring movements in sleep it is one of the most popular methods [281–284] despite the rather high computational complexity that

comes with it and its sensitivity to noise. The Eulerian Video Magnification approach described in [285] amplifies breathing motion in the video images for improved visualization of the subtle motion. No breathing parameters are automatically derived from the processed images. Subsequent optical flow motion estimation and sums of vertical or horizontal motion vectors are used to compute the breathing waveform in [279].

A recently published method by [286] computes the breathing waveform by performing cross-correlation of the current vertical profile with the vertical profile of earlier images. The vertical profiles are 1D vectors resulting from mean and standard deviation operations on the rows of an image.

[287] make use of an additional pattern projection device to enhance the subtle breathing motion in a video sequence. This approach is more costly in terms of hardware and requires two separate device entities as the pattern is best projected under an orthogonal angle to the camera viewing angle.

In this work, the breathing waveform is computed according to the approach illustrated in the top image of Fig. 5.3. Video chunks of 30 seconds are processed consecutively and independently. The breathing waveform is computed by automatically selecting a *reference frame* during an expiratory pause of the subject. Frames during expiratory pauses are more stable/alike than at other breathing time instances (including maximum inhale peaks) and can easily be reused to render a breathing waveform similar to the reference signal. The algorithmic steps of each 30-second video chunk are the following:

- Consecutive image dissimilarity: The correlation coefficient based on Pearson [288] between consecutive images is computed and subtracted from 1, yielding a 1D signal. A small dissimilarity indicates very small or no motion (which is the case during an expiratory pause and inhale/exhale transitions, i.e., all circles in the bottom image of Fig. 5.3).

- BP: A second order Butterworth band-pass filter removes the high frequencies and the DC component for further processing. A seven-year old child has a respiratory rate range from 18 to 30 respirations per minute and an adult from 12 to 20 respirations per minute [289]. Therefore, we decided to have a detectable breathing rate range between 10 breaths / min and 30 breaths / min.

  In order to be able to capture an expiratory pause, we aim at a worst case target accuracy of 1/6 of a resulting breathing waveform period, corresponding to 6 samples per breath. For the lower boundary of 10 breaths / min, we can determine the boundary frequency: one breathing cycle takes 6 s (60 s / 10 breaths = 6 s / breath), corresponding to 1 Hz ((6 samples / breath) / (6 s / breath) = 1 sample / s). For the upper boundary of 30 breaths / min, one breathing cycle takes 2 s (60 s / 30 breaths = 2 s / breath), corresponding to 3 Hz ((6 samples / breath) / (2 s / breath) = 3 samples / s). Thus, the cut-off frequencies for the band-pass filter are chosen at 1 Hz and 3 Hz.

- Local minima as reference frame candidates: Local minima serve as reference frame candidates since expiratory pauses are assumed to be a subset among the local minima. All circles in the bottom image of Fig. 5.3 are local minima although only the bold circles indicate an expiratory pause (minima in the final breathing signal). The dashed circles are minima resulting at times of the inhale/exhale transition (maxima in the final breathing signal).

- The computation of the expiratory pause reference frame (grey bold circle in bottom image of Fig. 5.3) is based on frame distance clusters, number of edges per image and cross correlation lag of number of edges per row. The reference frame candidates are clustered using the correlation coefficient between them. One of the clusters is likely to be the one with frames from expiratory pauses as the return to the expiratory pause position is likely to yield similar images. In an expiratory pause, the subject is completely exhaled resulting in a higher number of folds of blankets/clothes produced by the collapsing movement of foldable objects (i.e., blankets). Therefore, an edge detector should return more edges in the reference frame images. In an expiratory pause, the edges of the image should be lower than in images surrounding the candidate reference frame. This is computed with the cross correlation lag of number of edges per row. When the confidence of the reference frame history (RFH) is high (based on similarity of previous waveforms) also the current candidate frames should yield a continuation of the previous breathing waveform.

- Average frame of RFH and new reference frame average: In order to increase robustness, an average reference frame is computed as the average of the reference frame history and the new reference frame. When a large motion (movement not due to respiration) is detected, we assume that the subject changed position. The past reference frames are thus not anymore valid and the RFH is emptied.

- The image correlation coefficient between all images and the average reference frame returns the breathing waveform where minima correspond to expiratory pauses and maxima to inhale/exhale transitions.

5.4 Evaluation methods

In this section the methods for the evaluation of the three steps are described. First, the evaluation methods for segmentation of the PS are discussed in Section 5.4.1, followed by the evaluation methods for actigraphy and breathing in Section 5.4.2 and 5.4.3 respectively.

5.4.1 PS segmentation

A classifier has been designed to segment the PS in the video data from the background. The classifier has been created using a training algorithm, which uses a

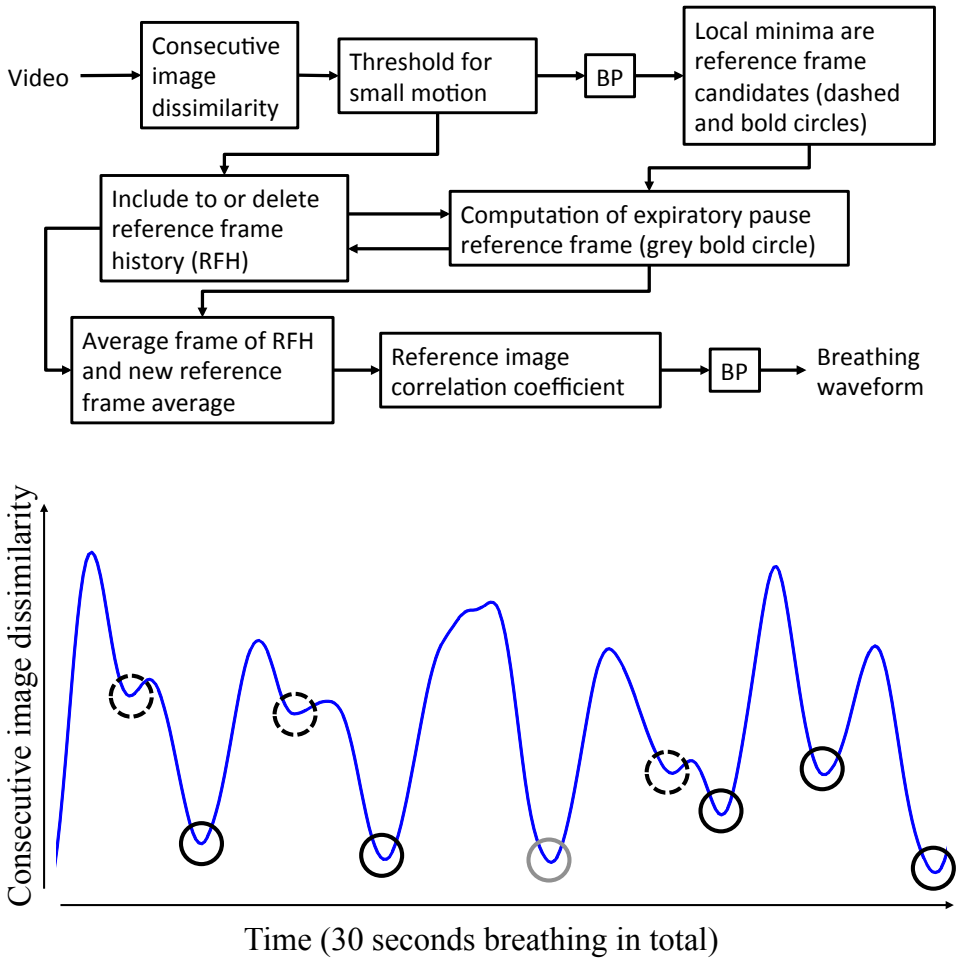


Figure 5.3: Top: Proposed breathing waveform computation method. Bottom: Intermediate output signal after computing the consecutive image dissimilarity. Minima in this consecutive image dissimilarity graph correspond either to expiratory pauses (minima) or inhale/exhale transitions (maxima) in the final breathing signal.

(manually) annotated training set. The annotation has been created using an annotation tool called ITK-SNAP [290] on 240 camera images yielding pixel accurate ground truth. The tool's user interface is shown in Fig. 5.4. The classifier is trained with training data and evaluated with a test data set that is different from the training data set. The classifier has been evaluated by means of K-fold cross-validation with $K = 4$, by determining the accuracy A and precision P according to Eq. (5.1) and Eq. (5.2) [291] based on the true positives (TP), the true negatives (TN), the false positives (FP) and the false negatives (FN). The classifier is to be trained only

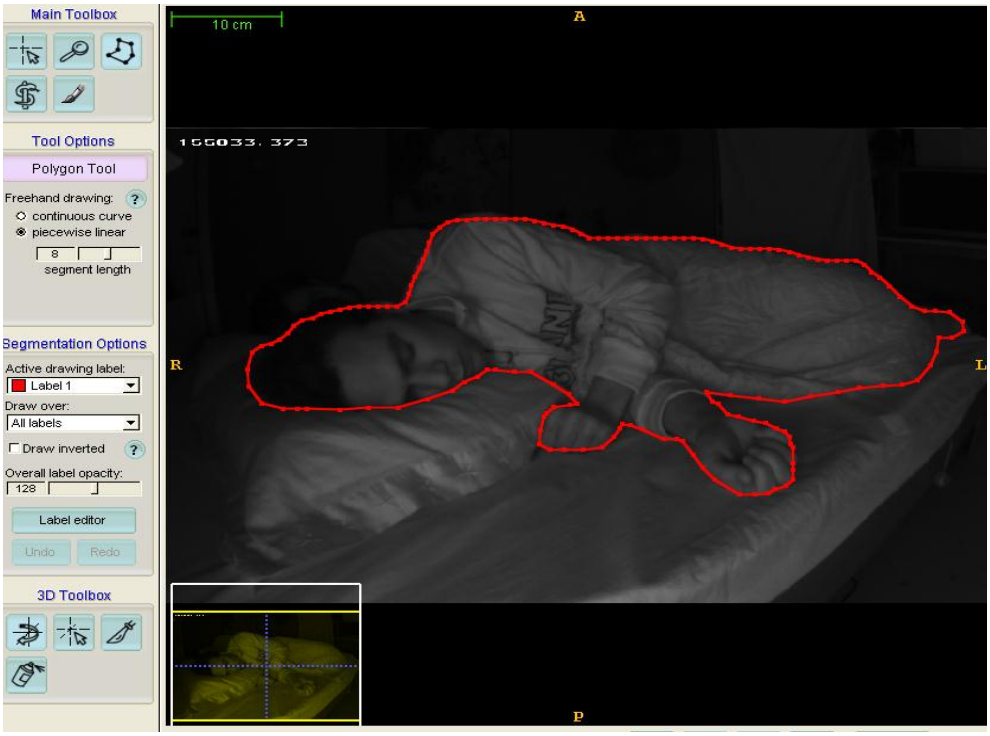


Figure 5.4: ITK-Snap annotation tool.

once with a large and diverse training data set and used for PS segmentation on all unseen data.

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{5.1}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5.2}$$

5.4.2 PS actigraphy

For the evaluation of the proposed video actigraphy algorithm in Section 5.3.2, ground truth annotations ('true' for PS movement, 'false' for no PS movement or bed partner movement) are compared with wrist and video actigraphy results.

The accuracy (see Eq. (5.1)) is determined for 40 short video clips of 5-7 frames. These 40 clips consist of representative movements (when both thin and thick blankets were used) with both small (single arm, single leg and head movements) and large movements (turning, moving both arms and legs) performed by the PS and the bed partner, one at a time and also simultaneously. If there is at least a single movement

magnitude above a threshold in the video actigraph or wrist actigraph signal, it is assumed to be a PS movement and is labeled positive (1). Otherwise, it is labeled negative (0).

5.4.3 PS breathing

Basic peak detectors (combination of findpeaks() and peakdet() in Matlab) are used to detect peaks from both the video and the two reference breathing waveforms. All the peaks in the reference signals are assumed correct and serve as ground truth. TPs, FNs and FPs of the video peaks are computed based on a tolerance interval of $\Delta t_i = \pm 1$s from the reference peak (see Fig. 5.5). The reference signals are obtained from inductive respiratory chest and abdominal belts as described in Section 5.2.

TP, see case (a) in Fig. 5.5: A video peak is detected within the tolerance interval of ± 1s (expected to be acceptable from experiments) from the reference peak.

FP, see case (b) in Fig. 5.5:

- Video peaks outside the tolerance intervals of the reference peaks.
- All video peaks detected within the same tolerance interval as an already detected video peak minus 1.

FN, see case (c) in Fig. 5.5: No video peak in the tolerance interval of a reference peak.

Based on the determined true positives, false positives, and false negatives, the sensitivity S (or recall rate), and precision P (or positive predictive value) are computed according to Eq. (5.3) and Eq. (5.2) (see [291]). The sensitivity measures the proportion of correctly detected peaks among all the actual peaks. The positive predictive value P measures the proportion of the correctly detected peaks among all the detected inhalation peaks.

$$S = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5.3}$$

The mean breathing rate correspondence (BR) is assessed based on the number of detected peaks within the epoch where v_p and r_p denote the breathing rate in one mode from the video signal v and the reference signal r, respectively. As shown in Eq. (5.4), the rate is computed as the ratio of the number of breathing cycles (number of peaks - 1) detected in an epoch to the time duration between the first and the last detected peak. The mean breathing rate correspondence should ideally approach the value of 1.

$$\mathrm{BR} = 1 - \frac{|v_p - r_p|}{r_p} \tag{5.4}$$

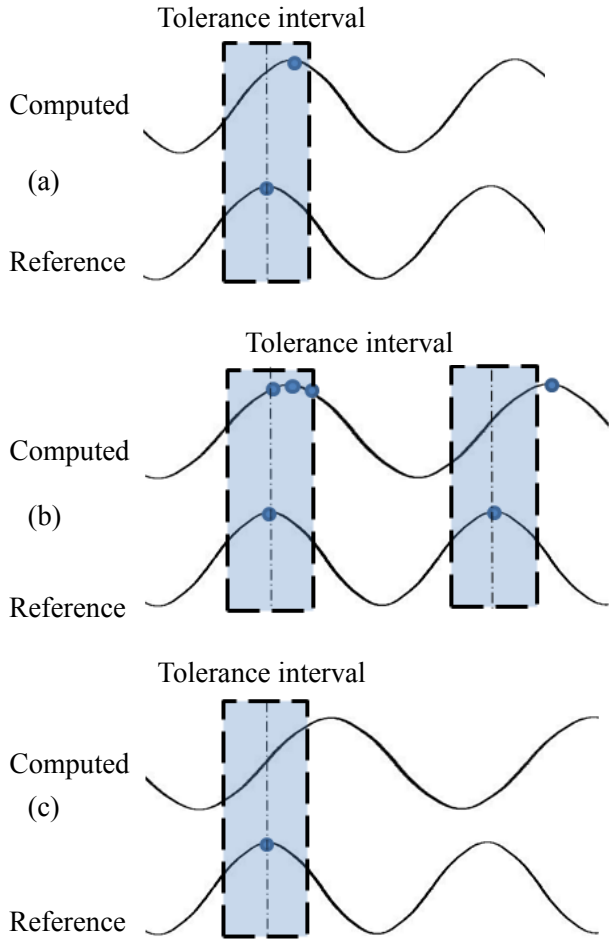


Figure 5.5: Computed breathing waveforms: (a) Presence of a TP. (b) Presence of 2 FPs around the first reference peak and 1 FP around the second reference peak. (c) Presence of a FN. The tolerance interval Δt_i is set to $\pm$1s.

5.5 Results and discussion

In this section the results of the separate experiments are presented. First, the PS segementation results are discussed in Section 5.5.1, followed by the evaluation results of the actigraph and breathing measurements in Sections 5.5.2 and 5.5.3.

5.5.1 Segmentation

The PS classifier performance has been determined by measuring the accuracy and precision as defined by Eq. (5.1) and Eq. (5.2) respectively, using K-fold cross-

Figure 5.6: Example frames with PS segmentation results. The boundary between PS and background has been highlighted with a white line.

validation on the data sets using $K = 4$. The AdaBoost classifier was trained 4 times using each time three of the four data sets (DATA Home1, DATA Home2, DATA Synt1, or DATA Synt2) and applying the classifier to the fourth. The evaluation results are summarized in Table 5.1. An accuracy between 87% and 90% is obtained

Training set	**Test set**	A [%]	P [%]
DATA Home2, DATA Synt1, DATA Synt2	DATA Home1	90	83
DATA Home1, DATA Synt1, DATA Synt2	DATA Home2	89	63
DATA Home1, DATA Home2, DATA Synt2	DATA Synt1	87	90
DATA Home1, DATA Home2, DATA Synt1	DATA Synt2	88	82
	Average	88	79

Table 5.1: Accuracy and precision results of the PS segmentation.

for these data sets, with a precision between 63% and 90%. Some examples of the segmentation results on separate frames are shown in Fig. 5.6. More qualitative insight in the classification accuracy has been obtained by determining the accumulated classification error at each pixel location. The normalized accumulation results are illustrated in Fig. 5.7, showing most classification errors occur at the border between background and PS.

The obtained average segmentation accuracy of 88% is promising for the considered application of home monitoring from a bedside table. Larger field tests, however, are required to obtain quantitative results for comparing with e.g., monitoring methods with ceiling mounted cameras. The segmentation of the PS should decrease the false positive rate induced by bed partner movements. Its impact on the actigraphy signal has been validated in Section 5.5.2.

5.5.2 Actigraphy

The comparison of the video actigraph with and without segmentation to the wrist actigraph are shown in Fig. 5.8. We have separately analyzed the influence of bed

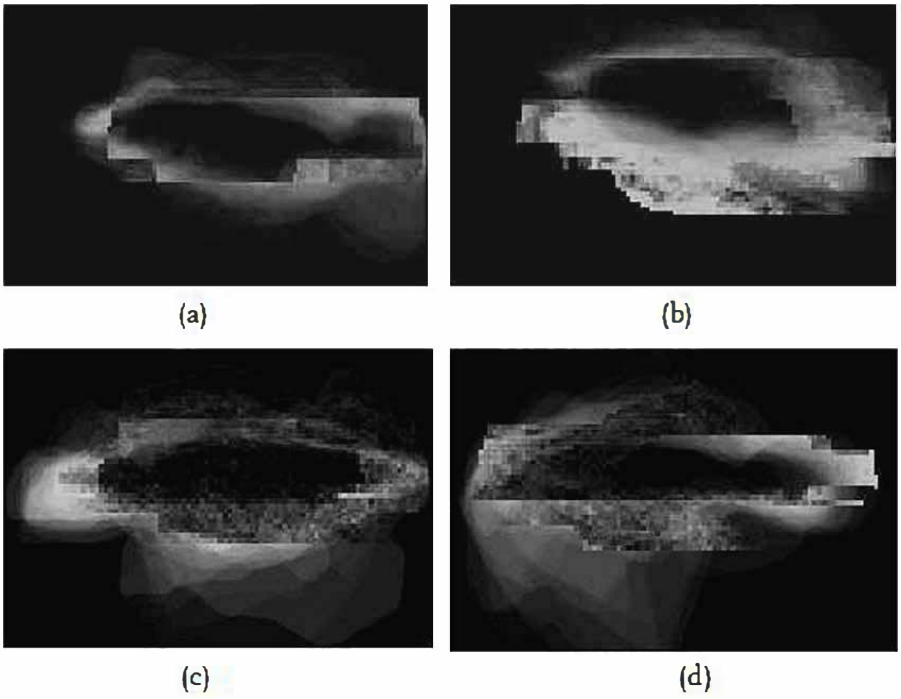


Figure 5.7: Accumulated normalized classification errors for each data set (a) DATA Home1 (b) DATA Home2 (c) DATA Synt1 (d) DATA Synt2. Brighter intensities indicate a higher average error.

partner motions by determining the video actigraph when only the bed partner is moving.

The top image in Fig. 5.8 shows the actigraphs generated from the video sequence where only the PS moves. It can be seen that roughly for all peaks in the wrist actigraph signal (shown in red) there are corresponding peaks in the video actigraph signal (blue). In other measurements, additional peaks in the video actigraph signal originate from moving body parts of the body other than the wrist, which makes video actigraphy more sensitive for the entire body than wrist actigraphy. At times, the wrist actigraph records large movements while the video actigraph only registers small motions. These differences are typically due to the large movements by the arm or wrist alone, compared to full body motions.

The bottom image in Fig. 5.8 shows a concatenation of PS motion, followed by bed partner motion, for the wrist and video actigraphs. It can be seen that there is no wrist actigraph signal when the bed partner is moving, while motion is visible in the video actigraph. This graph also indicates that motion due to bed partner motion is relatively low compared to PS motion and that segmentation of the PS helps to reduce the influence of bed partner movements further. By thresholding, as indicated in the

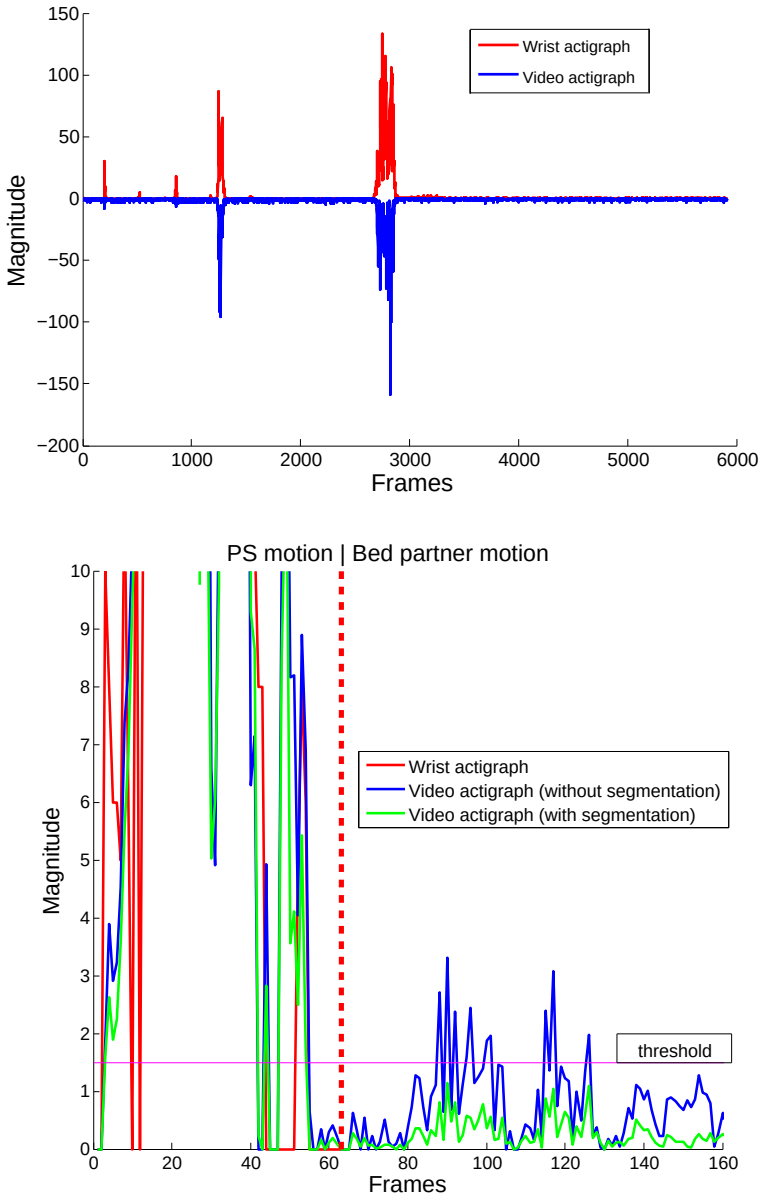


Figure 5.8: Comparison of wrist actigraph (a.u.) with video actigraph (a.u.) generated with and without segmentation of the PS. Top: There is a high correlation between the video actigraph (plot on the negative vertical axis) generated without segmentation to that of wrist actigraphy (plot on the positive vertical axis). Bottom: On the right of the red dashed line, the PS lies in supine position and the bed partner moves without any restrictions. There is no wrist actigraph signal as there is no PS movement. The segmentation yields a lower magnitude.

	WA	VA without segm.	VA with segm.
Avg. A [%]	75	70	85

Table 5.2: Average accuracy A of wrist actigraphy (WA) and video actigraphy (VA), with and without segmentation (segm.).

bottom image of Fig. 5.8, the generated video actigraph signal (with segmentation) could trigger only PS movements. For the example shown at the bottom of Fig. 5.8, a threshold value of 1.5 was empirically determined that would remove false triggers from the bed partner. From the 40 manually annotated video clips, an accuracy of 75% was found for the wrist actigraphy, while the video actigraphy without segmentation scored a 70% accuracy and the video actigraphy with segmentation a 85% accuracy (see Table 5.2).

In a shared bed it is common to share a blanket. We observed instances when the blanket covering the PS moved due to movements originating from the bed partner. This caused a false positive for the video actigraphy signal even after segmentation. An initial algorithm compensating for the blanket movement from the bed partner has been implemented. Due to the camera placement, a small directional movement to the top or left of the image was assigned to the bed partner movement instead of the PS. This first attempt increased the video actigraphy accuracy to 93%. More video clips consisting of blanket movement need to be selected and examined regarding robustness of this initial method.

These results show that the segmentation improves the video actigraphy when a bed partner is present and has a better performance compared to wrist actigraphy.

5.5.3 Breathing

The proposed video breathing (VB) algorithm is compared with the Horn-Schunck based optical flow (OF) implementation of [278] commonly used for video-based breathing analysis and the projection correlation (ProCor) method [286], a more recent method. The results obtained with the proposed video breathing algorithm, OF and ProCor are shown in Tables 5.3, 5.4, 5.5 and 5.6 with the chest belt as reference signal (similar results are obtained with the abdominal reference signal). Overall, the proposed VB method scores high, at around 90% for all three measures (see Table 5.3). The breathing rate correspondence achieves a correspondence of 93%. The BR between the two reference signals chest and abdomen are only slightly higher (98%). ProCor shows likewise high results for $S = 87\%$ with a clear drop in $P = 79\%$ and BR = 72%. In contrast, OF yields lower $S = 76\%$ and higher $P = 86\%$, with high scores for BR (91%). The main problems observed with the state-of-the-art methods are the at times indistinct breathing signals of OF and very noisy signals of ProCor (see Fig. 5.9).

