Inaugural - Dissertation

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-mathematischen Gesamtfakultät
der
Ruprecht - Karls - Universität
Heidelberg

vorgelegt von
Diplom-Ingenieur Mark Christoph Jäger
aus Stuttgart

Tag der mündlichen Prüfung: 24.05.2007

Time Series Analysis and Classification
with State-Space Models
for Industrial Processes
and the Life Sciences

Abstract

In this thesis the use of state-space models for analysis and classification of time series data, gathered from industrial manufacturing processes and the life sciences, is investigated. To overcome hitherto unsolved problems in both application domains the temporal behavior of the data is captured using state-space models.

Industrial laser welding processes are monitored with a high speed camera and the appearance of unusual events in the image sequences correlates with errors on the produced part. Thus, novel classification frameworks are developed to robustly detect these unusual events with a small false positive rate. For classifier learning, class labels are by default only available for the complete image sequence, since scanning the sequences for anomalies is expensive.

The first framework combines appearance based features and state-space models for the unusual event detection in image sequences. For the first time, ideas adapted from face recognition are used for the automatic dimension reduction of images recorded from laser welding processes. The state-space model is trained incrementally and can learn from erroneous sequences without the need of manually labeling the position of the error event within sequences. In addition, a second framework for the object-based detection of sputter events in laser welding processes is developed. The framework successfully combines for the first time temporal change detection, object tracking and trajectory classification for the detection of weak sputter events.

For the application in the life sciences the improvement and further development of data analysis methods for Single Molecule Fluorescence Spectroscopy (SMFS) is considered. SMFS experiments allow to study biochemical processes on a single molecule basis. The single molecule is excited with a laser and the photons which are emitted thereon by fluorescence contain important information about conformational changes of the molecule. Advanced statistical analysis techniques are necessary to infer state changes of the molecule from changes in the photon emissions. By using state-space models, it is possible to extract information from recorded photon streams which would be lost with traditional analysis techniques.

Zusammenfassung

In dieser Dissertation wird die Anwendung von Zustandsraum-Modellen zur Analyse und Klassifikation von Zeitreihen, die aus industriellen Produktionsprozessen und den Biowissenschaften stammen, untersucht. Um bisher ungelösten Problemen in beide Anwendungsgebiete beizukommen, wird eine Zustandsraum-Darstellung gewählt, die den zeitlichen Charakter der Daten erfassen kann.

Industrielle Laserschweißprozesse werden mit Hochgeschwindigkeitskameras überwacht, wobei das Auftreten von ungewöhnlichen Ereignissen in den Bildfolgen mit Fehlern am produzierten Bauteil korreliert. In dieser Arbeit werden neuartige Auswertesysteme entwickelt um diese ungewöhnlichen Ereignisse zuverlässig auffinden zu können. Für das Trainieren eines Klassifikationssystems stehen standardmässig nur "Label" für gesamte Bildfolgen zur Verfügung, da es sehr zeitaufwendig ist, jedes einzelne Bild der Folge einzeln auf Unregelmässigkeiten hin zu untersuchen.

Das erste Überwachungssystem kombiniert erscheinungsbasierte Merkmale und Zustandsraum-Modelle zum Auffinden ungewöhnlicher Ereignisse. Hierbei werden erstmals Ideen aus der Gesichtserkennung für die automatische Dimensionsreduktion der aufgenommen Bilder von Laserschweißprozessen verwendet. Das Zustandsraum-Modell wird inkrementell aufgebaut und kann den Informationsgehalt von fehlerhaften Sequenzen automatisch nutzen, ohne dass die genaue Fehlerposition manuell spezifiziert werden muss (schwach überwachtes Lernen). Zusätzlich wird ein objektbasiertes Klassifikationssystem zur Erkennung schwacher Schweißspritzer vorgestellt. Dabei werden zeitliche Änderungsdetektion, Objekverfolgungsalgorithmen und die Klassifikation von Trajektorien erstmals erfolgreich miteinander kombiniert, um schwache Schweißspritzer robust zu erkennen.

Der zweite Beitrag ist die Verbesserung und Weiterentwicklung von Datenauswerteverfahren für die Einzelmolekülfluoreszenzspektroskopie (SMFS). Bei SMFS Versuchen werden einzelne Moleküle mit einem Laser angeregt und die durch Fluoreszenz erzeugten Photonen mit hoher zeitlicher Auflösung aufgenommen. Die Aufgabe der Datenanalyse ist, aus Änderungen in der Photonintensität, auf unterschiedliche Zustände des Moleküls zu schließen. Durch die Benutzung von Zustandsraum-Modellen wird es möglich Informationen zu extrahieren, die mit klassischen Auswerteverfahren verloren gegangen wären.

Hiermit erkläre ich, Mark Christoph Jäger, dass ich die vorgelegte Dissertation selbst verfasst und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfen bedient habe.

Kornwestheim, den 10. April 2007

Preface

First and foremost I would like to express my gratitude to my doctoral advisor Professor Fred A. Hamprecht from the Interdisciplinary Center of Scientific Computing (IWR), University of Heidelberg, Germany for his support and encouragement over the last years and particularly for giving me the push forward to finish this thesis. His outstanding commitment to my work was a key factor for its success.

I gratefully acknowledge financial support of this thesis by the Research and Development Department of the Robert Bosch GmbH, Schwieberdingen, Germany. In particular, I would like to thank Walter Happold from the Robert Bosch GmbH for offering me the opportunity to gain experience solving challenging real-world problems. He always managed to combine the company's interest with giving me freedom to pursue scientific research. Thanks also go to Christian Knoll from the Robert Bosch GmbH, Schwieberdingen, Germany. We worked together on several projects and he was always committed to transferring my work from the research stage to real applications.

Many thanks also go to Alexander Kiel and Dirk-Peter Herten from the Institute of Physical Chemistry, University of Heidelberg, Germany who introduced me to the interesting world of single-molecule spectroscopy and provided the data for the work presented in chapter 7. The discussions with them have always been an indispensable source of inspiration.

I am indebted to my colleagues and my fellow doctoral students: Andreas, Jochen, Linus, Marco, Matthias, Stefan and Thomas. If work was fun, then it was due to them. I will definitely miss our daily "Ice Cream-Tours". I also want to express my gratitude to all members of the Multidimensional Image Processing Group at the IWR for their support and the nice time we had during several seminars. Special thanks go to Nadin, Andreas, Chris and Fred for comments on the final version.

Sincere thanks go to my family who continuously supported me during the long years of my studies. And last but not least I want to thank Nadin for her patience and support over the last years, without her I would not have finished this work.

Contents

Contents

Contents

x

1 Introduction

In this thesis, novel approaches to overcome up to now unsolved problems in the quality control of industrial manufacturing processes and the analysis of experimental data from the life sciences are considered. For both application domains the considered data sets are multivariate time series, whose temporal information is captured using state-space models.

1.1 State-Space Models

Sequential data arises in many different applications and state-space models are an advanced probabilistic modeling technique to perform classification, prediction or filtering of the data sequences. In a state-space model it is assumed that the observations are generated from an underlying process which cannot be observed and is thus "hidden". Generally speaking, we try to infer with state-space models the underlying cause behind what we see. The underlying process can reside in different hidden states which evolve in time and generate the observations accessible to the observer.

State-space models have been successfully employed for modeling sequential data in various fields [22]. Although the "standard" algorithms for state-space models are already well studied [60, 66], there is still active research on how to apply them to various real world problems. State-space models have been extensively used for object tracking [27] and are still an active field of research especially in the area of speech recognition [52, 78], video content analysis [3, 23, 83, 100, 101, 103] or the life sciences [4, 9, 21, 43, 45, 54].

1.2 Industrial Process Monitoring

The importance of computer vision systems for the automatic quality assurance of industrial manufacturing processes is increasing. Due to advances in sensor technology, the diversity and complexity of problems to be addressed grows dramatically. For instance the further development of high-speed cameras [64] or three dimensional imaging techniques [75] has led to more and more structured data. For quality sensitive manufacturing processes, the produced component parts can be inspected for defects after the manufacturing step or the manufacturing process itself can be monitored. In this thesis analysis and classification methods for the latter are developed. Process monitoring systems that are deployed in mass production environments have high requirements regarding their usability. Industrial image processing applications rely typically on supervised classification, requiring a high labeling effort, especially for large data sets. This stands in the way of a user-friendly system, in particular with

respect to the expected further growth of data in upcoming years. The goal of this thesis is to engineer frameworks which on the one hand meet the high performance requirements of industrial quality inspection, namely all erroneous parts have to be found, and on the other hand minimize the labeling costs necessary for classifier learning. For the validation of the presented algorithms the in-line quality control of laser welding processes is considered. For this application unusual events in the recorded image sequences correlate with faults on the produced weld seam.

In recent years laser welding has become a widespread and highly developed industrial manufacturing process [85]. However, laser welding is still a dynamic and chaotic process and thus vulnerable to process errors such as material sputters, weld break-ins or weld reinforcements. Although process errors occur rarely, it is vital to ensure that all faulty welds are detected, since errors can lead to a malfunction or a complete outage of the manufactured component part. The improvement of automatic diagnostic and control systems for laser welding processes are of ongoing interest [1, 15–17, 31, 39, 44, 59, 61, 63, 84].

When laser radiation interacts with the work piece, secondary radiation is generated. This radiation contains information about the process stability and can thus be used to detect process errors. For the regular process state the recorded radiation forms approximately a disc-shaped object, in the following referred to as the "melt pool". Weld seam errors appear as two different phenomena in the image sequences: deformations of the recorded radiation and additional moving objects, so called sputter events. For both error types different detection frameworks are developed which are introduced in the following.

Incremental Learning for Unusual Event Detection

To detect deformations of the melt pool, the feature space dimension is reduced by exploiting the intrinsic low dimensionality of the images using subspace methods [57], and temporal dynamics are encoded using Hidden Markov Models (HMMs). HMMs are state space models with a finite state space and the most widely used method for sequence classification. Class labels for training the automatic monitoring system are by default only available for the complete image sequence, since scanning sequences for erroneous frames is tedious. To efficiently use the information of the weakly labeled data an incremental learning procedure is used which works as follows:

- A generative model for the regular sequences is trained.

- The regular sequence model (RSM) is used to locate potentially unusual segments within error sequences by means of a change detection algorithm.

- Unusual segments are used to expand the RSM to an error sequence model (ESM). The complexity of the ESM is controlled by means of the Bayesian Information Criterion (BIC).

The likelihood ratio of the data given the ESM and the RSM is used for the classification decision. This ratio is close to one for sequences without error events

and increases for sequences containing error events. The developed incremental learning procedure can significantly reduce the user interaction and off-line tests on representative data sets as well as benchmarking the algorithms directly at production lines proved that sequences with error events can be found with a small false positive rate. The key finding of the empirical study was that weakly labeled data can improve the classification performance and that the classification system can fulfill demanding industrial requirements. The considered task is closely related to unusual event detection in video sequences which is, in particular due to the increasing commercial interest in surveillance systems, an active field of research [3, 23, 83, 100, 101, 103]. State-space models have been investigated for process monitoring applications and have proved to increase the robustness for fault detection in various applications [6, 8, 82, 98, 104].

Object Tracking of Sputter Events

To automatically classify sputter events, the fault detection system flags suspicious objects by means of a change detection algorithm and then tracks these objects using Kalman filters. Kalman filters are Gaussian linear state-space models. The temporal correlation of the recorded images is used to establish a correspondence between suspicious objects in different frames. Classification of the sputter events is performed using simple trajectory features such as track length or track start and end position of the followed uspicious objects. During classifier training the user only needs to mark suspicious sequences, but does not need to label individual objects. Hence, the key design goal to minimize the labeling costs is satisfied. The framework is tested on two challenging datasets from industrial welding processes. The classification results show that the material particles can be tracked accurately and a sustainable improvement with respect to classification performance and labeling costs compared to previous published techniques [31] can be achieved.

1.3 Time Series Analysis in the Life Sciences

State-space models are also a widespread data analysis technique in the life sciences.In this thesis the analysis of photon detections of Single Molecule Fluorescence Spectroscopy (SMFS) experiments using state-space models is considered. SMFS allows scientists to follow biochemical processes on a single molecule basis [51]. The molecules display strong fluctuations of their fluorescence signals when exited with a laser [50]. These fluctuations are random variations and a detailed statistical analysis is necessary to be able to infer microscopic information from the changes in photon emission. State-space models are of increasing interest for SMFS data analysis [4, 9, 21, 38, 43, 54], since they show no limitation regarding the temporal resolution. The state-space model considered here is a Markov Modulated Poisson Process (MMPP), which is a Poisson process whose rate is governed by an underlying Markov chain. The developed algorithmic framework is used for the automatic analysis of individual association and dissociation events of copper-ions (Cu^{2+}) with the bidentate

ligand 2,2'-bipyridine-4,4'-dicarboxic acid (*dcbpy*) in aqueous media.

The process of association and dissociation of individual Cu^{2+}-ions with SMFS was reported in [42] for the first time. Upon copper addition, the fluorescence signal starts fluctuating between two intensity levels which can be associated with the bound and unbound states of copper. Thus, the model with the lowest complexity to describe the recorded photon emissions is a two-state ON and OFF model. In the ON-state (OFF-state) the photon arrival rate is high (low), and the time interval between photon arrivals tends to be shorter (longer). In addition to the two-state model, higher model complexities (up to 4 states) are investigated in order to best describe the recorded data. The determined kinetic rates for association and dissociation of Cu^{2+}-ions are compared to results obtained with a classical analysis approach. It is found that the parameter estimates for a photon-by-photon analysis significantly depend on the correct model assumption. We demonstrate that the proposed framework can accurately detect fast fluctuations in the photon intensity. In particular, in the analyzed sequences from the formation and dissociation of Cu^{2+}-ions with *dcbpy*, rapid fluctuations occur occasionally. These fluctuations were first observed when analyzing the raw data with MMPPs, without prior binning.

1.4 A Guide to this Thesis

The thesis starts with a theoretical chapter about state-space models. Important methods such as state inference from observations or maximum likelihood parameter estimation are introduced in chapter 2.

In the following, state-space models are employed in two different application domains: The in-line process control of industrial laser welding processes in chapter 3 to chapter 6 and the analysis of photon detections from single-molecule spectroscopy experiments in chapter 7.

For the quality control of industrial laser welding processes a pattern recognition system based on state-space models is developed. Pattern recognition systems can be partitioned into a sequence of three different stages - sensing, segmentation & dimension reduction and classification - where from stage to stage the information is reduced. At the output of the classification stage the recorded image sequence is reduced to just a few bits representing the chosen category (a single bit for quality control applications). This structure is reflected in the arrangement of the following chapters: sensing is discussed in chapter 3, segmentation & dimension reduction is presented in chapter 4 and classification is described in chapter 5 and chapter 6.

In chapter 3 the imaging sensor concepts are presented along with a description of the recorded image sequences. Since the raw data is too high-dimensional to allow for an efficient learning, the recorded images have to be described with low-dimensional feature vectors. Methods for dimension reduction based on object and appearance-based features are described in chapter 4. The next two chapters describe different methods to exploit the temporal information content offered by monitoring the process with a high-speed camera. In chapter 5 incremental learning of a classification framework combining HMMs and appearance

based features for unusual event detection is considered. The framework is optimized and validated for fault detection in laser welding processes, but the presented strategies can also be applied to other manufacturing processes. In chapter 6 dimension reduction is performed using object features and the temporal information content is used to establish correspondence between objects detected in different frames. Detected objects are tracked with the Kalman Filter and trajectory features are computed. The system is optimized for the robust sputter detection in laser welding processes.

The application of state-space models for advanced data analysis methods in the life-sciences are discussed in chapter 7. The Markov Modulated Poisson Process is introduced and it is shown that for SMFS experiments it is able to detect patterns in the data which are lost with classical analysis techniques.

Lastly, chapter 8 offers some conclusions about the use of state-space models in both application domains considered.

1 Introduction

2 State-Space Models

Sequence learning is an important task in many application domains and several algorithms exist to capture temporal information [87]. In this chapter state-space models are introduced which are the theoretical framework for several popular algorithms for modeling temporal information such as Hidden Markov Models or Kalman Filters. There exist two important types of state-space models for which algorithms for inference are mathematically tractable: state space models with discrete state spaces referred to as Hidden Markov Models (HMM) [66] which are used extensively e.g. in speech recognition, and Gaussian linear state space models commonly called Kalman filter models (KFM) [22] which are the dominating algorithm for tracking applications.

This chapter is organized as follows: First a short introduction to general state space models is given, followed by a definition and brief explanation of Hidden Markov Models (HMMs). Parameter estimation for HMMs is considered in section 2.3. Discriminative training of HMMs has shown promising results in speech recognition applications, therefore a short introduction is given in section 2.5. In the last section state inference algorithms for Gaussian linear state models (Kalman Filter Models) are presented.

2.1 Introduction

State space models consist of a Markov chain (also called state sequence) X_k, where k is an integer with $1 \leq k \leq K$. The state vectors X_k are hidden (unobserved) and visible to the observer are only the vectors Y_k. The state vectors X_k govern the distribution of the corresponding observations Y_k [22]. It is assumed that X_k is the only variable of the Markov Chain that influences the distribution of Y_k, and that the variable X_k given X_{k-1} is independent of previous state variables (Markov property). State variables X_k and observation variables Y_k can be either discrete or continuous.

The dependence structure of state-space model is represented in Fig. 2.1. Nodes correspond to random variables and the edges define the structure of joint probability distributions. For each node the joint probability distributions can be computed as a product of the distributions from its *parents* nodes (those directly linking to the given node). The implied conditional independence assumptions significantly simplify the computations of the probability distributions.

If the states X_k are discrete (finite state space), the model is referred to as a Hidden Markov Model (HMM)[1]. The success of HMMs to solve classification

---

[1]In some publications, the terms state-space model and Hidden Markov Model are used interchangeably.

and analysis tasks in various application domains including speech recognition, handwritten character recognition, modeling of human activity, various applications for time series data from biology and many more has lead to an ever-increasing interest in improving the algorithms for HMMs. State-space models with continuous state spaces are used mainly in tracking and control applications. Kalman filter models are the most prominent example for this family of continuous state-space models. In tracking application the states typically represent position and velocity of an object and the observations are noisy measurements of the state variables.

Two types of inference algorithms are used for the state-space models in this thesis, namely filtering and smoothing[2]. Fig. 2.1 gives a graphical representation of the different inference algorithms [60]. Filtering is commonly used in online analysis, where the state X_k is estimated from the data gathered up to k; e.g. filtering is used in on-line object tracking applications. Smoothing is important for parameter learning and can be performed online (fixed-lag smoothing) or offline (fixed interval smoothing). In both cases the predicted state variable X_k is a state of the past, but for fixed-lag smoothing the state estimate is computed after waiting for a fixed time period, whereas for fixed interval smoothing, X_k is estimated after knowledge of the complete sequence is available.

2.2 Hidden Markov Models

A Hidden Markov Model (HMM) is a finite state automaton. In order to emphasize the difference between continuous state space models and HMMs, the state variable for HMMs is changed from X_k to Q_k and the observation variable from Y_k to O_k, where $1 < k < K$. This makes the variable definitions consistent with most of the publications in the speech processing community, where HMMs have been researched extensively. The variables X_k and Y_k are used for continuous state space models (see section 2.6). For a tutorial introduction to

---

[2]Control and prediction are additional interference techniques which are not considered.

Figure 2.1: Graphical representation of the dependence structure of a state-space model. Y_k is the observable process (filled circles) and X_k (empty circles) is the hidden chain. Missing arrows between variables represent conditional independence. Since the structure repeats, the model can be defined by showing just the first two slices.

Figure 2.2: Illustration of the different inference types for state-space models. The shaded box marks the time period for which data is available and the arrow marks the time instance for which the state X_k is predicted for the particular inference type.

HMMs see e.g. [11,66] and for a detailed explanation of various algorithms and different kinds of HMMs see [22,60].