Lying on the right when the chest faces the camera, is consistently the most

	Prop. method VB	ProCor	OF
Avg. S [%]	87	87	76
Avg. P [%]	90	79	86
Avg. BR [%]	93	72	91

Table 5.3: Overall average sensitivity S, precision P and breathing rate correspondence BR of the proposed method VB, ProCor and OF.

challenging position with generally the lowest evaluation results for VB, ProCor, and OF (see Table 5.4). VB performs above 90% for the other body positions in S, P and BR but achieves only 72%, 76% and 75% respectively in the 'on right' position. Similarly, ProCor generally achieves results above 80% for the 'prone', 'supine', 'on left' body positions, but suffers severe cuts to $S = 49\%$, $P = 37\%$ and BR $= 34\%$ for the 'on right' position. This is most likely due to the counter movements of head and chest/arms that become most apparent in this position.

Concerning the inter-subject variability (see Table 5.5), there is a slight tendency of subject 3 to perform worst for VB and ProCor. This is particularly visible in the sensitivity results of VB and the precision and breathing rate results of ProCor. As subject 3 has a BMI in the normal range we do not expect the video breathing algorithm to be sensitive to differences in BMI. The cause for the worse performance needs to be investigated in future work.

Regarding the results per breathing mode in Table 5.6, the respiration rate correspondence is high for VB throughout all breathing modes contrary to the state-of-the-art algorithms. BR is lower for ProCor and OF in the natural mode, the high breathing rate mode (23 bpm) and in the deep-sigh mode. Subjects' breathing could be very irregular at the beginning of the experiment when the subjects just transitioned to a rest phase from an active phase out of the bed. There, inconsistencies

	Left	Right	Prone	Supine
VB S [%]	95	72	90	92
ProCor S [%]	96	49	96	97
OF S [%]	90	49	70	94
VB P [%]	98	76	93	94
ProCor P [%]	88	37	85	97
OF P [%]	99	58	86	99
VB BR [%]	95	75	93	95
ProCor BR [%]	80	34	74	95
OF BR [%]	95	88	85	95

Table 5.4: Average sensitivity S, precision P and breathing rate correspondence BR of the proposed method VB, ProCor and OF for the different body positions lying on the left, lying on the right, prone and supine.

	S1	**S2**	**S3**	**S4**	**S5**
VB S [%]	94	94	78	87	84
ProCor S [%]	82	78	86	90	97
OF S [%]	69	67	75	83	85
VB P [%]	98	95	85	90	84
ProCor P [%]	78	75	64	83	93
OF P [%]	77	76	85	95	95
VB BR [%]	96	94	91	92	94
ProCor BR [%]	81	84	41	63	91
OF BR [%]	87	92	91	91	93

Table 5.5: Average sensitivity S, precision P and breathing rate correspondence BR of the proposed method VB, ProCor and OF for the different subjects S1, S2, S3, S4 and S5.

in the video breathing signal were observed. In the high breathing rate mode (23 bpm) some subjects found it difficult to follow the metronome. It was the first mode to follow and get used to with a fast pace making it more difficult to follow than a slower pace. This is also the cause for lower S and P values which all methods suffer from. In low breathing rates (12 and 10 breaths per minute), peaks may be found just outside the accepted one-second interval resulting in lower S and P values. Deep and shallow breathing can be both robustly captured by VB as seen from the positive results of '15 bpm deep' and '15 bpm shallow'. ProCor and OF had more difficulties with shallow breathing.

Most of the problems with VB arise due to a wrong selection of the reference frame. A frame during a breathing peak is then selected as a reference frame instead of selecting it during an expiratory pause. S and P due to the wrong peak locations can drop to 56% in such cases. Future work needs to investigate improvements regarding the computation of the reference frame.

When adapting ProCor to incorporate the band-pass filter and peak detector used in the proposed method VB, average results have improved to $S = 86\%$, $P = 87\%$, $\text{BR} = 97\%$.

5.6 Conclusions

A contactless video-based activity and respiration monitoring system for home sleep analysis is proposed. The aim is to perform video-based monitoring of activity and breathing of a Primary Subject (PS) in the common shared bed scenario. The camera is placed on the bedside table to make the use case as convenient as possible for the user. The use of a video camera and intelligent algorithms for segmenting the PS, for calculating the PS' activity level and for extracting the breathing signal is validated.

The segmentation algorithm has been benchmarked on subjects in both a natural unrestricted sleeping environment at home, and with a fixed movement protocol both at home and at a sleep laboratory. The benchmark was done using an annotation

	Natural	23 bpm	19 bpm	15 bpm shallow	15 bpm deep	12 bpm	10 bpm	3 deep sighs
VB S [%]	83	80	91	91	91	86	85	93
ProCor S [%]	80	89	90	86	90	87	86	84
OF S [%]	70	68	78	80	84	82	73	70
VB P [%]	83	91	96	96	94	89	85	90
ProCor P [%]	70	82	83	75	83	77	68	94
OF P [%]	81	89	90	85	89	85	76	89
VB BR [%]	91	89	94	95	96	96	94	92
ProCor BR [%]	59	85	82	70	85	70	48	80
OF BR [%]	89	81	92	95	99	97	99	73

Table 5.6: Average sensitivity S, precision P and breathing rate correspondence BR of the proposed method VB, ProCor and OF for the different breathing rates natural, 23 bpm, 19 bpm, 15 shallow bpm, 15 deep bpm, 12 bpm, 10 bpm and 3 deep sighs.

tool called ITK-SNAP which allows pixel accurate manual annotation. On average, the proposed algorithm gives an accuracy of 88% at a precision of 79%, with the most errors occurring at the PS and background border. In future work, by training the classifier once with large and diverse (sleeping poses, lighting conditions, texture of blankets, with and without blankets) training data set, the classifier is expected to generalize to any unseen data yielding significant PS segmentation results.

The proposed video actigraphy system can detect movements up to an accuracy of 85% while outperforming wrist actigraphy at 75% accuracy. An initial method compensating for blanket movement from the bed partner yielded an even higher accuracy of 93%. Further research would benefit from compensating for blanket movement.

Regarding the video breathing algorithm, data collection with 5 test subjects with different BMIs, in 4 body positions, was performed where two respiration belts (chest and abdominal as used in traditional polysomnography to obtain the respiration signal) were used as reference. The video-based breathing algorithms faced challenges such as the use of a blanket, different breathing frequencies, breathing depths, and body positions. The proposed algorithm outperformed state-of-the-art algorithms for video-based breathing analysis. An overall sensitivity of 87%, precision of 90%, and breathing rate correspondence of 93% were obtained. Given a 98% breathing rate correspondence between the reference signals, the results of the proposed video method are well in range.

In future work, larger field tests are required to obtain quantitative results for comparing with e.g., monitoring methods with ceiling mounted cameras.

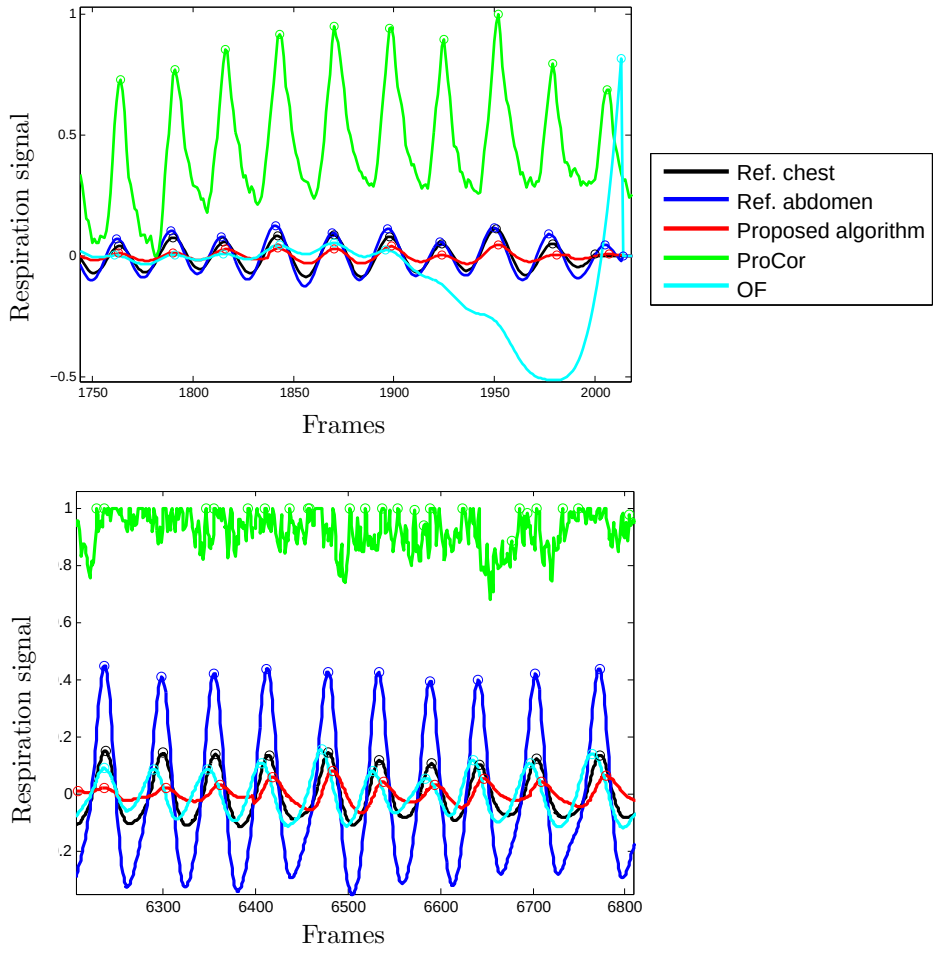


Figure 5.9: Computed breathing waveforms of two typical cases. Top: OF breathing signal becomes inaccurate at times, good alignment of VB and ProCor with reference signal. Bottom: Very noisy ProCor output, good alignment of VB and OF with reference signal.

Bibliography

[241] J.M. Fry, M.A. DiPhillipo, K. Curran, R. Goldberg, and A.S. Baran, "Full polysomnography in the home", *Sleep*, **vol. 21**, no. 6 pp. 635–642, Sep. 1998.

[242] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, *et al.*, "The role of actigraphy in the study of sleep and circadian rhythms", *Sleep*, **vol. 26**, no. 3 pp. 342–392, 2003.

[243] P. De Chazal, N. Fox, E. O'Hare, C. Heneghan, *et al.*, "Sleep/wake measurement using a non-contact biomotion sensor", *J Sleep Res*, **vol. 20**, no. 2 pp. 356–366, Jun 2011.

[244] T. Harada, A. Sakata, T. Mori, and T. Sato, "Sensor pillow system: monitoring respiration and body movement in sleep", in *Proceedings of 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov. 2000, vol. 1, pp. 351–356.

[245] L. Samy, Ming-Chun Huang, J.J. Liu, Wenyao Xu, and M. Sarrafzadeh, "Unobtrusive Sleep Stage Identification Using a Pressure-Sensitive Bed Sheet", *Sensors Journal, IEEE*, **vol. 14**, no. 7 pp. 2092–2101, July 2014.

[246] T. Willemen, B. Haex, J. Vander Sloten, and S. Van Huffel, "Biomechanics based analysis of sleep", in *Proc. of the 5th Dutch conference on Bio-medical engineering*, Jan. 2015, pp. 1–1.

[247] X.L. Aubert and A. Brauers, "Estimation of Vital Signs in Bed from a Single Unobtrusive Mechanical Sensor: Algorithms and Real-life Evaluation", in *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBS 2008.*, Aug. 2008.

[248] M. Littner, C.A. Kushida, W.M. Anderson, D. Bailey, *et al.*, "Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002", *Sleep*, **vol. 26**, no. 3 pp. 337–341, May 2003.

[249] R. Gardner and W.I. Grossman, "Normal motor patterns in sleep in man", *Advances in Sleep Research*, **vol. 2** pp. 67–107, 1976.

[250] J. Wilde-Frenz and H. Schulz, "Rate and distribution of body movements during sleep in humans", *Percept. Mot. Skills*, **vol. 56** pp. 275–283, 1983.

[251] S. Gori, G. Ficca, F. Giganti, I. Di Nasso, *et al.*, "Body movements during night sleep in healthy elderly subjects and their relationships with sleep stages", *Brain Research Bulletin*, **vol. 63**, no. 5 pp. 393–397, June 2004.

[252] A. Muzet, P. Naitoh, L. Johnson, and R. Townsend, "Body movements during sleep as a predictor of state change", *Psychon. Sci.*, **vol. 29** pp. 7–10, 1972.

[253] R.P. Allen and W.A. Hening, "Actigraph Assessment of Periodic Leg Movements and Restless Legs Syndrome", in *Restless Legs Syndrome*, W.B. Saunders, Philadelphia, pp. 142 – 149, 2009.

[254] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R.M. Aarts, "Respiration amplitude analysis for REM and NREM sleep classification", in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 5017–5020.

[255] Walter Karlen, Claudio Mattiussi, and Dario Floreano, "Improving actigraph sleep/wake classification with cardio-respiratory signals", in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, IEEE, 2008, pp. 5262–5265.

[256] Beddit, "Beddit", 2014, [Online; accessed 24-May-2014].
URL `http://www.beddit.com/`

[257] Withings, "Withings Aura Smart Sleep System", 2014, [Online; accessed 24-May-2014].
URL `http://withings.com/en/aura`

[258] Omron, "Sleepmeter HSL-101 and HSL-001 SleepDesign Lite", 2014, [Online; accessed 24-May-2014].
URL `http://www.healthcare.omron.co.jp/product/hsl/hsl-101.html`

[259] Gear4, "Renew SleepClock", 2014, [Online; accessed 24-May-2014].
URL `http://www.stage.gear4.com/product/_/426/renew-sleepclock/`

[260] National Sleep Foundation, "Sleep in America poll: Adult sleep habits and styles", 2005, [Online; accessed 5-October-2013].
URL `http://www.sleepfoundation.org/article/sleep-america-polls/2005-adult-sleep-habits-and-styles`

[261] F.-C. Yang, C.-H. Kuo, M.-Y. Tsai, and S.-C. Huang, "Image-based sleep motion recognition using artificial neural networks", in *Machine Learning and Cybernetics, 2003 International Conference on*, 2003, vol. 5, pp. 2775–2780 Vol.5.

[262] C.-W. Wang, A. Hunter, N. Gravill, and S. Matusiewicz, "Real time pose recognition of covered human for diagnosis of sleep apnoea", *Computerized Medical Imaging and Graphics*, **vol. 34**, no. 6 pp. 523 – 533, 2010.

[263] W.-H. Liao and J.H. Kuo, "Sleep monitoring system in real bedroom environment using texture-based background modeling approaches", *J. Ambient Intelligence and Humanized Computing*, **vol. 4**, no. 1 pp. 57–66, Feb. 2013.

[264] W.-H. Liao and C.-M. Yang, "Video-based activity and movement pattern analysis in overnight sleep studies", in *Int'l Conf. on Pattern Recognition*, 2008, pp. 1–4.

[265] TMSI, "Respiration effort module V6 and inductive respiration band, Adult L", 2013, [Online; accessed 5-October-2013].
URL `http://www.tmsi.com/images/stories/PDF-files/Accessorylist_2013.pdf`

[266] J. P. Vink and G. de Haan, "Comparison of machine learning techniques for target detection", *Artificial Intelligence Review*, pp. 1–15, Nov. 2012.

[267] D.-M. Tsai and H.-J. Wang, "Segmenting focused objects in complex visual images", *Pattern Recognition Letters*, **vol. 19**, no. 10 pp. 929 – 940, 1998.

[268] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd ed., 1992, ISBN 0201508036.

[269] G. de Haan and P.W.A.C Biezen, "Sub-pixel motion estimation with 3-D recursive search block-matching", *Signal Processing: Image Communication*, **vol. 6**, no. 3 pp. 229 – 239, 1994.

[270] A. Heinrich, C. Bartels, R.J. Van der Vleuten, C.N. Cordes, and G. de Haan, "Optimization of Hierarchical 3DRS Motion Estimators for Picture Rate Conversion", *IEEE Journal of Selected Topics in Signal Processing*, **vol. 5**, no. 2 pp. 262–274, Apr. 2011.

[271] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", in *In CVPR*, 2005, pp. 886–893.

[272] P. Viola and M. Jones, "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*, 2001, vol. 1, pp. I–511–I–518 vol.1.

[273] A. Heinrich, X. Aubert, and G. de Haan, "Body movement analysis during sleep based on video motion estimation", in *e-Health Networking Applications and Services (Healthcom), 2013. 15th IEEE International Conference on*, 2013.

[274] K. S. Tan, R. Saatchi, H. Elphick, and D. Burke, "Real-time vision based respiration monitoring system", in *Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th International Symposium on*, 2010, pp. 770–774.

[275] C.W. Wang, A. Ahmed, and A. Hunter, "Vision analysis in detecting abnormal breathing activity in application to diagnosis of obstructive sleep apnoea", in *Conf Proc IEEE Eng Med Biol Soc*, Sep. 2006, pp. 4469–4473.

[276] M. Frigola, J. Amat, and J. Pagès, "Vision based respiratory monitoring system", in *Proceedings of the 10th Mediterranean Conference on Control and Automation - MED2002 Lisbon, Portugal*, July 2002.

[277] K. Nakajima, A. Osa, T. Maekawa, and H. Miike, "Evaluation of Body Motion by Optical Flow Analysis", *Japanese Journal of Applied Physics*, **vol. 36**, no. Part 1, No. 5A pp. 2929–2937, 1997.

[278] K. Nakajima, Y. Matsumoto, and T. Tamura, "Development of real-time image sequence analysis for evaluating posture change and respiratory rate of a subject in bed", *Physiol. Meas.*, **vol. 22**, no. 3 pp. 21–28, 2001.

[279] N. Koolen, O. Decroupet, A. Dereymaeker, K. Jansen, *et al.*, "Automated Respiration Detection from Neonatal Video Data", in *Proc. of the 4th International*

Conference on Pattern Recognition Applications and Methods. (ICPRAM 2015), Jan. 2015, pp. 164–169.

[280] B. Horn and B. Schunck, "Determining Optical Flow", *Artificial Intelligence*, **vol. 17**, no. 1-3 pp. 185–203, 1981.

[281] K. Cuppens, L. Lagae, B. Ceulemans, S. Van Huffel, and B. Vanrumste, "Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy.", *Medical & Biological Engineering & Computing*, **vol. 48**, no. 9 pp. 923–931, Sep. 2010.

[282] N. B. Karayiannis, B. Varughese, Guozhi Tao, J. D. Frost, Jr., *et al.*, "Quantifying motion in video recordings of neonatal seizures by regularized optical flow methods", *Trans. Img. Proc.*, **vol. 14**, no. 7 pp. 890–903, Jul. 2005, ISSN 1057-7149.

[283] J.U. Bak, N. Giakoumidis, G. Kim, H. Dong, and N. Mavridis, "An Intelligent Sensing System for Sleep Motion and Stage Analysis", *Procedia Engineering*, **vol. 41**, no. 0 pp. 1128 – 1134, 2012, international Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012).

[284] S. Kalitzin, G. Petkov, D. Velis, B. Vledder, and F. Lopes da Silva, "Automatic Segmentation of Episodes Containing Epileptic Clonic Seizures in Video Sequences", *Biomedical Engineering, IEEE Transactions on*, **vol. 59**, no. 12 pp. 3379–3385, 2012.

[285] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, *et al.*, "Eulerian Video Magnification for Revealing Subtle Changes in the World", *ACM Trans. Graph.*, **vol. 31**, no. 4 pp. 65:1–65:8, July 2012.

[286] M. Bartula, T. Tigges, and J. Muehlsteff, "Camera-based system for contactless monitoring of respiration", in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 2672–2675.

[287] H. Aoki, K. Koshiji, H. Nakamura, Y. Takemura, and M. Nakajima, "Study on respiration monitoring method using near-infrared multiple slit-lights projection", in *Micro-NanoMechatronics and Human Science, 2005 IEEE International Symposium on*, 2005, pp. 291–296.

[288] S.M. Stigler, "Francis Galton's Account of the Invention of Correlation", *Statist. Sci.*, **vol. 4**, no. 2 pp. 73–79, 1989.

[289] W.Q. Lindh, M. Pooler, and C.D. Tamparo, *Delmar's Comprehensive Medical Assisting: Administrative and Clinical Competencies*, Delmar Thomson Learning, 2001, ISBN 9780766824188.

[290] Paul Yushkevich, "ITK-SNAP", 2012, [Online; accessed 5-October-2013]. URL `http://www.itksnap.org`

[291] S.S. Coughlin, B. Trock, M.H. Criqui, L.W. Pickle, *et al.*, "The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test", *Journal of Clinical Epidemiology*, **vol. 45**, no. 1 pp. 1 – 7, 1992.

Chapter 6

Robust and sensitive video motion detection for sleep analysis

Abstract

In this paper, we propose a camera-based system combining video motion detection, motion estimation and texture analysis with machine learning for sleep analysis. The system is robust to time-varying illumination conditions while using standard camera and infrared illumination hardware. We tested the system for Periodic Limb Movement (PLM) detection during sleep, using EMG-signals as a reference. We evaluated the motion detection performance both per frame and with respect to movement event classification relevant for PLM detection. The Matthews Correlation Coefficient (MCC) improved with a factor of 2, compared to a state-of-the-art motion detection method, while sensitivity and specificity increased with 45% and 15%, respectively. Movement event classification improved by a factor of 6 and 3 in constant and highly varying lighting conditions, respectively. On 11 PLM patient test sequences, the proposed system achieved a 100% accurate PLM index (PLMI) score with a slight temporal misalignment of the starting time (<1s) regarding one movement. We conclude that camera-based PLM detection during sleep is feasible and can give an indication of the PLMI score.

This chapter is published as: A. Heinrich, D. Geng, D. Znamenskiy, J. Vink, G. de Haan; Robust and sensitive video motion detection for sleep analysis, *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 1-9, May 2014.

6.1 Introduction

Body movements are an important behavioral aspect during sleep as shown in [292]. They can be associated to sleep states [293] and connected to sleep state transitions [294]. It was concluded in [295] that frequency and duration of body movements are important characteristics for sleep analysis. Moreover, various somatic and mental disorders are associated with movements during sleep and several clinical syndromes of pathological movements exist [296], like restless leg syndrome, periodic limb movement disorder, and movement disorders under parasomnias such as disorders of arousal and sleep-wake transitions.

The screening of sleep is typically obtrusive using contact sensors, like accelerometers, electrodes, wrist watches, headbands, and often requires visits to the sleep clinic to apply the various contact sensors. Also significant time and effort from sleep clinicians are involved, including working night shifts. Consequently, waiting lists for sleep screening are common, while possible skin irritations and disturbed sleep due to the detachment from the natural sleeping habitat frequently occur.

The use of a video camera to analyze movement patterns of a sleeping subject promises an attractive alternative, as it is inherently unobtrusive, and may be used in a home setting. State-of-the-art motion detection methods have been developed and are applied successfully in systems for surveillance and tracking, e.g., [297, 298].

Our main challenge was to design a system robust to dynamic illumination conditions. The generally low light conditions for sleep monitoring render the system sensitive to light pollution and changes. *Hardware* solutions for light robustness typically involve a short camera-integration and a synchronously flashing IR light source to overpower ambient illumination and shadows [299]. As we aim at using existing camera and lighting installations, as applied in sleep clinics, hospitals, and surveillance systems, we propose a video analysis *software* solution.

Dynamic illumination conditions can be manifested by global lighting changes and pose a smaller challenge than local dynamic illumination conditions. Of particular concern regarding local dynamic illumination are moving shadows, e.g., caused by a family member, outside tree branches or moving curtains (see Fig. 6.1) in an uncontrolled setting (due to e.g., street lighting, moonlight, sunlight, indoor lighting). Previous research imposed various restrictions on the camera [300], the shadow area properties related to the background characteristics [298, 301], or the background itself [302–304]. For the application targeted in this work, only a static camera is assumed, while further assumptions are relaxed. This is relevant, since the bed is typically non-planar due to the force on the beddings exhibited by the subject and beddings lying loosely on and around the subject. Moreover, the background changes dynamically, as subject movements change the folds' location and appearance of the beddings, while fold appearance (texture) varies with moving cast shadows. Local intensity changes of 25% between consecutive frames are observed due to cast shadows vs. a 10% change due to subject movement.

For the application of sleep monitoring including the interference of moving shad-

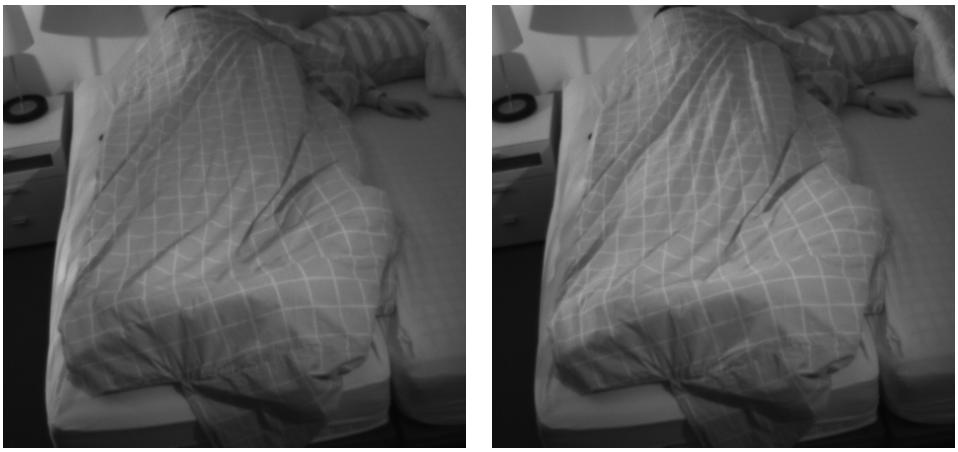


Figure 6.1: Lighting variations between two consecutive images due to moving cast shadows. Edges in bedding folds become visible in the bottom left quadrant of the bed.

ows, typically two illumination sources and conditions are present as is depicted in Fig. 6.2. The first illumination source is a constant near infrared light source to monitor sleeping subjects (object 1 in Fig. 6.2) both at daytime and nighttime in a constant lighting condition. The second illumination condition is variable due to interfering moving objects (object 2 in Fig. 6.2) in front of another artificial light source or sunlight producing moving cast shadows. The darker regions on the background of the objects consist of two parts, umbra and penumbra. Umbra is the centre part of the cast shadow which is not directly illuminated by any light source. Penumbra is the soft transition from dark to bright and partly illuminated by a light source. The length of this transition is referred to as edge width in [302].

In summary, we present a motion detector robust to moving cast shadows that can deal with fewer scene assumptions and limitations compared to other shadow detection methods and uses standard hardware. As a home sleep monitoring case study, we discuss the design of a near infrared video movement detection (VMD) system that can compete with on-body electromyography (EMG) movement detection for periodic limb movements in sleep (PLMS) and for periodic limb movement disorder (PLMD). To estimate true subject movement, the amplitude of the EMG signal is used for quantifying PLM as it indicates the electric activity in a muscle.

Section 6.2 discusses the state-of the-art motion detection methods, while our proposed VMD system robust to moving cast shadows is presented in Section 6.3. The experimental setup is described in Section 6.4 and the evaluation of the proposed VMD system is discussed in Section 6.5. Conclusions are drawn in Section 6.6.

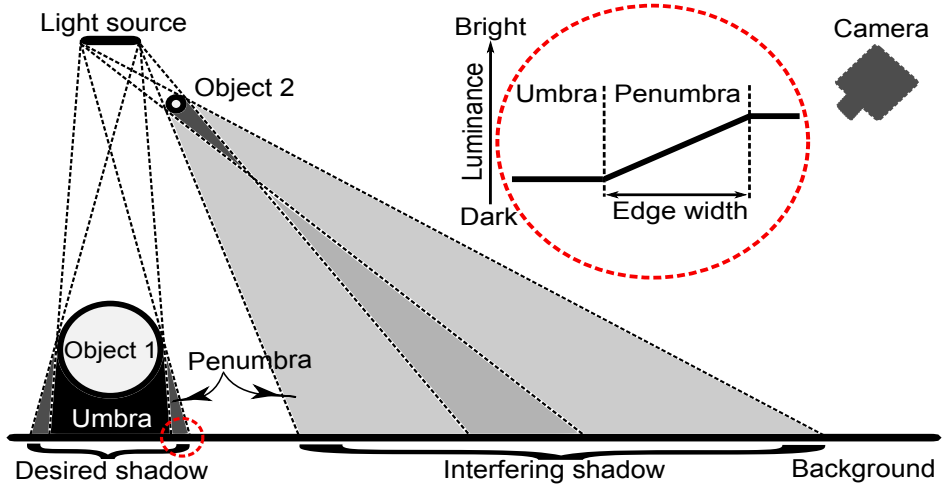


Figure 6.2: Formation of cast shadow with object 1 being close to the background and object 2 far away from the background.

6.2 Existing methods

The accumulation of absolute values of temporal frame differences is a widely used change detector, also applied to the movement analysis during sleep [305]. It is however not designed to be illumination insensitive and, more importantly, to distinguish between subject movements and a moving interfering shadow (IS). Different constant lighting levels are analyzed in [306]. In order to simulate lighting changes at dawn, global artificial light changes were added to the sleep videos in [307]. With the aim to detect the presence of shadows, the shading model

$$R_i(x,y) = \frac{E_{i-1}(x,y)}{E_i(x,y)} = \frac{I_{i-1}(x,y)S_{i-1}(x,y)}{I_i(x,y)S_i(x,y)} \tag{6.1}$$

is introduced by [308]. The underlying assumption is that the image luminance value $E_i(x,y)$ at pixel location (x,y) in frame i is the product of the irradiance $I_i(x,y)$ and the reflectance $S_i(x,y)$. The irradiance is the received light power per illuminated object surface and changes when an illumination change takes place, e.g., when a shadow moves over an object. When an object moves, the reflectance changes depending on object structure. It takes into account the surface material and the geometrical arrangement of camera, light source and object. For shadow detection, the reflectance remains constant in Eqn. (6.1) ($S_{i-1}(x,y) = S_i(x,y)$) and the luminance values between a current frame i and a reference or previous frame $i-1$ are compared with each other. In the case of a reference frame, typically a background and/or shadow model is computed and pixel values in the current image are compared with the corresponding pixel values in the reference model(s), e.g., [297, 307, 309]. This

is a very common approach confirmed in [300] and [303] where the majority of the 50 selected shadow detection methods computes a background model. A background and/or shadow model cannot be learned for the application discussed in this paper due to the dynamic and non-planar background yielding edges at different locations and at times larger intensity variations under the influence of moving cast shadows compared to subject movement.

The deterministic method described in [302] compares the previous and current frame by computing the local variance of the intensity ratios. The probability for shadow presence increases with small local variance V

$$V_i(b_x, b_y) = \frac{1}{n} \sum_{(s,t)\in B(b_x,b_y)} (R_i(s,t) - \mu_i(b_x, b_y))^2, \tag{6.2}$$

where (s,t) are the n pixels in block B with center coordinates (b_x, b_y) and mean μ defined as

$$\mu_i(b_x, b_y) = \frac{1}{n} \sum_{(s,t)\in B(b_x,b_y)} R_i(s,t). \tag{6.3}$$

One drawback of this approach is that uniform regions are considered as potential shadows. The same effect is observed in [310] using a similar approach. Stauder et al. [310] correctly stressed the assumption of textured objects and a planar background which does not hold for the application at hand. Folds in beddings and blankets yield a non-planar background. Texture differences that may be clearly visible in daylight may be greatly reduced in the infrared image (see Fig. 6.3). In order to deal with none or low-textured objects, an addition to the variance V is proposed in [311] and similarly used in [312]. The interior of appearing homogeneous objects on homogeneous background is on one hand detected for a relatively large luminance difference (i.e., when object motion is present) and is on the other hand illumination invariant for a relatively small luminance difference (i.e., in the case of moving cast shadows). We refer to it here as Variance Plus (VP) and later as VP orig. which is defined as

$$\mathrm{VP}_i(b_x, b_y) = \mathrm{VP}_i^+(b_x, b_y) + \mathrm{VP}_i^-(b_x, b_y), \tag{6.4}$$

where

$$\begin{aligned} \mathrm{VP}_i^+(b_x, b_y) &= \tfrac{1}{n} \sum_{(s,t)\in B(b_x,b_y)} \big(R_i^2(s,t) - R_i(s,t)\big) \\ &= \mathrm{V}_i(b_x, b_y) + \mu_i^2(b_x, b_y) - \mu_i(b_x, b_y) \end{aligned} \tag{6.5}$$

and

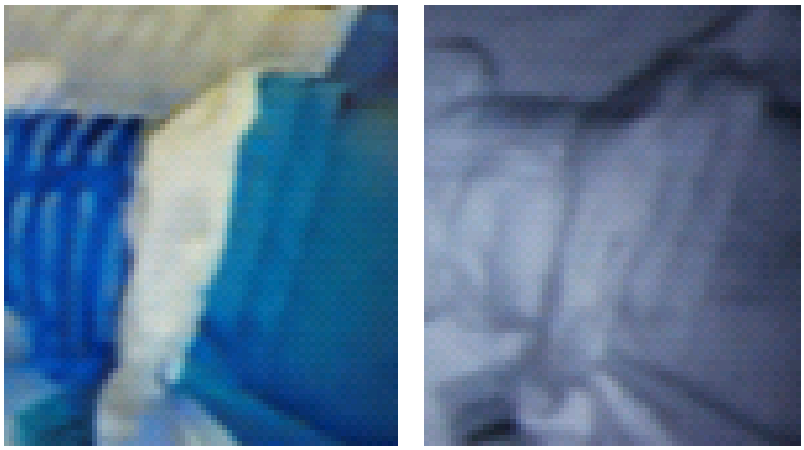


Figure 6.3: Loss of texture when the same scene is captured with an infrared camera (right) compared to a visible light camera in the presence of visible light (left).

$$\mathrm{VP}_i^-(b_x, b_y) = \frac{1}{n} \sum_{(s,t)\in B(b_x,b_y)} \left(R_i^{-2}(s,t) - R_i^{-1}(s,t)\right). \tag{6.6}$$

Motion is detected if the following criteria $\mathrm{VP}_i^+(b_x, b_y) > d_{\mathrm{VP}}$ or $\mathrm{VP}_i^-(b_x, b_y) > d_{\mathrm{VP}}$ are fulfilled where d_{VP} is a threshold determined in 6.3.4.

Substantial research efforts have been made in determining shadow areas based on edge and texture information [297, 298, 301, 304, 313]. It is assumed that textures in background and shadow areas show high resemblance and/or that a foreground object will have significant interior edges contrary to a shadow region. Tracking moving objects has proven to be beneficial for shadow detection in [297] and [309].

6.3 Proposed video movement detection system

Although competing methods as discussed in Section 6.2 are not expected to be robust enough in the varying lighting condition, they included important features such as spectral knowledge of the shadows, spatio-temporal and motion information, which we also incorporated in our proposed VMD method. We designed a motion detection cascade as illustrated in Fig. 6.4, where the spectral components typical for IS have been removed and we improved the VP method VP orig. discussed in Section 6.2. The resulting motion mask is a first selection of potential motion areas. Thereafter, spatio-temporal and motion compensated information is included in the underlying texture in the texture model (TM) and the motion compensated texture model (MCTM) to discern shadow movement from subject movement. At each stage of the cascade, the numbers of detected motion blocks from the motion mask are given as input to a machine learning classifier for the final subject motion detection result per video

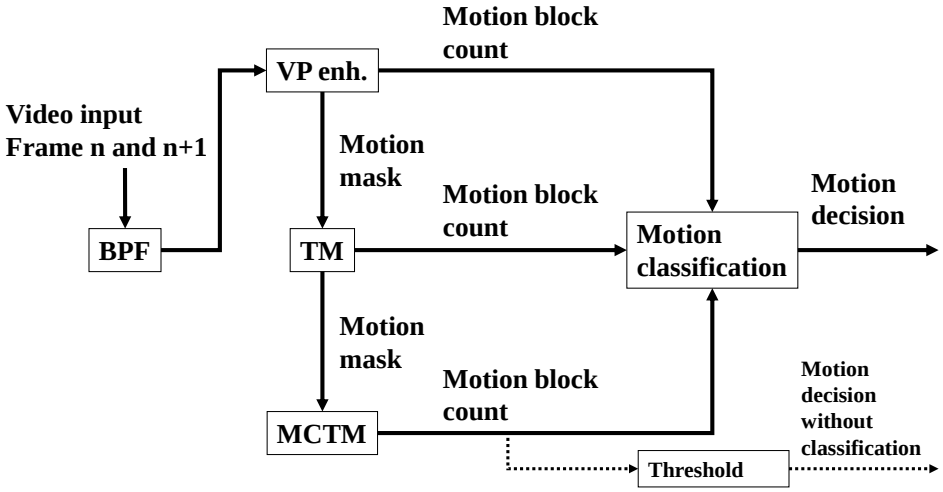


Figure 6.4: Block diagram of proposed VMD method, the dashed line refers to the feature selection method described in Section 6.3.4.

frame. The cascade structure is chosen in order to minimize computation time as the algorithmic complexity increases along the cascade. The run times per frame for the single blocks are 0.094s for VP enhanced (VP enh.), 1.012s for TM and 6.601s for MCTM (the methods are discussed in detail in the following sections).

We use a band-pass filter (BPF in Fig. 6.4) to remove the low-frequency edge of the IS and high-frequency noise while preserving the medium-frequency texture information of subjects and beddings. A Gaussian BPF is applied with a cut-off frequency at 100 cycles per picture-width (c/pw) and per picture-height (c/ph). The bandwidth is 180 c/pw and c/ph. Here we used an assumption that the size of the interferring shadow edge is determined by the size of the penumbra and the size of 3D features which modulate the specular reflections in the scene. For artificial light sources the lamp size is sufficiently large to produce a penumbra larger than the size of 3D features. Thus, for example, for a lamp with a diameter of 10 cm at a distance of 4 m and a moving object at a distance of 2 m the size of the cast shadow is 10 cm, while the background details like folds in the beddings are about 1 cm. If the camera monitors the scene with a width of 100 cm, the size of the 3D features are about 1% which motivates the use of a cut-off frequency at 100 cycles per picture-width. The bandwidth of 180 c/pw and c/p was chosen as a compromise beween two requirements: firstly, to keep sharp edges in the scene, and secondly, to reduce camera noise in low lighting conditions.

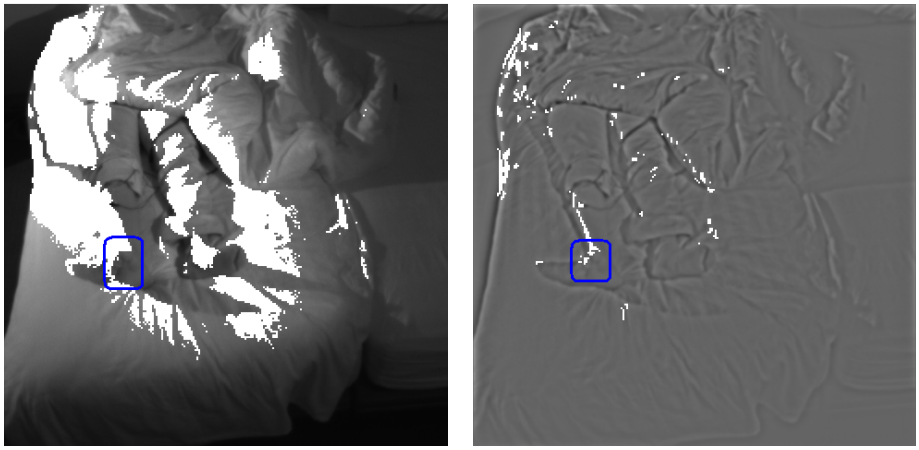


Figure 6.5: Motion mask of VP enhanced (VP enh.) without BPF (left) and with BPF (right) when the right toe is moving (blue square).

6.3.1 VP enhanced

We implemented an enhanced version of VP orig. by using an illumination adaptive threshold d_{VP} to increase robustness. If only illumination changes and noise occur, then

$$E_{i-1}(x,y) = \alpha E_i(x,y) + N, \tag{6.7}$$

where α indicates the illumination change and N random noise. Substituting $E_{i-1}(x,y)$ in Eqn. (6.1) and the obtained result in Eqn. (6.5) results in

$$d_{\mathrm{VP}}(x,y) = \left(\alpha + \frac{N}{E_i(x,y)} - \frac{1}{2}\right)^2 - \frac{1}{4}, \tag{6.8}$$

where $\alpha = 1$ corresponds to no illumination change. The value for N is determined in Section 6.3.4. The block size of VP enh. is set to 2×2 pixels.

The influence of the BPF on VP enh. is shown in Fig. 6.5 where the motion mask is indicated in white. The subject's right big toe is moving, thus only the white regions around the big toe and its shadows are actually 'true' subject movement. The rest is caused by the moving IS. It is clearly visible that the number of misdetected motion blocks due to the IS is reduced drastically by the BPF (right image) while preserving the true subject motion.

6.3.2 Texture model (TM)

The output of VP enh. is given as input to the texture model TM which eliminates misdetections in image areas where texture remains substantially stable over time.

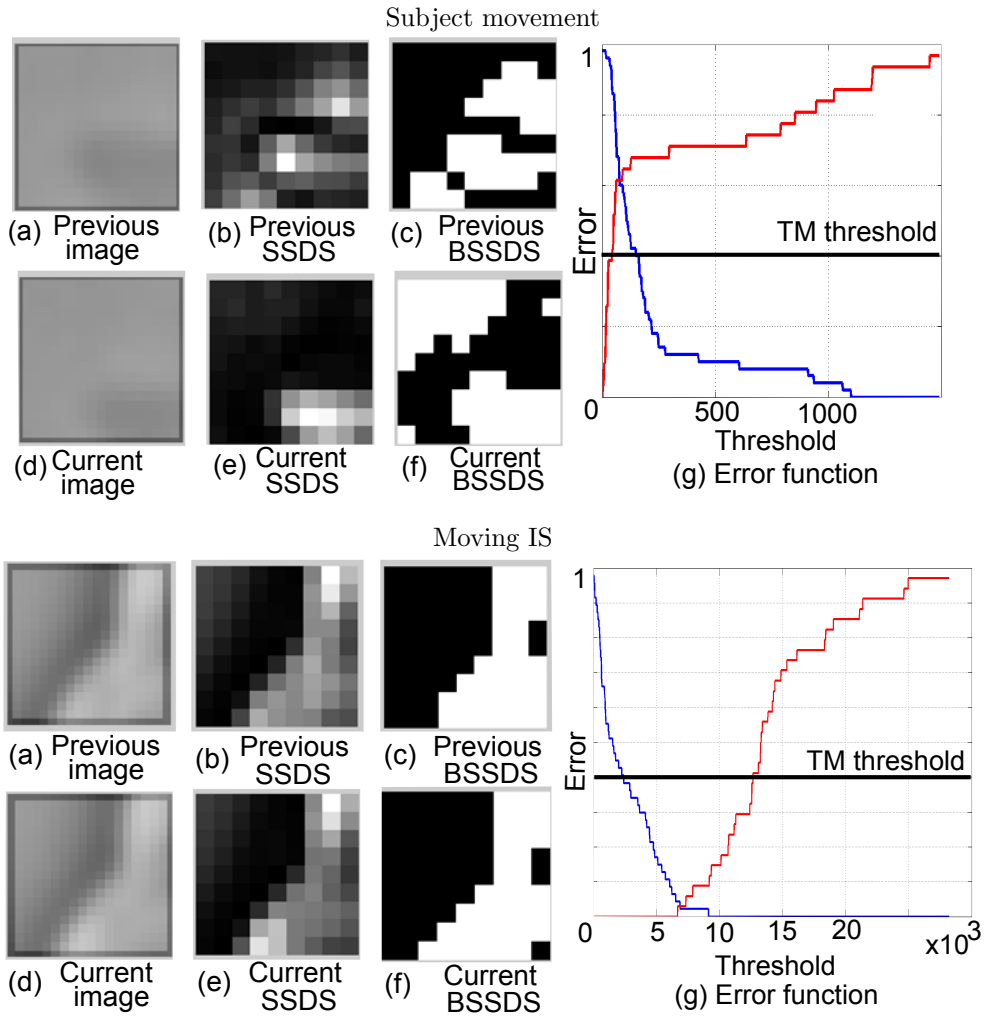


Figure 6.6: TM of subject movement (top) and moving IS (bottom). (a) and (d) are corresponding blocks in consecutive frames, (b) and (e) are the corresponding SSDSs, (c) and (f) the BSSDSs, (g) shows the error functions E_{b_i} and E_{w_i} in red and blue, respectively.

TM characterizes textures using sum of squared difference surfaces (SSDS) [314] and two textures are considered substantially unchanged if their corresponding binarized SSDS (BSSDS) differ less than the threshold value d_{TM} determined in 6.3.4. In [314], the SSDS is implemented to quantify the similarity across images or videos. It is created by computing the total sum of squared differences (SSD) between a given 2×2 block and the blocks in its surrounding region. Fig. 6.6 top (a) and (d) illustrate the corresponding 18×18 pixel blocks between two consecutive frames. The displacement of the dark object is caused by subject movement. The corresponding SSDSs are shown in Fig. 6.6 top (b) and (e) respectively. We compare textures using

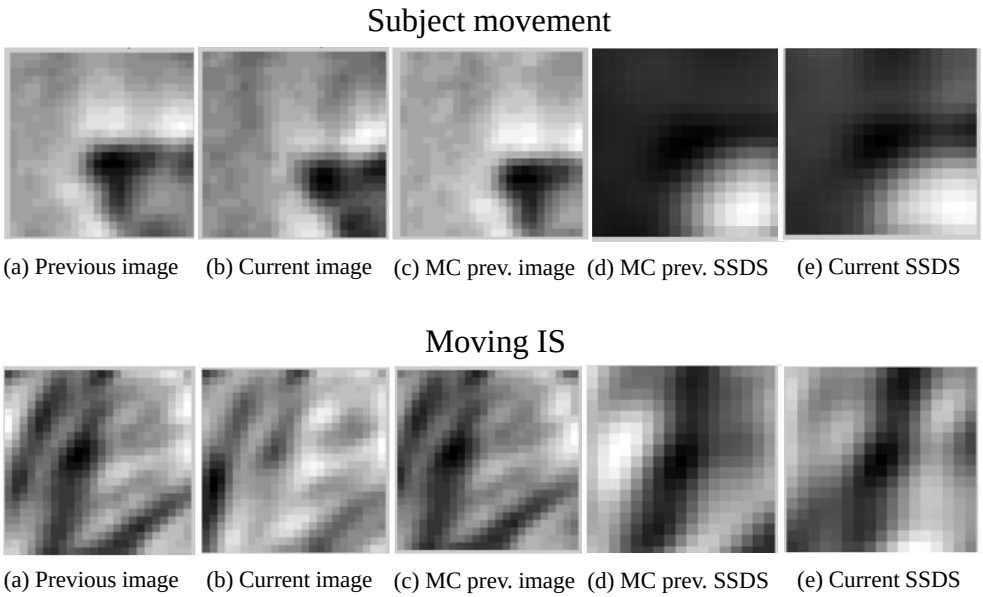


Figure 6.7: MCTM of subject movement (top) and moving IS (bottom). (a) and (b) are corresponding blocks in consecutive frames, (c) is the motion compensated block of (a), (d) and (e) are the corresponding SSDSs of (c) and (b).

binarized summed squared difference surfaces (BSSDS) with two different binarization threshold approaches. The initial threshold for BSSDS_{i-1} is the mean value of the corresponding SSDS. The BSSDS of the previous frame is shown in Fig. 6.6 (c). The second binarization threshold for BSSDS_i is selected such that the difference between BSSDS_{i-1} and BSSDS_i is minimized. The minimized difference E_{TM} is used as the measure for texture similarity. The cross point in Fig. 6.6 (g) of the monotonic error functions of the black block error E_{b_i} (red line) at frame i (see Eqn. (6.9)) and the white block error E_w (blue line) is regarded as the minimum error E_{TM}. If $E_{\mathrm{TM}}(b_x, b_y) > d_{\mathrm{TM}}$, the block (b_x, b_y) is classified as motion. The black block error E_b is computed as the conditional probability that a block is white in the current BSSDS given it is black in the previous BSSDS. The white block error E_w is computed analogueously. The number of blocks is denoted by n_b. Fig. 6.6 bottom shows the corresponding images for the case of an IS. The minimal error is clearly below the error threshold.

$$E_{b_i}(b_x, b_y) = \frac{n_b|\mathrm{BSSDS}_i(b_x, b_y) = 1 \cap \mathrm{BSSDS}_{i-1}(b_x, b_y) = 0}{n_b|\mathrm{BSSDS}_{i-1}(b_x, b_y) = 0} \tag{6.9}$$

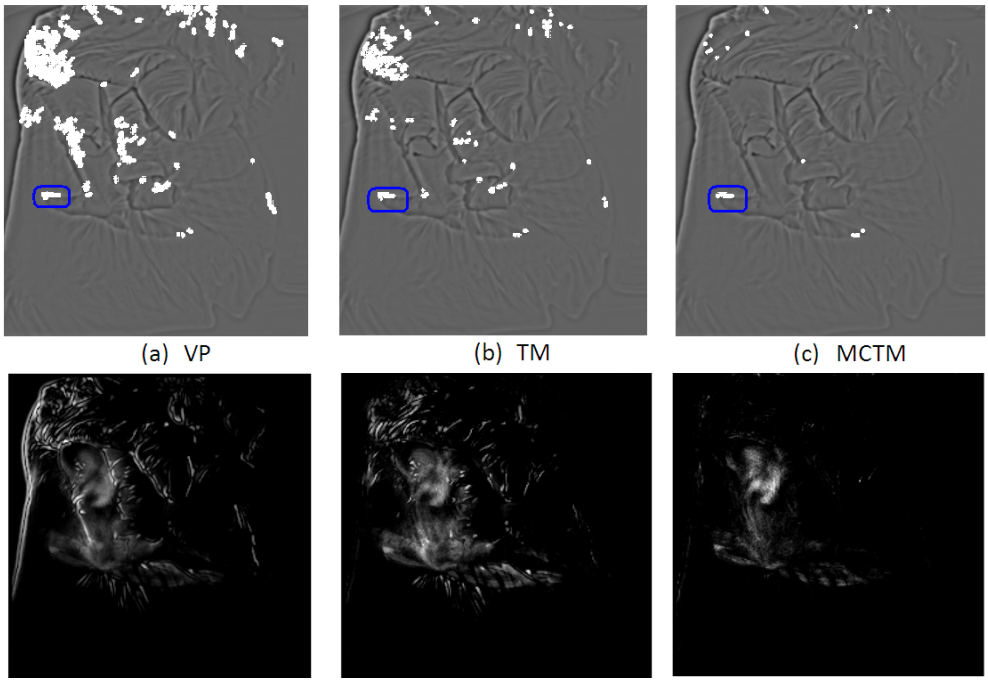


Figure 6.8: Motion masks (top) and accumulated motion masks (bottom) after applying VP enh. (a), TM (b), MCTM (c) when the right toe is moving (blue square).

6.3.3 Motion compensated texture model (MCTM)

TM is able to remove the false positive detections where the textures of the corresponding regions remain unchanged. However, the false positives along the sharp edges due to the folds in the background are barely reduced. When the IS moves to another image area and discloses texture which has previously been occluded due to the non-planar background, there may be sudden texture changes. Therefore, we introduce the motion compensated texture model (MCTM) which eliminates false detections where the texture changes over time in an unpredictable manner. The method is based on the assumption that when subject movement occurs, a motion vector exists which renders a small difference between SSDS_i and the motion compensated SSDS_{i-1}. The motion vectors within a search window of 24×24 pixels are estimated with the full search block matching technique on 8×8 pixel blocks and a search step of 2×2 pixels (the block sizes are empirically determined). Fig. 6.7 (top) shows an example of MCTM when subject movement is present. Using the motion vector a motion compensated image for the previous frame (a) can be calculated. Fig. 6.7 (d) and (e) show that the corresponding SSDSs are similar. Fig. 6.7 (bottom) shows an example of MCTM when there is only moving shadow. The textures in (a) and (b) are different due to the moving shadow. The texture of Fig. 6.7 (bottom) (e) shifts to the right compared to the texture in (d). To determine the

similarity between the corresponding SSDSs Eqn. (6.9) is reused. If the minimum error $E_{\mathrm{MCTM}}(b_x, b_y) < d_{\mathrm{MCTM}}$ determined in 6.3.4, the block (b_x, b_y) is classified as motion.

The white regions of (a), (b) and (c) in the top row of Fig. 6.8 are motion masks of VP enh., TM and MCTM respectively. At this moment, the IS is moving from bottom left to top right and the subject's right big toe is moving upwards. From (a) to (c) it can be clearly seen that the number of false positives due to the IS is greatly reduced at the cost of a limited reduction of true positive detections.