2.2.1 Definition

At any discrete time instance k the HMM exists in one of a finite set of states $Q_k = i$ with $1 \leq i \leq N_Q$. Without loss of generality the states are numbered from 1 to N_Q, where N_Q is the total number of states of the HMM. Unless otherwise stated, a homogeneous HMM is assumed, therefore stochastic transitions between states are governed by a time-independent transition probability matrix A with elements $a_{ij} = P(Q_k = j | Q_{k-1} = i)$. Each state Q_k that could be visited at time instance k could emit a single observation vector o_k according to a probability distribution that is specific to that state. The probability of a particular, continuous observation vector o_k in state $Q_k = i$ is given by $b_i(o_k)$. The initial state probabilities for time instance $k = 1$ are described with $\pi_i = P(Q_1 = i)$.

Gaussian Mixture Model (GMM)

A popular approach is to represent $P(O_k = o_k | Q_k = i)$ for continuous observation vectors o_k with a mixture of multivariate Gaussian distributions for each state Q_k. Gaussian mixture models (GMM) are computationally inexpensive and a well-understood statistical model. A GMM for state i with parameters $\lambda_i = \{w_{im}, \mu_{im}, \Sigma_{im}\}$ is defined by:

$$b_i(o_k) = P(O_k = o_k | Q_k = i) = \sum_{w=1}^{M} w_{im} P(o_k | \mu_{im}, \Sigma_{im}), \qquad (2.1)$$

where M is the total number of mixture components, w_{im} is the weight of the mth mixture component of state i, and $P(O_k = o_k | Q_k = i)$ specifies a multivariate normal distribution with mean vector μ_{im} and covariance matrix Σ_{im}.

The multivariate Gaussian distribution for the mth mixture element is given by:

$$P(o_k|\mu_{im}, \Sigma_{im}) = \frac{1}{(2\pi)^{D/2}|\Sigma_{im}|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_i)^T \Sigma_{im}^{-1}(x - \mu_i)\right\}, \quad (2.2)$$

where M is the total number of mixture components. The mixture weights must satisfy

$$\sum_{w=1}^{M} w_{im} = 1 \qquad (2.3)$$

to make $b_i(o_k)$ a valid probability density function for each state i. The form of the covariance matrix Σ_{im} can be either full or diagonal. In speech recognition it was found empirically that diagonal covariance matrices outperform full covariance matrices. The probability density which can be described with a full covariance GMM can equally be achieved using larger order diagonal GMMs [68]. Diagonal GMMs with $M > 1$ are capable of modeling correlated elements of feature vectors. In this thesis the number of mixture elements is the same for all states within one model. The complete set of HMM parameters for a particular model is given by $\Psi = \{\pi, \Lambda, A\}$, where $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{N_Q}\}$ is a vector containing the GMM parameters λ_i for each state i. Automatic complexity controls are employed to find the optimal number of states N_Q and mixture elements M [12].

2.2.2 HMM Topology

For fully connected or ergodic HMMs every state can be reached from every other state within a single time step. An example for an ergodic HMM is presented in Fig. 2.3(a) and the corresponding state transition matrix A is given by:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad (2.4)$$

Prior knowledge can be used to restrict the allowed state transition and achieve a HMM topology which might be better suited to solve a particular problem. For sequence classification left-to-right models are an important and often employed topology. The Markov Chain starts in an initial state, traverses a number of intermediate states, where it is not possible to go backwards, and ends in a final state. These topologies are well suited to model the sequential properties of time series. The state transition coefficients have to fulfill the property

$$a_{ij} = 0 \text{ for } j < i \qquad (2.5)$$

and the initial state probabilities π for $k = 1$ are

$$\pi = \begin{cases} 1 & \text{for } Q_1 = 1 \\ 0 & \text{otherwise} \end{cases} \qquad (2.6)$$

(a)   (b)

Figure 2.3: Illustration of a ergodic, fully connected three state HMM topology
in (a) and a five state left-to-right model with a maximum allowed
state change ΔQ of two in (b).

When representing the HMM topology with a state-transition diagram, arcs belonging to state transitions for $a_{ij} = 0$ are removed from the drawing. Normally additional constraints are imposed to avoid large changes in the state indices within a single step. As an example the topology of a left-to-right model with a maximum state change of $\Delta Q = 2$ is shown in Fig. 2.3(b). The state transition matrix belonging to Fig. 2.3(b) is given by:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 \\ 0 & a_{22} & a_{23} & a_{24} & 0 \\ 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & 0 & 0 & a_{44} & a_{45} \\ 0 & 0 & 0 & 0 & a_{55} \end{bmatrix} \tag{2.7}$$

For left-to-right models the number of states is usually large and a sequence of random length is produced. In contrast, for fully connected HMMs the underlying Markov chain is ergodic and admits an infinitely long sequence of output and also allows for periodicity in the data. The number of states of ergodic HMMs is typically much smaller than for a left-to-right model.

2.3 Learning

Assuming a parametric form of the class-conditional probability densities reduces the learning task to finding the parameters Ψ of the considered model. The maximum-likelihood (ML) method seeks to find the model parameters $\widehat{\Psi}$ which are best supported by the training data. Thus, the objective function which has to be maximized is given by

$$\sum_{u=1}^{U} \log P\left(\mathbf{O}^u \,|\, \Psi\right), \tag{2.8}$$

where the summation is over different observation sequences $\mathbf{O}^u$. The ML estimate of the parameters is found by:

$$\widehat{\Psi} = \arg\max_{\Psi} \sum_{u=1}^{U} \log P\left(\mathbf{O}^u \,|\, \Psi\right). \qquad (2.9)$$

2.3.1 Maximum Likelihood Estimation

HMMs have hidden state variables Q_k and a ML estimation technique which can handle incomplete or missing values if required is necessary. If the state sequence of a HMM were known, ML parameter optimization would just reduce to counting relative transitions and estimating sufficient statistics for the assumed parameter model [14]. However since the state sequence is unobserved another approach is adopted, namely the expectation maximization (EM) algorithm. The EM is an iterative procedure that consists of two alternating steps: the expectation step (E-step) and the maximization step (M-step). The E-step calculates the expected posterior state distribution given the observed data and current parameterization, i.e. the E-step marginalizes over the unobserved state variables. For HMMs the E-step can be efficiently calculated using the forward-backward algorithm presented in section 2.3.2. The re-estimation of the parameters on the basis of the posterior state distribution computed in the E-step is performed in the M-step. The parameters in the M-step are computed such that the posterior log-likelihood in the E-step is maximized. Each iteration of the EM is guaranteed to increase the log-likelihood, but the estimation procedure can get stuck in local optima, requiring for example multiple initializations in order to overcome the problem. For HMMs the EM algorithm is also referred to as the Baum-Welch algorithm. It is the most basic learning algorithm for parameter estimation for HMMs. The main advantages are that no free parameters have to be tuned for its convergence, it is easily scalable to large data sets, and compared to other methods the computation time is short [78]. To improve the recognition rates for speech recognition, advanced objective functions have been considered, which rely on discriminative learning criteria. It has been shown for various applications in particular in the field of speech recognition that discriminative learning approaches outperform ML estimation [40]. Rather than learning true data probability distributions, discriminative learning techniques focus on the end goal and try to directly reduce the error rate. A short overview of state of the art discriminative learning methods for HMMs is presented in section 2.5.

2.3.2 Baum-Welch Algorithm

The Baum-Welch algorithm is the sequence version of the EM. In the E-step, the forward-backward procedure is applied to find the state posteriors probabilities $P\left(Q_k = i \,|\, \mathbf{O}, \Psi^{(n)}\right)$ given the observation sequence $\mathbf{O} = \{o_1, \ldots, o_k, \ldots, o_K\}$ and parameters $\Psi^{(n)}$ of the nth EM iteration. Rather than making a hard decision to which state each observation belongs, a soft segmentation of the

sequence is performed. The posterior probabilities in the E-step are computed using the forward-backward algorithm, which can be best understood by representing the HMM in a trellis structure [66]. In the forward procedure, the forward probability:

$$\alpha_{i,k} = P\left(o_1, \ldots, o_k, Q_k = i \,\middle|\, \Psi^{(n)}\right) \qquad (2.10)$$

is computed, where the forward probability $\alpha_{i,k}$ specifies the probability of observing the partial sequence $o_1, ..., o_k$ and ending up in state i at time k. Due to the conditional independence assumptions of the Markov chain the forward probabilities can be iteratively computed for $2 < k < K$ with:

$$\alpha_{j,k+1} = \left[\sum_{i=1}^{N_Q} \alpha_{i,k} a_{ij}\right] b_j(o_{k+1}) \qquad (2.11)$$

where for $k = 1$ the probability is initialized with $\alpha_{i,1} = \pi_i b_i(o_1)$. Using the forward probabilities the posterior likelihood of the observation sequence under the assumed model parameters $\Psi^{(n)}$ can be computed with:

$$P\left(\mathbf{O} \,\middle|\, \Psi^{(n)}\right) = \sum_{i=1}^{N_Q} \alpha_{i,K} \qquad (2.12)$$

For the backward iteration, the computation is started at time instance $k = K$ and goes back in time to $k = 1$ to compute the backward probabilities:

$$\beta_{i,k} = P\left(o_{k+1}, ..., o_K \,\middle|\, Q_k = i, \Psi^{(n)}\right). \qquad (2.13)$$

The backward probability $\beta_{i,k}$ is defined as the probability of observing the partial sequence $o_{k+1}, ..., o_K$ having started in state i at time k: The induction steps for $K - 1 > k > 1$ are given by

$$\beta_{i,k} = \sum_{j=1}^{N_Q} a_{ij} b_j(o_{k+1}) \beta_{j,k+1} \qquad (2.14)$$

where for $k = K$ $\beta_{i,K} = 1$. For long or even moderate sequence length the posterior probabilities become very small and the forward-backward procedure gets numerically instable. This problem is well known and can be solved by either performing the calculations in the log-domain or using a scaling procedure as suggested in [66][3].

Using the forward and backward probabilities, several important a posteriori probabilities can be computed. In the following only the important results are stated; detailed derivations can e.g. be found in [14]. The posterior state

---

[3]For the computation in this thesis the scaling approach is used to avoid numeric underflow of the forward-backward algorithm.

probability $\gamma_{ik} = P\left(Q_k = i \,\middle|\, \mathbf{O}, \Psi^{(n)}\right)$ is given by:

$$\gamma_{i,k} = \frac{\alpha_{i,k}\beta_{i,k}}{P\left(\mathbf{O} \,\middle|\, \Psi^{(n)}\right)}. \tag{2.15}$$

Similarly, the probability $\xi_{ij,k} = P\left(Q_k = i, Q_{k+1} = j \,\middle|\, \mathbf{O}, \Psi^{(n)}\right)$ of being in state i at time k and in state j at time $k+1$ for the detected data sequence $\mathbf{O}$ can be computed with:

$$\xi_{ij,k} = \frac{\alpha_{i,k}a_{ij}b_j\left(o_{k+1}\right)\beta_{j,k+1}}{P\left(\mathbf{O} \,\middle|\, \Psi^{(n)}\right)}. \tag{2.16}$$

The expected number of times in state i is given by

$$\sum_{k=1}^{K} \gamma_{i,k} \tag{2.17}$$

and the expected number of transitions from state i to state j by

$$\sum_{k=1}^{K-1} \xi_{ij,k} \tag{2.18}$$

The entries of the transition matrix A can then be estimated with:

$$\widehat{a}_{ij} = \frac{\sum_{k=1}^{K-1} \xi_{ij,k}}{\sum_{k=1}^{K-1} \gamma_{i,k}} \tag{2.19}$$

where the quantity is the expected number of transitions from state i to state j divided by the total number of expected times the Markov Chain has been in state i.

2.3.3 Gaussian Mixture Models

To be able to estimate the parameters $\widehat{\Lambda}$ of the GMM an additional quantity $\gamma_{i,l,k}$ has to be defined:

$$\gamma_{i,l,k} = \gamma_{ik}\frac{w_{i,l}b_{i,l}(o_k)}{b_i(o_k)}, \tag{2.20}$$

which specifies the probability that the lth mixture component associated with state i generated observation o_k ($b_{i,l}(o_k)$ is the lth Gaussian component associated with state i evaluated at o_k). With this at hand, the ML estimates for the mixture weight $w_{i,l}$ are given by:

$$\widehat{w}_{i,l} = \frac{\sum_{k=1}^{K-1} \gamma_{i,l,k}}{\sum_{k=1}^{K-1} \gamma_{ik}} \tag{2.21}$$

and the re-estimation formulas for the parameters of the lth Gaussian mixture associated with state i are given by:

$$\widehat{\mu}_{i,l} = \frac{\sum_{k=1}^{K-1} \gamma_{i,l,k} o_k}{\sum_{k=1}^{K-1} \gamma_{i,l,k}} \tag{2.22}$$

$$\widehat{\Sigma}_{i,l} = \frac{\sum_{k=1}^{K-1} \gamma_{i,l,k} \left(o_k - \widehat{\mu}_{i,l}\right) \left(o_k - \widehat{\mu}_{i,l}\right)^T}{\sum_{k=1}^{K-1} \gamma_{i,l,k}} \tag{2.23}$$

Note that equations (2.22) and (2.23) are only slight modifications of the well known formulas for ML estimation of multivariate Gaussian distributions. The new set of model parameters is in the next iteration step used as the current set $\widehat{\Psi} = \{\widehat{\pi}, \widehat{\Lambda}, \widehat{A}\} \to \Psi^{(n+1)}$.

2.3.4 Computational Complexity

For N_Q discrete states the complexity for the forward pass takes $O(N_Q^2)$ operations per time step, since we must perform a matrix-vector multiplication at each time step. The complete forward-backward procedure therefore has a time complexity of $O(N_Q^2 K)$ [24]. The backward pass requires to store α_k for $k = 1, \ldots, K$ until we do the backward pass therefore the space consumption is $O(N_Q K)$. Hence, time/space complexity increase linearly with the sequence length K and linearly/quadratically with the number of discrete states N_Q.

2.4 Classification

2.4.1 Sequence Classification

It is assumed that an observation sequence $\mathbf{O}^u$ belongs to one of N_ω classes and thus a class label ω_i with $\omega_i \in \{\omega_1, \omega_2, \ldots, \omega_{N_\omega}\}$ can be associated with each observation sequence. During learning, for each sequence class ω_i a generative HMM with model parameters Ψ_i is learned from the training sequences. Hence, for each unlabeled observation sequence the class likelihood $P(\mathbf{O}|\omega_i)$ can be computed using the forward pass (see eq. (2.12)). The Maximum A Posteriori (MAP) decoder assigns an unlabeled sequence to the class ω_{MAP} with the highest posterior probability, i.e. the class that best "explains" the sequence:

$$
\begin{aligned}
\omega_{MAP}(\mathbf{O}^u) &= \arg\max_{\omega_i} \log P(\omega_i|\mathbf{O}^u) & (2.24)\\
&= \arg\max_{\omega_i} \log \frac{P(\mathbf{O}^u|\omega_i)P(\omega_i)}{\sum_{j=1}^{N_\omega} P(\mathbf{O}^u|\omega_j)P(\omega_j)} & (2.25)
\end{aligned}
$$

where $P(\omega_i)$ are the class priors. The MAP decoder minimize the probability of error and is also known as an optimal minimum Bayes classifier.

2.4.2 Viterbi Decoding

The most likely sequence of hidden states $\widehat{\mathbf{q}} = \{\widehat{q}_1, \widehat{q}_2, \ldots, \widehat{q}_K\}$ given the data and a trained HMM with parameters Ψ can be obtained with Viterbi decoding. By Bellman's principle of optimality, the single most likely path to reach state Q_k consists of the most likely path to some state Q_{k-1} at time instance $k-1$, followed by a transition to Q_k [58]. Hence, not all possible paths through the trellis have to be considered to find the most likely, and the complexity only increases linearly with time. Indeed, the forward pass for Viterbi Decoding is the same as the one presented in eq. 2.11, except that the sum is replaced by the max operation:

$$\widetilde{\alpha}_{j,k+1} = \max_i \left[\widetilde{\alpha}_{i,k} a_{ij} \right] b_j(o_{k+1}), \tag{2.26}$$

where $\tilde{\alpha}_{i,0} = \pi$. In addition during the forward pass one keeps track of the most likely predecessor to each state:

$$\psi_k = \arg \max_i \left(a_{ij} \widetilde{\alpha}_{i,k-1} \right) \tag{2.27}$$

In the backward pass the most likely state sequence $\widehat{\mathbf{q}}$ is computed recursively:

$$\widehat{q}_k = \psi_{k+1} \left(\widehat{q}_{k+1} \right) \tag{2.28}$$

Note that the Viterbi path is different from finding the most likely state at time k which can be computed using the Bahl-Cocke-Jelinek-Raviv (BCJR) algorithm [7].

2.4.3 Performance Metrics

The data sets considered in this thesis are strongly imbalanced (see chap. 3), i.e. although a large data set is gathered for at least one class only a few examples are available. For imbalanced datasets accuracy is not an appropriate measure of performance, since a simple classifier which predicts every case as the majority class can still have a high accuracy. The performance metrics used are the false positive (FP) rate, false negative (FN) rate, and recall and precision which are based on the notation introduced in Table 2.1:

$$\text{FP rate} = \frac{\text{FP}}{\text{FP} + \text{TP}} \tag{2.29}$$

$$\text{FN rate} = \frac{\text{FN}}{\text{FN} + \text{TN}} \tag{2.30}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.31}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.32}$$

Between the FP rate and FN rate a trade off has to be found. For process monitoring systems a FN rate of 0% has to be achieved on the test data set (zeror-error-policy). Hence, the FP rate for a FN rate of 0% $FP|_{FN=0}$, is the

Table 2.1: Confusion Matrix

Truth/Prediction Truth	Positive Class	Negative Class
Positive Class	TP (true positive)	FN (false negative)
Negative Class	FP (false positive)	TN (true negative)

most important quality measure for the considered classification frameworks. A zero-error-policy corrsponds to a recall of one. The trade off between the FP rate and FN rate for any possible cut off can be graphically presented in a receiver operator charateristic (ROC); see e.g. [26] for an introduction. Commonly the area under the ROC curve, referred to as AUC (area under the curve), is used the compare the performace of different classifcation systems using a single scalar value.

2.5 Advanced Learning Methods for HMMs

There exist many techniques which can improve the classification performance of HMMs using advanced objective functions for parameter estimation rather than the basic ML estimation. This section briefly compares generative and discriminative classification models and reviews the different optimization criteria for discriminative learning of HMMs.

2.5.1 Comparison of Generative and Discriminative Learning

In a supervised classification scenario, several observation sequences $\mathcal{O} = \{\mathbf{O}^1, ..., \mathbf{O}^U\}$ together with corresponding class labels $\mathcal{C} = \{C^1, ..., C^U\}$ are given. The task of the classification framework is to predict the class label $\widehat{C}$ for an unlabeled observation sequence $\mathbf{O}$ by using the information of the labeled sequences. HMMs belong to the family of generative models. In generative models the joint distribution $P(\mathcal{O}, \mathcal{C})$ over both input observations and class labels is modeled. If the generative model fits the "reality" perfectly, no other method can improve the performance, but for real world applications it is (normally) not possible to find the correct model. Indeed, empirically it was found that if labeled data is plentiful, discriminative models that directly estimate the decision boundary can outperform generative models [47].

2.5.2 Weakly Labeled Data

Nevertheless, in case of weakly labeled data sets generative models have desirable features. A weakly labeled data set is a data set where only partial information is provided for the training data labels [69]. In contrast, for a fully labeled data set there exists for each data point a class label. An example for weakly labeled data in the context of unusual event detection in image sequences is the setting where we are told that the unusual event is present in the

sequence, but no information regarding the location within the image sequence nor the length of the unusual event is given. A beneficial property of weakly labeled data is that it can be recorded very cost efficiently. Generative models can make use of the information contained in the unlabeled data by semi-supervised learning techniques. In contrast, discriminative techniques cannot benefit from unlabeled data [93]. Therefore in [47] a framework for blending generative and discriminative models for partially labeled data is proposed. By optimizing a parameter, depending on the ratio of labeled to unlabeled data a trade-off between a generative and a discriminative model is found which optimizes the classification performance.