The bottom row of Fig. 6.8 shows the accumulated motion masks over the whole period of a training sequence using VP enh., TM and MCTM. In Fig. 6.8 (a), the detected motion regions are spread over a large area, indicating that the specificity of VP enh. is low. TM is good at classifying the image variations along the bed because the textures in these regions are unchanged. Fig. 6.8 (c) depicts that MCTM can eliminate the misdetections on the quilt. The remaining white areas in (c) represent the subject movements best.

6.3.4 Subject motion classification and feature selection

A motion classifier determines, using the number of features, if a frame is detected as a *motion frame*. To avoid suboptimal, manual selection of these features, we propose to use supervised learning. Based on the extensive literature study and benchmark in [315], cascaded AdaBoost was selected due to its potential in terms of achieved performance and computational complexity. More information regarding the specific implementation of the used cascaded AdaBoost can be found in [316].

We apply cascaded AdaBoost using the intermediate output of VP enh., TM and MCTM sub-methods as features. Each sub-method of the cascade (VP enh., TM, MCTM) detects 'motion' blocks using internal thresholds d_{VP}, d_{TM}, d_{MCTM}. The methods of the cascade are executed several times with different threshold combinations, and the numbers of motion blocks per frame detected by VP enh., TM and MCTM are used as features. A complete set of motion block thresholds is shown in Table 6.1 where the specific values of the thresholds were found by analysing the influence of the sub-methods on the throughput of the cascade. First, for every method, a minimal (maximal) threshold was found which effectively minimized (maximized) the number of motion blocks passing to the next level of the cascade. Where the number of motion blocks varied highly, a denser distribution of the thresholds was selected (see regions with steep slope in Fig. 6.9). Thus, 10 thresholds for the VP enh. method correspond to 10 VP enh. features. Since TM considers only 'motion' blocks detected by VP enh., TM features correspond to combinations of 10 thresholds of VP enh. and 9 thresholds of TM, which gives us another $10 \cdot 9 = 90$ TM features. Similarly, MCTM features correspond to triples of VP enh., TM, and MCTM thresholds, where we sub-sampled the threshold set in order to reduce the number of combinations (bold values in Table 6.1). We selected thresholds which gave promising results when the pure cascade, without AdaBoost classification, was considered as the motion detection

method, see dashed line in Fig. 6.4.

The performance of the pure cascade with respect to the ground truth was evaluated on basis of Matthews Correlation Coefficient (MCC) [317] - used in machine learning for assessing the performance of classification algorithms [318]. Only 6 VP enh., 6 TM and 5 MCTM thresholds, giving a reasonable performance of the cascade, were selected resulting in $6 \cdot 6 \cdot 5 = 180$ MCTM features.

Then cascaded AdaBoost is applied to determine the most relevant features based on $6 \cdot 4 \cdot 16 = 384$ training frames (six test sequences, four 'subject motion'/'shadow motion' combinations, 16 frames per combination). Six feature combinations were extracted for VMD (see Table 6.2). Consequently, these features were computed for each test sequence. For the purpose of benchmarking, cascaded AdaBoost was also applied to VP orig., using the initial set of thresholds given in the bottom row of Table 6.1.

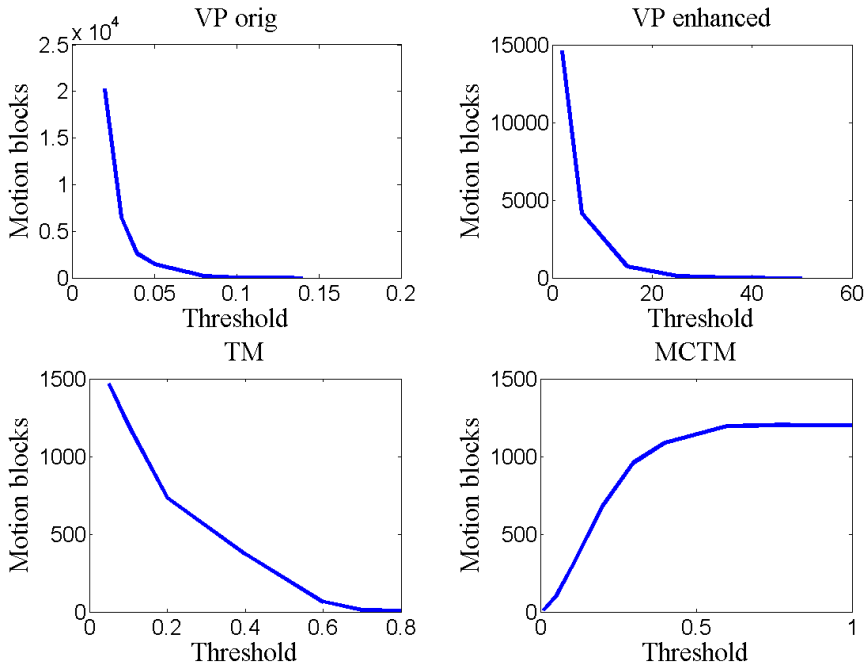


Figure 6.9: Motion block variation with different thresholds for VP orig., VP enh., TM, MCTM.

6.4 Experimental setup

Six video sequences were recorded with six different subjects simulating PLM activity in bed. All subjects followed the motion scheme in Fig. 6.10 after being trained with video clips from the 'Video Guide for PLM' [319]. The duration of each test sequence was around 30 minutes where the volunteers repeated the same movement program

Method	Set of thresholds for AdaBoost training
VP enh.	**2, 3, 3.5, 4, 4.5, 5**, 6, 12, 25, 40
TM	0.1, 0.15, **0.2, 0.3, 0.35, 0.4, 0.45, 0.5**, 0.6
MCTM	0.01, **0.05, 0.1, 0.15, 0.2, 0.25**, 0.3, 0.45
VP orig	0.01, 0.015, **0.02**, 0.022, 0.025, 0.027, 0.03, 0.035, **0.04**, 0.045, 0.05, 0.06, 0.08, **0.1**, 0.12, 0.125, 0.13, 0.135, 0.14, 0.145, 0.15, 0.155, **0.16**

Table 6.1: VP enh., TM, MCTM, and VP orig. thresholds for AdaBoost training. Bold thresholds are in final selection.

twice: first, with constant NIR illumination only, and a second time with different interfering light sources and moving shadows. Sunlight and street light were taken into account by opening the curtains. The secondary ambient illumination was modeled with a standing lamp. In front of the lamp, plastic plants were moved to simulate a waving tree branch and its effects. Curtains were manually opened and closed to simulate moving curtains. Legs were covered and uncovered randomly in between the two specified time instances indicated in gray in Fig. 6.10.

The scene was registered with a monochrome camera equipped with a NIR-pass filter and an auxiliary NIR illumination source, positioned about 0.75 m above the bed surface at the foot end of the bed. While a low frame rate can make the individual frames more distinctive and therefore facilitate the detection of the slow motion, a relatively high frame rate is necessary to capture fast limb movements. Human voluntary movements are typically limited to 200/min (3.3 Hz) [320], however, human involuntary movements (as can be the case of PLMS) may be faster [321]. Movements in a frequency range larger than 5/s are caused by tremor and shivering [322]. In order to disregard shivering movements, a frame rate of 10.9 Hz was chosen (the possible camera setting larger than twice the acceptable movement frequency). In our trial experiments, we found that the selected frame rate provided a sufficiently good temporal resolution to capture subject movements.

For a direct comparison between the developed VMD method and a standard PLMI score on a group of PLM patients, eleven video clips from [319] were processed.

For the first data set, classifier parameters were determined by performing leave-one-subject-out cross-validation on the six test sequences. For the second data set

	F1	F2	F3	F4	F5	F6
VP enh.	2	2	3	25	4	5
TM		0.45	0.2	0.4	0.45	0.2
MCTM		0.1	0.05		0.15	0.25

Table 6.2: Selected set of thresholds returning the VMD features.

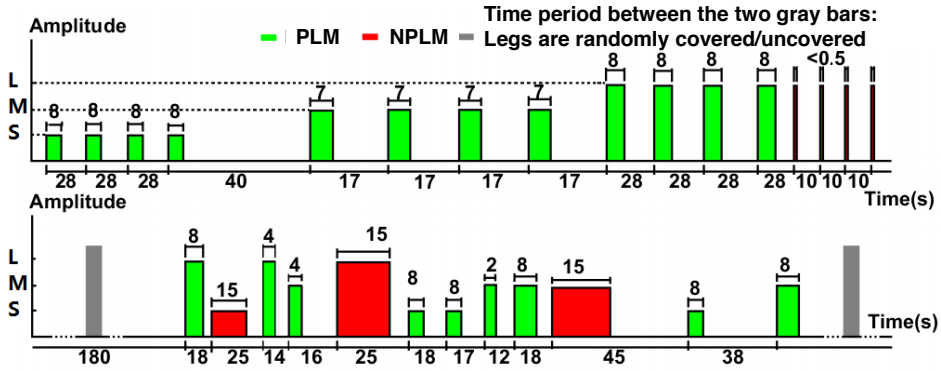


Figure 6.10: Program for limb movements to simulate PLMS.

where it is unknown whether one subject is involved in multiple test sequences we perform leave-one-sequence-out cross-validation on the eleven test sequences.

6.5 Results and discussion

6.5.1 Frame based motion detection

Given the conditions of the current application, a universal state-of-the-art motion detector robust to moving cast shadows is found in VP orig. [311] which is successfully applied in [312].

The motion decision output of VP orig. and the proposed VMD method for each frame are shown in Fig. 6.11, along with the ground truth, in red, green and black respectively. The black curve represents the ground truth that has been extracted from the EMG signal, (a) denotes the first half of the experiment where no IS was present, (b) the second half with IS. Fig. 6.11 shows that VP orig. may underdetect subject movement at times (e.g., 5(a)) and may at other times be highly sensitive to moving IS (e.g., 2(b)). Most motion events are marked (with a large number of classified motion frames in a motion event period) as motion by the proposed method VMD. It is less sensitive to the moving IS and can preserve a larger number of true positives (TP). The improved performance in distinguishing between subject and shadow motion is illustrated well in subject 2. Even when the limbs are covered (shaded areas in Fig. 6.11) VMD can detect several movements. Subject 4 presents the largest subject motion challenge where the motion decisions between 620s and 820s in 4(a) show that both algorithms are unable to detect subject movements when the subject's legs were covered.

Regarding the quantitative results for frame-based motion detection, Table 6.3 shows the MCC, sensitivity and specificity scores of VP orig. and the proposed VMD method. All the average scores are improved with VMD (MCC by factor 2, sensitivity

	MCC	Sens.	Spec.
VP orig			
Mean	0.3	47.4%	84.2%
StD	0.2	30.8	10.7
VMD			
Mean	0.7	68.5%	96.5%
StD	0.1	15.6	1.7

Table 6.3: MCC, sensitivity and specificity of VP orig and VMD.

by 45%, specificity by 15%) resulting in a MCC of 0.7, sensitivity of 68.5%, specificity of 96.5%. Sequence 1 represents an outlier for VMD as the results are significantly worse.

6.5.2 Event based PLMS detection on simulated data

This section evaluates the motion event classification with the investigated methods VP orig. and the proposed VMD on an event basis. In this study, motion and non-motion events are computed for the detection of PLM. The severity of PLM can be quantified using the so-called periodic limb movement index (PLMI), see [323]. PLMI is the total number of PLMS events divided by the total sleep time in hours. The 'gold standard' and the corresponding five rules for scoring PLMS events are defined in [323] in terms of limb movement duration (LMD) and inter-movement interval (IMI), which can be naturally identified in the EMG signal. The VMD signal can naturally substitute the EMG signal in the detection and counting of the PLMS events, leading to video-based PLMI (VPLMI).

Both video signals were filtered with a m-tap median filter to provide the minimal possible root mean square (RMS) error with the EMG based movement events. Table 6.4 shows the average performance for VP orig. with $m = 13$ and VMD with $m = 9$ on six sub-sequences with constant illumination and with IS. One can see that VMD scores better than VP orig. in both lighting conditions with an improvement factor of 6 of the normalized average error in constant lighting and a factor of 4 in challenging lighting while keeping the standard deviation small. Similarly, the RMS error count is reduced from 9.7 to 1.8 and from 9.8 to 3.7. A comparison of the motion detection output is given in Fig. 6.12.

6.5.3 Event based PLMS detection on patient data

Regarding the validation on PLM patient sequences, the same six features presented in Table 6.2 are computed. Due to the (heavy) compression artifacts, the classifier parameters were redetermined by performing cross-validation on the eleven test sequences. A 3-tap median filter returned the minimal RMS error. As the availabe video

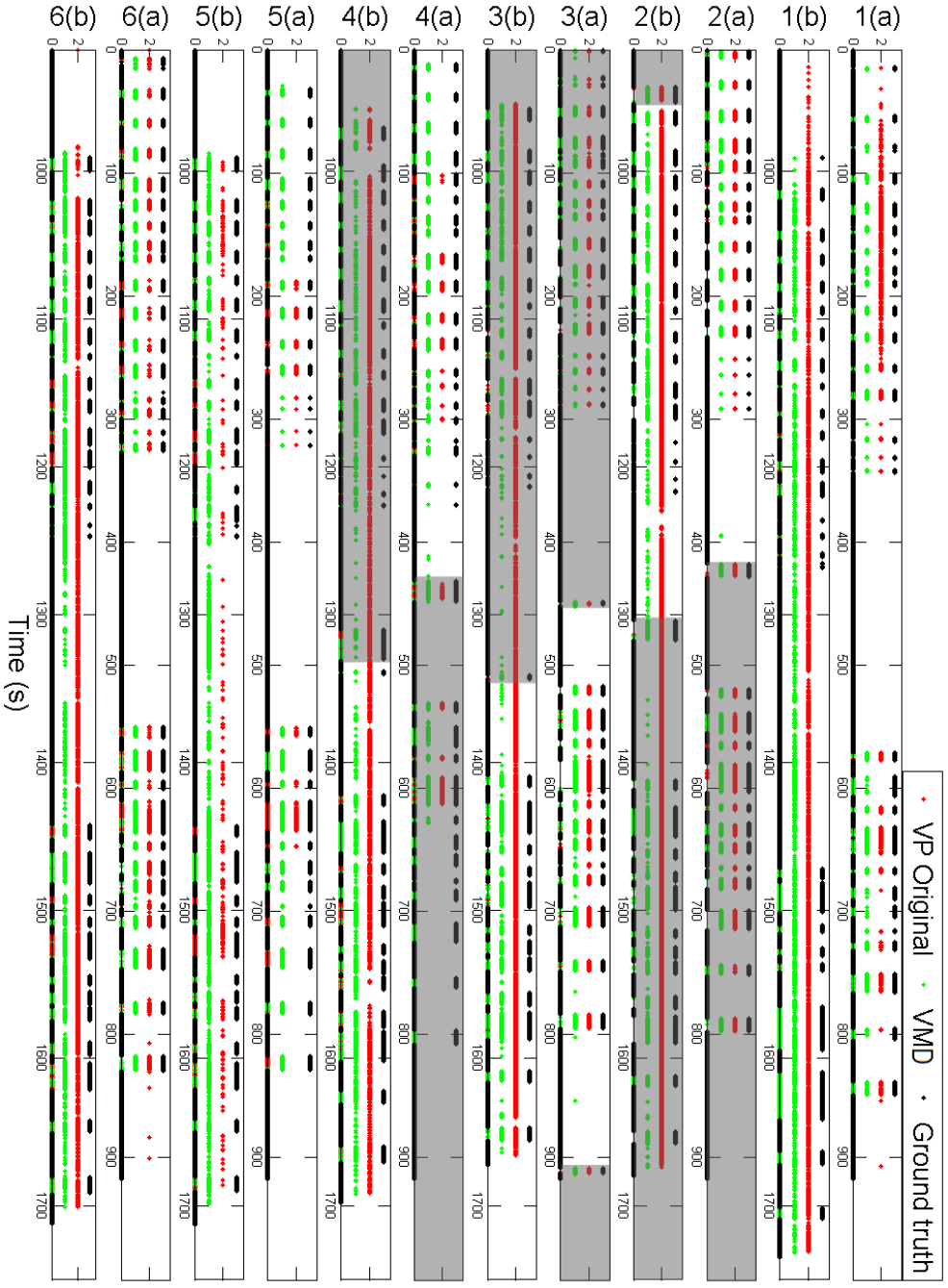


Figure 6.11: Frame based motion detection output of VP orig. (red) and proposed VMD (green), groundtruth (black). (a) and (b) indicate the periods without and with moving IS respectively. The shaded areas indicate the periods when the subject's legs are covered.

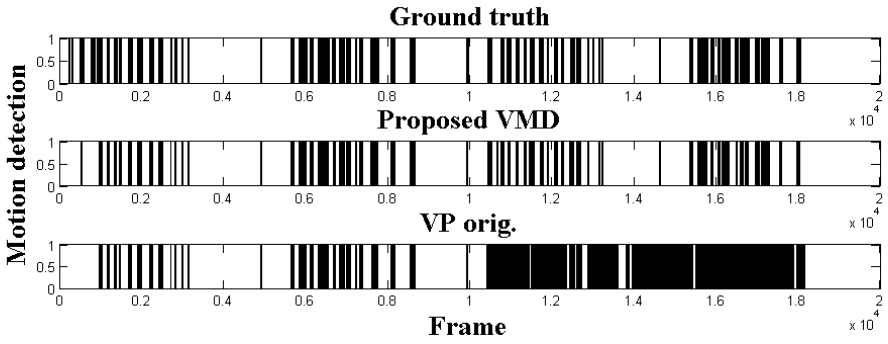


Figure 6.12: Comparison of filtered motion detection results for subject 3.

data consists of cut fragments that were subsequently concatenated, the gold standard IMI requirement [323] of 4s was relaxed to 2s. Due to the absence of the EMG signal, ground truth was annotated according to frame based movement observation. Similarly, an actigraphy based approach for detecting PLM is suggested and found reliable and correct in [324] and [325]. Some problems with EMG measurements may even be avoided with actigraphy sensors according to [324]. The results are presented in Table 6.5. A total of 56 PLM movements were counted both with VMD and the ground truth data. Ten out of eleven sequences accurately detected the timing of the movement. One sequence showed a slight temporal misalignment (<1s) regarding the start of the movement. The often uncovered and well visible body parts in the video clips were favorable for our algorithm. Future work should include the validation of the proposed VMD algorithm on several nights of multiple patients.

	Mean $\frac{\lvert\text{VPLMI-PLMI}\rvert}{\text{PLMI}}$	**StD** $\frac{\lvert\text{VPLMI-PLMI}\rvert}{\text{PLMI}}$	**RMS** error
No IS			
VP orig	0.46	0.37	9.7
VMD	0.08	0.1	1.8
IS			
VP orig	0.61	0.44	9.8
VMD	0.21	0.17	3.7

Table 6.4: Average PLMI performance difference of VP orig and VMD methods in constant lighting conditions and with IS.

	Mean $\frac{\lvert\text{VPLMI-PLMI}\rvert}{\text{PLMI}}$	**StD** $\frac{\lvert\text{VPLMI-PLMI}\rvert}{\text{PLMI}}$	**RMS** error
VMD	0	0	0.0373

Table 6.5: Average PLMI performance of proposed VMD method for compressed videos of 11 PLM patients.

6.6 Conclusions

We proposed a video movement detection (VMD) system that can discriminate movements caused by a sleeping subject from motion due to cast shadows on a non-planar, dynamic background. The video processing integrates motion detection, motion estimation and texture analysis. The proposed cascade includes the variance plus (VP enh.) motion detector, the texture model (TM) and motion compensated texture model (MCTM) features efficiently aggregated in a strong classifier using cascaded AdaBoost.

The proposed robust motion detector has been evaluated in a scenario combining illumination changes and subject motions for the purpose of periodic limb movement detection in sleep (PLMS). Simulated PLMS movements have been recorded on video along with a reference EMG signal.

The motion detection performance per frame of the proposed system provides an MCC improvement of a factor 2, an increase of 45% in sensitivity and 15% in specificity, on six test sequences compared to a state-of-the-art method for shadow robust motion detection.

Movement event classification for periodic limb movement detection improved with the proposed VMD method by a factor of 6 and by a factor of 3 in constant and highly varying lighting conditions, respectively, compared to state-of-the-art. A 100% accurate PLMI score is obtained on eleven PLM patient test sequences due to favorable viewing conditions. Only for one movement a slight misalignment (<1s) in starting time was observed between the ground truth and VMD. We conclude that video-based PLMS is feasible and can provide a coarse indication of the PLMI score.

Bibliography

[292] R. Gardner and W.I. Grossman, "Normal motor patterns in sleep in man", *Advances in Sleep Research*, **vol. 2** pp. 67–107, 1976.

[293] J. Wilde-Frenz and H. Schulz, "Rate and distribution of body movements during sleep in humans", *Percept. Mot. Skills*, **vol. 56** pp. 275–283, 1983.

[294] A. Muzet, P. Naitoh, L.C. Johnson, and R.E. Townsend, "Body movements during sleep as a predictor of state change", *Psychon. Sci.*, **vol. 29** pp. 7–10, 1972.

[295] S. Gori, G. Ficca, Di Nasso, L. I. Murri, and P. Salzarulo, "Body movements during night sleep in healthy elderly subjects and their relationships with sleep stages", *Brain Research Bulletin*, **vol. 63**, no. 5 pp. 393–397, June 2004.

[296] T. Etzioni and G. Pillar, "Movement disorders in sleep", *Harefuah*, **vol. 146**, no. 7 pp. 544–548, July 2007.

[297] L. Unzueta, M. Nieto, A. Cortes, J. Barandiaran, *et al.*, "Adaptive Multicue Background Subtraction for Robust Vehicle Counting and Classification", *Intelligent Transportation Systems, IEEE Transactions on*, **vol. 13**, no. 2 pp. 527–540, June 2012.

[298] A. Leone and C. Distante, "Shadow detection for moving objects based on texture analysis", *Pattern Recognition*, **vol. 40**, no. 4 pp. 1222–1233, Apr. 2007.

[299] "http://www.smarteye.se", in *accessed online 5-October-2013*.

[300] A. Prati, I. Mikic, M.M. Trivedi, and R. Cucchiara, "Detecting Moving Shadows: Algorithms and Evaluation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **vol. 25**, no. 7 pp. 918–923, July 2003.

[301] W. Zhang, X.Z. Fang, X.K. Yang, and Q.M.J. Wu, "Moving Cast Shadows Detection Using Ratio Edge", *Multimedia, IEEE Transactions on*, **vol. 9**, no. 6 pp. 1202–1214, Oct. 2007.

[302] J. Stander, R. Mech, and J. Ostermann, "Detection of moving cast shadows for object segmentation", *IEEE Transactions on Multimedia*, **vol. 1**, no. 1 pp. 65–76, Mar. 1999.

[303] A. Sanin, C. Sanderson, and B.C. Lovell, "Shadow detection: A survey and comparative evaluation of recent methods", *Pattern Recogn.*, **vol. 45**, no. 4 pp. 1684–1695, Apr. 2012.

[304] J.V. Panicker, "Detection of moving cast shadows using edge information", in *Int'l Conf. on Computer and Automation Engineering*, 2010.

[305] W.-H. Liao and C.-M. Yang, "Video-based activity and movement pattern analysis in overnight sleep studies.", in *Int'l Conf. on Pattern Recognition*, 2008, pp. 1–4.

[306] K. Cuppens, L. Lagae, B. Ceulemans, S. Van Huffel, and B. Vanrumste, "Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy", *Med. Biol. Engineering and Computing*, **vol. 48**, no. 9 pp. 923–931, Sep. 2010.

[307] W.-H. Liao and J.H. Kuo, "Sleep monitoring system in real bedroom environment using texture-based background modeling approaches", *J. Ambient Intelligence and Humanized Computing*, **vol. 4**, no. 1 pp. 57–66, Feb. 2013.

[308] K. Skifstad and K. Jain, "Illumination independent change detection for real world image sequences", *CVIP*, **vol. 46**, no. 3 pp. 387–399, June 1989.

[309] Z. Liu, K. Huang, and T. Tan, "Cast Shadow Removal in a Hierarchical Manner Using MRF", *Circuits and Systems for Video Technology, IEEE Transactions on*, **vol. 22**, no. 1 pp. 56–66, 2012.

[310] E. Kermani and D. Asemani, "A New Illumination-Invariant Method of Moving Object Detection for Video Surveillance Systems", in *Iranian Conf. on Machine Vision and Image Processing*, 2011, pp. 1–5.

[311] E. Durucan and T. Ebrahimi, "Change detection and background extraction by linear algebra", *Proceedings of the IEEE*, **vol. 89**, no. 10 pp. 1368–1381, Oct. 2001.

[312] R. Liao, "2-D/3-D registration of C-ARM CT volumes with fluoroscopic images by spines for patient movement correction during electrophysiology", in *Proc. of the Int'l Conf. on Biomedical Imaging*, 2010, pp. 1213–1216.

[313] J. Zhu, K.G.G. Samuel, S.Z. Masood, and M.F. Tappen, "Learning to recognize shadows in monochromatic natural images", in *Computer Vision and Pattern Recognition, 2010 IEEE Conference on*, 2010, pp. 223–230.

[314] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos", in *IEEE Conf. on Computer Vision and Pattern Recognition 2007*, 2007.

[315] J.P. Vink and G. de Haan, "Comparison of machine learning techniques for target detection", *Artificial Intelligence Review*, pp. 1–15, Nov. 2012.

[316] J.P. Vink, M.B. van Leeuwen, C.H.M. van Deurzen, and G. de Haan, "Efficient nucleus detector in histopathology images", *Journal of Microscopy*, **vol. 249**, no. 2 pp. 124–135, Feb. 2013.

[317] B.W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", *Biochim. Biophys. Acta*, **vol. 405**, no. 2 pp. 442–451, Oct. 1975.

[318] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview", *Bioinformatics*, **vol. 16**, no. 5 pp. 412–424, Feb. 2000.

[319] B. Högl, M. Zucconi, and F. Provini, "RLS, PLM, and their differential diagnosis – a video guide.", *Mov Disord*, **vol. 22 Suppl 18** pp. S414–9, 2007.

[320] F.G. Foster, R.J. McPartland, and D.J. Kupfer, "Motion sensors in medicine, Part I. A report on reliability and validity", *Journal of Inter-American Medicine*, **vol. 3** pp. 4–8, 1978.

[321] E.J.W. Van Someren, R.H. Lazeron, and B.F.M. Vonk, "Wrist acceleration and consequences for actigraphic rest-activity registration in young and elderly subjects", in *Sleep-Wake Research in the Netherlands*, 1995, pp. 6123–6125.

[322] D.P. Redmond and F.W. Hegg, "Observations on the design and specification of a wrist-worn human activity monitoring system", *Behavior Research Methods, Instruments, & Computers*, **vol. 17**, no. 6 pp. 659–669, Nov. 1985.

[323] M. Zucconi, R. Ferri, R. Allen, P. C. Baier, *et al.*, "The official World Association of Sleep Medicine (WASM) standards for recording and scoring periodic leg movements in sleep (PLMS) and wakeful-ness (PLMW) developed in collaboration with a task force from the International Restless Legs Syndrome Study Group (IRLSSG)", *Sleep Medicine*, **vol. 7**, no. 2 pp. 175–183, Mar. 1972.

[324] R.P. Allen and W.A. Hening, "Actigraph Assessment of Periodic Leg Movements and Restless Legs Syndrome", in *Restless Legs Syndrome*, W.B. Saunders, Philadelphia, pp. 142 – 149, 2009.

[325] E. Sforza, M. Johannes, and B. Claudio, "The PAM-RL ambulatory device for detection of periodic leg movements: a validation study.", *Sleep Med*, **vol. 6**, no. 5 pp. 407–13, Sep. 2005.

Chapter 7

Lifestyle applications from sleep research

Abstract

Most of the research performed in the area of movement analysis of sleeping subjects has been targeted at sleep stage classification or monitoring of sleep disorders. In this paper, we present an innovative approach and show how movement analysis of sleeping subjects can be used to enable new lifestyle related applications. The first application we propose targets an intelligent baby monitor that informs parents about changes of their baby's pose in its sleep. The second application shows how a sleeping subject's movement pattern can be used to build an intelligent wake-up light system. For the two proposed systems, we present design considerations and initial results showing the potential of camera-based movement analysis in sleep related applications outside the common interest.

This chapter is published as: A. Heinrich, V. Jeanne, X. Zhao; Lifestyle applications from sleep research, *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, no. 6, pp. 829-842, Apr. 2014.

7.1 Introduction

Body movements are an important behavioral aspect during sleep as shown in [326]. They can be associated to sleep states [327] and connected to sleep state transitions [328]. It was concluded in [329] that frequency and duration of body movements are important characteristics for sleep analysis. Moreover, various somatic and mental disorders are associated with movements during sleep and several clinical syndromes of pathological movements exist [330], like restless leg syndrome, periodic limb movement disorder, and movement disorders under parasomnias such as disorders of arousal and sleep-wake transitions. To this end, the movements of sleeping subjects are typically analyzed with the aim to perform sleep/wake classification [331] or to screen for diseases characterized by particular movement patterns [332].

Traditional sleep screening and sleep classification approaches are performed in polysomnography (PSG) studies in sleep clinics. In the last decade, quite some research efforts have been carried out to monitor sleeping subjects at home, in the natural sleeping environment, while employing sensors that typically suffer from reduced accuracy, therefore offer more unobtrusiveness and easy installation at home. These less obtrusive sensors include amongst others wrist actigraphy [331], pressure sensors in the pillow [333], piezoelectric sensors, strain-gauge and electret foil sensors [334]. A promising sensor gaining more and more attention in the recent years is the video camera [335–337]. Attempts have been made to use the video camera and novel video processing algorithms for sleep/wake classification [335, 338], sleep movement disorder detection [339] and sleep breathing disorder detection [340].

In this paper, we propose to use the movement information for a different application focus. Two lifestyle applications make use of the movement information recorded with a camera sensor and processed with two novel algorithms. The first application describes an intelligent baby monitor that informs the parents when their baby is turning in its sleep and when it is sleeping on its belly (parents are currently advised to have their baby sleep on its back as it is believed to be safer). The second application describes an intelligent wake-up light adapting the lighting settings dependent on the sleeping subject's movement pattern.

Section 7.2 describes the intelligent baby monitor, Section 7.3 the intelligent wake-up light system and Section 7.4 concludes this paper.

7.2 Intelligent baby monitor

We propose to analyze body movement information for baby monitoring to inform parents when their baby has turned to his/her belly. This solution is interesting for parents as 4500 infants in the United States alone die annually of sudden infant death syndrome (SIDS) [341]. In order to reduce the risk for SIDS, parents are advised to put their babies to sleep on their back (supine) and not on their stomach (prone).

Several baby monitors exist on the market that transmit the video image of the sleeping baby to the parent unit. These devices do not perform any automatic analysis

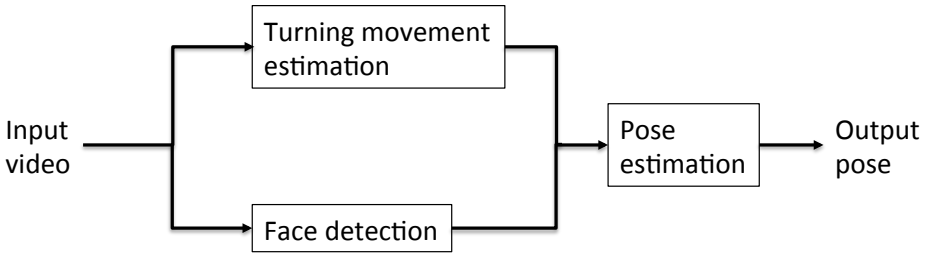


Figure 7.1: Block diagram of proposed method to determine the infant's sleeping pose.

of the available video data. In order to assist parents in unobtrusively monitoring their baby with familiar technology and giving them a more reassured feeling when they are not looking at the transmitted video, we propose an algorithm to detect turning movements of the baby and the body position (prone or supine) the baby has turned to. The system could then notify the parents when their baby has turned to a more unsafe sleeping position.

A solution for recognizing the pose of a sleeping subject is proposed in [342] where the authors distinguish between lying on the side and lying in a supine/prone position. However, their proposed method is not capable of distinguishing between supine and prone. Reflective markers are attached to the baby's sleeping bag in [343] which makes it easier to monitor the movements and pose of a baby. Our goal is to stick to existing consumer product environments with infrared baby monitors without the need for markers in order to be most forthcoming to parents.

Traditional computer vision methods for face detection have been used for infants in [344]. We found them ineffective in our setting due to the at times unfavorable face pose/angle towards the camera (often non-frontal), low contrast in the infrared images and the rather low quality video images. Therefore, we propose a solution adopting information from both motion and face for infant pose detection. This approach is illustrated in Fig. 7.1 where both turning movement estimation (described in detail in Section 7.2.1) and face detection (described in Section 7.2.2) are used to determine the infant's sleeping pose (Section 7.2.3). The evaluation methods and user study are outlined in Section 7.2.4 and Section 7.2.5, respectively. Results are presented in Section 7.2.6.

7.2.1 Turning movement estimation

A turning movement can be regarded as a rotation around the body axis. Hence, the proposed turning movement estimation relies heavily on body axis calculation and motion estimation returning a motion vector per image block/pixel. An overview of the turning movement estimation method is given in Fig. 7.2.

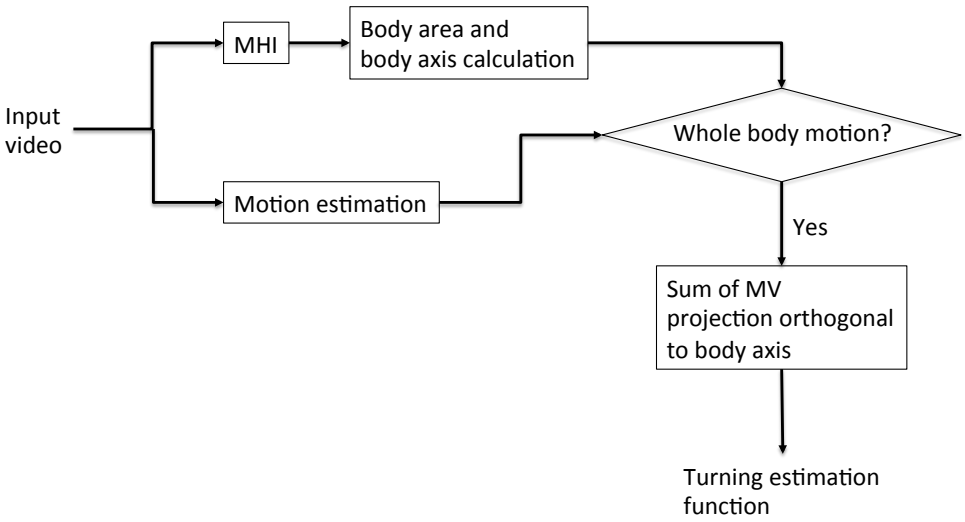


Figure 7.2: Block diagram of proposed turning movement estimation method.

Body axis calculation

The body axis is computed by first, detecting the body area, and second, fitting a line to this area by applying the least squares method. The part of the image the body occupies is determined by the motion history image (MHI) method, also already applied for video sleep analysis [337]. With MHI, detected movement pixels from an image-time volume are compressed into a single image where more recent motion pixels are assigned a larger value. MHI is adequate for large motion analysis in the sense that temporally consistent motion areas can be determined and movements occurring only in one frame discarded. Fig. 7.3 illustrates the detection of the majority of the body with MHI versus few silhouette pixels with simple frame differencing.

We propose to fit a line on the MHI by applying the least squares method. The fit line yields the body axis (see red line in the second and third row of Fig. 7.3).

Movement projection orthogonal to body axis

A turning movement is detected when the motion vectors between two consecutive images yield a movement in the orthogonal direction of the body axis. Frequently occurring movements such as kicking and waving arms may also occur in the said direction and need to be neglected for the turning movement detection. In a turning movement, the entire body is involved, namely also the infant's trunk. Thus, two conditions for a potential turning movement are required:

- A large movement area (number of motion blocks $n_m > d_m$ with the threshold d_m),

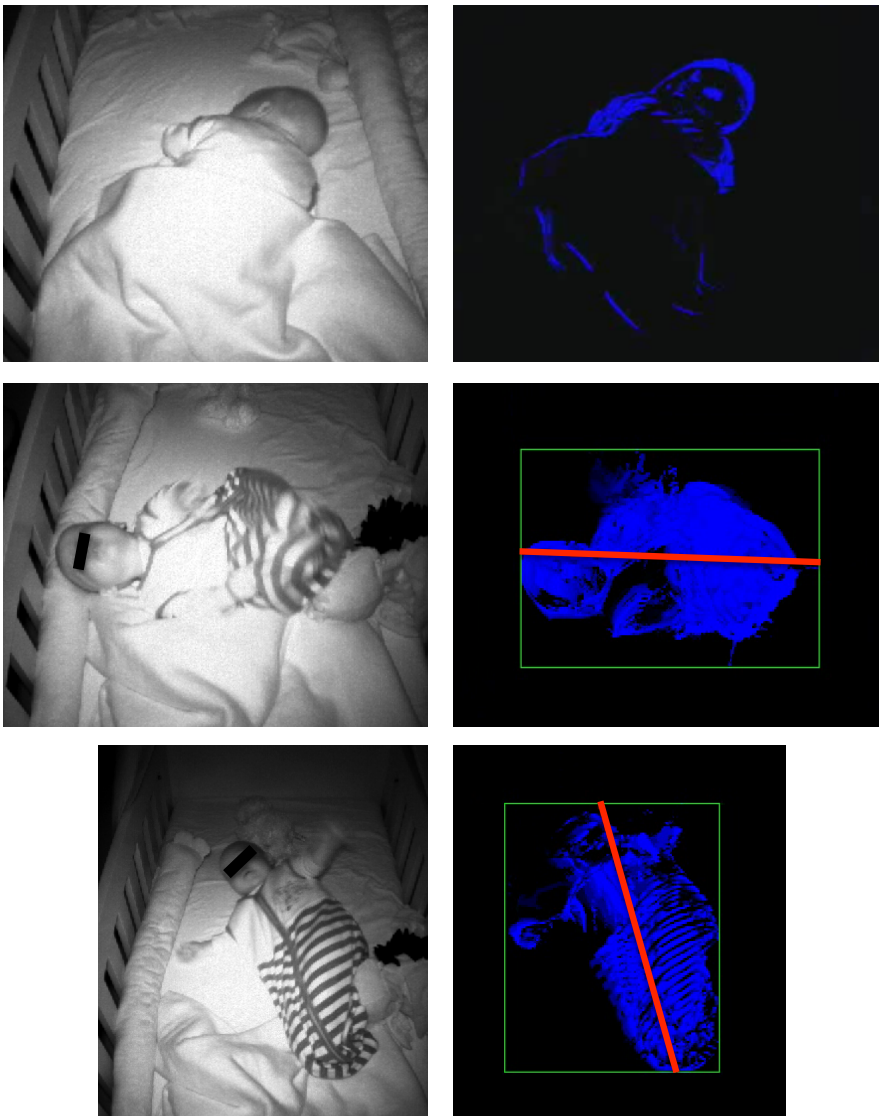


Figure 7.3: Detected motion pixels (blue) with frame differencing method (top row) and MHI (2nd and 3rd row). The red bold line marks the calculated body axis.

- involvement of the upper body in the movement. The upper body is set to the upper 40% of the body area.

The motion vectors are computed for each 8×8 image block by employing the motion estimation method developed for sleep movement analysis in [338].

When the two conditions for a potential turning movement are satisfied, all the motion vectors in the image are projected to an orthogonal vector (the vector with the direction orthogonal to the body axis). When the infant rolls over, most vectors would have a large projection of vectors orthogonal to the body axis (from hereon called orthogonal vector). The summation of the projections on the orthogonal vector is used as a measure for the extent that the body moves along this orthogonal direction. This is illustrated in Fig. 7.4 where the yellow arrows are the computed motion vectors, the green arrow is the orthogonal vector, the thin red arrows are the orthogonal projections and the bold red arrow represents the summation of projections on the orthogonal vector.

The low-pass filtered (with a 5-tap moving average filter with equal coefficient values of 1) accumulation T over time of the sum of these orthogonal projections is depicted in Fig. 7.5. Local minima and maxima represent supine and prone positions, respectively. Depending on the infant's initial pose, the positions associated with minima and maxima could just as well be reversed. Therefore, including the information of face detection is beneficial.

7.2.2 Face detection

Face detection is a popular technique in computer vision allowing fast and reliable detection of human faces in a natural environment. One requirement of the proposed application is that the face has to be detected in any possible orientation since babies move heavily during sleep. This complicates the task of performing robust face detection since most face detectors will allow the detection of a face only when it is more or less facing upward or when the baby is lying on its tummy and looking to the side. To fulfill our requirements, one could apply a multi-view face detection algorithm [345]. However the complexity of such an algorithm reduces significantly its attractiveness. For this reason, we propose to use a dedicated framework, based on Viola-Jones face detector in [346], allowing fast and efficient face detection regardless of the baby pose.

The proposed framework, illustrated in Fig. 7.6, is composed of the following main elements: an automatic region of interest (ROI) selection based on the computed body axis and body area, an image rotation module and face detector module. The basic flow is as follows: the proposed framework firstly focuses on a specific part of the image (ROI), then performs the face detection process for several image orientations, and stops when the face detection module produces satisfactory results or when all possible image orientations have been scanned. This reduction of the image to a relevant ROI allows the system to remain efficient by running several face detectors on small images.

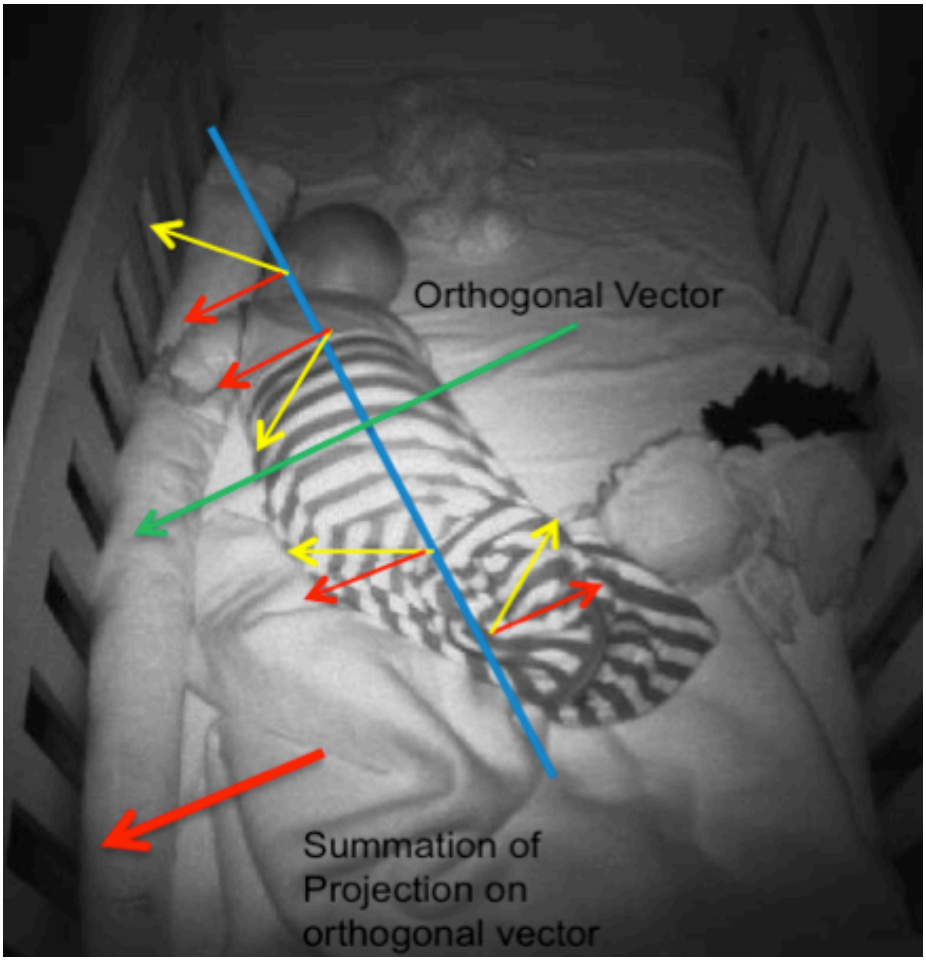


Figure 7.4: Sum (bold red arrow) of motion vector (yellow arrows) projections (thin red arrows) orthogonal (green arrow) to body axis (blue line).

ROI selection

The body area information from the MHI image is used to select automatically the relevant ROI. The body area information is divided into two parts: the upper body and lower body. Each part accounts respectively for 40% and 60% of the body area, see example in Fig. 7.7 top right.

The upper body region is selected as ROI. Doing so, the detection speed of the face detector is tremendously increased. Furthermore, by focusing only on the relevant ROI, the number of false detections significantly decreases.

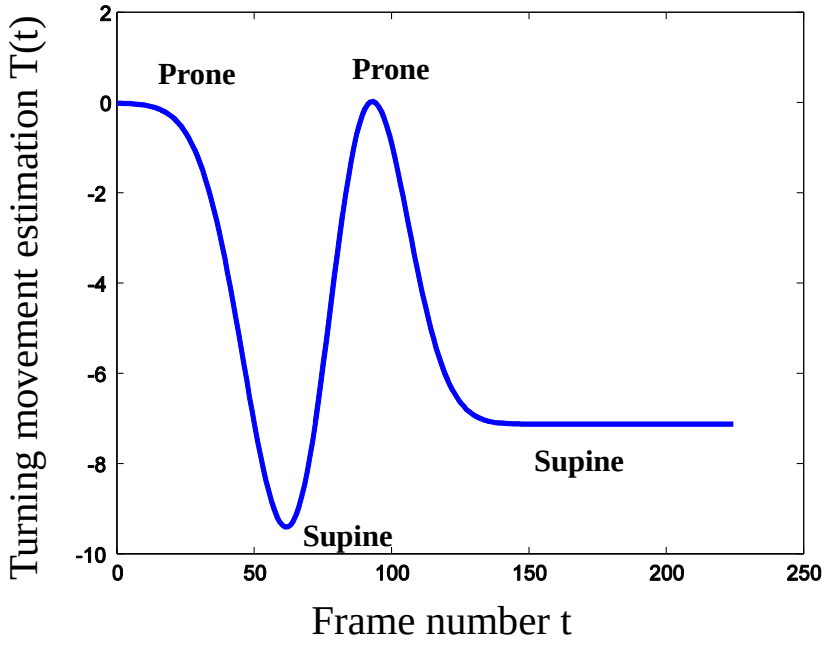


Figure 7.5: Low-pass filtered accumulation T over time of the sum of orthogonal projections.

Image rotation

The original face detector [346] ensures optimal performance for frontal face detection. To be able to detect faces in several orientations, our framework rotates the above defined ROI to ensure that the face is present in an upright position.

In an initial rotation the face is rotated into a favorable orientation by rotating the ROI according to the body axis as defined in Section 7.2.1. However, due to inherent inaccuracies of the body axis calculation the face might not be exactly in an upright position after this initial rotation. Therefore, extra rotations are needed to maximize the detection performance. These rotations are defined over a $\pm 15°$range, centered on the body axis angle, with a step of $3°$. These values have been found empirically and showed the best performance in our framework.

Figure 7.7 shows an example input image with the baby lying in an arbitrary position (top left), the body axis and relevant ROI based on the MHI calculation (top right), the ROI after the initial rotation (bottom left) and finally the optimized ROI after the rotation iterations.

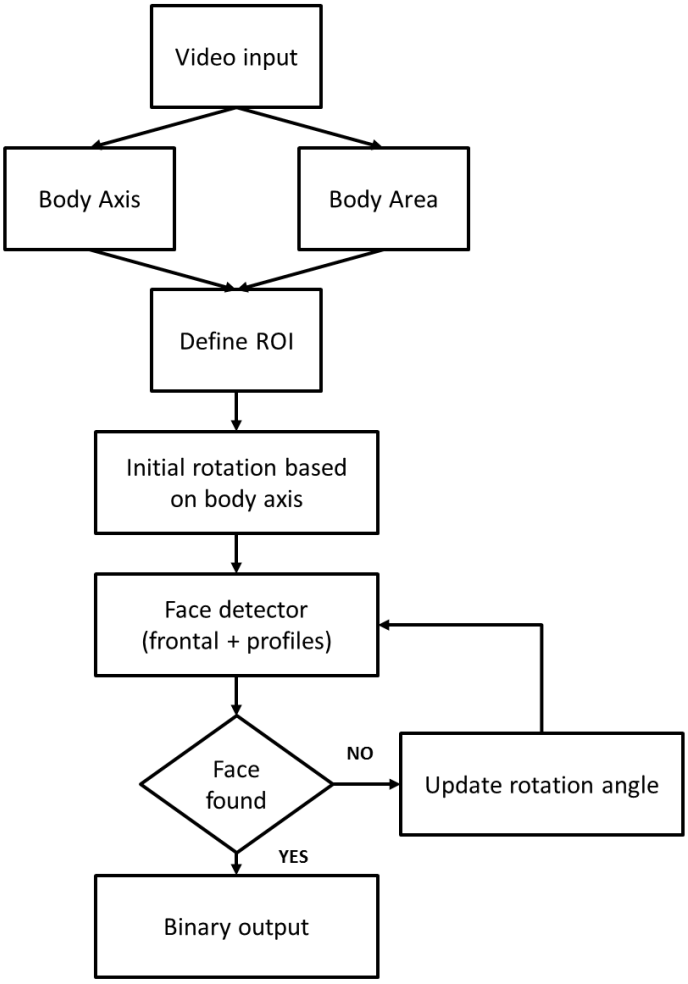


Figure 7.6: Face detection framework.

Face detection

The proposed framework uses the OpenCV [347] implementation of the Viola-Jones face detector. This choice is first based on the availability of this open source library and also on the very fast and efficient implementation of the Viola-Jones method.

To provide more robustness regarding face pose, we use two Viola-Jones cascades, one to cope with frontal faces and another one to cope with left profiles. In order to detect faces showing a right profile, the ROI is mirrored by an image flip operation.

This step provides a binary output per incoming image (0 being the absence of a face in the ROI). To remove jitter due to false alarm or no detection, a smoothing function similar to the one introduced in Section 7.2.1 is applied to this output. This

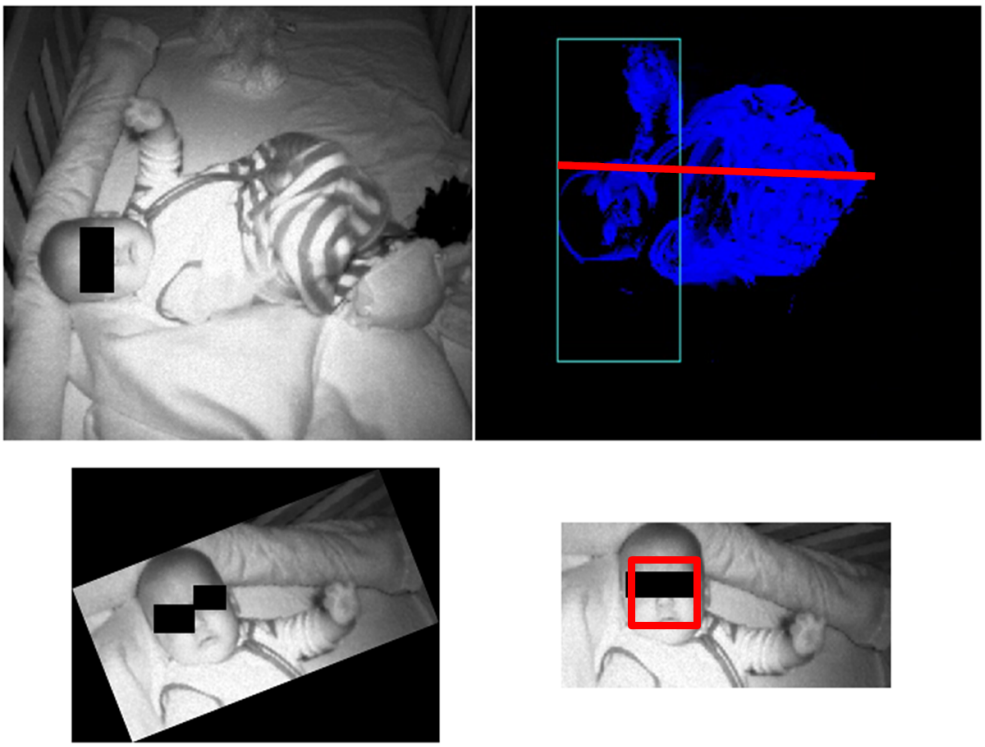


Figure 7.7: Subsequent steps in the face detection process.

smoothed output is referred to as the face function *F(t)*.