2.5.3 Discriminative Learning Criteria

Over the last years it has been observed for speech recognition that the performance can be improved by discriminatively training the generative HMM [13, 40]. Various objective functions have been proposed and tested to optimize the classification performance of HMMs. The following section provides an overview of the different training criteria.

Conditional Maximum Likelihood (CML) and Maximum Mutual Information (MMI)

For Conditional Maximum Likelihood (CML) training the parameter estimate is found by:

$$\Psi^{CML} = \arg\max_\psi \sum_u P(C^u|\mathbf{O}^u). \tag{2.33}$$

The class labels are conditioned on the input vectors instead of modeling their distribution. In CML the parameters are chosen such that the likelihood gap between correct labellings and incorrect labellings is increased [78]. Maximum Mutual Information (MMI) is derived from information theory rather than decision theory [52] and is closely related to CML. The MMI estimator is given by

$$\Psi^{MMI} = \arg\max_\psi \sum_n \log \frac{P\left(C^n|\mathbf{O}^n\right)}{P\left(C^n\right)P\left(\mathbf{O}^n\right)}. \tag{2.34}$$

If the probabilities $P(C^n)$ are held fixed the same results are obtained as for CML, thus $\Psi^{CML} \equiv \Psi^{MMI}$.

Minimum classification error

Minimum classification error (MCE) training is aimed directly at minimizing the number of sequence misclassification N_{err}:

$$N_{err} = \sum_{u=1}^{U} \text{sign}\left[-\log P\left(\mathbf{O}^u, C^u\right) + \max_{S \neq \mathbf{O}^u} \log P\left(\mathbf{O}^u, S\right)\right] \tag{2.35}$$

where $\text{sign}[z] = 1$ for $z > 0$ and $\text{sign}[z] = 1$ for $z \leq 0$ and S are all possible incorrect labellings for training utterance $\mathbf{O}^u$. The function of the training set accuracy N_{err} in eq. (2.35) is non-differentiable due to the sign and max functions and has therefore to be approximated by a continuous differentiable function in order to use optimization algorithms. N_{err} is commonly approximated with a sigmoid function.

Large Margin Training

Large Margin Training is a novel framework for discriminative learning based on the idea of margin maximization inspired by support vector machines. The framework tries to maximize the distance between labeled examples and the decision boundaries that separate different classes [78]. On sample datasets the method outperformed the above mentioned methods.

Kernel Methods

An alternative technique to boost the classification performance of HMMs is to combine them with discriminative methods, like support vector machines (SVM). SVM employ a kernel mapping function to transform input features to a high-dimensional feature space in which the data points can be better separated than in the input space. For sequence classification it is necessary to use Kernels that can map variable-length sequences to a fixed-dimensional feature space. The first framework which combined the HMMs ability to process variable-length sequences and the generalization performance of kernel based machines was presented in [36]. To improve the discrimination ability of generative models the examples are mapped into the log-likelihood gradient-space, known as the Fisher score-space Φ^F:

$$\Phi^F(O|\Psi) = \nabla_\Psi \log P\left(\mathbf{O} \,|\Psi\right) \tag{2.36}$$

where $P\left(\mathbf{O} \,|\Psi\right)$ is the probability that the observation sequence $\mathbf{O}$ was generated with the HMM with model parameters Ψ. In eq. (2.36) only one common base model is trained for all observations independent of the class label. To further improve the classification performance separate HMMs can be trained for each class and also the log-likelihood ratio can be included in the score space [81]. For a two class problem with class specific HMM parameters Ψ_1 and Ψ_2 the generative score space is given by:

$$\Phi^{LL}\left(\mathbf{O} \,|\Psi_1, \Psi_2\right) = \begin{bmatrix} \log P\left(\mathbf{O} \,|\Psi_1\right) - \log P\left(\mathbf{O} \,|\Psi_2\right) \\ \nabla_{\Psi_1} \log P\left(\mathbf{O} \,|\Psi_1\right) \\ \nabla_{\Psi_2} \log P\left(\mathbf{O} \,|\Psi_2\right) \end{bmatrix} \tag{2.37}$$

It was shown that the score space can increase the performances if only few data samples are available and a simple generative model had to be trained [81].

2.6 Gaussian Linear State-Space Models

Besides HMMs with their finite state space, Gaussian linear state-space models form another important class of state-space models for which algorithms for state inference are mathematically tractable. The transition and observation functions are linear Gaussians:

$$X_{k+1} \,|\, X_k = x_k, U_k = u_k \quad \sim \quad \mathcal{N}\left(x_{k+1}; A_k x_k, \Sigma_k^U\right) \tag{2.38}$$

$$Y_k \,|\, X_k = x_k, V_k = v_k \quad \sim \quad \mathcal{N}\left(y_k; B_k x_k, \Sigma_k^V\right). \tag{2.39}$$

where A_k is referred to as system model, B_k as observation model and the random vectors U_k and V_k are called state evolution and observation noise, respectively. Both noise sources are assumed to be Gaussian distributed with zero mean and covariance matrices Σ_k^U and Σ_k^V, respectively. The functional representation of a Gaussian linear state space is given by:

$$X_{k+1} \quad = \quad A_k X_k + U_k \tag{2.40}$$

$$Y_k \quad = \quad B_k X_k + V_k \tag{2.41}$$

The state variable X_0 for $k = 0$ is also assumed to be Gaussian distributed with zero mean and covariance Σ_0 and independent of the noise processes U_k and V_k. Note that A_k corresponds to the state transition probability matrix A with elements $a_{ij} = P(Q_{k+1} = j | Q_k = i)$ for HMMs and B_k is equivalent to the parametric or discrete probability distribution of the observations $b_i(o_k)$ given state $Q_k = i$ for HMMs. As for HMMs we assume in the following that A_k, B_k, Σ_k^U and Σ_k^V are time-independent.

2.6.1 Kalman Filter

In order to estimate the state X_k from the sequence of noisy observations Y_k for a Gaussian linear state-space model, the Kalman Filter is employed. In an analogous manner to the forward equations for the HMM presented in section 2.3, the Kalman filter equations can be obtained [60][4]. The forward computations can be split in two parts which are evaluated iteratively. Firstly, the state prediction which uses the information of the system model:

$$P\left(x_{k+1} \,|\, y_1, \ldots, y_k\right) = \int P\left(x_{k+1} \,|\, x_k\right) P\left(x_k \,|\, y_1, \ldots, y_k\right) \mathrm{d}x_k \tag{2.42}$$

and the measurement update which uses the information of the observation model:

$$P\left(x_{k+1} \,|\, y_1, \ldots, y_{k+1}\right) = \frac{P\left(y_{k+1} | x_{k+1}\right)}{P\left(y_{k+1} \,|\, y_1, \ldots, y_k\right)} P\left(x_{k+1} \,|\, y_1, \ldots, y_k\right). \tag{2.43}$$

Since any distribution at any time step is Gaussian for linear Gaussian models, the probability distributions can be parametrized simply by their mean and

---

[4]From the backward operator the Kalman smoother is derived.

covariance matrix and the Kalman filter recursions described with eqs. (2.42) and (2.43) simplify to:

Prediction

$$\hat{x}_{k+1|k} \quad = \quad A\hat{x}_{k|k} \tag{2.44}$$

$$P_{k+1|k} \quad = \quad AP_{k|k}A^T + \Sigma_k^U \tag{2.45}$$

Kalman Gain

$$K_{k+1} \quad = \quad P_{k+1|k}B_{k+1}^T \left(B_{k+1}P_{k+1|k}B_{k+1}^T + \Sigma_k^V \right) \tag{2.46}$$

Update

$$\hat{x}_{k+1|k+1} \quad = \quad \hat{x}_{k+1|k} + K_{k+1} \left(y_{k+1} - B_{k+1}\hat{x}_{k+1|k} \right) \tag{2.47}$$

$$P_{k+1|k+1} \quad = \quad (I - K_{k+1}B_{k+1}) P_{k+1|k} \tag{2.48}$$

where $\hat{x}_{k+1|k}$ (eq. (2.44)) and $P_{k+1|k}$ (eq. (2.45)) are the predicted mean and co-variance matrix of the Gaussian distributed state vector for time instance $k+1$ using the information of the state vector X_k and the system model A, respectively. K_{k+1} (eq. (2.46)) is called the Kalman gain and $\hat{x}_{k+1|k+1}$ (eq. (2.47)) and $P_{k+1|k+1}$ (eq. (2.48)) are the updated mean and covariance matrix of the state vector X_{k+1} using the information of the observation vector y_{k+1}. The expected belief $\hat{x}_{k+1|k+1}$ is equal to the predicted value $\hat{x}_{k+1|k}$ plus a weighted error term $K_{k+1} \left(y_{k+1} - B_{k+1}\hat{x}_{k+1|k} \right)$, where the Kalman gain K_{k+1} depends on the ratio of prior uncertainty to the uncertainty in the measurements. For a detailed derivation of the results see e.g. [58].

2.6.2 Tracking Algorithms Overview

Kalman filters have been used extensively in tracking and control applications. If the highly restrictive assumptions (linearity and Gaussianity) hold, no other algorithm can do better than the Kalman filter. The major advantage of a Kalman filter is its ease of computation and memory efficiency; both the prediction and update can be implemented with matrix multiplications (see eqs. (2.44) to (2.48)). Kalman Filters are most favorable if the state uncertainty is not too high, meaning that accurate sensors or senors with high update rates are available [27]. One possibility to handle tracking for non-linear systems is to approximate non-linear state prediction and measurement equations by a Taylor expansion. After linearization the regular Kalman Filter can be used; the combination of linearization and tracking is referred to as the extended Kalman Filter.

The main disadvantage of the regular and extended Kalman Filter is that the belief (posterior probability distribution of the state vector) is unimodal. Presence of complex motion, self-occlusions or multiple objects during tracking results in multi-modal state distributions which cannot be handled with a single Kalman Filter. A classical technique to overcome this shortcoming is multi hypothesis tracking (MHT) which models the beliefs with mixtures of Gaussians. MHT maintains a bank of Kalman Filters, where each hypothesis belongs to a separate Kalman filter. Then elaborate techniques or heuristics are needed to

decide when to add or delete a hypothesis and to match the different hypothesis to trajectories (correspondence problem). Due to the correspondence problem, MHT is computationally much more expensive than a regular Kalman Filter and in general prior assumptions are made in order to narrow down the number of possible assignments of hypotheses to trajectories.

The Particle Filter is the most general tracking approach and a very flexible tool. Particle filters approximate the density directly as a finite number of samples using sequential importance sampling with re-sampling [5]. Hence, they are not limited to a specific parametric belief distribution and can handle non-linear systems. The worst-case complexity of particle tracking grows exponentially with the dimension of the state-vector and hence, high-dimensional problems should be tackled with caution.

3 Quality Monitoring of Industrial Laser Welding Processes

The aim of this chapter is to give a brief introduction to industrial laser welding and camera-based monitoring systems for their in-line quality control. Laser welding is an important and widely used industrial manufacturing process [85]. Although process errors occur rarely, it is vital to ensure that all faulty welds are detected, since errors can lead to a malfunction or a complete outage of the manufactured component part. Therefore quality sensitive welding processes have to be monitored. One possibility to automatically detect welding errors is to monitor the laser welding process with a high-speed camera [44, 59]. When laser radiation interacts with the work piece, secondary radiation is generated. This radiation contains information about the process stability and can thus be used to detect process errors.

The chapter is organized as follows: The industrial needs for diagnostics systems are reviewed in section 3.1. In section 3.2 a brief introduction to laser welding and camera-based monitoring concepts is given. A description of the gathered image sequences with the camera-based sensor system is given in section 3.3.

3.1 Prerequisites for Industrial Process Control

In pattern classification a trade-off between the false positive (FP) rate and false negative (FN) rate has to be made. Depending on the application, the severeness of a FP or FN can vary. Usually this circumstance is accounted for by assigning different costs to a FP and FN and then optimizing the decision threshold by minimizing the assumed cost function [31]. Typically in industrial applications the zero-error policy is enforced which means that a FN is by far more critical than a FP. Fault detection applications share some common design criteria which have to be fulfilled in order to have a monitoring system suitable for a mass production environment [96]:

- Computational Efficiency: The allowed response time of the classification system is limited, i.e. the decision has to be made within a rapid clock cycle.

- Robustness: The system has to be robust to various noise sources and uncertainties. If the noise increases the system should gradually degrade, instead of failing abruptly. Again a trade-off between robustness and performance has to be taken into account.

- Novelty Identifiability & Generalization: A challenging characteristic of process monitoring is that sufficient data is available from the regular process state, but only a few data patterns from abnormal behavior. Hence, the classification framework must have a high generalization ability to be able to learn from the few training error patterns and predict similar future patterns. In addition the system has also to be able to recognize the occurrence of novel, unforeseen fault patterns.

- Adaptability: The process operating conditions can change not only due to disturbances but also due to changing environmental conditions e.g. changes in quality of the raw material. The fault detection system must have the ability to adapt to these changing circumstances. In addition it must be able to incorporate additional knowledge during operation for example from newly detected error events.

- Explanation potential: The outcome of the classification system must be justifiable and interpretable so that the operator can act accordingly.

- Design costs: The training and application of the classification system should be as fast and easy as possible to allow users without expert knowledge to utilize the fault detection system.

3.2 Laser Welding

In recent years, industries increasingly substitute conventional welding processes with laser welding units. The major benefits of laser welding applications are that there is no mechanical contact with the work piece, that a high energy concentration can be achieved and that a high degree of automation is possible [85]. The energy required for welding is generated either with *Nd:YAG* lasers (solid state lasers) operating at a wavelength of 1.06 μm or *CO2* lasers (gas lasers) operating at a wavelength of 10.6 μm. Since the operating wavelength of *Nd:YAG* lasers is in the near infrared spectrum, it is possible to guide the laser beam through a glass fiber cable. The functional principal of laser welding is presented in Fig. 3.1(a). The laser beam is focused onto the two workpieces that have to be joined, to a tiny spot of only tenths of a millimeter achieving a power density of up to 10^6 W/cm^2. In consequence the material of the workpieces is evaporated and a capillary is built, which is referred to as keyhole. Due to the keyhole the laser can penetrate deeper into the material and depending on the used laser power, deep or narrow weld seams can be achieved. Welding lasers can be operated either in pulse mode where short and high powered laser pulses are used (pulse laser welding) or in continuous wave mode (continuous wave (cw) welding). *CO2* lasers are primarily used in the cw mode for applications where a high output and high processing speed is necessary; *Nd:YAG* lasers are employed in both operating modes (pulsed and cw) and are the dominating laser type in the automobile industry.

Despite the advanced development of laser welding applications, the process is still highly dynamic and chaotic and thus vulnerable to process errors such
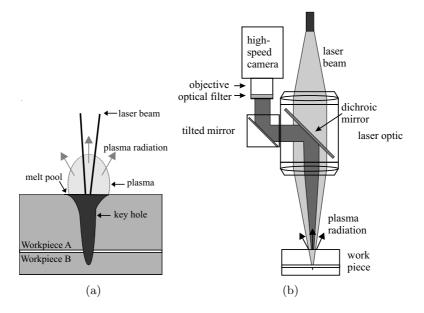
Figure 3.1: Principal of laser welding in (a) and a schematic of a coaxial monitoring setup with a high-speed camera for laser welding processes in (b).

as material sputters, weld break-ins or weld reinforcements. Although process errors occur rarely, it is vital to ensure that all faulty welds are detected, since errors can lead to a malfunction or a complete outage of the manufactured component part. Therefore quality sensitive welding processes have to be monitored. The interaction between the laser beam and the work piece leads to the generation of plasma (see Fig. 3.1(a)). It has been observed that the radiation of the laser induced plasma contain information about the process stability and can thus be used to identify process errors [85]. Several groups presented algorithms and monitoring setups to automatically classify or control laser welding processes based on the laser induced plasma emissions [1, 15–17, 31, 39, 44, 59, 61, 63, 84]. A commonly used constellation to on-line monitor the laser welding process is to record the radiation of the plasma with high speed CMOS (Complementary Metal Oxide Semiconductor) imaging sensors [44, 59]. A high frame rate of the imaging sensor is necessary for monitoring laser welding processes, since changes in the plasma can last only a few milliseconds. A general review of different sensor concepts for on-line monitoring of laser welding can be found in [79, 85]

3.2.1 Monitoring Setup

Fig. 3.1(b) presents a coaxial monitoring setup for laser welding. The laser induced plasma radiation is focused onto the imaging senor of the high-speed camera using standard optical components. An optical filter is used in front of the CMOS camera to block out possibly back reflected laser light from the component part. During the welding process, the cylindrical part (work piece)

is rotated under the laser beam and thus the position of the radiation of the laser induced plasma in the recorded frames does not change over time. A typical image of the plasma radiation in case of a regular welding process is shown in Fig. 3.2(a). To simplify the terminology the recorded radiation from the plasma is referred to as the *melt pool*, although this is physically not correct, since the laser-induced plasma radiation does not coincide with the melt pool[1]. Compared to integral sensors, like photo diodes, spatially resolved sensors have the advantage that possible error occurrences such as melt pool deformations (see Fig. 3.2(c)) or material sputters (see Fig. 6.1) can be better recognized due to the higher information content of the recorded observations.

The high-speed camera in the coaxial setup presented in Fig. 3.1 records the plasma radiation from the same direction as the laser beam is directed onto the part. Other possibilities are to monitor the process from the opposite direction or any other off axis position if structurally possible. Combining the output of various sensors can increase the fault detection performance but with respect to hardware, engineering and maintenance costs the number of sensors should be kept as low as possible [86].

3.2.2 Sensor Optimization

Examinations for a welding process, where process monitoring from the opposite direction was possible, showed that the image sequences gathered with a coaxial setup have approximately a 43% higher information content to detect errors than image sequences gathered with the camera from the opposite direction. Compared to an integral sensor (photo diode) the image sequences of a high speed camera have approximately a 30% higher information content[2]. Although this result is based on an empirical study for only one laser welding process and the exact numbers are always application depedent, the overall trend coincides with the experience from other welding processes.

3.2.3 Technical Details

For the experimental data presented in this thesis the laser welding processes are monitored with a fast PhotonFocus^{TM} CMOS camera. State of the art CMOS cameras offer a frame rate of up to 150 frames-per-second (fps) with a frame size of 1024×1024 pixels [64]. CMOS cameras allow to read out only part of the camera image, enabling to trade spatial for temporal resolution; for instance, when restricting the region-of-interest (ROI) to 40×40 pixels, up to 40000 fps can be acquired with the above mentioned performance. A detailed introduction to CMOS imaging sensors can be e.g. found in [28]. In most welding applications a spatial resolution between 40×40 and 64×64 pixels

---

[1]A direct observation of the melt pool by means of the laser-induced plasma radiation is for example possible in a narrow wavelength band in a coaxially aligned sensor setup [59].

[2]To ensure an algorithm independent evaluation of the different sensor settings, the observations were screened by an expert and evaluated if and how clearly the irregularities can be observed in the image sequences on a scale from 0 (no error observable) to 3 (error event spatially and temporally distinct in raw sequence) [86].