Fig. 7.8 shows the output result of the face detection process on a test sequence containing 450 frames. The reference ground truth data is shown in red (based on manual annotation) versus the output produced by our framework (green). Higher values indicate the presence of a face. The main observation from Fig. 7.8 is that the proposed face detection mechanism is not robust in all situations. We noticed that different kinds of artifacts affect the performance. These artifacts originate mostly from objects surrounding the baby (e.g., face of a stuffed animal or doll). This shows that face detection, as proposed, cannot be used on its own to provide accurate measures on the baby pose. In Section 7.2.3 we investigate the combination of this output with the turning movement estimation.

7.2.3 Pose estimation

From the time a face is reliably detected, the supine position is used as a starting point and the following turning movement events are used to estimate the lying pose of the infant (i.e., after the first turning event the prone position is returned, thereafter the supine position, etc.). The face detection output is only used in the following pose

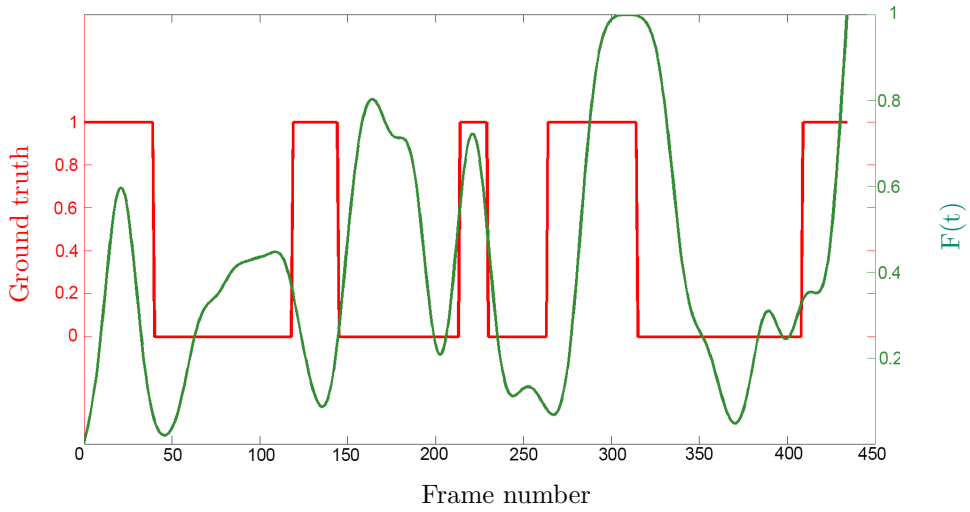


Figure 7.8: Face detection results on a test sequence. The reference ground truth data is shown in red (based on manual annotation) versus the output produced by our framework (green). Higher values indicate the presence of a face.

estimation periods when it is reliable enough as the Viola-Jones face detector is rather unstable as mentioned in Section 7.2.2. A reliable detection of a face is obtained when the face function $F(t)$ exceeds the threshold $d_f = 0.9$. The pose estimation function $P(t)$ for frame t is calculated as

$$P(t) = \begin{cases} 1, & \text{if } F(t) > d_f \\ |P(t-1) - 1|, & \text{if local maximum/minimum} \\ & \text{detected in } T(t) \end{cases} \tag{7.1}$$

where 1 indicates a supine position and 0 a prone lying pose.

7.2.4 Evaluation methods

A quantitative evaluation regarding turning movement detection and pose estimation is formulated as follows. The detection rate of total turning movements d_T is computed as

$$d_T = \frac{n_T}{n_G}, \tag{7.2}$$

where n_T corresponds to the number of detected turning movements and n_G to the number of annotated turning movements (ground truth). The accuracy rate of the pose estimation method a_P is calculated by

	Seq. 1	Seq. 2	Seq. 3	Seq. 4	Avg., StD
d_T	100%	100%	100%	100%	100% ±0
a_P	100%	53.5%	100%	74.1%	81.9% ±22.5
a_{P_F}	87.9%	67.4%	53.9%	73.7%	70.7% ±14.1

Table 7.1: Turning detection rate d_T and pose estimation accuracy rate a_P of the proposed method and a_{P_F} of a face detection method suggested in [344] on four video clips. The proposed pose estimation method combines turning movement detection and face detection results.

$$a_P = \sum_{t=s_t}^{n_t} \frac{|p_d(t) - p_g(t)|}{n_t - s_t + 1}, \tag{7.3}$$

where t denotes the frame number, s_t the starting frame of reliable face detection, n_t the total number of frames, p_d the detected pose and p_g the annotated ground truth pose.

7.2.5 User study

Six 1 to 2 hour recordings (approved by the internal ethics committee of Philips Research) were made of 4 babies aged 3 months to 1 year. All the videos were recorded by the infants' parents at home using a near infrared camera (monochrome uEye UI-1220SE from IDS Imaging) and light source. The camera was positioned at the foot end of the bed and videos at 5 frames per second and a resolution of 600×400 pixels were shot. In three recordings, the infants did not change their sleeping pose. In the end, only two different recording days with the same infant could be used. Segments from these recordings were used as training data for developing the algorithm. Four different turning movement clips were cropped out of the two remaining recordings and were used to evaluate the proposed pose estimation algorithm.

7.2.6 Results

Quantitative results regarding turning movement detection and pose estimation are given in Table 7.1. The proposed pose estimation method combines turning movement detection and face detection results. A competing method based only on face detection is described in [344] and is denoted as a_{P_F} in Table 7.1.

From Table 7.1, we can see that all turning movements are correctly detected (when the time delay is neglected). The average accuracy rate of the proposed pose estimation method amounts to 82%. It is high in sequences 1 and 3 (100%), but drops in sequences 2 (54%) and 4 (74%). When comparing with the face detection solution proposed in [344], an average performance improvement of 11% is observed with our proposed method.

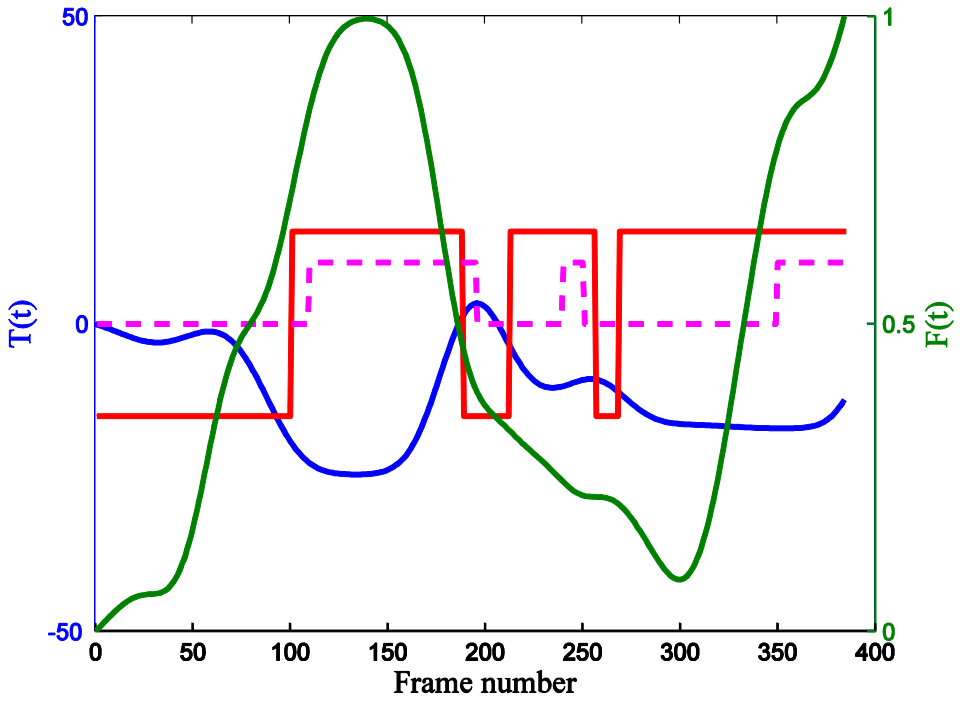


Figure 7.9: Pose estimation output (magenta) with higher value indicating supine, lower value prone position. Turning movement estimation T(t) (blue), face detection function F(t) (green). The red line represents the ground truth (higher value indicates supine, lower value prone position).

Fig. 7.9 shows a pose estimation output of sequence 2 along with the turning movement estimation, the face detection function and the ground truth signal. From the time a face is reliably detected (around frame 112), the turning movement estimation shows correct turning events. A delay can occur as is the case around frame 230 and frame 350. This results in a delay in the pose estimation function. Currently, a turning event is detected when a local maximum or minimum is computed. This should be replaced with a more sophisticated turning detection function which would highly increase the accuracy rate of the pose estimation method. A similar behavior is observed in sequence 4. Moreover, future work should investigate a more robust face detection method sensitive to baby faces with a high specificity (i.e., insensitive to face-like features in stuffed animals and dolls). This would be of advantage for decreasing the unstable phase at the beginning of the video clip and increase robustness in the pose estimation function thereafter.

7.3 Intelligent wake-up light

The second application in this paper also makes use of the body movement information of a sleeping subject. An important factor influencing the subjective sleep quality is the wake-up experience. Sleep inertia is a common phenomenon when waking up. It can be characterized by confusion, disorientation, sleepiness and grogginess, with reduced cognitive and physical performances [348]. The use of a so-called artificial dawn wake-up light (e.g., Philips HF3490, [349]) that aims at waking up a person in a gradual manner has been found to result in a better wake-up experience [348, 350–352] with reduced sleep inertia. In these studies, existing wake-up light products were used that start shining around 30 minutes before the set wake-up time and continuously increase their brightness. Subjective sleepiness decreased while subjective activation and alertness increased. Benefits have been found in terms of wake-up quality, easy rising, energetic feeling, the mood after waking up, social interactions, concentration and productivity.

An individual passes through sleep cycles which consist of five sleep stages. One of these sleep stages is referred to as Rapid Eye Movement (REM). If a person is woken up after the REM sleep stage, he/she feels more conscious and alert than when woken up from a deeper sleep stage [353, 354]. From sleep studies, an increased rate of body movements is observed in the lighter sleep phases and REM sleep [327]. These insights are used in [355] where an intelligent alarm clock wakes up the sleeping subject when he/she transitions through the light sleep phase within 30 minutes before the set wake-up time. When the subject does not pass through the light sleep phase in the mentioned time frame, the intelligent alarm clock behaves as any traditional alarm clock. In that case, the subject is still woken up in an abrupt manner by the alarm clock. A similar approach is implemented in the 'iwaku' product [356]. As a user-friendly enhancement of current propositions and products [349, 355, 356] where the subject still gets most of his/her sleeping time, we propose to guide the sleeping subject into a light sleep phase starting 30 minutes before the set wake-up time and let the subject wake up only at the preferred wake-up time. This is achieved by adapting the exposed light intensity level according to the sleeping subject's unconscious reaction to it. We have not come across any similar propositions in literature or product implementations.

Wake-up light products follow a fixed luminance curve in the 30 minutes prior to the set wake-up time, which means the lamp is not responsive to the person's reaction. If the sleeping subject is very sensitive to light, he/she would be woken up by the light earlier than his/her planned wake-up time, and this situation is not desired. By monitoring the sleeping person's activity level, a system is proposed and designed to adapt the lamp's luminance value to it. Furthermore, we suggest adding more colors to the wake-up light, for the reason that certain colors (e.g., blue) have more potential to make people feel alert [357].

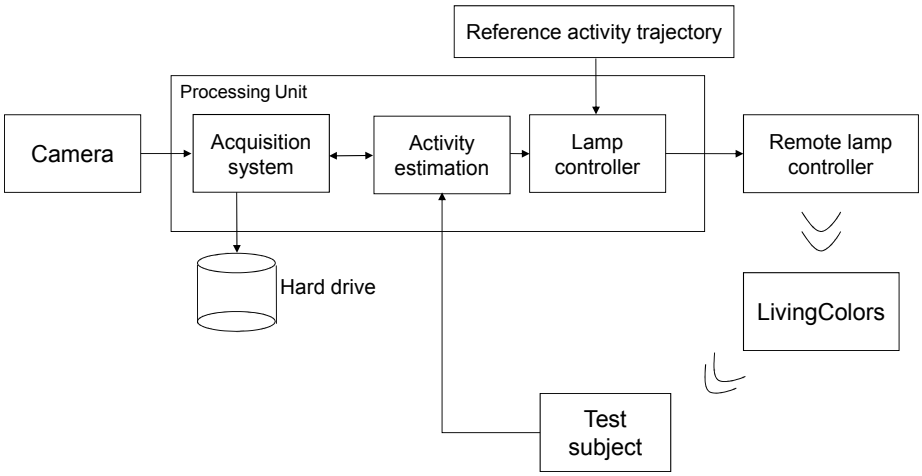


Figure 7.10: System architecture of the proposed wake-up light system.

7.3.1 Proposed wake-up light system

We propose a personalized wake-up system that exposes the sleeping subject to a colored light that is increasing in intensity over time according to the subject's measured activity level. Therefore, a system which can detect the sleeping person's activity, process this data, and control the light needs to be designed.

The proposed intelligent wake-up light system is shown in Fig. 7.10. It is composed of three parts: a camera, a processing unit and remote LivingColors lamps (including their controller). Inside the processing unit, there are three blocks. The acquisition system block communicates with the camera. It reads in the picture data at 10 frames per second from the infrared camera UI-1220SE-M, sends this picture data to the activity estimation block when requested, and at the same time, the video capturing block stores all the picture data on the hard disk for further offline analysis. The activity estimation block reads in picture data from the video capturing block, does the activity estimation process based on the method proposed in [338] and then sends this outcome to the lamp controller. For the reason that the employed motion estimation version usually takes more time than the interval between two frames, a two way communication is set up between these two blocks. Only after the activity estimation block has finished its processing work will it request for the next frame. The lamp controller is implemented based on a proportional-integral-derivative (PID) controller [358] discussed in Section 7.3.3. PID controllers are widely used in industrial control systems. A PID controller calculates an 'error' value as the difference between a measured process variable and a desired setpoint. The controller attempts to minimize the error by adjusting the process control inputs. In this project, the setpoint is not a

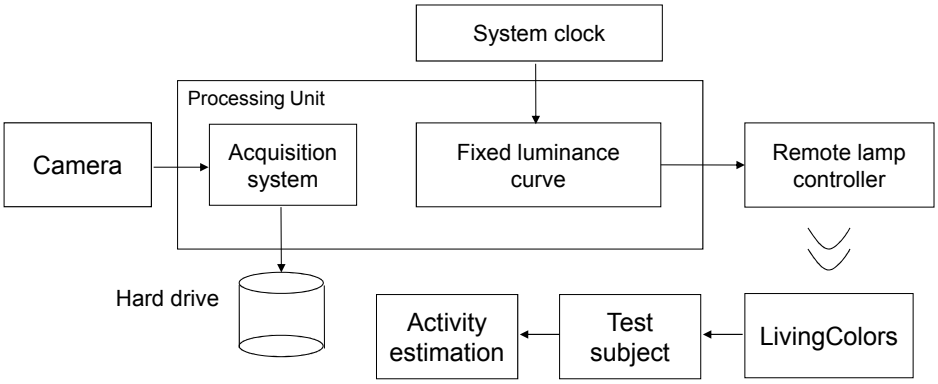


Figure 7.11: System architecture of the open loop perception test to determine reference activity trajectory.

fixed number but a desired wake-up activity trajectory or reference activity trajectory over time. A reference activity trajectory is presented in Section 7.3.2 based on a user study involving four participants. A more personalized reference could be computed in practice by having the same user evaluate his/her wake-up experience at more instances and during a longer period of time.

The experiments and choices made to come to an intelligent and personalized wake-up light illustrate one of many design possibilities of the wake-up light. This work is meant to show a creative and meaningful sleep enhancement solution. The authors would like to stress the potential that comes with the activity level information of a sleeping subject.

In Section 7.3.2, a reference activity trajectory is computed which is used in the controller for regulating the light exposure towards the sleeping subject. The feedback control design is described in Section 7.3.3 and results are discussed in Section 7.3.4.

7.3.2 Reference activity trajectory

The aim of this section is to determine a wake-up activity trajectory that corresponds to a good wake-up experience. The system as shown in Fig. 7.11 is designed in order to find the reference activity curve for the PID controller to follow. Contrary to the system architecture shown in Fig. 7.10, we use this open loop system without a lamp controller block. The acquisition system records the sleeping person's video data over night from which the activity levels are derived (dependent variable). The 'Fixed luminance curve' program is triggered by the system clock when it reaches 30 minutes before the set wake-up time. It controls the remote lamps' controller and lets them follow the luminance curve of the wake-up light which is proven to improve the average wake-up experience [348].

Different light colors may have different impact on the user. Therefore, different colors were included in the user study, however, not with the goal to find one color working best for all users. A good wake-up experience can be evoked with one color for user A, whereas another color works best for user B. We increased the chance to provide a good wake-up experience for each user by exposing the same user to different colors.

Three important physiological factors and a cultural factor play important roles in the color perception process. These are

- the closed eyelids filtering the amount of light to pass dependent on the light's wavelength [359],
- the number of cone photoreceptors being activated depends on the exposed light spectrum [360],
- the melanopsin receptors influencing melatonin production dependent on the light's wavelength [361–363],
- the different culture conventions resulting in various preferences for colors and associations with feeling relaxed [364].

Closed eyelids act as an attenuator for blue and green light (estimated transmission is 0.3%). Red light is transmitted through the eyelids at a level of 5.6% [359]. Cone photoreceptors are photoreceptor cells in the retina of the eye with a varying sensitivity depending on the frequency of the incoming light. Human beings have three types of cone cells with maximal absorption efficiency at wavelengths of 420 nm (blue, S cones), 530 nm (green, M cones) and 560 nm (orange, L cones) [360]. Each of the three cone classes has an absorption ability across a range of wavelengths which may have considerable overlap. Additionally, the three cone classes are not equally represented. The L cones make up 63%, the M cones 31% and the S cones only 6%. Human beings are thus more responsive to green, yellow and red, and less sensitive to blue light. The third physiological factor influencing the color perception is given by the melanopsin receptors which mediate the production of melatonin. Melatonin is a hormone promoting sleepiness [361]. It has been proven in [362] that blue light is the most potent color for melatonin suppression. Not neglectable is the cultural impact on the color preference. Experiments have shown that people have different preferences for colors they like and colors making them feel relaxed [364]. From the above it becomes apparent that there is no clear winner for the wake-up color.

Experiment

We conducted a within-subjects design user study to determine the desired wake-up activity trajectory or reference activity trajectory. Four participants, 3 males and 1 female, not color blind nor suffering from any sleep disorders, were selected for this home study. Their age ranged from 24 to 29. As discussed in the previous paragraph,

it is not uncommon for people to have different associations with the same color. Since we aim for a good wake-up experience and since the color of the light is the first thing the subjects see when they wake up, we let the test subjects choose in advance the wake-up light colors (independent variables) that they perceived as alerting and soothing, respectively. Similarly to [348], every test person did the experiment for 3 or 4 mornings (Tuesday, Wednesday, Thursday and Friday, when people's circadian rhythms are more regular). One morning the subject was exposed to the soothing color, another morning to the alerting color, another morning to the blue color (peak wavelength at 462 nm) due to its melatonin suppression feature (if it has not yet been selected) and one morning no light was used. The order of preference was left up to the test subject. As a wake-up light in this experiment, two Philips LivingColors (model 6914360PH) lamps were used due to the possibility to set different colors and to control its luminance output from a laptop. The LivingColors lamps were put on each side of the bed in order to make sure the light can reach the subject's face. 30 minutes before the set wake-up time, the lamps would turn on and follow the intensity curve of a Philips Wake-up Light HF3490. Upon wake-up, a commonly used rating scale (Karolinska Sleepiness Scale (KSS) according to [365], dependent variable) had to be completed to self-assess the subjective level of sleepiness at 1 minute, 15, 30, 60, 90 minutes after waking up, respectively, evaluating the sleepiness level subjectively.

A wake-up experience is classified as 'good' when the test subject reports to feel alert (as a criterion grade 3 on the KSS scale was set, where 1 qualifies as extremely alert and 9 as very sleepy) 30 minutes after waking up and does not decrease in alertness thereafter. Among the 15 recorded nights, five wake-up experiences satisfied these criteria. The low-pass filtered activity levels of a test subject's entire night are given in Fig. 7.12. An averaging filter with a large window size (6000, corresponding to approximately 10 minutes) has been applied in order to understand the low-frequent behavior. It would be rather difficult to generate a high-frequent behavior at the subject's side based on light exposure during wake-up. The test subject reacted quite soon after the lamps turned on at 6.45 a.m. and even more at 6.56 a.m. Approximately 10 minutes later, the activity level drops clearly. This behavior is a typical example of other test subjects in our experiment and is similarly observed in [348] according to which people tend to fall asleep again 10 minutes after the artificial light has turned on. From the corresponding activity curves, an average reference activity trajectory (see left most column in Fig. 7.16) was computed by applying a mean filter with a window size of 10 minutes.

The subjects showed a clear preference for selecting the red light as wake-up color, sometimes as soothing, sometimes as alerting color. Nevertheless, there was no dominating color in the pool of 'good' wake-up experiences.

From the obtained results, we conclude that the wake-up light controller should regulate the light output such that an initial activity phase is stimulated in the sleeping subject, followed by a 'quiet' phase before increasing the light output to a level that produces rather high activity levels for the definite wake-up.

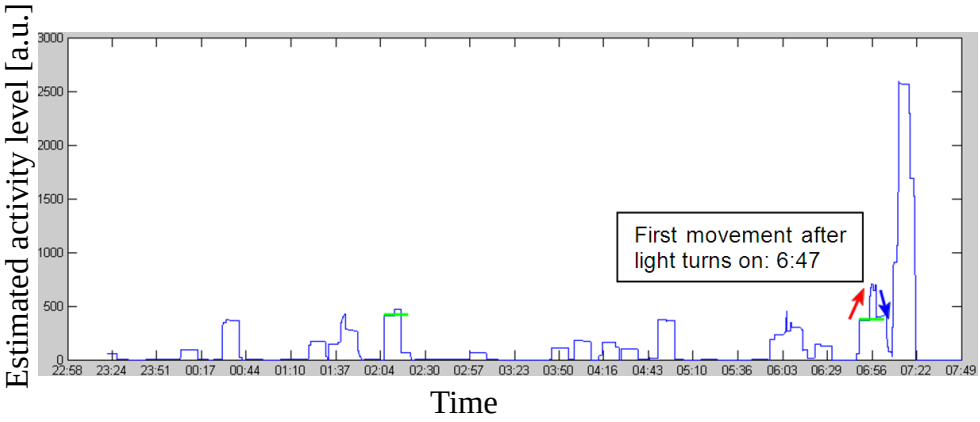


Figure 7.12: Low-pass filtered activity level of a test subject's entire night. An averaging filter with window size of approximately 10 minutes has been applied in order to understand the low-frequent behavior. The wake-up lamps are turned on at 6.45 a.m.

7.3.3 Feedback control design

The same gradual changes in light intensity are not perceived in the same way for low light intensities as for high light intensities. Thus, different reactions are expected when a subject is exposed to a light increase in low light intensities than to the same increase in high light intensities. It is crucial for the controller to have a model of the perceived light so that the lamps are steered with the appropriate intensities. Therefore, a perception test with five test subjects has been carried out where the perceived strength is measured of red light (due the strong preference of red light among the test subjects in the reference activity trajectory user study) gradually intensifying in a dark room. The eyes were closed while the light intensity was gradually increased and the test subjects graded the perceived strength of the light from 1 to 9. The resulting perception model shows that increases up to one third of the total intensity is felt strongly. Later differences are not perceived that strongly. This is in accordance with the response function of the human visual system [366]. For the controller, a reaction model in Fig. 7.13 is deduced from the perceived light strength. It consists of two linear functions where the first linear graph ends at 67% chance of reaction at one third of the total luminance, and the second linear graph with a less steep slope continues up to 100% chance of reaction at full light luminance.

The lamp controller in Fig. 7.10 is implemented as a PID controller where the differential gain is neglected as its primary purpose is to take influence on short-term changes (we do not expect the light to have immediate effect on the user). The PI controller computes the 'error' between the reference activity trajectory and the measured activity level which is given as feedback signal to the controller. The controller then attempts to decrease this error by adjusting the light intensity output.

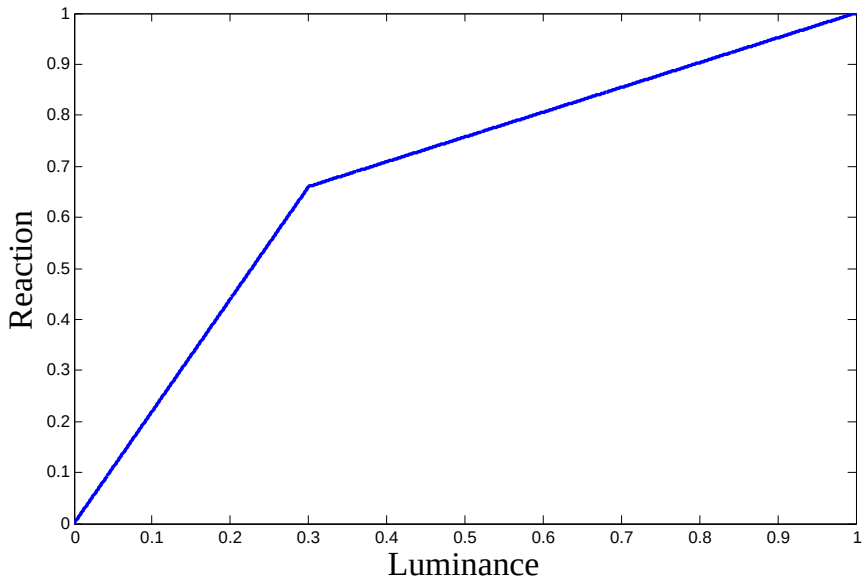


Figure 7.13: Expected reaction model to increasing luminance where 1 indicates the strongest reaction.

The light intensity output is determined based on the current error (P) and the accumulation of past errors (I). The designed control loop model is given in Fig. 7.14 and in the first three blocks in Fig. 7.15. The first non-linear gain block in Fig. 7.15 determines the luminance increase/decrease based on the input error, which is added in the integral block 1/s to the current luminance. The saturation block ensures a luminance output between 0 and 1 (maximum normalized luminance for the lamp). These three blocks form the actuator. Originally, a stair case gain was envisioned instead of the first non-linear block. However, the staircase block is highly non-linear and can contribute much more to instability. Therefore, we approximated it with a non-linear gain where the normalized luminance gain is 0.002 if the error is smaller than 1 and 0.01 when it is larger than 1.

7.3.4 Results

A stability analysis of the control system and simulation results are given in this section. Therefore, a process model needs to be included in the control loop as is done with two rightmost blocks in Fig. 7.15. As a simplified model of a sleeping subject's reaction, the reaction probability model in Fig. 7.13 is used in the non-linear gain block following the delay block of 5 seconds. The delay is empirically determined and accounts for the delayed reaction time due to the sleeping state of

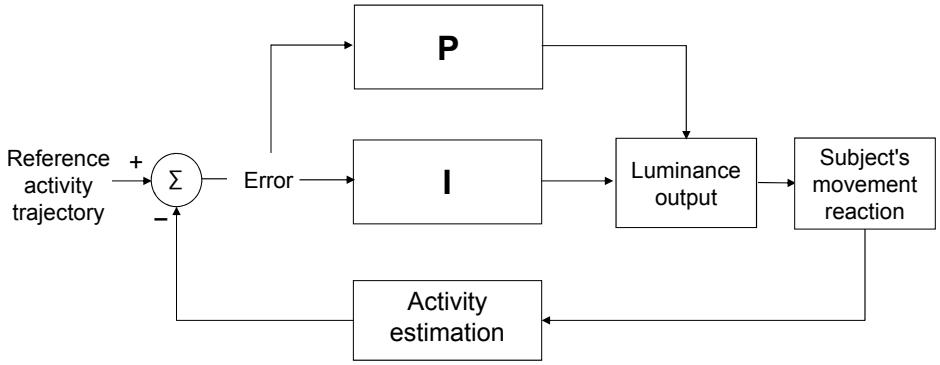


Figure 7.14: Control loop diagram.

the subjects.

The control loop model is not a linear time-invariant (LTI) system. For the stability analysis, the non-linear blocks are approximated with a linear gain. The behavior of the feedback system is determined by the closed loop transfer function. The transfer function $H(s)$ is expressed as the ratio of the Laplace transform of the output variable to the Laplace transform of the input variable of the system [358]. The transfer function of a system such as in Fig. 7.15 with an actuator A_c and a process model P_m is described as

$$H(s) = \frac{A_c(s) \cdot P_m(s)}{1 + A_c(s) \cdot P_m(s)}. \tag{7.4}$$

Incorporating the linear gain, an integral part and a delay block leads to

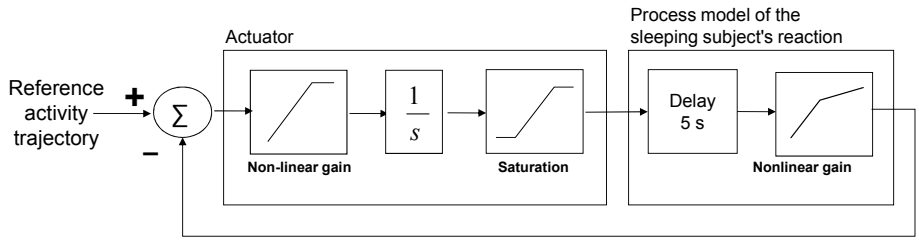


Figure 7.15: Control loop model where the first three blocks form the actuator and the last two the process model of the sleeping subject's reaction.

$$H(s) = \frac{0.01 \cdot \frac{1}{s} \cdot 1 \cdot e^{-5s} \cdot 2}{1 + 0.01 \cdot \frac{1}{s} \cdot 1 \cdot e^{-5s} \cdot 2} = \frac{A \cdot \frac{1}{s} \cdot e^{-5s}}{1 + A \cdot \frac{1}{s} \cdot e^{-5s}} = \frac{A \cdot e^{-5s}}{s + A \cdot e^{-5s}}, \tag{7.5}$$

with the linear gain $A = 0.02$ and a delay of 5 s. It consists of the highest gain in the first non-linear block in Fig. 7.15 (0.01), a gain of 1 from the saturation block and a maximum gain of 2 from the second non-linear gain block. There are two poles, both in the left half plane (-0.0224 and -0.7154) indicating a stable system.

Simulation results of the control loop model in Fig. 7.15 are shown in Fig. 7.16. Three conditions are simulated (one condition per row). The first condition assumes a sleeping subject to react as given by the process model. The second condition represents a subject not reacting to the light influence at all, and the third condition incorporates a person with an initially quick reaction. The columns show leftmost the reference activity trajectory or set points, in the middle the measured activity level of the sleeping subject and rightmost the luminance output of the lamps. The simulation results show a system behaving in the envisioned way for the three conditions. Contrary to the alternative attempt to use a staircase block instead of the first non-linear gain block, this system does not oscillate in the first condition (see Fig. 7.17).

The final wake-up light system shows promising initial results with a stable control system that exposes sleeping subjects to the wake-up light with different intensity progressions depending on their personal wake-up state. A real-time implementation of the final wake-up light system needs to be validated with test subjects to confirm its benefits. Based on self-rated experiences, new personalized reference activity trajectories can be computed building on multiple nights of the same subject.

7.4 Conclusions

In this paper, we proposed to use movement information of sleeping subjects for a different application focus than is commonly done. Two lifestyle applications making use of movement information recorded with a camera sensor and processed with two novel algorithms are presented. Design considerations as well as algorithmic details are presented.

The first application illustrates an intelligent baby monitor that informs the parents when their baby is turning in its sleep and when it is sleeping on its belly. Such an application enables real-time feedback to parents about the sleeping pose of their baby which can be critical regarding SIDS. Using advanced computer vision techniques and motion analysis we determine the infant's sleeping pose. The average accuracy rate of the proposed pose estimation method amounted to 82%, 11% better than a method based solely on face detection. Our future work will focus on reducing false detections mostly caused by delayed/instable face detections as the face detector

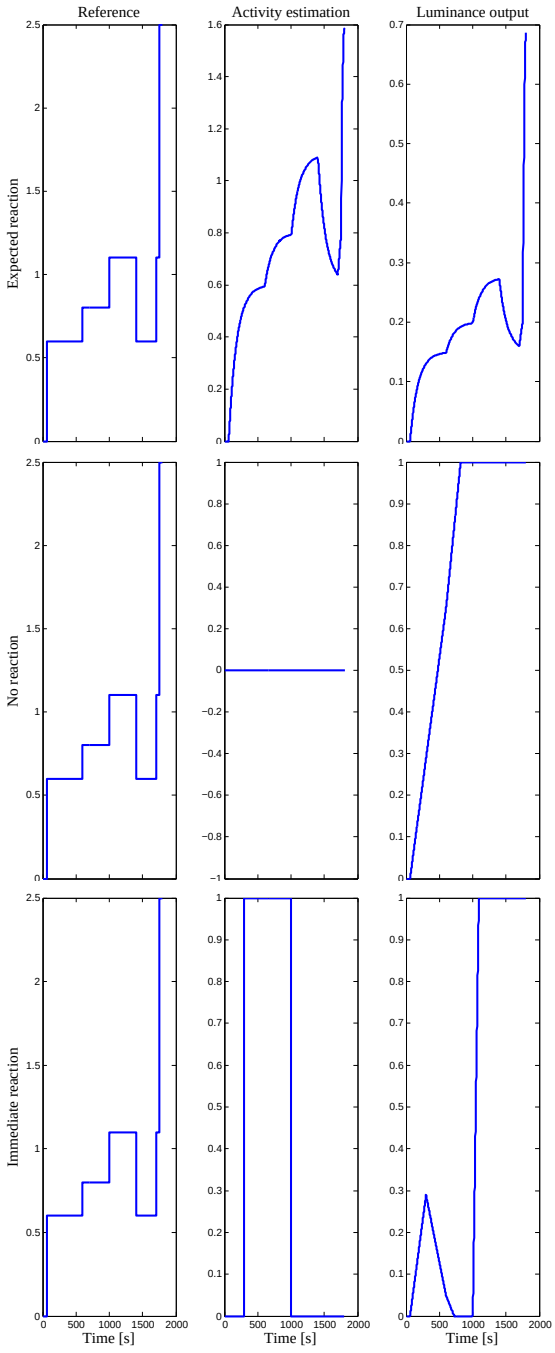


Figure 7.16: Rows: Simulation results of three test conditions. (1) Subject reacts as computed by the process model, (2) subject does not react, (3) subject reacts immediately. Columns: Set points (reference activity trajectory), measured activity level of the sleeping subject, luminance output of the wake-up light.

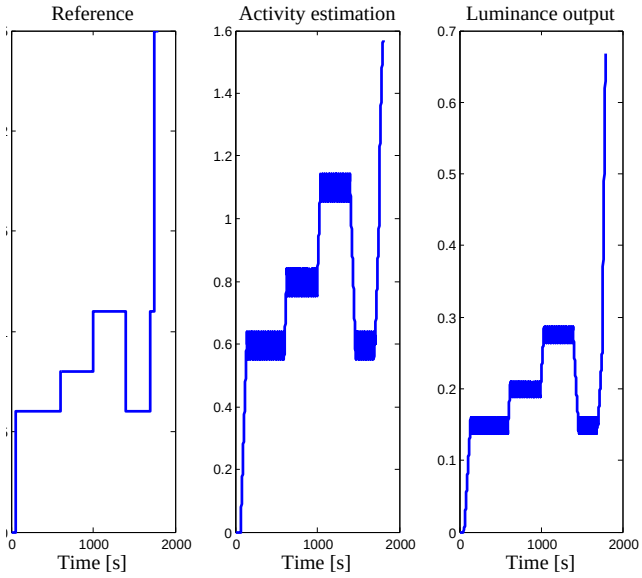


Figure 7.17: Simulation result of the alternative model with the staircase block and the subject reacting as predicted by the process model. Note the oscillating luminance output of the wake-up light.

reacted sensitive to objects with face-like properties such as stuffed animals and dolls in bed.

The second application describes an intelligent wake-up light adapting the lighting settings based on the sleeping subject's movement pattern. Our main contributions are the introduction of a reference activity trajectory and the design of a controller for a personalized wake-up experience. The proposed system shows promising initial results where sleeping subjects are exposed to different wake-up light intensity progressions depending on their personal wake-up state. In a next step, a real-time implementation of the system needs to be validated with test subjects to confirm its benefits. Based on self-rated experiences, new personalized reference activity trajectories can be computed building on multiple nights of the same subject.

Bibliography

[326] R. Gardner and W.I. Grossman, "Normal motor patterns in sleep in man", *Advances in Sleep Research*, **vol. 2** pp. 67–107, 1976.

[327] J. Wilde-Frenz and H. Schulz, "Rate and distribution of body movements during sleep in humans", *Percept. Mot. Skills*, **vol. 56**, no. 1 pp. 275–283, Feb. 1983.

[328] A. Muzet, P. Naitoh, R. E. Townsend, and L. C. Johnson, "Body movements during sleep as a predictor of state change", *Psychon. Sci.*, **vol. 29** pp. 7–10, 1972.

[329] S. Gori, G. Ficca, F. Giganti, *et al.*, "Body movements during night sleep in healthy elderly subjects and their relationships with sleep stages", *Brain Research Bulletin*, **vol. 63**, no. 5 pp. 393–397, June 2004.

[330] T. Etzioni and G. Pillar, "Movement disorders in sleep", *Harefuah*, **vol. 146**, no. 7 pp. 544–548, July 2007.

[331] S. Ancoli-Israel, R. Cole, C. Alessi, *et al.*, "The role of actigraphy in the study of sleep and circadian rhythms", *Sleep*, **vol. 26**, no. 3 pp. 342–392, 2003.

[332] R.P. Allen and W.A. Hening, "Actigraph Assessment of Periodic Leg Movements and Restless Legs Syndrome", in *Restless Legs Syndrome*, W.B. Saunders, Philadelphia, pp. 142 – 149, 2009.

[333] T. Harada, A. Sakata, T. Mori, and T. Sato, "Sensor pillow system: monitoring respiration and body movement in sleep", *Proceedings of 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems*, **vol. 1** pp. 351–356, 2000.

[334] X.L. Aubert and A. Brauers, "Estimation of Vital Signs in Bed from a Single Unobtrusive Mechanical Sensor: Algorithms and Real-life Evaluation", in *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBS 2008.*, Aug. 2008.

[335] W.-H. Liao and C.-M. Yang, "Video-based activity and movement pattern analysis in overnight sleep studies.", in *Int'l Conf. on Pattern Recognition*, 2008, pp. 1–4.

[336] K. Cuppens, L. Lagae, B. Ceulemans, *et al.*, "Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy.", *Med. Biol. Engineering and Computing*, **vol. 48**, no. 9 pp. 923–931, Sep. 2010.

[337] W.-H. Liao and J.H. Kuo, "Sleep monitoring system in real bedroom environment using texture-based background modeling approaches", *J. Ambient Intelligence and Humanized Computing*, **vol. 4**, no. 1 pp. 57–66, Feb. 2013.

[338] A. Heinrich, X. Aubert, and G. de Haan, "Body movement analysis during sleep based on video motion estimation", in *e-Health Networking Applications and Services (Healthcom), 2013. 15th IEEE International Conference on*, 2013.

[339] A. Heinrich, D. Geng, D. Znamenskiy, J.P. Vink, and G. de Haan, "Robust and sensitive video motion detection for sleep analysis", *IEEE Journal of Biomedical and Health Informatics (J-BHI)*, **vol. 18**, no. 3 pp. 790–798, May 2014.

[340] C.W. Wang, A. Ahmed, and A. Hunter, "Vision analysis in detecting abnormal breathing activity in application to diagnosis of obstructive sleep apnoea", in *Conf Proc IEEE Eng Med Biol Soc*, Sep. 2006, pp. 4469–4473.

[341] S. Parmet, A.E. Burke, and R.M. Golub, "Sudden infant death syndrome", *JAMA*, **vol. 307**, no. 16 p. 1766, 2012.

[342] C.-W. Wang, A. Hunter, N. Gravill, and S. Matusiewicz, "Real time pose recognition of covered human for diagnosis of sleep apnoea", *Computerized Medical Imaging and Graphics*, **vol. 34**, no. 6 pp. 523 – 533, 2010.

[343] A. Jeung, H. Mostafavi, M.L. Riaziat, *et al.*, "Method and system for monitoring breathing activity of a subject", *Patent US 7403638 B2*, July 2008.

[344] M.N. Mansor, M.N. Rejab, S.H.-F.S.A. Jamil, A.H.-F.S.A. Jamil, *et al.*, "Fast infant pain detection method", in *Computer and Communication Engineering (ICCCE), 2012 International Conference on*, July 2012, pp. 918–921.

[345] C. Huang, H. Ai, Y. Li, and S. Lao, "High-performance rotation invariant multiview face detection", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **vol. 29**, no. 4 pp. 671–686, 2007.

[346] P. Viola and M.J. Jones, "Robust real-time face detection", *International journal of computer vision*, **vol. 57**, no. 2 pp. 137–154, 2004.

[347] G. Bradski, "The OpenCV Library", *Dr. Dobb's Journal of Software Tools*, 2000.

[348] M. Van De Werken, M.C. Gimenez, B. De Vries, D.G. Beersma, *et al.*, "Effects of artificial dawn on sleep inertia, skin temperature, and the awakening cortisol response", *J Sleep Res*, **vol. 19**, no. 3 pp. 425–435, Sep. 2010.

[349] Philips, "Philips HF3490 Wake-up Light", http://www.philips.co.uk/c-p/HF3490_01, 2014, Accessed 20 March 2014.

[350] M.C. Gimenez, M. Hessels, M. van de Werken, B. de Vries, *et al.*, "Effects of artificial dawn on subjective ratings of sleep inertia and dim light melatonin onset", *Chronobiol. Int.*, **vol. 27**, no. 6 pp. 1219–1241, July 2010.

[351] L. Thorn, F. Hucklebridge, A. Esgate, P. Evans, and A. Clow, "The effect of dawn simulation on the cortisol response to awakening in healthy participants", *Psychoneuroendocrinology*, **vol. 29**, no. 7 pp. 925–930, Aug. 2004.

[352] E. Fromm, C. Horlebein, A. Meergans, M. Niesner, and C. Randler, "Evaluation of a dawn simulator in children and adolescents", *Biological Rhythm Research*, **vol. 42**, no. 5 pp. 417–425, Apr. 2011.

[353] M. Ferrara and L. De Gennaro, "The sleep inertia phenomenon during the sleep-wake transition: theoretical and operational issues", *Aviat Space Environ Med*, **vol. 71**, no. 8 pp. 843–848, Aug 2000.

[354] P. Tassi and A. Muzet, "Sleep inertia", *Sleep Med Rev*, **vol. 4**, no. 4 pp. 341–353, Aug. 2000.

[355] W.-H. Liao, J.-H. Kuo, C.-M. Yang, and I.Y. Chen, "iWakeUp: A video-based alarm clock for smart bedrooms", *Journal of the Chinese Institute of Engineers*, **vol. 33**, no. 5 pp. 661–668, 2010.

[356] iwaku, "iwaku wake-up light", http://www.iwaku.com/product/item13, 2014, Accessed 20 March 2014.

[357] S. Lehrl, K. Gerstmeyer, J. H. Jacob, H. Frieling, *et al.*, "Blue light improves cognitive performance", *J Neural Transm*, **vol. 114**, no. 4 pp. 457–460, 2007.

[358] G.F. Franklin, D.J. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 4th ed., 2001, ISBN 0130323934.

[359] K. Ando and D.F. Kripke, "Light attenuation by the human eyelid", *Biological Psychiatry*, **vol. 39**, no. 1 pp. 22 – 25, 1996, ISSN 0006-3223.

[360] G. Wyszecki and W.S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae (Wiley Series in Pure and Applied Optics)*, Wiley-Interscience, New York, USA, 2nd ed., Aug. 2000, ISBN 0471399183.

[361] A. Cagnacci, J.A. Elliott, and S.S. Yen, "Melatonin: a major regulator of the circadian rhythm of core temperature in humans", *J. Clin. Endocrinol. Metab.*, **vol. 75**, no. 2 pp. 447–452, Aug 1992.

[362] G.C. Brainard, J.P. Hanifin, J.M. Greeson, B. Byrne, *et al.*, "Action spectrum for melatonin regulation in humans: evidence for a novel circadian photoreceptor", *J. Neurosci.*, **vol. 21**, no. 16 pp. 6405–6412, Aug 2001.

[363] D.M. Berson, F.A. Dunn, and M. Takao, "Phototransduction by Retinal Ganglion Cells That Set the Circadian Clock", *Science*, **vol. 295**, no. 5557 pp. 1070–1073, 2002.

[364] L.-C. Ou, M.R. Luo, A. Woodcock, and A. Wright, "A study of colour emotion and colour preference. Part I: Colour emotions for single colours", *Color Research & Application*, **vol. 29**, no. 3 pp. 232–240, 2004.

[365] T. Åkerstedt and M. Gillberg, "Subjective and Objective Sleepiness in the Active Individual", *International Journal of Neuroscience*, **vol. 52** pp. 29–37, 1990.

[366] M.D. Fairchild, *Color Appearance Models*, John Wiley & Sons, Chichester, UK, 2005.

Chapter 8

Conclusion

8.1 Main conclusions

This thesis has addressed video motion analysis advances for the fields of TV picture enhancement and sleep analysis. The conclusions for each research question RQ are discussed in the following.

RQ1: *How can a large range of parameter values, parameter types, and interaction of different parameter choices be taken into account in the optimization of motion estimators for picture rate conversion?*

We developed a method to efficiently analyze and validate different parameter combinations in motion estimation methods. On the example of hierarchical 3-Dimensional Recursive Search (3DRS), we explored the extensive parameter space of 13000 motion estimators and provided insights with respect to the importance and the influence of the individual parameters. The variations ranged from parameters related to different scaling factors and block sizes, parameters influencing spatial smoothness and convergence speed of vector fields, different candidate structures, up to multi-scale vs. multi-grid motion estimation. We found that the motion estimators optimized with the proposed validation scheme are superior to multiple existing techniques as well as standard 3DRS with regard to performance at a low computational complexity. The proposed hierarchical motion estimators achieve a complexity reduction of 38% while outperforming 3DRS on average by 0.7 dB. This holds particularly for the unconverged state with an improvement of more than 1 dB and 7% in consistency. Even a sophisticated motion estimation method with 3-picture estimates is surpassed in the unconverged state (peak signal-to-noise ratio (PSNR) difference of 0.9 dB). However, in the steady state, this more complex motion estimator shows a clearly better performance than any of the hierarchical motion estimators. The benchmark further showed that the non-predictive search methods such as Full Search are generally unsuitable for picture rate conversion. As these methods purely optimize for minimal 'residue' in the match criterion, they produce highly inconsistent vector fields. The predictive search methods generally perform better, as they (implicitly) enforce vector field consistency, with the methods

Enhanced Predictive Zonal Search (EPZS) and Multiresolution-Spatio-Temporal Correlations (MRST) achieving the best steady-state PSNR performance (slightly below the proposed MEs). Among these, when taking the computational complexity into account, EPZS is identified as the ME achieving the best compromise between performance and complexity. Yet, its spatial inconsistency (SI) is more than 50% larger than the SI values of the proposed MEs, and this has a large impact on the perceived picture quality.

RQ2: *Can high quality robust motion estimators be identified while applying metrics with limited validity?*

Since no objective criterion has been accepted in the literature as a standard metric for evaluating ME methods, we developed a motion estimator design methodology that can operate with performance measures that model the human perception poorly. Still, a good correspondence with subjectively perceived picture quality is achieved. The conducted perception test with subjective scoring of motion compensated test sequences confirmed that the components of the proposed methodology are well chosen and yield motion estimators with a good picture rate conversion performance. We have analyzed sub-parts of the methodology and gained the following insights. Viewers rated motion estimators that were located within the computed attractive segment ('well performing' range) according to the proposed methodology significantly higher on quality than the motion estimators outside the attractive segment. Additionally, the results of the perception test demonstrated the mismatch between the performance measure scores and the perceived picture quality. No significant difference in perceived picture quality has been found between motion estimators on the contour line (the curve with optimal motion estimators according to two performance measures) and other motion estimators.

In the emerging application area of video-based sleep analysis, we have explored the feasibility of using a camera as a remote and reliable sensor. To analyze movements during sleep in different situations, we developed several algorithms.

Test data for evaluating the designed methods in this thesis are mostly limited to a handful of test subjects. Where possible, we have recorded new data (i.e., for movement monitoring compared with actigraphy, shared-bed home monitoring of movements and breathing, adaptive wake-up light, and infant pose detection). Besides the video signal, we measured the accepted reference signal(s) for sleep monitoring or made use of available databases containing both the video recording and the corresponding reference signal (i.e., recordings of periodic limb movement patients and night recordings for sleep efficiency computations with video and wrist actigraphy).

The employed evaluation methods are taken either from accepted evaluation methods in the literature in the sleep research field or from similar type problems described in other research areas (e.g., computer vision, machine learning). In some instances, we were forced to design our own evaluation method. When this was

the case, we included expert annotations to cross-check the quantitative objective validation.

RQ3: *How does video actigraphy compare to wrist actigraphy and is it more sensitive to movements originating from body parts other than the wrist?*

We designed a contactless, off-body video actigraphy system to monitor a sleeping subject's movements. With the aim to analyze competitiveness with wrist actigraphy, we conducted a differentiated comparison of video and wrist actigraphy on 188 movement events (zero, small, medium, large movements) taken from five nights of sleep (one night of each participant). The two different methods were compared in a Bland-Altman plot. Furthermore, the sleep efficiency measure computed with video, wrist actigraphy and PSG were compared with each other.

The Bland-Altman analysis demonstrated that different types of motion from different body parts were detected with video actigraphy and correspond to the reference wrist actigraphy results in 60% of the cases. The difference in the remaining movements is for a large part caused by an advantage in the video system, that can discern more movements than wrist actigraphy. Visual inspection shows that video actigraphy contains more comprehensive information and is generally more sensitive than wrist actigraphy. This is especially true for the 20% of body motions correctly detected by the video actigraphy system and missed by the reference wrist sensor. Five participants installed the video monitoring system individually at home. On the sleep dataset of these five participants we found that the video actigraphy method is sensitive to movements originating from under the blanket, is robust to various sleeping positions, different illumination conditions, viewing angles, and types of beds and blankets.

Concerning the estimation of sleep efficiency from activity patterns, the average PSG to video-based sleep efficiency error amounts to 5.8% vs. 4.5% with wrist actigraphy. These results can be regarded as quite positive, given that the sleep-wake classification algorithm has been optimized for wrist actigraphy. In [367], a video-based sleep monitoring technique is proposed based on frame differencing and compared with both PSG and wrist actigraphy. Liao and Yang report an 8% performance difference between video and PSG and a 1% performance difference between wrist actigraphy and video actigraphy. The achieved performance scores with our proposed video actigraphy system are in the same range (6% error between PSG and video, 1% error difference between video and wrist actigraphy). Besides contributing with a competing video system to state-of-the-art video systems for sleep analysis, our framework provides the opportunity to reuse motion estimation results for other sleep analysis solutions. Such potential applications were discussed in this thesis, e.g., body part movement analysis, periodic limb movement detection, segmentation of person of interest in a shared-bed environment, infant pose and turning movement estimation.

RQ4: *What are meaningful descriptors for motion vector clustering given the large variety of motion vector field properties produced by the varying recording conditions in sleep monitoring?*

In order to discriminate movements of different body parts, we developed an enhanced K-Means clustering approach for motion vectors. Our multi-distance clustering algorithm is not only based on spatial distances between data points but also on motion vector angle and length. When performing ME on sleep sequences, large environmental variations between recording situations such as viewing angles, blanket types, zoom factors and illumination conditions, can yield different motion vector fields for similar movements. Therefore, for each video, an automatic content-dependent optimization of the dissimilarity measure (based on spatial distances, and motion vector angle and length) was applied to optimally balance between the different descriptors. We were able to produce several dissimilarity measures with promising first clustering results where motion vectors of one body part movement are assigned to the same cluster. We have performed the evaluation according to three aspects, namely the clustering ability (inspired by [368]) among different dissimilarity measures, and the quantitative and qualitative agreement with expert cluster annotations. The clustering ability counts motion vector pairs not belonging to the same cluster. With a satisfactory clustering algorithm, a high occurrence rate is expected for large dissimilarities and a low occurrence rate for small dissimilarities (as these would ideally occur in the same cluster). As the type of quantitative evaluation is new to this field, we stress the necessity of qualitative expert ground truth annotations. Although still at an early stage, this adaptive clustering method can enable multiple applications in sleep research (e.g., body part segmentation, person segmentation, sleep disorder analysis).