<table>
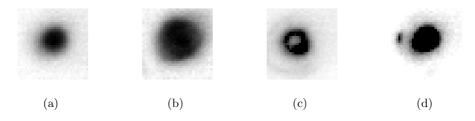<tr><td>(a)</td><td>(b)</td><td>(c)</td><td>(d)</td></tr>
</table>

Figure 3.2: Example images of the welding process as recorded by a CMOS camera (inverted colormap): Recorded laser induced plasma radiation (melt pool) of a regular welding process in (a) and recorded intensity distributions of error events (b)-(d).

is sufficient and the temporal resolution is then limited by the intensity of the back scattered radiation from the laser induced plasma which prevents exposure times below a certain limit. An additional advantage of CMOS cameras is the high dynamic range which can be up to 120 dB using a logarithmic characteristic curve. The temperature gradient of the plasma during welding is approximately 10000 K/mm, resulting in high intensity differences between the pixels located in the middle of the melt pool and pixels located outside.

3.2.4 Sensory Gap

Naturally, there exists a difference between the physical scene and the observation of the scene made by the sensor system [29]. This so called sensory gap can also complicate the camera based quality control of laser welding processes. For instance a CMOS camera is sensitive in the spectral range of approximately 350 nm to 1000 nm. Yet, this spectral range is too narrow to capture all the information of the laser-induced plasma radiation. and only limited information about the physical process is obtained. Due to the limited information it can happen that some weld seam errors are not clearly observable in the recorded image sequences. The consequences are that the classification system must react more sensitive to irregularities and this in turn reduces the precision of the system. The use of several sensors, where each sensor monitors different aspects of a welding process is one possibility to minimize the sensory gap [86].

3.3 Data Description

Error events can be distinguished from regular frames (see Fig. 3.2(a)) in terms of their spatial and temporal appearance. Typical deformations which indicate weld seam faults for a coaxial setup are presented in Fig. 3.2(b) to Fig. 3.2(d). Class labels are only available for the complete sequence (sequence labels) and specify the quality outcome of the produced part (either error free or erroneous). The quality of the produced part is established by visual examination and the part can be associated to the corresponding sequence by an unique identification number. The visual examination of the parts by experts is regarded as the

ground truth. It is time consuming to obtain labels for each individual frame within a sequence (frame labels) due to the large amount of recorded data. In cw laser welding applications, an image sequence comprises typically between 4000 to 12000 images.

4 Dimension Reduction

The images I_k of the recorded sequences can be considered as high-dimensional feature vectors. Thus, the number of features is much larger than the number of observations making it necessary to compress the relevant information of each recorded image I_k to a m-dimensional observation vector o_k before training a classifier. The goal in dimension reduction is to describe the high-dimensional input feature vectors I_k such that the resulting low-dimensional feature vectors o_k are similar for images from the same class and are very different for images from different classes [24]. Dimension reduction is often followed by a feature selection step, since in practice relevant features which can best distinguish between different classes are not known a priori. In this chapter object- and appearance-based features for the quality control of laser welding processes are introduced. While object features are a widely employed technique for dimension reduction in laser welding application, it is the first time that appearance-based features using subspace methods adapted from face recognition are employed.

The chapter is organized as follows: After a short introduction in section 4.1, object features and appearance-based for laser welding processes are presented in section 4.2 and section 4.3, respectively. The application of subspaces methods to images from laser welding and practical issues are discussed in section 4.4. A summary is given in section 4.5.

4.1 Introduction

A widely used approach to dimension reduction of images is to isolate relevant objects from the image background and compute features reflecting meaningful properties of the segmented objects e.g. using geometric-based or moment-based features. The challenge in the dimension reduction step for images recorded from laser welding processes is that the welding process is highly dynamic and possible erroneous deformations of the recorded melt pools are manifold. Furthermore error events are rare, typically true negative rates of 0.1% to 0.5% can be expected on the long run and it is thus not guaranteed that all possible error occurrences have been observed during the training phase. Under these prerequisites it is difficult to engineer proper features.

To improve the classification performance and to make the detection system more independent from human intuition, ideas from face recognition are adapted to laser welding in this chapter. Relevant characteristics of the recorded images are found automatically by machine learning from a large training set of images from regular welding sequences. Principal component analysis (PCA) is performed directly on the recorded pixel intensities and the computed prin-

cipal component loadings (PCs) are used as features for a subsequent sequence classification. The subspace is computed only from melt pool images belonging to regular welding sequences. The variability of the data, which cannot be explained with the training data set, can then be used as an additional feature. With this feature it is possible to detect deformations which have not been observed during training but significantly deviate from the regular appearances of melt pools (measure of novelty).

4.2 Object Features

Typically dimension reduction in laser welding applications is performed using geometric information about detected objects. The objects are either segmented from the noisy background of the recorded images from the laser-induced radiation (e.g. the melt pool) [59] or the image are preprocessed with a temporal change detection algorithm to focus the attention on the occurrence of abnormal objects (e.g. sputter events) [17].

Object Detection

Segmentation of objects on the recorded images is simply performed using an adaptive threshold depending on the mean gray value of the complete image sequence. To focus the attention on suspicious objects, a temporal change detection algorithm can be used. In [17] a change detection algorithm is proposed for laser welding sequences which pre-processes each pixel $I_k(x_p, y_p)$ of image I_k with:

$$I_{k,norm}(x,y) = \frac{|I_k(x,y) - I_{med}(x,y)|}{I_{mad}(x,y)}, \qquad (4.1)$$

where x_p and y_p specify the pixel positions in frame I_k. The robust median operator "med" is used to compute $I_{med}(x_p, y_p)$:

$$I_{med}(x_p, y_p) = med\{I_1(x_p, y_p), I_2(x_p, y_p), \ldots, I_K(x_p, y_p)\} \qquad (4.2)$$

and the robust variation measure "mad" (median absolute deviation) is used to determine $I_{mad}(x_p, y_p)$:

$$I_{mad}(x_p, y_p) = med\left\{|\bar{I}_1(x_p, y_p)|, |\bar{I}_2(x_p, y_p)|, \ldots, |\bar{I}_K(x_p, y_p)|\right\}. \qquad (4.3)$$

where $\bar{I}_k(x_p, y_p) = I_k(x_p, y_p) - I_{med}(x_p, y_p)$ and K is the total number of images belonging to one sequence. Robust estimation of the mean and variance image are beneficial due to outliers caused by error events. The normalization in eq. (4.1) helps avoid the use of absolute intensities and thus makes the system more robust to variations in the weld materials, sensor degradation etc. A binary image is obtained from $I_{k,norm}(x_p, y_p)$ by thresholding.

$$I_{bin,k}(x,y) = \begin{cases} 0 & \text{for } I_{k,norm}(x,y) < th_{bin} \\ 1 & \text{for } I_{k,norm}(x,y) \geq th_{bin} \end{cases} \qquad (4.4)$$

where th_{bin} specifies the applied threshold. Detailed information of change detection toward the detection of sputter events can be found in chapter 6.

Geometric Feature-Based Approach

The found objects are characterized with simple features such as solidity, compactness ς, area A, perimeter p or mean grayvalue GV_m[1]. The advantage of geometric-based features for the use in industrial applications is that they are easily interpretable and domain knowledge can be incorporated without the need for large train data sets. But human intuition is necessary in order to narrow down the huge amount of theoretically possible features to a subset from which a classification system can automatically select the relevant ones. Automatic feature selection for laser welding applications with shape-based and moment-based features for objects flagged with a change detection algorithm is considered in [31].

For instance to detect melt pool deformations the necessary features can be divided in two groups regarding their ability to detect different irregularities: i) features describing changes in the size of the melt pool (such as area or mean brightness) see Fig. 3.2(b) for an error type of this kind and ii) features which can detect deformations (such as solidity or compactness) see Fig. 3.2(c) for an error type of this kind. For example the feature compactness ς is defined as:

$$\varsigma = \frac{(perimeter)^2}{4\pi\,(area)} \tag{4.5}$$

and has its minimum value for a disc ($\varsigma = 1$), for discrepancies from a disc, i.e. deformations of the disc-shaped melt pool, ς increases.

4.3 Appearance-Based Features

Over the last years appearance-based approaches which exploit the intrinsic low-dimensionality of images have become a standard tool for object detection and recognition. The methods work directly on the image-based representation i.e. pixel intensity array I_k without the need of segmentation. It has been first found in face recognition application that subspace methods outperform traditional techniques which used geometric information to describe facial characteristics [80, 92]. Due to the disc-shape of the melt pools, feature extraction in laser welding applications can be related to spot detection applications. Appearance based spot detection has been for instance successfully employed for the detection of ancient settlement mounds [55] or for rain drop detection for driver assistance applications [46].

---

[1]See e.g. [37] for an overview on possible geometric-based features for feature extraction.

4.3.1 Preprocessing

Since the observed gray values are directly used for subspace computation, subspace methods are sensitive to translations and rotations of the recorded melt pools. To achieve translation invariance a sub-window is extracted from the recorded images. The melt pool can be fully observed in this sub-window and, since only part of the recorded image is used, minor translations can be compensated. Within one sequence the melt pool is not expected to change its position (unless in case of an error event), but between different sequences position invariance has to be ensured[2]. The recorded melt pools are generally rotation invariant and no major changes are expected in their size or brightness during a welding sequence unless in case of an error event. Laser parameters can slightly change in laser welding applications (for example if the laser focus has to be slightly adjusted) and this also effects the intensity of the recorded images. The complete image sequence are therefore normalized to zero mean and unit variance. The mean and variance are estimated over all extracted sub windows belonging to one welding sequence. Normalization of each individual image is not possible, since this might result in false negatives.

4.3.2 Subspace Computation

Principal component analysis (PCA) is one of the most widely used tools for data reduction [49]. It decomposes the high-dimensional feature space into a principal subspace F and an orthogonal complementary space $\overline{F}$, as shown in Fig. 4.1(a). The residual error ϵ, also called distance from feature space (DFFS) in the context of the work with *eigenfaces*, is the Euclidean distance of a point in the high-dimensional feature space from the principal subspace F. The DFFS signal ϵ increases for images that are "far" from the images used to compute the principal subspace and is therefore a measure of novelty [57].

The training set consists of several images I_k sampled from different error free welding sequences. From the training images I_k, vectors I_k' are formed by lexicographic ordering of the pixel elements of each image. The result is matrix $\mathcal{I}$, where each column represents an image vector I_k'. The PCA method finds its directions ϕ_i with $i = 1, 2, \ldots, N_I$ where $N_I = \dim(I_k')$ in $\mathcal{I}$ space that have high variance [95]. The first PCA ϕ_1 direction

$$\phi_1 = \arg\max_{\|\phi\|=1} \text{var}\left(\mathcal{I}\phi\right),\tag{4.6}$$

is a linear combination of $\mathcal{I}$ that has maximum variance; the second PCA direction ϕ_2 is orthogonal to the first one and has the second highest variance and so on. The complete matrix of PCA directions Φ can be found by eigenvalue decomposition:

$$\Lambda = \Phi^T C \Phi\tag{4.7}$$

---

[2]If the melt pool is not stationary during a welding sequence, it can be easily tracked and shift invariance between different images can be achieved. No advanced tracking algorithm is necessary to perform the task, since the melt pool can be defined simply as the largest segmented object in each frame.

where Φ^T is the transpose of matrix Φ, $C = \mathcal{I}^T \mathcal{I}$ is the covariance matrix and Λ is the matrix of eigenvalues λ_i corresponding to direction ϕ_i.

To obtain the principal component feature vector $y_{k,pc}$ the high-dimensional image vector I'_k is projected onto the eigenvector directions. To reduce the dimension the images are projected on the $M_{pc} < N_I$ eigenvectors corresponding to the M_{pc} largest eigenvalues, only:

$$y_{k,pc} = \Phi^T_{M_{pc}} \widetilde{I}'_k \tag{4.8}$$

where $\Phi_{M_{pc}}$ is a submatrix of Φ containing the M_{pc} largest-eigenvalue eigenvectors and $\widetilde{I}'_k = I'_k - \overline{I'}$ is the center image vector. The mean vector $\overline{I'}$ has been obtained from the training data set. The eigenvalue corresponding to each eigenvector is equal to the variance of the data, when the data is projected onto the one-dimensional space corresponding to that eigenvector. Thus the amount of variance which is captured by the M_{pc}-dimensional subspace is equal to the sum of the M_{pc} highest eigenvalues. A typical spectrum of eigenvalues for the dynamic of a regular welding process is presented in Fig. 4.1(b). Approximately 2.1% of the PCs ($M_{pc} = 20$, $N_I = 961$) capture

$$\frac{\sum_{m=1}^{M_{pc}} \lambda_m}{\sum_{m=1}^{N_I} \lambda_m} = 98.5\% \tag{4.9}$$

of the data variance. The first three PCs already cover approximately 81% of the data variance. Since the subspace is computed from regular melt pool appearances only, the residual reconstruction error ϵ can be interpreted as a measure of novelty:

$$\epsilon^2(\widetilde{I}'_k) = \left\| \widetilde{I}'_k \right\|^2 - \left\| \Phi^T_{M_{pc}} \widetilde{I}'_k \right\|^2. \tag{4.10}$$

4.3.3 Limitations and Extensions of PCA

PCA approximates the data in terms of a single multivariate Gaussian distribution, hence only one and second order statistical dependencies of the pixels can be considered. For complex objects such as faces it is often not be possible to capture the important information for recognition or discrimination with the covariance matrix and therefore extensions of PCA have been researched [49]. Independent Component Analysis (ICA) is one possible method to take into account higher order pixel dependencies. Like PCA, ICA linearly projects the data set onto a basis, but the basis functions are not limited to Gaussian distributions. A further possibility to capture non-linearity is to employ Kernel methods [74]. Prior to the component analysis the input images are non-linearly mapped to a high-dimensional space and efficient computation is enabled using the Kernel trick. ICA has been tested on melt pool images and no significant difference could be observed compared to the results of PCA. It seems that second order statistics are sufficient to describe the properties of the disc-shaped
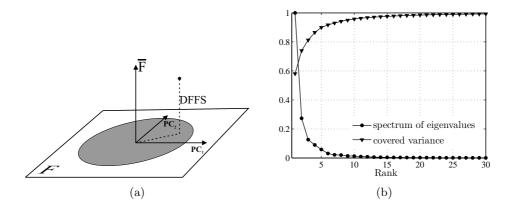
Figure 4.1: Decomposition of the feature space in the principal subspace F and an orthogonal complementary space $\overline{F}$ in (a) and spectrum of eigenvalues λ and percentage of captured variance for a regular welding sequence in (b).

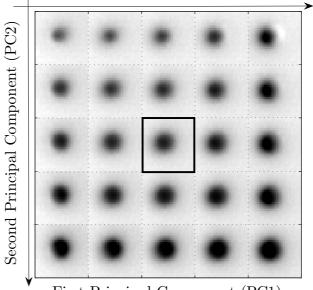melt pools which are compared to faces simple objects.

A second limitation of PCA is that it is an unsupervised method. In general this limits the classification system since not all available information is employed for direction choosing. A prominent supervised subspace method is Fisher's linear discriminant analysis (LDA). Fisher's LDA tries to find the direction such that the ratio of the between-class scatter and the within-class scatter is maximized [24]. Fisher's LDA can improve the performance of face recognition systems in particular under varying lightning directions or facial expressions [10]. For process monitoring applications it is beneficial to restrict the subspace computation to negative examples and utilize an unsupervised technique. It is then not necessary to gather a large number of positive examples which is difficult due to the unequal class proportions.

4.4 Eigen-MeltPools

The computed eigenvectors from the training data set are, following the notation in face recognition, referred to as *Eigen-MeltPools*. The *Eigen-MeltPools*, that correspond to the first four largest eigenvalue-eigenvectors for a typical welding process, are shown in Fig. 4.2. The *Eigen-MeltPools* span the principal subspace F, which is called the *spotspace*. A representation of the *spotspace* for the dynamic of regular melt pool appearances is presented in Fig. 4.3. The figure visualizes the possible melt pool variations captured with the first two principal components. The images are generated by varying the values of the first two PCs and setting all other PC components to zero. The first two PCs describe the allowed increase and decrease of the melt pool and represent.

Figure 4.2: Mean image $\bar{I}$ of the melt pool of the training data set, followed by the first 4 eigenvectors ϕ_i ("Eigen-MeltPools") describing the principal deformations of a regular weld sequence.

Figure 4.3: Representation of the *spotspace*. Melt pools are generated by varying the values of the first and second principal components (PC). The values for the principal components increase from top to bottom ($PC1$) and from left to right ($PC2$). The boxed melt pool represents simply the mean image with $PC1 = PC2 = 0$.

4.4.1 Practical Issues

If the number of *Eigen-MeltPools* M_{pc} is chosen too high, the reconstruction error ϵ does not carry any information since abnormal deformations are also captured with the *Eigen-MeltPools*. If M_{pc} is chosen too low the dynamic of the regular welding process is not completely captured and the reconstruction error ϵ is also high for regular sequences. The right choice of M_{pc} preserves the major linear correlations in the data and discards the minor ones. It was found empirically that optimal values for M_{pc} vary between 10 and 20 capturing approximately 98% to 99% of the variance of regular welding sequences [67]. In Fig. 4.4 the decompositions of one error free and two erroneous images into the principal and residual subspace are presented. The residual images correctly detect the abnormal deviations for the erroneous images whereas the error free image can be correctly reconstructed (low reconstruction error).

In addition to the DFFS it is necessary to use the information of the PCs in order to detect abnormalities within the *spotspace* (see experimental results presented in section 5.7.2). For example a possible welding error might be that the melt pool size decreases during welding without significant deformations. This is a normal behavior at the beginning and the end of a welding process (therefore it cannot be found with the DFFS) but during the welding this indicates an error event.

4.5 Summary

We note that in this chapter for the first time the *Eigenface* method from face recognition was proposed for automatic, dimension reduction of images recorded from laser welding processes. Results for experimental data and a comparison to geometric-features are presented in section 5.7.5.

The DFFS signal can detect a wide range of deformations and is thus a very good choice for a feature in a hierarchical classification system. The DFFS can be used to focus the attention to suspicious frames. For flagged frames computationally more demanding features can be calculated to decide weather an error occurred or not.

The disadvantage of the PCA method compared to the use of geometric information for dimension reduction is that it is less general and also more sensitive to variations of the welding process. For example if major changes in the laser parameters are undertaken, the *Eigen-MeltPools* have to be recomputed in order to obtain a good performance, but this can be automatic.

Figure 4.4: First row, images (a) to (c): Original melt pools; Second row, images (d) to (f): Projected melt pool in the *spotspace*; Third row, images (g) to (i): Projected melt pool in the residual $\overline{F} - Space$. First column: Error free melt pool; Second column: *Donut* shaped melt pool (erroneous weld); Third column: Deformation of melt pool (erroneous weld). For the third row dark pixel values indicate a high deviation from the original image. The brightness scale of the images is equal for all images in one row.

5 Weakly Supervised Learning of a Classifier for Unusual Event Detection

For efficient industrial manufacturing processes errors occur only rarely. This lack of negative examples complicates the training of automated diagnostic systems that rely on statistical learning. In this chapter a novel pattern recognition framework to detect anomalies in image sequences recorded from industrial laser welding processes with high speed cameras is presented. Although the framework is optimized for laser welding, the introduced concepts can be applied for the quality control of other manufacturing processes.

The classification of the recorded image sequences is based on state-space models which are used to capture the temporal behavior of the dynamic manufacturing process. Sensor observations and prior knowledge of the problem are the main sources of information and both can be combined by using state-space models [33]. Besides the imbalanced data set, another challenge in training a classification system for the considered application is that by default only labels for image sequences are available. Further information about the position or length of error events within sequences is not available.

The chapter is organized as follows: Section 5.1 presents related work for unusual event detection in the field of video content retrieval and previous work on state-space models for industrial process monitoring. The incremental learning framework for unusual event detection which can handle weakly labeled data is introduced in section 5.2 and implementation details are presented in section 5.3 to section 5.5. Experimental results on representative datasets from real world manufacturing processes are presented in section 5.7 and discussions are offered in section 5.8.