RQ5: *Is a user-convenient camera placement on the bedside table acceptable for measuring actigraphy and breathing of a sleeping subject in a shared bed?*

To realize an easy-to-install system for the end user, we investigated an installation where the camera is not mounted high up on the wall or ceiling overlooking the bed. Contrary to other existing methods, the camera is more conveniently placed on the bedside table of the primary subject who is to be monitored. It can be a standalone device or integrated in existing bedside table products, such as a wake-up light. The developed shared bed video system has an average accuracy of 88% at a precision of 79% for segmentation of the primary subject. It can detect movements with an accuracy of 85% while outperforming wrist actigraphy at 75% accuracy. Hence, even in the presence of a bed partner and with a more challenging camera viewing angle, we were able to reconfirm the findings of the earlier research question RQ3 that video actigraphy accomplishes comparable results to wrist actigraphy. We designed a method sensitive to small scale movements so that not only activity levels are monitored but also the respiratory waveform can be computed. Since also tidal

volume [369] and respiratory waveform regularity [370] are important characteristics for sleep, we included the measures sensitivity and positive predictive value for the peak locations besides the mean breathing rate correspondence. Therefore, the number and location of the inhalation peaks in the video and reference breathing waveforms were compared. Our breathing analysis method performed admirably with an overall sensitivity of 87%, precision of 90%, and a breathing rate correspondence of 93%, surpassing the results of state-of-the-art video based breathing algorithms. Given the variation in the reference signals themselves (breathing rate correspondence of 98%), the results of the proposed video method are promising.

RQ6: *How can subject movement be distinguished from moving cast shadows for the case of periodic limb movements?*

To discriminate small movements of a subject from moving cast shadows on a non-planar and dynamic background, our video processing method integrates motion detection, motion estimation and texture analysis, efficiently aggregated in a strong classifier using cascaded AdaBoost. The designed motion detector can deal with fewer scene assumptions and limitations compared to other shadow detection methods while using standard hardware. Our proposed approach incorporates a motion detection cascade which is inspired by moving object tracking and the resemblance between background and shadow areas.

By enhancing an existing motion detection method fewer false positives are detected in the first stage of the cascade. The output of this first stage is given as input to the second stage which eliminates misdetections in image areas where texture remains substantially stable over time. However, the false positives along the sharp edges due to the folds in the background are barely reduced. When the shadow moves to another image area and discloses texture which has previously been occluded due to the non-planar background, there may be sudden texture changes. Therefore, a motion compensated texture model is introduced which eliminates false detections where the texture changes over time in an unpredictable manner. The method is based on the assumption that when subject movement occurs, a motion vector exists which renders a small difference between an image patch and the corresponding motion compensated image patch in the previous frame.

The subject motion detection system has been evaluated in a scenario combining illumination changes and subject motions for the purpose of periodic limb movement (PLM) detection in sleep (PLMS). The motion detection performance per frame of the proposed system provides a Matthews Correlation Coefficient improvement of a factor 2, an increase of 45% in sensitivity and 15% in specificity, on six test sequences compared to a state-of-the-art method for shadow robust motion detection [371]. Movement event classification for periodic limb movement detection improved with the proposed method by a factor of 6 and by a factor of 3 in constant and highly varying lighting conditions, respectively, compared to state-of-the-art. A 100% accurate PLM index score was obtained on all eleven PLM patient test sequences.

Although lighting variations were limited in the PLM patient test sequences, we infer from these results that video-based PLMS is feasible and can provide a coarse indication of the PLMI score.

RQ7: *How can we detect when a baby is turning from the supine to the prone sleeping pose?*

One of the investigated lifestyle applications is an intelligent baby monitor that warns parents when their baby is turning in its sleep to its belly. Such an application enables real-time feedback to parents about the sleeping pose of their baby which can be critical regarding sudden infant death syndrome (SIDS). By designing a turning movement detector and combining its information with face detection, we determined the infant's sleeping pose. Motion estimation has been useful in detecting a turning movement. Motion vectors between two consecutive images yield movements in the orthogonal direction of the body axis. Summation of the projections on the orthogonal vector was used as a measure for the extent that the body moves along this orthogonal direction. We combined this information with a Viola-Jones face detector which is designed to work fast and efficiently on high quality images. By automatically reducing the video images to a relevant region of interest, we were able to design a working system by running the face detection algorithm designed for high quality images on multiple rotations of our small, low quality images of the region of interest.

We computed the accuracy rate per frame with the annotated ground truth in four infant test sequences recorded at home. Additionally, we benchmarked our developed method with a face detection solution to understand if robustness is improved by combining turning movement analysis with face detection. The average accuracy rate of the proposed pose estimation method amounted to 82%, 11% better than a method based solely on face detection. In particular, this argues for using information from both motion and face for infant pose detection. Issues were observed with an at times instable face detection output (faces of dolls and stuffed toys were falsely detected as infant faces).

RQ8: *How can we create a good wake-up experience with light?*

A personalized wake-up system is envisioned that exposes the sleeping subject to light that is increasing in intensity over time according to the subject's measured activity level. Therefore, we developed a system which can detect the sleeping person's activity and controls the light output such that the subject's behavior corresponds to an activity trajectory of a favorable wake-up experience. Wake-up experiences were classified as favorable based on high subjective scoring of multiple wake-up experiences in an open-loop wake-up light system. Based on the favorable wake-up experiences, we computed a reference activity trajectory for the wake-up light controller to follow. We found that it corresponds well with reactions during a good wake-up experience found by another independent study [372]. The proportional-integral (PI) [373] lamp

controller computes the 'error' between the reference activity trajectory and the measured activity level which is given as feedback signal to the controller. The controller then attempts to decrease this error by adjusting the light intensity output. The light intensity output is determined based on the current error (P) and the accumulation of past errors (I). In the control loop model we approximated the actuator and a process model of the sleeping subject's reaction. We managed to design a stable control system with all the poles in the left half plane according to the stability criteria for linear time-invariant systems [373]. Also simulations were conducted to validate the choice of the control system elements.

8.2 Future work

In the following sections, we discuss future work for both TV picture enhancement and sleep analysis.

TV picture enhancement

Besides hierarchical 3DRS, alternative motion estimation method combinations can be considered. We expect promising results from blending the sophisticated motion estimator with 3-picture estimates [374] with hierarchical 3DRS or optical flow motion estimation with hierarchical 3DRS. Performance-critical optical flow parameters that are not too costly could yield a low-cost motion estimation method with superior performance compared to current motion estimators.

In the methodology for robust motion estimation design, more weight is assigned to the PSNR measure than to the inconsistency measure. A higher impact of the PSNR measure compared to the inconsistency measure is also observed in the conducted user study. Regression models for the quality as a function of 1/PSNR and Inconsistency SI were estimated. These models explained 59%-68% of the variance in quality. For MEs with lower PSNR (sequences A and C in Chapter 3), the PSNR measure plays the only role in assessing the ME quality. Users may not see the difference in inconsistency when the PSNR is too low. For sequence B with the highest PSNR MEs, a slight influence of the spatial inconsistency (SI) measure became visible (6% improvement compared to PSNR only at 60% explained variance). A clear degradation in performance is observed for motion estimators with large SI values. These should be discarded and therefore, the vertical cut-off line limiting the attractive segment should be well chosen. Still, the regression models explained only up to 68% of the variance in quality. Other performance measures should be investigated in future work to increase the explained variance in the ME quality.

Sleep analysis

The feasibility studies conducted for sleep analysis would need to be followed up

with larger user validation studies. This would serve the verification of the proposed methods and insight would be gained into the acceptance level of the proposed system.

The employed sleep-wake classification algorithm has been optimized for wrist actigraphy. Future work can include the design of a dedicated video sleep-wake classifier and improve the activity-count method as the amplitude dynamics are clearly larger than for wrist actigraphy and the time resolution much higher than the usual minute or 30-second epochs.

Additional descriptors are worth examining for an enhanced multi-distance motion vector clustering algorithm. Promising descriptors are body part labels and a temporal descriptor where the cluster history is taken into account. Additionally, a comparison with a non-adaptive clustering method should be performed to verify the benefit of the designed content-dependent method.

Regarding the new lifestyle applications, future work for the intelligent baby monitor should focus on reducing false detections mostly caused by delayed/instable face detections as the face detector reacted sensitively to objects with face-like properties such as stuffed animals and dolls in bed. The intelligent wake-up light system needs to be validated in a closed-loop real-time setting with test subjects to confirm its benefits. Based on self-rated experiences, new personalized reference activity trajectories can be computed building on multiple nights of the same test subject.

Video-based sleep analysis solutions are rather new and therefore the performance levels to accept such a video system are still undefined. Where possible, we compared with performance levels of accepted on-body sensors, like wrist actigraphy, respiratory belts and EMG sensors. More knowledge on acceptance levels for e.g., infant sleeping pose monitoring, would be beneficial.

With the aim to support the diagnosis of sleep disorders that are expressed with particular body movement patterns, the approaches described in this thesis can be extended and tailored to detect relevant and abnormal movement behavior. A video-based motion analysis system could be especially interesting for parasomnias. Some parasomnias have a high prevalence in (young) children, such as rhythmic movement disorder (up to 59%) [375], sleepwalking (up to 17%) [376], and sleep terrors (up to 35%) [376]. The advent of a video camera integrated in baby monitors poses an opportunity to process the video of the sleeping child. Features helping in the differentiation of parasomnias from nocturnal frontal lobe epilepsy [377, 378] can be computed, such as number of attacks per night, time of episodes during sleep, duration of an episode, stereotyped vs. complex motor patterns [379, 380]. In an expert clinical interview, the history is carefully obtained [380], yet, patient and witness reports may be limited due to the nocturnal occurrences [380, 381]. The inter-observer reliability in the diagnosis of parasomnias has been investigated in [381]. The diagnoses are based solely on the interviews according to the minimal diagnostic criteria provided by the American Academy of Sleep Medicine in the International Classification of Sleep Disorders-Revised [382]. The parasomnias sleepwalking, sleep terrors, nightmares, and REM sleep behavior disorder scored an unsatisfactory inter-observer reliability. These parasomnias have the expression of complex movement behaviors in

common where video analysis can aid in the detection of specific features visible in the recorded images. Moreover, the video-based approach has the potential to assist in the detection of sleep-related movement disorders [379], such as restless leg syndrome and periodic limb movement disorder.

Besides reducing manual work of sleep technicians with the automatic analysis of recorded video streams in the sleep lab, the video system offers a long-term home monitoring solution that allows for a better understanding of a subject's sleeping behavior. The use of home video recordings for an improved diagnosis has been suggested in [379, 383] since complex movement episodes are less likely to occur in the sleep laboratory (as observed in e.g. sleep walking in [383] and [379]). Similarly to the Somnolyzer [384], a clinically validated automatic sleep scoring system, the automatic home video system could report deviations from the patient's sleep profile with respect to movements and breathing. The home monitoring system can analyze the sleep habits with regard to circadian rhythm disorders. Times of going to bed and rising can be linked to delayed sleep-phase syndrome and advanced sleep-phase syndrome [376].

The proposed TV methodology for optimizing motion estimators may provide an inspiration for similarly complex multi-dimensional optimization tasks with suboptimal objective metrics in videos of sleep. This may form the next milestone for video-based sleep algorithms where also the optimization of motion algorithms may become more important as research matures in this area. By replacing the performance measures with those relevant for sleep applications, a similar parameter optimization methodology can be applied to motion estimation for sleep applications.

The above insights and uncertainties lead to research questions that could be addressed in future work:

Does a low-cost motion estimation method with superior performance exist when combining the sophisticated HRNM motion estimator or optical flow with hierarchical 3DRS?

Can the insights for the robust motion estimation design methodology be confirmed by a larger user study?

Can other performance measures than the ones used in this thesis for TV picture rate conversion increase the explained variance in ME quality?

Would a dedicated sleep/wake classifier based on activity levels derived from the video signal improve the current sleep/wake classification performance?

Can the multi-distance clustering algorithm be successfully enhanced by additional descriptors, such as body part identification and/or temporal descriptors for the cluster history?

How does the sleep/wake classification performance change when respiration features extracted from the video signal are added to the video actigraphy features?

What is the benefit of the wake-up light control system when tested on a larger user group?

Do personalized reference activity trajectories based on multiple night recordings improve the wake-up experience with the intelligent wake-up light?

Can a face detection method be designed with high specificity for objects with face-like properties while being sensitive to infant faces, rotations and partial occlusions thereof by utilizing photoplethysmography-signal recognition (e.g., [385])?

Can a similar optimization method be applied to the motion analysis algorithms for sleep as is successfully done for TV?

What are the user requirements and acceptance levels of a video monitoring system for the different challenges (e.g., baby pose monitoring, PLM detection, sleep disorder detection, sleep/wake classification in the home setting)?

Bibliography

[367] W.-H. Liao and C.-M. Yang, "Video-based activity and movement pattern analysis in overnight sleep studies.", in *Int'l Conf. on Pattern Recognition*, 2008, pp. 1–4.

[368] R. Brunelli and O. Mich, "Histograms analysis for image retrieval", *Pattern Recognition*, **vol. 34**, no. 8 pp. 1625–1637, Aug. 2001.

[369] A. Xie, "Effect of sleep on breathing - Why recurrent apneas are only seen", *Journal of Thoracic Disease*, **vol. 4**, no. 2, 2011.

[370] S.A Immanuel, Y. Pamula, M. Kohler, D.A Saint, and M. Baumert, "Characterizing respiratory waveform regularity and associated thoraco-abdominal asynchrony during sleep using respiratory inductive plethysmography", in *Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on*, April 2013, pp. 329–332.

[371] E. Durucan and T. Ebrahimi, "Change detection and background extraction by linear algebra", *Proceedings of the IEEE*, **vol. 89**, no. 10 pp. 1368–1381, Oct. 2001.

[372] M. Van De Werken, M.C. Gimenez, B. De Vries, D.G. Beersma, *et al.*, "Effects of artificial dawn on sleep inertia, skin temperature, and the awakening cortisol response", *J Sleep Res*, **vol. 19**, no. 3 pp. 425–435, Sep. 2010.

[373] G.F. Franklin, D.J. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 4th ed., 2001, ISBN 0130323934.

[374] E. Bellers *et al.*, "Solving occlusion in Frame-Rate up-Conversion", in *Digest of the ICCE*, Jan. 2007, pp. 1–2.

[375] K.N. Anderson, I.E. Smith, and J.M. Shneerson, "Rhythmic movement disorder (head banging) in an adult during rapid eye movement sleep", *Movement disorders*, **vol. 21**, no. 6 pp. 866–867, 2006.

[376] M.W. Mahowald and C.H. Schenck, "Insights from studying human sleep disorders", *Nature*, **vol. 437**, no. 7063 pp. 1279–1285, 2005.

[377] C.P. Derry, M. Davey, M. Johns, K. Kron, *et al.*, "Distinguishing sleep disorders from seizures: diagnosing bumps in the night", *Archives of neurology*, **vol. 63**, no. 5 pp. 705–709, 2006.

[378] M. Zucconi and L. Ferini-Strambi, "NREM parasomnias: arousal disorders and differentiation from nocturnal frontal lobe epilepsy", *Clinical Neurophysiology*, **vol. 111** pp. S129–S135, 2000.

[379] P. Tinuper, F. Provini, F. Bisulli, L. Vignatelli, *et al.*, "Movement disorders in sleep: guidelines for differentiating epileptic from non-epileptic motor phenomena arising from sleep", *Sleep medicine reviews*, **vol. 11**, no. 4 pp. 255–267, 2007.

[380] C. Derry, "Nocturnal frontal lobe epilepsy vs parasomnias", *Current treatment options in neurology*, **vol. 14**, no. 5 pp. 451–463, 2012.

[381] L. Vignatelli, F. Bisulli, A. Zaniboni, I. Naldi, *et al.*, "Interobserver reliability of ICSD–R minimal diagnostic criteria for the parasomnias", *Journal of neurology*, **vol. 252**, no. 6 pp. 712–717, 2005.

[382] American Academy of Sleep Medicine, *International classification of sleep disorders, revised: Diagnostic and going manual*, American Academy of Sleep Medicine, 2001, ISBN "0-9657220-1-5".

[383] B. Mwenge, A. Brion, G. Uguccioni, and I. Arnulf, "Sleepwalking: long-term home video monitoring", *Sleep medicine*, **vol. 14**, no. 11 pp. 1226–1228, 2013.

[384] P. Anderer, G. Gruber, S. Parapatics, M. Woertz, *et al.*, "An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24× 7 utilizing the Siesta database", *Neuropsychobiology*, **vol. 51**, no. 3 pp. 115–133, 2005.

[385] C. Park and H.J. Choi, "Motion artifact reduction in PPG signals from face: Face tracking amp; stochastic state space modeling approach", in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, Aug. 2014, pp. 3280–3283.

Acknowledgements

When I started working at Philips Research, I associated the pursuit of a PhD degree with a lot of solitary work. Little did I know then that much larger steps can be taken with fruitful collaborations. Therefore, I would like to acknowledge and express my heartfelt gratitude to everyone who has accompanied me along parts of this journey.

A special thank you to my promotor, Gerard de Haan, for his honest and critical discussions, and to my co-promotor, Reinder Haakma, for his support and guiding advice. I highly appreciate the time and effort it cost the committee members to read and comment on my thesis. Thank you Sabine Van Huffel, Gari Clifford, Peter de With, Ronald Aarts, Sebastiaan Overeem, and Ton Backx.

I would like to acknowledge the co-authors and students who have significantly shaped this work. It was a pleasure working with you: Chris Bartels, René van der Vleuten, Nico Cordes, Gerard de Haan, Tim Nederhoff, Xavier Aubert, Xin Zhao, Frank van Heesch (for both your video signal processing expertise and the cover design!), Bhargava Puvvula, Mukul Rocque, Dmitry Znamenskiy, Di Geng, Jelte Vink, and Vincent Jeanne.

My sincere gratitude to Philips Research, to the TU/e, and to my previous and current group leaders Hans Huiberts, Marc Op de Beeck, and Guido Volleberg, who gave me the opportunity to pursue a PhD degree.

I was fortunate to have worked with experts with highly varied backgrounds. Thank you to the members of the TV and Sleep projects who have inspired me in developing the approaches described in this thesis. From TV, these include Dmitry Znamenskiy, Chris Bartels, Nico Cordes, Harm Belt, Ingrid Heynderickx, and Gerard de Haan. From Sleep, I would like to thank Igor Berezhnyy, Henriette van Vugt, Tim Weysen, Pedro Fonseca, Reinder Haakma, Dmitri Chestakov, DuJia, Gary Garcia, Jan Tatousek, Boris de Ruyter, Roy Raymann, Henning Maass, René Derkx, Xi Long, Xavier Aubert, Els Møst, Yingrong Xie, Shakith Fernando (TU/e), Jan van Dalfsen (TU/e), and Sidarto Bambang Oetomo (Máxima Medisch Centrum Veldhoven).

For the balance between fun activities, technical discussions and explorations, a big thanks to the members of the former Video and Image Processing group.

Finally, I would like to express a deep sense of gratitude to my family and friends, both in Eindhoven and around the globe. I thank you, my friends, for being good listeners and for your understanding of my reduced social interaction the past months. Van harte bedankt, Theo en Ineke, voor jullie steun, de extra oppasuren, en voor de

gezellige brunches en diners samen met Eefke, Sebas, Yinthe en Thije. I am grateful to my parents for their encouragement and honesty in all matters. And last but not least, special thanks to Bram and our sons, Thomas and Nils, for giving me the time and energy to complete this dissertation.

Adrienne Heinrich
July 2015

List of publications

This thesis has resulted in the following publications:

1. A. Heinrich, G. de Haan, C.N. Cordes; A Novel Performance Measure for Picture Rate Conversion Methods, *IEEE Digest of Technical Papers of the ICCE*, Las Vegas, pp. 255-256, 2008.

2. L. An, A. Heinrich, C.N. Cordes, G. de Haan, Improved picture-rate conversion using classification-based LMS-filters, *IS&T/SPIE Electronic Imaging; Visual Communications and Image Processing*, San Jose, Paper 7257-56, 2009.

3. A. Heinrich, H. van Vugt; A new video actigraphy method for non-contact analysis of body movement during sleep, *Journal of Sleep Research*, vol. 19:S283, 2010.

4. A. Heinrich, C. Bartels, R.J. van der Vleuten, G. de Haan; Robust motion estimation design methodology, *Proc. of CVMP 2010, 7th European Conference on Visual Media*, pp. 49-57, 2010.

5. A. Heinrich, C. Bartels, R.J. van der Vleuten, C.N. Cordes, G. de Haan; Optimization of hierarchical 3DRS motion estimators for picture rate conversion, *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 2, pp. 262-274, Mar. 2011.

6. A. Heinrich, I. J. Berezhnoy, G. de Haan; Towards contactless screening of sleep-disordered breathing, *Sleep Medicine*, vol. 12:S116, Sep. 2011.

7. F.J. de Bruijn, A. Heinrich, R. Vlutters; Respiratory motion detection apparatus, *Patent application*, WO2011132118 A2, Nov. 2011.

8. J. Xuyuan, A. Heinrich, C. Shan, G. de Haan; Shared-bed person segmentation based on motion estimation, *IEEE International Conference on Image Processing (ICIP)*, pp. 137-140, 2012.

9. A. Heinrich, R.J.E.M. Raymann, H.C. van Vugt; Contactless sleep disorder screening system, *Patent application*, WO2012131589 A2, Oct. 2012.

10. J. Du, M.G.N. Garcia, A. Heinrich, H. Maass, H.C. van Vugt; Light therapy device, *Patent application*, WO2012085805 A3, Nov. 2012.

11. H.C. van Vugt, A. Heinrich; Method and apparatus for monitoring movement and breathing of multiple subjects in a common bed, *Patent application*, WO2012164453 A1, Dec. 2012.

12. A. Heinrich, H.C. van Vugt, R.M.M. Derkx, M.G.N. Garcia, J. Du; Apparatus and method for the detection of the body position while sleeping, *Patent application*, WO2012164482 A1, Dec. 2012.

13. D. Geng, A. Heinrich, D.N. Znamenskiy, G. de Haan; A motion detection method for a video processing system, *Patent application*, WO2013093731 A1, June 2013.

14. H.C. van Vugt, R.M.M. Derkx, A. Heinrich; Improved detection of breathing in the bedroom, *Patent application*, WO2012095783 A1, July 2013.

15. A. Heinrich, X. Aubert, G. de Haan; Body movement analysis during sleep based on video motion estimation, *IEEE International Conference on e-Health Networking, Applications & Services (Healthcom)*, pp. 539-543, Oct. 2013.

16. A. Heinrich, X. Zhao, G. de Haan; Multi-distance motion vector clustering algorithm for video-based sleep analysis, *IEEE International Conference on e-Health Networking, Applications & Services (Healthcom)*, pp. 223-227, Oct. 2013.

17. P.M. Fonseca, A. Heinrich, I. Berezhnyy, R. Haakma, R. de Bruijn; Separating cardiac signal and respiratory signal from vital signs, *Patent application*, WO2013179189 A1, Dec. 2013.

18. S. Fernando, A. Heinrich, S. Bambang Oetomo, G. de Haan, H. Corporaal; Clinical video recording methodology for contactless vital sign monitoring of neonates, *IEEE SBE symposium on Advancing Healthcare*, p. 25, Feb. 2014.

19. A. Heinrich, D. Geng, D. Znamenskiy, J. Vink, G. de Haan; Robust and sensitive video motion detection for sleep analysis, *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 790-798, May 2014.

20. A. Heinrich, V. Jeanne, X. Zhao; Lifestyle applications from sleep research, *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, no. 6, pp. 829-842, Apr. 2014.

21. A. Heinrich, R.J. van der Vleuten, G. de Haan; Perception-oriented methodology for robust motion estimation design, *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 463-474, June 2014.

22. A. Heinrich, F. van Heesch, B. Puvvula, M. Rocque; Video based actigraphy and breathing monitoring of shared beds from the bedside table, *Journal of Ambient Intelligence and Humanized Computing*, vol. 6, no. 1, pp. 107-120, Feb. 2015.

The author has also contributed to a number of publications outside the scope of this thesis:

1. C. Setz (1st author), A. Heinrich (1st author), P. Rostalski, G. Papafotiou, M. Morari; Application of Model Predictive Control to a Cascade of River Power Plants, *Proceedings of the 17th. IFAC World Congress*, Seoul, Korea, vol. 17, Part 1, Paper 11876-11983, 2008.

2. J. F. Alcon, C. Ciuhu, W. ten Kate, A. Heinrich, N. Uzunbajakava, G. Krekels, D. Siem, G. de Haan; Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis, *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 1, pp. 14-25, Feb. 2009.

3. A. Heinrich, F.H. van Heesch, B. Varghese, N.E. Uzunbajakava; Hair treatment device having a light-based hair detector, *Patent application*, WO2012164441 A1, Dec. 2012.

4. S. de Waele, A. Heinrich, F.J. de Bruijn; Method and apparatus for determining anatomic properties of a patient, *Patent application*, WO2013057627 A1, Apr. 2013.

5. E.M. van der Heide, A. Heinrich, T. Falck; Monitoring system for monitoring a patient and detecting delirium of a patient, *Patent application*, WO2013050912 A1, Apr. 2013.

6. T. Gritti, A. Heinrich, G. de Haan; Method and apparatus for registering a scene, *Patent application*, WO2013072807 A1, May 2013.

Curriculum Vitae

Adrienne Heinrich was born in Zurich, Switzerland, on August 9, 1980. She finished gymnasium with matura type B (Humanistic with Latin) at the Literargymnasium Rämibühl, in Zurich, Switzerland, in 2000. Adrienne took Electrical Engineering and Information Technology at the ETH Zurich in Switzerland. In 2006, she received her M.Sc. degree on the thesis entitled "Application of Model Predictive Control to a Cascade of River Power Plants". This work was performed at the Automatic Control Laboratory of the ETHZ in collaboration with SCIETEC and E.ON Wasserkraft.

After completing her master's degree in 2006, Adrienne joined the Video Processing & Analysis group at Philips Research Laboratories in Eindhoven, the Netherlands, as a research scientist. Initially, her research concentrated on the area of motion estimation and picture rate conversion for TV picture quality improvement. It later evolved into her current research interests that include the design of video analysis algorithms in the areas of unobtrusive sleep monitoring and ICU patient monitoring. Her work in these fields garnered her honours: six peer-reviewed papers that were published in journals, accepted papers and presentations at eight international conferences, and twelve patent applications.

Adrienne's ambition lies in identifying and developing new business opportunities by combining her scientific skills with market insights.