5.1 Related Work

In-line process monitoring of manufacturing processes with imaging sensors is closely related to the detection of unusual activity in video sequences. In particular, due to the increasing commercial interest in surveillance applications, unusual event detection in video sequences is a topic of active research [3, 103]. In the following state of the art techniques for unusual event detection with HMMs are reviewed and the similarities and differences of finding anomalies in video sequences describing human activities and image sequences recorded from a manufacturing process are identified. In addition previous work on industrial process monitoring with state-space models is discussed.

5.1.1 Unusual Event Detection in Video Sequences

Typical properties of unusual event detection in online surveillance applications are that the data set is large, i.e. cannot easily be screened by a human, that unusual events are rare and that it is hard to describe them a priori, since they are rare and in addition can be manifold in their appearance. These are all properties which are equally true for industrial process monitoring applications.

A widespread approach to detect unusual activity in video sequences is to employ probabilistic models, like HMMs [3, 23, 83, 100, 101, 103]. In order to detect unusual events, the video sequences are matched against regular models and those sequences which do not agree with the model are defined as unusual. Model-based approaches can be quite effective where the regular activity can be described well [102]. This is not or rarely the case for real-life video sequences, where the normal activity can vary even more than unusual events, but agrees well with monitoring applications for the in-line quality control of manufacturing processes. For manufacturing processes images of the regular process state belong to a narrower image domain than e.g. video sequences of usual human activity. Possible variations can be caused by parameter changes or changes in the used materials, but overall these variation are relatively small since mass production manufacturing processes must be highly repeatable.

Since labeled data sets are expensive and weakly labeled and particularly unlabeled data is cheap, unsupervised and semi-supervised methods for unusual event detection using HMMs have been considered. General techniques for the unsupervised discovery of structure in unlabeled sequence data based on HMMs are addressed in [2, 35, 56]. In [102] a framework for the unsupervised detection of unusual events adopting ideas from document-keyword analysis is presented. The video sequences are split into overlapping time segments, a similarity measure between the segments based on a co-occurrence matrix is defined and outlier clusters are considered as unusual events. The work which is conceptually closest to the work presented in this thesis is [100, 101], where a semi-supervised learning technique for unusual event detection in the context of audio-visual meeting analysis is proposed. The training procedure is iterative, where in each iteration the observations of the test event with the lowest likelihood under the regular model is used to add an additional state to the regular model. The result of the training procedure is a tree-structured state-space model. The work described in this thesis is related but uses a different selection metric and different stopping criteria to add additional states during the incremental learning procedure. Here, the selection metric is based on a change detection algorithm on the posterior frame probabilities in order to be independent from absolute likelihood values. In addition, a discriminative model selection criteria is used which only adds states to the model for unusual events if they help to discriminate between a regular sequence and a sequence containing an unusual event. The study of an optimal selection metric for semi-supervised learning with weakly labeled data in object detection is also considered in [70, 77].

5.1.2 Industrial Process Control

HMMs for the detection of error states in industrial manufacturing processes have been considered e.g. in [6, 8, 82, 98, 104]. [82] addresses the problem of novelty detection of unforeseen events in process control applications. The precision of the system could be increased by modeling the temporal correlations of the sensor signals with HMMs. [6, 98] consider to the task of finding sudden changes in tool wear for machining processes like drilling or milling with HMMs. In [104] principal component analysis and HMM are combined for on-line fault detection for industrial processes and the main contribution of [8] is to use trained HMMs not only for an automatic diagnosis of machining processes but also for their prognostics. The papers do not show a connection to the more general work of unusual event detection and assume fully labeled data sets (except [82] for the detection of unforeseen events).

5.2 System Overview

It can be assumed that the sensor system which monitors the industrial manufacturing process offers a high temporal correlation between the recorded images I_k and thus it is reasonable to capture short-term temporal dynamics of the (manufacturing) process using Hidden Markov Models (HMMs). A major drawback of HMMs for speech recognition is the independence assumption. For HMMs it is assumed that a certain observation vector is independent of previous or following observations vectors given the state [71]. Due to the physical attributes of speech this assumption is not valid for speech recognition. In contrast manufacturing process are less deterministic than speech and the independence assumption is less critical. In general, observation vectors belonging to the same state are not expected to have a strong temporal correlation due to the high dynamic of the process.

The general ideas of the incremental learning framework are explained in the following, and a schematic of the proposed incremental learning system is presented in Fig. 5.1; the implementation details follow in sections 5.3 and 5.4.

5.2.1 Incremental Learning with Weakly Labeled, Imbalanced Data Sets

We are confronted with a highly imbalanced data set, where the regular sequences (negative examples, members of class ω_R) form the overwhelming majority and the sequences containing error events (positive examples, members of class ω_E) constitute the minority class[1]. A block diagram of the training procedure is presented in Fig. 5.1(a). Sequence labels are available before starting

---

[1]Note that the sequences from the error class ω_E can be further subdivided into different error types and for each error type a separate HMM can be trained. To simplify the following exposition, the two-class problem is considered; but it is straightforward to extend the approach to multiple error classes due to the chosen generative classification approach.

(a)

(b)

Figure 5.1: Schematic overview of the training procedure of the classification system (a) and the application to new sequences (b). A generative model for the regular sequences is trained and is used in the next step to locate potentially unusual segments within error sequences by means of a change detection algorithm (outlier detection). The unusual segments found are then used to expand the RSM to an error sequence model (ESM). The log-likelihood ratio of the data given the ESM and RSM is used for the classification decision.

the training procedure and dimension reduction is performed using appearance-based features as described in chapter 4. A subspace is computed from images belonging to regular sequences and the images are projected onto the found directions.

After dimension reduction the image sequences are represented as a series of low-dimensional feature vectors. HMMs with diagonal GMMs are employed to represent the temporal dynamics of the underlying (manufacturing) process and to achieve invariance to varying sequence length. A generative learning approach is opted, since only weak labels (sequence labels) are provided. One can assume that a sequence should contain an error event, if it is put in class ω_E, but further specifics like the position or length of the error event are unknown. Without this additional information the sequences cannot be used for discriminative learning methods [93]. The discriminative classification of complete sequences would be possible, but this would result in a very high-dimensional classification problem and due to limited training data and the curse of dimensionality this approach would not be feasible in practice. A direct estimation of the class boundary, as in discriminative methods, is not possible on a single frame basis without labels or prior knowledge, e.g. the expected fraction of outliers. In contrast, HMMs allow to detect unusual frames in terms of their exceptionally low likelihood given the trained model.

Regular Sequence Model

First a model for the regular sequences with class label ω_R is trained. The number of states of HMMs is an important design parameter; enough states are needed to represent the underlying process. However, in a family of models of increasing complexity, the ones with more parameters (more states) will always allow for a better fit of the data. To avoid over-fitting, it is advisable to verify whether an increase in the number of model parameters is really justified by the improvement of the fit. The HMM parameters are determined via EM and the trained model with the optimal number of parameters (found with an automatic complexity control mechanism) is referred to as regular sequence model (RSM).

Outlier Detection and Model Expansion

The sequences belonging to class ω_E are used to construct an error sequence model (ESM), which is obtained by expanding the RSM with additional states. The state expansion procedure is based on an outlier detection and works basically as follows: The error sequences ω_E are partitioned into non-overlapping segments of fixed size, and the segments are tested for compatibility with the RSM. The test statistic is the computed posterior log-likelihood of a segment given the RSM. Segments which provoke a significant drop in the log-likelihood are marked as outliers. The parameters of the additional states are trained based on these outlier segments. The number of additional states added to the RSM has to be controlled to avoid unnecessarily large models. If adding additional states to the ESM does not increase the log-likelihood ratio between sequences from the regular class ω_R and sequences from the error class ω_E,

the training is stopped. Because further additional states would describe the sequences from both classes ω_R and ω_E equally well and would not enhance the discriminative power.

5.2.2 Classification

A MAP decoder (see section 2.4) is used to assign an unlabeled sequence $\mathbf{O}$ either to the regular sequence class ω_R or error sequence class ω_E. The classification decision is based on the posterior log-likelihood ratio of the data between the ESM and RSM (see Fig. 5.1(b)):

$$\rho(\mathbf{O}) = \log\frac{P(\omega_E|\mathbf{O})}{P(\omega_R|\mathbf{O})} + \log\left[\frac{\chi_E}{\chi_R}\frac{P(\omega_E)}{P(\omega_R)}\right] \tag{5.1}$$

$$= \log\frac{P(\omega_E|\mathbf{O})}{P(\omega_R|\mathbf{O})} + \Delta_{LL} \tag{5.2}$$

where

$$\Delta_{LL} = \log\frac{\chi_E}{\chi_R}\frac{P(\omega_E)}{P(\omega_R)} \tag{5.3}$$

is the ratio of model priors weighted with the non-symmetric cost factors χ_E and χ_R for the sequences containing errors and the error free sequences. If $\rho(\mathbf{O}) > 0$, the MAP decoder predicts class $\omega_{MAP} = \omega_E$, otherwise it predicts $\omega_{MAP} = \omega_R$. The weights χ_E and χ_R are introduced to enable a trade off between the FP rate and the FN rate. Since $P(\omega_E)/P(\omega_R) << 1$, the MAP decoder would almost always select ω_R in order to achieve an overall minimum probability of error. However, a FN is much more severe than a FP in fault detection applications. Note that the ESM is only an extension of the RSM and hence the threshold Δ_{LL} can be interpreted as a measure of the severity of an error event. If the log-likelihood ratio is above an empirical threshold Δ_{LL}, the sequence is marked as erroneous. In the following the implementation of the incremental learning approach is described in more detail.

5.3 Training of the Regular Sequence Model

Instead of immediately training a separate HMM with parameters Ψ_i for each class ω_i, first a model is learned for the regular sequence class ω_R, only. Due to the large number of available negative examples, the parameters Ψ_R for the regular sequence model (RSM) can be estimated with high precision. The feature vector o_k at time k consists of appearance based features, the principal component vector $y_{PC,k}$ and the residual error ϵ_k, as described in section 4.3.The HMM parameters Ψ_R are estimated with the EM algorithm and the number of different states N_Q is optimized using the Bayesian Information Criterion (BIC) [48]:

$$BIC(\widehat{\Psi}) = \log P(\mathbf{O}|\widehat{\Psi}) - \frac{N_P}{2}\log K_D \tag{5.4}$$

where N_P is the number of free model parameters and K_D is the size of the data set. The first term in eq. (5.4) is the likelihood of the data given the model and the second term a penalty for the model complexity. Thus, the BIC criterion tries to select the simplest permissible model, among competing complexities, which still fits the data well (Occams razor). The BIC is estimated using a 5-fold cross validation from the training data. The covariance matrices $\Sigma_{i,m}$ are assumed diagonal and all state transitions are allowed (fully connected or ergodic HMM). For speech recognition, left-to-right models are beneficial since speech has sequential properties. For process monitoring this sequential property cannot be assumed in general, since manufacturing processes show no overall sequential behavior. There are parts like the beginning or end of a process which have clearly a sequential property but in the regular state a sequential property cannot be assumed.

5.4 Expansion of the RSM to the ESM by adding additional states

A block diagram of the ESM training is presented in Fig. 5.3. The parameter training consists in total of three stages: outlier detection, parameter estimation and parameter update.

The first stage in the incremental learning procedure automatically selects the outlier data which is then used to find a first parameter estimate $\widehat{Psi}_{ESM}$ for the additional error states. The selection is based on a temporal change detection algorithm. It can be assumed, that the unusual event is present in only part of the sequence, therefore a significant change in the log-likelihood can be expected (see Fig.5.2(a)). Each weakly labeled sequence is partitioned into non-overlapping, segments of constant size and for each segment $\mathcal{S}_p$ the log-likelihood $L_{\mathcal{S},p}$ is approximated using the forward probability α (see eq. 2.10), where p is an integer with $1 \leq p \leq N_{\mathcal{S}}$ and $N_{\mathcal{S}}$ is the total number of segments within one sequence $\mathbf{O}$. The log-likelihood $L_{\mathcal{S},p}$ for segment $\mathcal{S}_p$ can be approximated with:

$$L_{\mathcal{S},p} = \sum_{k \in \mathcal{S}_p} \sum_{j=1}^{N_Q} \alpha_{j,k} \left(\mathbf{O} \,|\, \Psi_R \right) \tag{5.5}$$

where $k \in \mathcal{S}_p$ specifies the time instances k belonging to segment $\mathcal{S}_p$. A robust temporal change detection is used to flag those segments within a sequence that show an abnormal change in the log-likelihood (see Fig.5.2(b)):

$$\widetilde{L}_{\mathcal{S},p} = \frac{|L_{\mathcal{S},p} - \mathrm{med}_p(L_{\mathcal{S},p})|}{\mathrm{med}_q \left(|L_{\mathcal{S},q} - \mathrm{med}_p(L_{\mathcal{S},p})| \right)}, \tag{5.6}$$

where $\mathrm{med}_p(\mathcal{S},p) = \mathrm{med}\{L_{\mathcal{S},1}, L_{\mathcal{S},2}, \ldots, L_{\mathcal{S},N_{\mathcal{S}}}\}$. Outlier segments are found with an empirical fixed threshold $th_{L_{\mathcal{S}}}$. If $\widetilde{L}_{\mathcal{S},p} > th_{L_{\mathcal{S}}}$, $\mathcal{S}_p$ is identified as incompatible with the RSM and used for the parameter estimation of the error states (see Fig.5.2(b)).
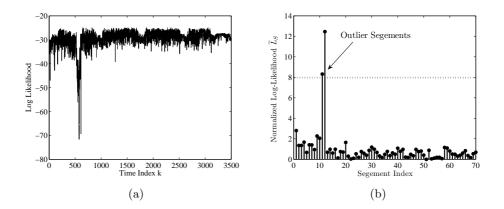
(a)   (b)

Figure 5.2: (a) Logarithmic values of the forward probability α for each frame within a sequence. The dip in the probability around frame 600 is caused by an unusual (error) event. In (b) the normalized log-likelihood $\widetilde{L}_S$ of the segments computed with eq. (5.6) is presented.

The outlier segments are used to estimate the GMM parameters of and transition probabilities a_{ij} between the newly added (error) states. Next, the complete error sequences (including the error free part) are used to estimate the transition probabilities a_{ij} from the newly added states and the states of the RSM. In addition, the parameters of the error states are updated in this second training phase. This enables, that information from unusual segments which have not be flagged in the first stage, can now be incorporate in estimating the model parameters due the semi-supervised learning capabilities of the the EM algorithm [62]. During the training procedure of the ESM, the parameters of the well trained regular states remain unchanged. State transitions are allowed from all states of the RSM to the newly added (error) states, and vice versa. The transition probabilities from the states of the RSM to the newly added (error) states are set to small constant values. Since sequences with unusual events are rare, the transition probabilities a_{ij} are overestimated if they are directly determined from the erroneous sequences. As an alternative to empirically chosen constant values, the transition probabilities estimated by EM can be multiplied with the expected ratio of sequences containing error events to sequences without error events. The transition probabilities from the error states to the states of the RSM describe the mean duration of an error event and can be determined directly from the training sequences.

The number of necessary additional error states $N_{Q,\text{add}}$ is optimized by maximizing the log-likelihood ratio between the data from the sequences containing errors and the regular sequences when using the ESM. Discriminative model selection is studied in [12] and the criterion is referred to as the Discriminative
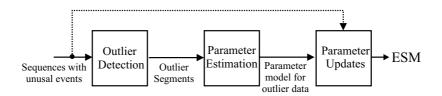
Figure 5.3: Schematic overview of parameter estimation for the error sequence model (ESM) from weakly labeled sequences containing error events.

Information Criterion (DIC):

$$
\begin{aligned}
DIC(\Psi_E) &= \frac{1}{N_E} \sum_{u=1}^{N_E} \log p(\mathbf{O}_E^u | \Psi_{ESM}) - \frac{1}{N_R} \sum_{v=1}^{N_R} \log p(\mathbf{O}_R^v | \Psi_{ESM}) \\
&- \frac{N_{P,\mathrm{add}}}{2} \log K_D
\end{aligned}
\tag{5.7}
$$

where $N_{P,\mathrm{add}}$ is the number of additional parameters necessary to specify the additional states of the ESM and $\mathbf{O}_E^u / \mathbf{O}_R^v$ is the uth/v^{th} training sequence and N_E/N_R the total number of training sequences with/without error events. The DIC ensures that an increase in the number of states for the error model ω_E is not accompanied by an increase of the likelihood for the regular sequences. The DIC decreases if the additional states with their parameters N_add increase the likelihood of both the regular and error sequences to a similar extent. In this case, the new states contain no information that is useful for the discrimination of the two classes. A 5-fold cross-validation is used to estimate the DIC from the training data.

5.5 Fast Feature Selection

For pattern recognition applications feature selection is necessary to optimize the classification performance. During feature selection for each possible feature subset a measure of quality is determined and the feature set with the highest quality measure is chosen. In order to reduce the computation time, it is highly beneficial before starting a feature selection based on wrapper approach to reduce the number of considered features e.g. using a priori knowledge. But often this can only be done to a certain extent. Instead of maximizing a quality measure often a cost function is minimized. The cost function is estimated from a test set for a given feature set. The obtained number of false positives and false negatives from the test set are weighted with different cost scores (reflecting the severity of obtaining FPs and FNs, respectively) and summed up to a cost function. For many applications the time complexity for a complete feature selection is too high, since the training set and test set must be large in order to obtain reliable results. In particular this is valid for HMMs. Thus, an approximation is introduced which can estimate the cost function for each feature set in less time and does not consider all possible feature subsets.

The optimal but computationally most expensive procedure for feature selec-

tion is to determine all possible feature subsets, train for each subset a HMM using the training data set and predict a cost function from the test set. To reduce the processing time the following procedure is applied: ESM and RSM models with all considered features are computed from the training data set. It is assumed that a correct generative model has been found with the used features, but some features are not useful for the classification of unlabeled sequences because they carry no or only little discriminative information. In order to optimize the classification performance and processing time (additional features increase the complexity), those features have to be found and dropped. During feature selection the determined HMM parameters are not re-trained and backward selection is performed to find the best feature set. Backward selection is an iterative procedure, where in each iteration features are removed one after the other from the feature set and the cost function is recomputed for each reduced feature set. Only the reduced feature set with lowest cost is used in the next iteration. The procedure is stopped if the cost function can no longer be reduced by removing features.

The drawback of backward selection is that a dropped feature might reduce the cost function if added to a reduced feature set. To overcome this drawback, feature selection methods exist which combine forward (adding features) and backward selection but these are not considered in this thesis. A complete search of all possible feature subsets is too time consuming. The HMMs are retrained with the optimized feature set since it was observed empirically that the above explained approximation overestimates the cost function.

5.6 Application of Incremental Learning for the Quality Control of Laser Welding Processes

The incremental classification framework is used to detect unusual events in the recorded sequences of laser welding processes as described in chapter 3. Unusual events correlate with faults on the produced weld seam. The experimental data was gathered over a 3 month period from a production line. The welding process was monitored with a high-speed CMOS camera with a rate of 7915 frames per second and a region of interest (ROI) of 64×64 pixel. The recorded weld images correspond to a field of view of approximately $0.9 \times 0.9 \ mm^2$. The manufactured weld seams were visually inspected by experts and matched to the corresponding sequences using an identification number. In total, 96 parts with weld errors were collected and classified in 3 different error classes ω_{E1} to ω_{E3} by visual inspection:

- ω_{E1}: Annealing material particle

- ω_{E2}: Weld reinforcement / weld break-in [2]

- ω_{E3}: General irregularity

---

[2]Since the welding process is controlled, these have similar appearance in the sequences.

The extent of the error on the manufactured part was rated between 1 (weak) to 3 (strong) by visual examination[3]. In addition around 1000 sequences from error free weld processes were collected. The sequences were uniformly sampled over the observation period in order to capture the regular process fluctuations.

Approximately 40% and 80% of the data from regular sequences and erroneous sequences were used to estimate HMM parameters, respectively. The remaining data was reserved for test purposes. The framework was trained using appearance based features and the subspace directions were automatically found from the test data of the regular image squences (majority class). The high-dimensional images were reduced to a time series of only four-dimensional feature vectors which consist of the first three principal components, covering approximately 81% of the data variance of regular processes, and the DFFS signal. The DFFS signal was computed with eq. (4.10) using the $M_{pc} = 20$ first principal components capturing 99% of the variance. The time series of feature vectors were normalized over time for each sequence and for each feature separately, in order to compensate for regular, long-term process drifts. The mean and variance for normalization were estimated from the data between the first and third quartile, to reduce the effect of outliers (which should be detected).

5.6.1 Regular Sequence Model

The obtained BIC values for different model complexities for the training data set are compared in Fig. 5.4. Only for a low number of states ($N_Q < 3$) the additional mixture elements offer a significant increase in the BIC values; for $N_Q < 3$ relevant information has to be modeled with mixture elements. For $N_Q > 4$ additional mixture elements offer no significant increase in the BIC. It is interesting to note that for $M = 1$ the performance cannot be improved by adding additional states. Since diagonal covariance matrices are used, for $M = 1$ correlation between the elements of the feature vectors cannot be modeled. For the current application model complexity was severely penalized to avoid overfitting and high computation times. An additional state was added only if it increased the BIC for more than 5% compared to the BIC for the model without the additional state. The optimal number of states was found to be $N_Q = 4$ and the optimal number of mixture components $M = 2$. The low number of Gaussian mixture components ensures that temporal dynamics are modeled with different states instead of different mixture elements.

5.6.2 Error Sequence Model

So far only information of the regular class was used. In the next step, the information offered by the sequence labels was used to extend the RSM to different ESMs. A separate HMM was trained for each error class ω_{Ei} with $i = 1, 2, 3$. The segment length for the change detection algorithm was set to 50 frames and the threshold to indicate outlier segments according to eq. (5.6) was

---

[3]Each sequence was screened to ensure that the position of the fault on the weld seam and the irregularity in the raw sequence coincide.
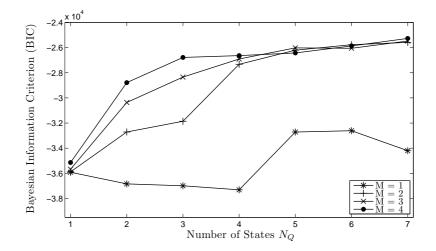
Figure 5.4: Topology optimization for the Regular Sequence Model (RSM) with the Bayesian Information Criterion (BIC).

set to $th_{L_S} = 8$. This is a rather conservative choice and ensures that in the first step of parameter estimation only error events which clearly deviate from regular sequences are used. The ESMs were trained with 80% of the gathered error sequences. The number of additional states $N_{Q,\text{add}}$, optimized with the DIC (see eq. (5.7)) varied between 1 and 2 for the different error classes.

The functional principle of the classification system is shown in Fig. 5.5, where the likelihood of a recorded sequence under the RSM and ESM are presented for a sequence containing an error event. The RSM cannot describe the error event, therefore the log-likelihood values drop, whereas the likelihood values for the ESM decrease less. The difference in likelihood is accumulated and used for the classification decision and the larger the difference in the log-likelihood, the more distinct the error event is. Outside the error event, the likelihood values for the ESM and RSM coincide.

Fig. 5.6 shows a sequence containing an unusual event from error class ω_{E1} (annealing material particle) along with the computed features (PC and DFFS) and the a posteriori log-likelihood ratio $L_{F,k}$:

$$L_{F,k}(\mathbf{O}) = \log \frac{\frac{1}{N_E} \sum_{j \in Q_E} P(Q_k = j | \mathbf{O}, \Psi_E)}{\frac{1}{N_R} \sum_{j \in Q_R} P(Q_k = j | \mathbf{O}, \Psi_R)}, \tag{5.8}$$

where Q_R/N_R and Q_E/N_E are the state index/number of states from an ESM describing the regular part and erroneous part, respectively. The sign of the posterior likelihood ratio indicates if the frame belongs to an error event (positive value) or not (negative value), and the absolute value indicates the confidence in this decision.
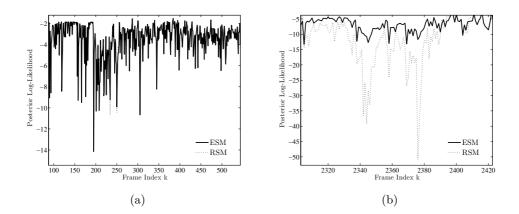
(a)

(b)

Figure 5.5: Frame log-likelihood values of part of a regular sequence (a) and of an error event under the RSM (dashed line) and ESM (solid line) in (b). For the regular part of the sequence, the log-likelihood values for both models coincide. Negative peaks in the log-likelihood under the RSM can be seen in (a); these negative peaks are caused by normal process fluctuations. The ESM follows these peaks and therefore they do not contribute to the classification decision. In (b) the log-likelihood values for the RSM decrease significantly whereas the values for the ESM only show a slight decrease, with this difference indicating an error event.
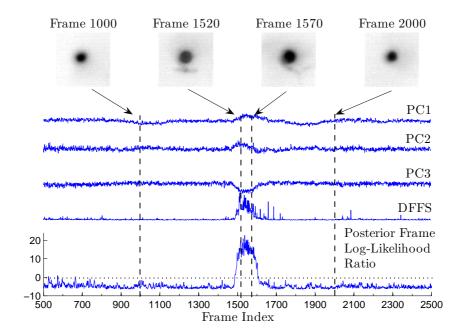
Figure 5.6: Part of a sequence containing an error event (annealing material particle) with some recorded images, used features (PC and DFFS) and computed posterior frame log-likelihood ratio. Frames 1000 and 2000 belong to the error free part of the sequence. The images at frames 1520 and 1570 pertain to the error event and show a burning material particle.

5.6.3 Applicability of Advanced HMM Training Techniques

In section 2.5 it has been reported that the classification performance can be improved by discriminative training methods such as CML, MMI, MCE or score spaces. But in contrast to speech recognition applications reviewed in section 2.5, we have a highly imbalanced data set with only weak class labels. These boundary conditions make discriminative objective functions less attractive. The above described incremental training and model adaptation approach should be more robust to the limited training data for erroneous sequences than directly applying discriminative methods. Score spaces seemed a promising technique to improve the classification performance, since for the ESM only a simple HMM can be trained. Thus, the ESM might be too simplistic and further improvement could be gained from score spaces. Thus after model training, score spaces for the ESMs were constructed. Yet, the benefit in the classification results obtained with score spaces and a SVM compared to a simple MAP decoder was negligible.

5.7 Performance Validation

In this section the classification performance of the above trained and optimized incremental framework for the quality control of laser welding processes is tested. The performance is benchmarked with respect to the following criteria:

- Benefit of using information from erroneous sequences compared to one-class classification (section 5.7.1).

- Comparison of the information content of the principal and the residual subspace (section 5.7.2).

- Comparison to an alternative, discriminative classification approach where temporal information is modeled with a low-pass filter instead of a HMM (section 5.7.3).

- Correlation between the computed likelihood-ratio and the error severity on the produced weld seam (section 5.7.4).

- Comparison of the classification performance between appearance-based and geometric-based features (section 5.7.5).

- Suitability for use in a production environment (section 5.7.6).

The ROC are presented in Figs. 5.7 and 5.9 and the FP rate $FP|_{FN=0}$ and the AUCs are summarized in table 5.1 and for 5.2.

5.7.1 One-Class versus Two-Class Classification

The ROCs presented in Fig. 5.7 show that the classification performance can be improved by using the weak labels. Using only the RSM (one-class classification) yields a FP rate $FP|_{FN=0} \approx 4.8\%$ and the AUC is 0.998. Two-class classification with the ESMs reduces the FP Rate to $FP|_{FN=0} \approx 1.8\%$ and increases the AUC to 0.999.

The two-class classification approach which evaluates the log-likelihood ratios instead of absolute log-likelihood values as for the one-class classification improves the detection of weak pronounced anomalies (temporally short and/or only minor deviations in the feature values from normal sequences). Without learning the character of the weak anomalies, it is more difficult to distinguish them from normal process variations. For more pronounced anomalies the performance for one and two-class classification coincides.

5.7.2 Information Content of Principal and Residual Subspace

The ROCs in Fig. 5.7 demonstrate that only the combination of features of the principal subspace F (PC1,2,3) and residual error (DFFS) enable a satisfactory classification performance. The features of the principal subspace detect changes in the overall brightness and translations of the recorded melt pools, whereas the DFFS detects deformations which have not been observed in the
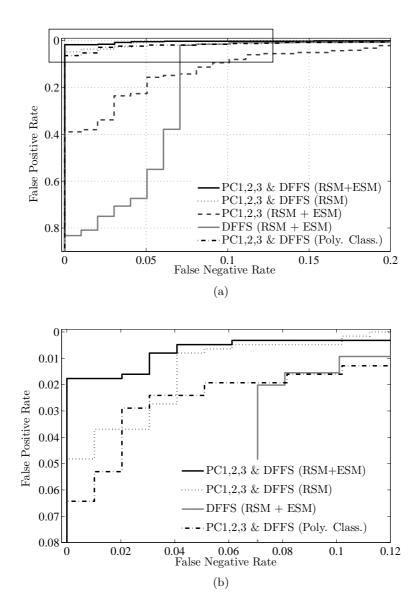
(a)

(b)

Figure 5.7: Comparison of the ROC curves for different classification approaches and feature combinations for the test data set. The ROCs are labeled with the used features and the employed classification approach. The ROCs demonstrate that the features of the principal and residual subspace have to be used together to achieve a good classification performance, and that the use of ESMs (two-class classification) improves the classification performance compared to one-class classification with a RSM. The HMM approach also outperforms a classification framework for industry processes presented in [31] ("Poly. Class"), which uses time averaging to take into account sequence information. (b) shows a magnification of the boxed part in (a).

Table 5.1: Summary of FP rates $FP|_{FN=0}$ for the test data (all erroneous sequences are found), and the area under the ROCs (AUC) for different classification approaches using appearance based features.

	$FP|_{FN=0}$	AUC
RSM (one class learning)	4.8%	0.998
RSM&ESM (weakly supervised two-class learning)	1.8%	0.999
Polynomial Classifier & Time Averaging [31]	6.4%	0.994

training data set from several regular melt pool appearances. For example, the ROC in Fig. 5.7 for the classification based on the DFFS only, clearly demonstrates that the DFFS cannot detect all errors which in turn results in a high FP rate $FP|_{FN=0}$. The irregularities in some erroneous sequences are only observable in the principal subspace and therefore the DFFS can hardly recognize them. In turn this is valid for a classification with the features of the principal subspace only.

5.7.3 Comparison with a Discriminative Classification Approach

The HMM classification system is compared with a two-stage, supervised, discriminative classification approach for industrial processes [31]. The classification system in [31] evaluates each individual frame using a polynomial classifier. The classification scores from consecutive frames are then aggregated with a temporal low pass filter. In its training phase, this approach requires a label for each individual frame. The Viterbi path for each sequence, computed using the trained ESMs, is employed to obtain frame labels. These frame labels are used for the training of the parameters of the polynomial classifiers. For the training, 80% of the erroneous sequences and 40% of the regular sequences are used. The optimum polynomial degree was 8 and the filter length 125. The classification performance is presented in Fig. 5.7. It can be seen that the performance decreases compared to the HMM approach. The ROC curve for the discriminative approach is always below the ROC for the HMM. It seams that a HMM can increase the performance of the classification system due to its superior capability of modeling temporal dynamics. Note that the HMM approach was used to get the frame labels for the discriminative approach. Without the HMM as a preprocessing step, fully labeled sequences would have been necessary.

5.7.4 Correlation between Classification Outcome and Error Severity

In Fig. 5.8 the log-likelihood ratios are sorted according to the error severity of the error on the part (established by visual inspection). A statistically non-significant correlation between the error severity on the part and the output of
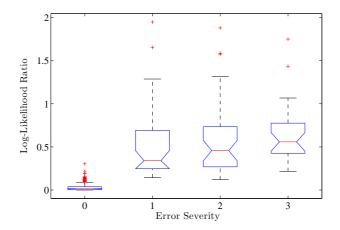
Figure 5.8: The computed log-likelihood ratio for 1014/35/38/26 parts of error severity 0/1/2/3. By definition the error severity of zero corresponds to error free parts. A non-significant correlation between the error severity and the computed log-likelihood ratio is visible, indicating the correct interpretation of the recorded sequences with the used classification method.

the classification system can be seen: the median log-likelihood ratio increases with higher error severity. The boxes in Fig. 5.8 are presented together with notches. For a box-to-box comparison the notches serve as a robust estimate of the uncertainty about the medians; the medians of boxes whose notches do not overlap are significantly different at a 5% significance level [89]. The notches for the box belonging to *Error Severity 1* and *Error Severity 3* are at the boundary of non-overlapping; the notches of the box belonging to *Error Severity 2* overlap with the other two error severity stages. The results indicate that the classification system correctly interprets the recorded image sequences and that the log-likelihood ratio can be used as an indication of the error severity. The non-significance of the correlation must be also attributed to the sensor system, not only the algorithmic interpretation of the sequences: In some sequences, the error on the weld seam is more severe than would be expected from its appearance in the sequences (sensory gap).

5.7.5 Comparison between Geometric and Appearance-based features

Fig. 5.9 presents the ROCs for geometric-based and appearance-based features for the test data. The obtained FP rate for a FN rate of 0% and the AUC, which gives an integral measure of the discriminative power of the classification system, are summarizied in table 5.2. The ROCs and the computed FP rates $FP|_{FN=0}$ clearly indicate that the appearance based features computed directly on the pixel intensities outperform the geometrical features for the considered application. The FP rate $FP|_{FN=0}$ is the key figure which has to be optimized.

Table 5.2: Summary of FP rates $FP|_{FN=0}$ for the test data (all erroneous sequences are found), and the area under the ROCs (AUC) for comparison of appearance-based and geometric-based features.

	$FP|_{FN=0}$	AUC
Appearance Based Features (PC + DFFS)	1.6%	0.999
Geometrical Features (all)	33.1%	0.918
Geometrical Features (after feature selection)	14.8%	0.993

The ROC graphs and AUC values further demonstrate that a better FP rate $FP|_{FN=0}$ is not solely based on very few samples of the positive class but is an overall trend.

Feature Selection

For geometric-based features, feature selection is necessary to optimize the classification performance. The HMM was first trained with seven hand-selected geometric-based features promising a good performance. The features were selected from a pool of expert-crafted features which have been successfully employed in various welding applications. The best feature combination after feature selection (as explained in section 5.5) reduced the FP rate $FP|_{FN=0}$ by a factor of two compared to the HMM trained with all features. The optimized feature set consisted of three simple geometric-based features: object area, compactness (ratio perimeter squared to object area) and the ratio of the bounding box area to the object area. The geometric feature area can detect changes in the size of the melt pool, whereas the other two features detect deformations. The dropped features in the feature selection stage where highly correlated with the chosen feature area such as mean gray value or melt pool perimeter.

Classification Results

The FP rate of approximately $FP|_{FN=0} \approx 15\%$ for the optimized geometric feature set is unacceptable for an automatic industrial quality control system. The subspace method clearly outperforms the geometric features with a FP rate of approximately $FP|_{FN=0} \approx 1.6\%$[4].

---

[4]The $FP|_{FN=0}$ rate varies slightly from the results in table 5.1, since a new RSM and ESM have been trained. But yet the results show that the performance is reproducible.
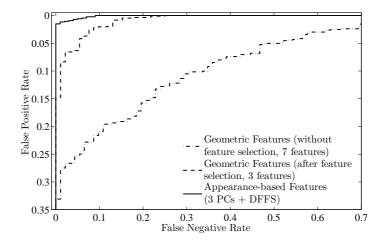
Figure 5.9: Comparison of the ROC curves for geometrical and appearance based feature combinations for the test data set. The ROCs are labeled with the used features.

5.7.6 Endurance Test

A detection system based on the ideas presented above was benchmarked directly at a production line to verify the suitability for use in a production environment [20]. 5000 parts were inspected over a one month time period by the automatic quality control system and for validation purposes also by visual inspection which was regarded as the ground truth. On several days during the test period the quality control system was activated for several hours; the total running time was approximately 60 hours. A false negative rate of 0%, a true positive rate of 1.6% and a false positive rate of 0.6% was achieved by the quality control system. This corresponds to a recall of 100% and a precision of 72% which are very good values for an industrial process monitoring system and prove the applicability of the proposed framework to manufacturing processes[5].

5.8 Discussion

An automatic classification system for the detection of rare events in image sequences has been presented. The system can analyze a large amount of weakly labeled data with strongly unequal class proportions with minimal user interaction. In the considered application, sequence labels are relatively cheap, whereas the marking of error events within sequences is tedious and expensive.

---

[5]The laser power was stabilized with a closed-loop control by another sensor system and the control parameters of the laser-power control system were also used for process monitoring. The performance was determined for the monitoring system consisting of a camera based control and the close-loop power control. But only approximately 15% of the true positives can be found exclusively by monitoring the parameters of the closed loop laser power control.

Starting from a RSM, ESMs are built by using a temporal change detection algorithm to select outlier segments. The usefullness of the classification system was validated on industrial data from laser welding processes. In an industrial environment, it is imperative that classification systems can be trained with as little user interaction as possible. For the analyzed data set, all sequences containing unusual events can be detected with a small false positive rate of approximately 1.6% to 1.8%. The use of HMMs allows to take temporal dependencies of the features into account. This modeling capability improves the classification performance compared to approaches which merely use temporal smoothing of classification scores obtained from indvidual frames.

For the challenging dataset the appearance based features reduced the FP rates $FP|_{FN=0}$ by an order of magnitude compared to the classically used geometric features. The geometric information was not sufficient to achieve an acceptable classification performance for the considered application. In general, subspace methods seem to outperform geometric-based features if distortions of the melt pool occur without significant changes in its size. Geometric features such as compactness or solidity can also detect deformations, but strongly depend on a segmentation threshold. It would be necessary to analyze the melt pools at several segmentation thresholds to overcome this problem or use more advanced segmentation methods.

The observations in section 5.7.1 also lead to the following practical use of the classification framework for monitoring industry processes:

- Initialization: First, a RSM is trained from regular sequences and a conservative classification threshold is used to flag erroneous parts, resulting in a high FP rate. The parts corresponding to the flagged sequences are then appraised by an expert. The knowledge of which sequences correspond to real errors on the produced part can be used to train ESMs and reduce the FP rate in the following.

- Classification & Optimization: The RSM (one-class classification) is used in combination with the RSM&ESM (two-class classification) approach. On the one hand it is possible to detect fault states which were not accounted for during the training procedure and on the other hand the available information of anomalies has been included to reduce the FP rate. The threshold for the one-class classification with the RSM is chosen such that only strongly pronounced anomalies are found in order to avoid an increase in the FP rate. Once more error sequences are available, the parameters of the error states of the ESM can be updated. ML parameters estimates Ψ^{new} from the newly collected training data are computed and the updated parameters Ψ are found with a weighted sum between Ψ^{new} and the parameters of the present model Ψ^{old} [68]:

$$\Psi = \xi \Psi^{old} + (1 - \xi)\Psi^{new} \qquad (5.9)$$

  where ξ is a parameter that controls the balance between the new and present parameter estimates.

The proposed approach would combine the benefits of both one- and two-class classification.

6 Object Tracking Algorithms for the Robust Detection of Weak Sputter Events

Besides melt pool irregularities an additional major source of errors in laser welding applications are sputter events. Sputter events can lead to pores in the produced weld seam and appear as additional moving objects in the recorded image sequences. A promising technique to robustly detect sputter events is to flag suspicious objects with a change detection algorithm, track the suspicious objects and describe the resulting trajectories with features which can be used for a classification decision.

The chapter is organized as follows: In section 6.1 an introduction to the detection of sputter events for laser welding processes is given. On overview of the algorithmic framework is given in section 6.2 and the implementation details are explained in section 6.3. The performance of the tracking algorithm is validated using two different datasets in section 6.4. Lastly, a summary is offered in section 6.5.

6.1 Introduction and Related Work

In Fig. 6.1 an example for a sputter event is shown. The sputter object appears besides the melt pool as a moving object in the recorded image sequences. Thus, the first step toward automated sputter recognition is the detection of suspicious objects by means of change detection algorithms. Change detection algorithms mark moving objects in image sequences by analyzing the gray-level difference of successive frames. Change detection algorithms applied to image sequences recorded from laser welding processes allow to find even weak brightness changes, but have the disadvantage that many false positives are found, as shown in Fig. 6.2 [16,17]. A high sensitivity of the change detection is necessary, since the material sputters can be small in size and low in intensity and for quality monitoring purposes it is necessary to avoid false negatives. It can be assumed that the sputter events are sufficiently long-lived and the frame rate of the camera is sufficiently high, respectively, that sputter events are observed in several consecutive frames of the recorded sequences.

In [31], a two-stage algorithmic framework for process monitoring is introduced, the so-called Two Stage Image Sequence Classification-Method or short *TISC*-method. It classifies suspicious events found in the change detection stage on a frame-per-frame basis and then aggregates these classification scores from consecutive frames. The algorithm does not establish whether two suspi-
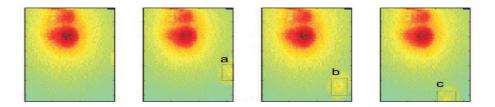
Figure 6.1: Slowly moving particle. Given an appropriate probability threshold, the Kalman Filter tracks the object correctly.
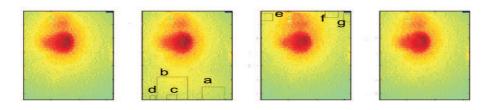

Figure 6.2: Suspicious objects flagged by the change detection in consecutive frames. The brightness fluctuations found are not due to an error event.

cious events from consecutive frames are related. Since random fluctuations of the brightness of background pixels can occur in consecutive frames due to the dynamic behavior of a welding process[1], a large number of false positives may result when applying a change detection algorithm.

In contrast, the approach proposed and tested in this chapter uses a Kalman filter to establish whether or not slightly suspicious objects that are flagged in different frames by the change detection algorithm may be related. Depending on this decision, the objects may either be discarded as harmless fluctuations, or be considered as indicative of a welding error. An introduction to the Kalman Filter and an overview of different tracking algorithms is presented in section 2.6. Since the tracking algorithm has to be deployed in a mass production environment with a rapid clock cycle, its computational efficiency is of great importance. With state of the art CMOS cameras it is possible to achieve a very high update rate (see section 3.2.3). The regular Kalman Filter with a linear system model and Gaussian noise has proven to be a suitable choice for the system under investigation [34].

6.2 System Description

A block diagram for the *TISC*-method is shown in Fig. 6.3(a) and for the tracking algorithm in Fig. 6.3(b). For both approaches, suspicious objects are flagged by the change detection algorithm. The *TISC*-method rates the detected suspicious objects in each frame using a polynomial classifier. The classifier outputs

---

[1]The recorded melt pool of laser welding processes varies in its size and brightness. These brightness changes also effect the background pixels.
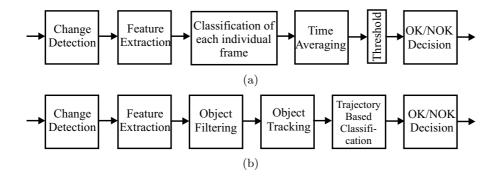
Figure 6.3: Block diagrams for the *TISC*-approach (a) and object tracking and subsequent classification proposed here (b).

are then smoothed temporally and if the result exceeds a certain threshold, the sequence is marked as erroneous.

In the tracking approach, the change detection and feature extraction stages are followed by an object filter which rejects objects that are not relevant for the tracking procedure[2]. The object filtering is followed by the object tracking step (section 6.3). The output of the object tracking are trajectories which are summarized with simple and rapidly computable features such as direction and length of trajectory, or the average area of the tracked object. The decision whether or not a sputter really occurred is made in the trajectory feature space. Thus a single label is obtained for an entire welding process.

A disadvantage of the the *TISC*-approach is that it requires labels on a frame-per-frame basis. All suspicious objects in each frame have to be labeled manually if they belong to a sputter event during the training procedure. In contrast the tracking framework proposed here requires only a single label for an entire welding sequence. Labeled data is often scarce in practice. The tracking framework increases the computation time, but is an approach, which does not need object labels for training. In essence, these two approaches trade human for computational effort.

6.3 Outline of Object Tracking Algorithm

The complete image sequence $\mathbf{I}$ consisting of K images I_k with $k \in 1, ..., K$ is available before starting the tracking procedure. The assumed system model A describes a constantly accelerated motion and does not change over time:

$$A = \begin{pmatrix} 1 & 1 & 0.5 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0.5 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad (6.1)$$

---

[2]The object filtering allows to reduce the computation time if a priori knowledge is available.

Table 6.1: Variables in the Kalman algorithm

A	System model, describing a constantly accelerated motion
$y_{k,i}$	Measurement vector of the ith object in frame I_k
$\hat{x}_{k+1|k,i}$	Prediction of the state vector using the system model A for the ith object of frame I_{k-1} in frame I_k
$\hat{x}_{k+1|k+1,i}$	Estimated state vector of the ith object in frame I_k after the measurement update
$P_{k+1|k,i}$	Covariance matrix of $\hat{x}_{k+1|k,i}$
$P_{k|k,i}$	Covariance matrix of $\hat{x}_{k|k,i}$

and the state vector $x_{k,i}$, for the ith object in frame I_k, is given by:

$$\mathbf{x}_{k,i} = (px, vx, ax, py, vy, ay)^T, \qquad (6.2)$$

i.e. the coordinates for the center of mass, its velocity and acceleration in x and y-directions, respectively. Velocity and acceleration can only be estimated if the object is tracked at least over two and three consecutive frames, respectively. The system model A is used to predict the state vector $\hat{x}_{k+1|k,i}$ for the ith object in frame I_{k+1} from $\hat{x}_{k|k,i}$

$$\hat{x}_{k+1|k,i} = A\hat{x}_{k|k,i}. \qquad (6.3)$$

The notation for the variables used in the Kalman algorithm is summarized in Table 6.1.

A block diagram of the tracking framework is shown in Fig. 6.4. In the *Preprocessing* stage, the change detection algorithm is applied to the entire recorded image sequence, outliers are marked in each frame and the positions of the segmented objects are determined. The state vectors of the objects in the previous frame I_{k-1} are predicted for I_k in the *Time Update* step. It is possible that several objects are detected in a single frame, such that more than one sputter event has to be tracked. Since the Kalman Filter can only track a single object, multi-hypothesis tracking (MHT) is used to track several objects simultaneously [27]. MHT treats each hypothesis with a separate Kalman Filter. Their outcomes are combined in the *Hypothesis Management* step. MHT is computationally exponential in both memory and time [27]. Here however the number of simultaneously occurring sputter events is small (normally ≤ 3) and thus the computation time is not increased significantly. In addition, we assume that an observation can only be assigned to a single track and a track can only be the source of a single observation per frame. This reduces the complexity of the hypothesis management step. After the *Object Assignment* the velocity and acceleration of the sputters can be calculated and the predicted values can be reconciled with the observation using the Kalman gain. In the following, each step of the block diagram in Fig. 6.4 is explained in more detail.
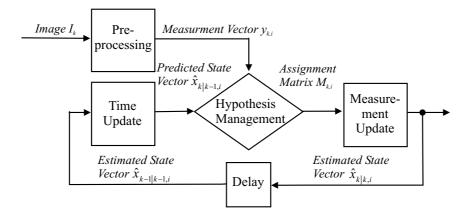
Figure 6.4: Block diagram for the object tracking algorithm.

6.3.1 Preprocessing

Following the change detection algorithm described in section 4.2 to detect suspicious objects, each pixel $I_k(x_p, y_p)$ is normalized with:

$$I_{k,norm}(x_p, y_p) = \frac{|I_k(x_p, y_p) - I_{med}(x_p, y_p)|}{I_{mad}(x_p, y_p)}, \tag{6.4}$$

where x_p and y_p specify the pixel positions in each frame. A binary image is obtained by thresholding $I_{k,norm}(x_p, y_p)$ with th_{bin}. For the tracking approach, it is important to choose th_{bin} such that all sputter objects are flagged as suspicious objects. If th_{bin} is chosen too low, the processing time increases along with the number of suspicious objects.

6.3.2 Time Update

In this step, the state vectors for the objects found in I_{k-1} are predicted for I_k. Since it is possible that objects are lost due to the pre-processing step, a track continuation function is implemented. The track continuation function predicts the state vector of trajectories in the absence of observations. The parameter *recall* specifies the allowed number of consecutive frames in which it has not been possible to find an object for a given trajectory. If, after *recall* consecutive frames, no further object for the trajectory is found, the track is terminated. If an object appears for the first time, the predicted value $\hat{\mathbf{x}}_{k,i}$ is missing. Therefore $\mathbf{x}_{k,i}$ is initialized with the observed state vector $\mathbf{y}_{k,i}$ and the covariance matrix $P_{k,i}$ is set to high values. The covariance matrices for the uncertainty of the system model and measurement model are kept constant during the tracking procedure.

6.3.3 Hypothesis Management

The trajectories describe the motion of a tracked object from the point at which it is detected until it is no longer observable. Objects belonging to the same trajectories are described with the assignment matrix $M_{k,i}$, where $M_{k,i}$ with $i = 1, ..., N_D(k)$ and $M_{k,i} \in \{1, 2, ..., N_M\}$ specifies the index number of the trajectory to which the ith detected object in frame k belongs to and N_M is the total number of found trajectories. For MHT, it is necessary to define an appropriate metric for the distance between the state vector for the currently observed and for the predicted object. In order to take the information of the covariance matrix into account the following probability assignment matrix is calculated:

$$
\begin{aligned}
P_{ij}^M \quad = \quad & \frac{1}{(2\pi)^d |P_{k|k-1,i}|^{1/2}} \\
& \exp\left(-\frac{1}{2} \left(y_{k,i} - \hat{x}_{k|k-1,j} \right)^T P_{k|k-1,j}^{-1} \left(y_{k,i} - \hat{x}_{k|k-1,j} \right) \right), \quad (6.5)
\end{aligned}
$$

where i and j are varied over all measured and predicted objects in frame I_k, respectively. Only the values for the object positions are used for the calculation of eq. (6.5). For the measured state vector, the velocity and acceleration are not yet known; thus $d = 2$ and $y_{k,i} = (px, py)$. The observed objects are assigned to the trajectories in a bijective manner in order of decreasing P_{ij}^M. Measured objects for which the maximum probability entries of P_{ij}^M are below a certain threshold p_{th} are not assigned to an already existing trajectory and initialize a new trajectory.

6.3.4 Measurement Update

After matching the observed objects in I_k with the trajectories, the missing values for the velocity and acceleration in the measured state vectors can be calculated from the difference in the position and velocity of objects belonging to the same trajectory in the previous frame. Now the predicted and measured values can be reconciled by using the Kalman Gain to form the final estimate for the state vector $\hat{x}_{k|k,i}$.

6.4 Experiments

The tracking algorithm is evaluated on two different datasets recorded from real welding processes. *Dataset A* was recorded with a CMOS camera with 1000 frames per second (fps). In total, 129 erroneous sequences containing sputter events were obtained. The dataset is used to validate how accurately the Kalman filter can track sputter events. It was tested on challenging error events, where several particles had to be tracked simultaneously over a short period of time. In general, these events occur rarely. Please note that *Dataset A* is not challenging with respect to fault detection, but with respect to matching

the suspicious objects to the correct trajectory [3].

The frame rate of the CMOS camera for *Dataset A* was relatively low. Higher frame rates increase the temporal correlation and thus allow for better results with the object tracking algorithm. Therefore a second dataset - *Dataset B* - was investigated, which was obtained with a CMOS camera with a frame rate of 8000 fps. *Dataset B* is challenging with respect to the process control, since the sputter events are only weak and it is difficult to robustly distinguish them against normal process fluctuations. This dataset was compared with the procedure presented in [31] for quality monitoring of laser welding processes. In this application the sputters were caused by material remains from a preprocessing step; the material is burned upon interaction with the laser beam and this leads to errors on the weld seam.

6.4.1 Comparison between Manual and Automatic Object Tracking

The first step in system validation was to compare the calculated trajectories of the proposed tracking procedure with the ones obtained by visual examination. The manual determination of object trajectories is regarded as ground truth in the following. In the 129 erroneous sequences of *Dataset A*, 2220 suspicious objects were found by means of the change detection algorithm (in a single sequence normally between $10 - 20$). Since the suspicious objects are found by temporal brightness changes, not all of the objects can be seen in the raw data (see e.g. Fig. 6.2). If the brightness of background pixels is relatively constant (small variance) at a certain position of the frame over time, small brightness changes are already sufficient to detect an imaginary object.

In Table 6.2 the results for the manual and automatic determination of the assignment matrix for the sequence shown in Fig. 6.1 are shown. Visual inspection suggests that the objects in the three frames belong together. Hence, all objects have the same trajectory index number. The C-value in table 6.2 is a measure of how well the results from the algorithm correspond to the manual assignments. A low C-value corresponds to a high agreement between the ground truth labels and the outcome of the algorithm. The value of C is increased by one whenever an object is assigned to a different trajectory by the algorithm compared to the manual assignment. Therefore C has the value of 2 in Table 6.2 for $p_{th} = 10^{-8}$. Suspicious objects **b** and **c** are both incorrectly assigned to a new trajectory. If, after assigning the object to a wrong trajectory no further errors occur C is not increased. For instance, for $p_{th} = 10^{-16}$, the number of wrong assignments is $C = 1$. The object **b** is incorrectly considered as the beginning of a new trajectory, thus C is increased by one; but object **c** is then correctly assigned to the same trajectory as object **b** and C is not increased any further.

Fig. 6.5 shows an example of a challenging tracking scenario. Close to the melt pool irregularities are found which do not have sharp boundaries, as e.g. the

---

[3]The sputters of *Dataset A* are high in intensity, relatively large and also many sputters occur simultaneously. For this dataset the TISC approach would be also sufficient for a satisfactory process control.

Table 6.2: Assignment Matrix for the sequence shown in Fig. 6.1 and for different probability thresholds p_{th}. The number in the columns below a,b and c specifies the index of the trajectory to which the object is assigned. The manual assignment in the last row is considered as the ground truth. A low C-value corresponds to a high agreement between the ground truth (manual) labels and the outcome of the algorithm.

	p_{th}	a	b	c	C
	10^{-8}	1	2	3	2
Kalman	10^{-16}	1	2	2	1
	10^{-20}	1	1	1	0
Manual	-	1	1	1	-

objects in Fig. 6.1. This makes it difficult to obtain the ground truth labels. In such a case, manual assignments tend to be strongly subjective. Table 6.3 compares the assignments from the algorithm with the ground truth. The visual examination considers the objects **b**, **d** and **f** as a single particle that moves downwards very quickly. The tracking algorithm, bases its decision only upon the measured coordinates of the objects and groups objects which are close together. In contrast, a visual examiner takes into account the shape and brightness of the moving particle. In addition, the tracking algorithm uses only the information from previous but not from upcoming frames. Since the complete sequence is available before starting the tracking algorithm, a Kalman Smoother could be used to overcome this shortcoming. A Kalman Smoother is a non-causal filter which bases the correspondence decision on the frames up to time index k and the upcoming frames (e.g. [58]).

Fig. 6.6 shows the recorded frames of an explosion. Many material particles are moving away from the melt pool. Table 6.4 compares the assignments for this event. The visual examination assigns the objects **c**, **g**, **l** and **e**, **i**, **j** to two separate trajectories. The track **c**, **g**, **l** is also automatically detected by the algorithm for all used parameters. Objects **i** and **j** are assigned to one trajectory for p_{th} smaller than 10^{-16}. Object **e** is always regarded as the start of a new trajectory, since the distance to object **i** is too large. Fig. 6.7 illustrates the effects of preprocessing errors on the tracking algorithm. The material particles are too close together, so that the threshold segmentation cannot detect them as separate objects. The two objects move from the melt pool to the lower right corner. In the first frame, they overlap completely and then travel with different velocities and thus separate in the following frames. The Kalman Filter is not capable of tracking the objects correctly due to the preprocessing errors. The tracking algorithm finds the trajectory **a**, **d**, **e** for $p_{th} = 10^{-20}$.

Fig. 6.8 evaluates the performance of the tracking algorithm for different probability thresholds between 10^{-16} and 10^{-3} for all of the 129 erroneous sequences. In general, the number of wrong matches decreases for lower probability thresholds. For a high probability threshold, the algorithm tends to assign
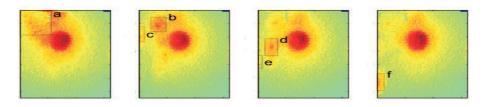
Figure 6.5: Sputter event with ambiguous object assignment.

Table 6.3: Assignment matrix for the sequence shown in Fig. 6.5

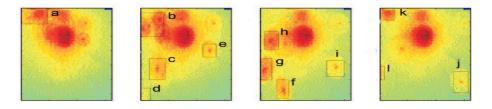|	p_{th}	a	b	c	d	e	f	C
	$1 \cdot 10^{-8}$	1	1	2	3	4	5	3
Kalman	$1 \cdot 10^{-16}$	1	1	2	2	3	4	3
	$1 \cdot 10^{-20}$	1	1	2	2	3	4	3
Manual	-	1	2	3	2	4	2	-

Figure 6.6: Recorded image sequence of an explosion; many objects occure.

Table 6.4: Assignment matrix for the sequence shown in Fig. 6.6

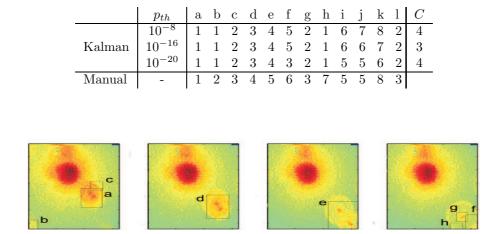|	p_{th}	a	b	c	d	e	f	g	h	i	j	k	l	C
	10^{-8}	1	1	2	3	4	5	2	1	6	7	8	2	4
Kalman	10^{-16}	1	1	2	3	4	5	2	1	6	6	7	2	3
	10^{-20}	1	1	2	3	4	3	2	1	5	5	6	2	4
Manual	-	1	2	3	4	5	6	3	7	5	5	8	3	

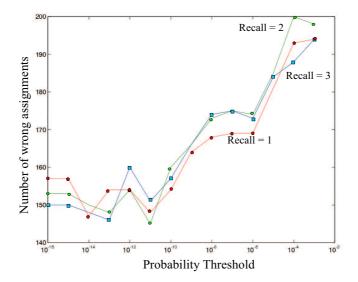Figure 6.7: Sequence with problems in the segmentation stage.

Figure 6.8: Number of wrong assignments C of the tracking algorithm compared to the ground truth labels obtained from visual inspection, for dataset containing 2200 suspicious objects. 150 wrong assignments correspond to approximately 93% correct object assignments

objects to new trajectories even though they should belong to the same, whereas for smaller thresholds the algorithm groups the objects. From $p_{th} < 10^{-10}$, no further improvement can be obtained. The number of wrong assignments C varies between 145 and 160. This corresponds to approximately 93% correct object assignments. The wrong assignments occure mainly if several sputters are present, since it is diffcult (even for a human) to match the objects to the correct trajectory. But since the overall goal is to use the tracking system for the process control the assigment errors in case of multiple sputters are acceptable. The main benefit of the object tracking framework is its ability to disdinguish weak pronounced sputter events from regular process fluctuations. This main criteria is verified with *Dataset B*. In addition, Fig. 6.8 shows the influence of the *recall* parameter on the tracking performance. It is varied between values 1 and 3. The graph shows that the influence of the *recall* parameter on the performance of the tracking algorithm is negligible for *Dataset A*, since the sputter events are strong in intensity and it is unlikely that the pre-processing steps misses an object during tracking.

6.4.2 Object Tracking for Automatic Quality Control

Dataset B is obtained from a welding process with a larger dynamic range. So the number of detected suspicious objects is an order of magnitude higher (on average 0.1 suspicious objects are detected per frame) than for the *Dataset A* (on average 0.01 suspicious objects per frame). The frame rate of the CMOS camera was 8000 fps and the duration of the welding process was approximately $0.4s$. Fig. 6.9 presents a sputter event out of *Dataset B*; the intensity of the sputters is much lower than in *Dataset A*. In the recorded sequence of the laser-induced plasma radiations in Fig. 6.9(a) the sputter event cannot be clearly seen. Only in the image sequence processed with the change detection algorithm the sputter object is visible (see Fig. 6.9(b)).

The data set consisted of 44 sputter sequences and approximately 500 welding sequences without process errors. The error sequences were gathered from a real production process over a time period of 3 months. In the recorded laser welding sequences, a large number of false positive objects are found. In Fig. 6.10, the feature space without using any sequence information is shown. The suspicious objects are manually labeled as false positives (circle) or sputter events (cross). Fig. 6.10 clearly shows the high overlap between the classes.

Both approaches sketched in Fig. 6.3 were compared. To ensure a fair comparison, the polynomial classifier was used both times. In both approaches, the feature selection was performed with a wrapper approach and the classifiers were trained using a 10 fold cross-validation so as to avoid false negatives (all parts with errors should be found). In total, 13 object based features were cal-
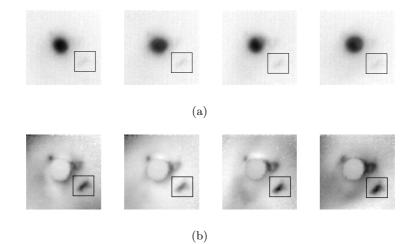

(a)

(b)

Figure 6.9: Recorded image sequence from a laser welding process containing a weak sputter event in (a) and the corresponding output of the change detection algorithm in (b). The boxed object is the sputter event. The displacement of the sputter object from frame to frame is small, due to the high update rate (the time difference between two consecutive frames is only $125\mu s$).
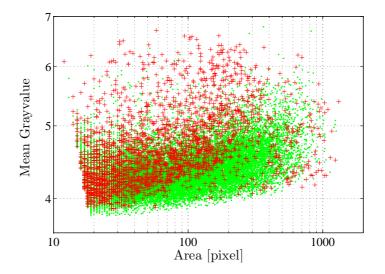
Figure 6.10: Feature space of suspicious objects found by the change detection algorithm. As an example the mean gray value and the area of suspicious objects are presented. The false positives are marked with a circle, the sputter objects with a cross. The feature space shows a high degree of overlap between the two classes.

culated for the *TISC*-approach. The best feature combination of the first stage consisted simply of the mean gray value of a suspicious object, the polynomial degree was one and the optimal temporal filter length of the second stage was two.

For the tracking approach, 11 trajectory features were calculated. The best feature combination consisted of three features (the aggregated track length and the start and stop position of the trajectory) and the optimal polynomial degree in the trajectory feature space was three. The trajectory feature space of the tracking approach is shown in Fig. 6.11, which shows a clear separation between the normal and erroneous welding sequences.

The results show that the tracking approach can reduce the false positive (FP) rate from 30.1% for the *TISC* approach to 2.7% for the tracking and subsequent classification. The error rates were both determined with a 10-fold cross-validation for a false negative rate of 0%. The primary reason for the performance improvement is that due to the large number of false positive objects, the object based classification in the first stage of the *TISC*-approach shows poor results which cannot be compensated with the temporal filter in the second stage. Only the information on the correspondence of objects enables a separation. For comparison five sequences that showed only weak sputter events were eliminated from the dataset and the classification was performed again. In this case, the FP rate was 9.2% for the *TISC* approach and 1.2% for the tracking and subsequent classification. This shows that even when NOK borderline cases are eliminated, the tracking approach outperforms the *TISC* approach by approximately an order of magnitude.
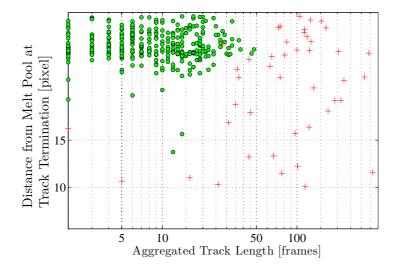
Figure 6.11: Trajectory feature space with the erroneous sequences marked with crosses and the normal sequences with circles. The feature on the x - axis describes the aggregated track length (sum of the length of all detected trajectories within one sequence) and the feature on the y - axis shows the distance of the suspicious objects from the weld pool at the track termination. The overlap between the classes is reduced substantially in comparison to Fig. 6.10.

Both approaches were implemented in Heurisko [25], a digital image processing software. The total processing time increased from 5s for the *TISC* approach to 13s for the object tracking on a Pentium 4 personal computer. Reasons for the large increase in the processing time are that the tracking approach could not be implemented without "*for*" loops that have a high execution time in Heurisko, and that a large number of false positives are found. Since the clock cycle time was met by the implementation in Heurisko, no further improvements were considered. A further advantage of the tracking approach is that entire trajectories and not individual objects as in the *TISC* approach have to be labeled in the training stage. Therefore the labeling costs can be decreased dramatically. Table 6.5 summarizes the results of the two approaches.

6.5 Summary

We proposed a tracking framework for the automatic online detection of sputter events in welding processes that can lead to critical pores in the weld seam. The training procedure avoids a tedious labeling of individual suspicious objects. We have compared the performance of the tracking algorithm with respect to ground truth labels obtained by visual examination of the sequences. With proper parameter settings, the tracking algorithm makes correct object assignments in up to 93% of all cases. We have also compared our method with the previously developed *TISC* [31] method for the quality monitoring of

Table 6.5: Summary of the results of the *TISC* approach and the sputter tracking and subsequent classification. The false positive (FP) rate is determined such that all erroneous sequences are found.

	FP rate	FP rate w/o border NOK cases	Computation Times	Labelling Costs
TISC-approach	30.1%	9.2%	≈ 5 s	High
Tracking	2.7%	1.2%	≈ 13 s	Low

industrial processes. We were able to reduce the false positive rate by an order of magnitude to 2.7% for the given dataset.

An obvious limitation of the proposed approach is the representation of each suspicious object merely in terms of its position, velocity and acceleration, i.e. the loose coupling of the model with the underlying observations. Yet, experiments to include further object features such as the mean grayvalue or area in the system model only led to a slight improvement of the tracking performance. The main reason is that due to the change detection algorithm in the preprocessing stage, more detailed object features vary strongly from frame to frame for one object. Thus, no accurate system model for the Kalman filter could be found and a high uncertainty for the system covariance matrix had to be assumed.

7 Analysis of Single Molecule Spectroscopy Data with Markov Modulated Poisson Processes

Single Molecule Fluorescence Spectroscopy (SMFS) is the study of single molecules in the focus of a confocal microscope to reveal interesting properties of the molecule, such as movement, binding or state changes. The single molecules display strong fluctuations of their fluorescence signals once exited with a laser [50]. The observed fluctuations are random variations and a detailed statistical analysis is necessary to be able to interpret the recorded optical signals and obtain microscopic information of the objects under investigation. In this chapter a fully automatic photon-by-photon analysis framework for the evaluation of data from SMFS experiments using a Markov Modulated Poisson Process (MMPP) is presented.

The chapter is organized as follows: After motivating the use of advanced data analysis techniques for SMFS experiments in section 7.1, the state of the art in data analysis for SMFS experiments is presented in section 7.2. In sections 7.3 and 7.4 the algorithmic framework based on state-space models is introduced. The framework is validate for data recorded from the complex formation and dissociation of Cu^{2+}-ions with *dcbpy* in section 7.5. A discussion is offered in section 7.6 and conclusions in section 7.7.

7.1 Motivation

Single Molecule Spectroscopy (SMS) offers richer chemical information on the dynamics of molecules than the traditional ensemble-averaged experiments [43]. Since SMS has found widespread use for investigating the structure and dynamics of single molecules, the development of data analysis algorithms to improve the interpretation of the experiments is of increasing interest [4, 53, 88].

Classical data analysis approaches for the photon traces have some general limitations which can be overcome with advanced techniques based on state-space models. Although analysis techniques based on state-space models have been proposed already by other groups and have been validated on simulated data sets [4, 43] the empirical study of the techniques on real experimental data sets and the comparison of the results with classical techniques is limited [50]. In this chapter a fully automatic analysis framework is presented. For the first time the benefit of the advanced techniques based on state-space models on experimental data could be clearly proven. State-space models are a powerful technique to interpret the data but have the disadvantage that they force the

data into a model, which might lead to oversimplification [50]. The presented study proves that model selection is a crucial step to obtain a correct interpretation of the photon streams. Once the correct model is selected, information can be extracted from the photon stream which would be lost with other analysis techniques.

7.2 State of the Art

Different approaches are discussed in literature to evaluate the photon emissions from SMFS experiments. The following section gives an overview of the most prominent ones.

7.2.1 Binning

For the analysis it is assumed that the molecule can reside in two different states, the ON and OFF states with high and low fluorescence levels, respectively. The received photons are binned in time bins of fixed length Δt_{bin} and possible ON and OFF states are found from the binned data by using an empirically found threshold [50]. Bins with photon counts below a certain threshold are associated to the OFF state and in turn bins with photon counts above the threshold to the ON-state. The ON/OFF durations are collected in histograms, normalized to probability density functions and a single-exponential model is fitted to the density functions to obtained the kinetics of the investigated molecule [42]. The major disadvantage of the binning approach is that it entails a significant loss in temporal resolution [43]. Possible molecular dynamics on the time scale of the binning window are lost and the shortest observable state duration is limited by the binning size Δt_{bin}. In addition the correct choices on the used bin size and applied threshold is not easy and strongly influence the results [50]. For intense signals segmentation is easy whereas in most practical cases segmentation is difficult due to a poor signal/background ratio or experimental imperfections.

7.2.2 Fluorescence Correlation Spectroscopy

Fluorescence correlation spectroscopy (FCS) is a powerful technique to analyze SMS data. Compared to binning FCS has the advantage that it can retain the temporal information of the raw data, but the information about the points at which interesting events occur are lost. The autocorrelation function $AC(\tau)$ for a give lag time τ is given by:

$$AC(\tau) = \int_t I_p(t)I_p(t+\tau)dt, \tag{7.1}$$

where $I_p(t)$ is the photon detection intensity at a given time t. In [97] a computationally efficient technique with complexity $O(K)$, where K is the total number of received photons, is presented to compute from a photon stream the corresponding correlation curve on a photon-by-photon basis. The shape of the correlation curve reveals interesting properties about the molecule.

7.2.3 Hidden Markov Models

The state-space view on the problem is the following: The molecular state cannot be observed directly but is encoded in the photon stream emitted from the molecule. In principal, the task of the analysis algorithm is to estimate the most likely state sequence from the photon stream. Instead of determining the state sequence indirectly through a non-parametric approach (like binning), it is advantageous to find the most likely state sequence on a photon-by-photon basis, without information loss. Since the molecular state is unobserved (hidden), only indirectly manifests itself trough recorded photons and is not expected to have a long-term memory, a natural way of modeling the data is to employ Hidden Markov Models (HMMs).

To enable an efficient analysis, the recorded photon stream is processed with a Markov Modulated Poisson Process (MMPP), introduced in detail below. MMPPs are useful in a variety of applications where events occurring in continuous time exhibit periods of differing intensity. In particular they have been successfully applied in communication systems as network traffic models [76,99]. A MMPP is a Poisson process whose rate λ is governed by an underlying Markov chain S. MMPPs exhibit, like analyzing the photon emissions with FCS, no limitation regarding the temporal resolution and have also a computational complexity which is linearly increasing with the number of recorded photons K. But in addition the MMPP can resolve the position within a photon stream where interesting events happen.

7.2.4 Previous Work on Hidden Markov Models

Hidden Markov Models for the analysis of experimental data from single molecule spectroscopy are a subject of active research. In [4,43] individual photon arrivals are used as an input to the models and Markov Chain Monte Carlo (MCMC) sampling methods are employed to determine the model parameters. In both papers, models with two and four distinct states, motivated by [73], are considered. In [4] the position of the molecule is assumed to be fixed under the microscope and model validation is performed on simulated data, only. In contrast, [43] in addition allow the molecule to diffuse in the focal volume and the algorithms are validated on experimental data sets. In addition, a continuous state model, that was also previously described in [73], is validated for analysis purposes in [43].

On the contrary to [4, 43], we infer the parameters of the MMPP from the recorded data with an Expectation-Maximization (EM) algorithm. This parameter estimation method is also chosen by [21], but there the HMM is used to analyze trajectories of photon counts and not individual photons. Recently a HMM approach for fluorescence resonance energy transfer (FRET) trajectories was proposed [54] and validated on experimental data [38]. This approach employs the most widely used HMM type, with discrete transmission probabilities and Gaussian distributed emission probabilities; the parameter estimates are not found by EM but using Brent's algorithm, a standard optimization algorithm. In [9] a standard HMM (discrete transmission probabilities and

Gaussian distributed emission probabilities) is modified to be able to handle diffusive motion.

One of the main contributions of this thesis is to link MMPP algorithms to standard HMM algorithms, which have been extensively researched and optimized in other communities (e.g. in speech recognition)[1]. Thus we have access to fast, well-understood algorithms which have be proven to work efficiently on large datasets (e.g. in speech processing). Hence, the algorithms can handle the large amount of data usually recorded in SMFS experiments. In addition we demonstrate very clearly the high importance of model selection on the parameter estimates for the kinetic rates. State-space models are a powerful technique to interpret the data but have the disadvantage that they force the data into a model, which might lead to oversimplification [50].

7.3 Hidden Markov Models for SMFS Experiments

Following the notation in section 2.2 the probability model for a HMM is defined as follows: At any discrete time instance k the system is assumed to be in one of a finite set of N_Q states. Associated with each state Q_k is a fixed probability distribution of observations o_k that could be emitted by that state. The probability of a particular observation o_k given state $Q_k = i$ is $b_i(o_k) = p(O_k = o_k|Q_k = i)$. State transitions take place according to a transition probability matrix A' with elements $a'_{ij} = P(Q_{k+1} = j|Q_k = i)$, where a'_{ij} describes the transition probability from state $Q_k = i$ at time instance k to state $Q_{k+1} = j$ at time instance $k + 1$.

The top drawing in Fig. 7.2 presents a possible HMM to model the recorded data. The HMM consists of $N_Q = 2$ states, the OFF and ON state, corresponding to bound and unbound state of copper, with discrete emission probabilities for each state (the discrete density distributions are shown below the states). The photon arrival process is modeled with a discrete sampling time Δt_s (e.g. the time resolution of the photon detector). At each sampling instance t'_k a photon can be detected and the HMM can undergo a state change. The sampling time Δt_s is much shorter than the mean time interval $\overline{\Delta t}$ between recorded photons. It can be assumed that the probability of detecting more than one photon in a time interval is approximately zero and the probability $P(\text{one photon detected}) = p_{h\mu} = \lambda_P \Delta t_s$ of detecting exactly one photon in the time interval Δt_s is proportional to Δt_s, where λ_P is the proportionality constant. Thus an appropriate model for the arrival process is a Poisson distribution with parameter λ_P [30]. The probability $p_{h\mu}$ is higher in the ON state than in the OFF state, but $p_{h\mu}$ is small for both states. Most of the time, no photon will be detected during the time interval Δt_s, resulting in a high computational overhead when applying the above described model for the data analysis. Instead of using the discrete probabilities $p_{h\mu}$ and $1 - p_{h\mu}$ of observing a photon or not in a sampling period, an equivalent formulation for a Poisson

---

[1]The link between HMMs and MMPPs is mentioned in other communities [76], but not in HMM papers on SMFS.
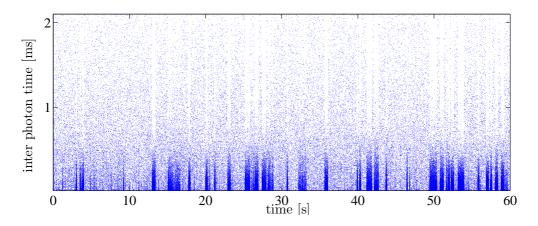
Figure 7.1: Data from a SMFS experiment. On the x-axis the absolute photon arrival time t is plotted and on the y-axis the time interval between occurring events, Δt.

process is to model the time between photon arrivals Δt. It can be shown (see e.g. [30]) that the time differences $\Delta t_k = t_k - t_{k-1}$ follow an exponential distribution with parameter $\lambda = \lambda_P$. The alternative model is referred to as the Markov Modulated Poisson Process (MMPP) and is presented in the bottom drawing in Fig. 7.2. The sequence of observations **O** for the MMPP are the time differences Δt_k between recorded photons:

$$o_k = \Delta t_k = t_k - t_{k-1}. \tag{7.2}$$

for $k = 1, 2, ..., K$. The observations o_k reveal information about the hidden process. The obtained data is visualized in Fig. 7.1. In the ON state, Δt_k is small whereas in the OFF state Δt_k is larger. In mathematical terms, it may be assumed that the distribution of O depends on a Hidden Markov Chain S, namely the state of the molecule. The next section gives an introduction to MMPPs and compares it with the HMM.

7.4 Markov Modulated Poisson Processes

Like the HMM, the MMPP has a finite number of discrete states N_Q. Let $\Psi = \{G, \Lambda\}$ denote the parameters of the MMPP, where $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_M\}$ holds the different Poisson rates for each state $i = 1, \ldots, N_Q$ and G is a $N_Q \times N_Q$ generator matrix, which defines the state transition rates. In the MMPP the time interval Δt between observations and associated possible state changes is variable, whereas in the HMM the time between observations Δt_s is constant. For the HMM the transition probabilities can be defined with the time-independent transition probability matrix A'. From the definition of HMMs follows that state durations are exponentially distributed. The longer the process is in a certain state, the higher the probability for a state change. This must also be true for the MMPP, thus the transition probabilities have to depend
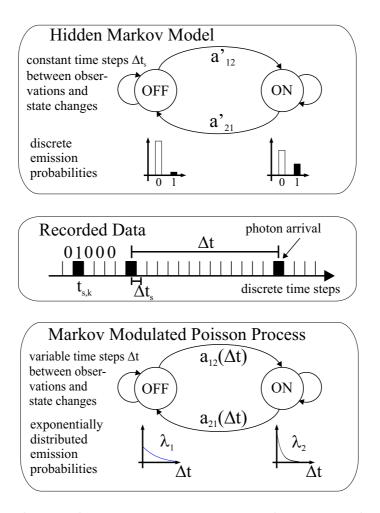
Figure 7.2: The top drawing presents a HMM with 2 states, the ON and OFF state, to model the experimental data. The HMM has time-independent transition probabilities and discrete emission probabilities, where in the ON state the probability to detect a photon is higher than in the OFF state. An equivalent model to describe the observations is presented in the bottom drawing. The MMPP has time-dependent transition probabilities and exponentially distributed emission probabilities. It requires fewer iterations to model the same data.

exponentially on the inter photon time Δt. A time-independent G is defined for the MMPP with elements

$$g_{ij} = \begin{cases} \text{the rate of transition from state } i \text{ to state } j & \text{for } i \neq j \\ -\sum_{i \neq j} g_{ij} = -g_i & \text{for } i = j \end{cases} \qquad (7.3)$$

where g_i is the parameter of the exponential distribution of the dwelling time in state i. The generator matrix G for the HMM can be computed with $G = \lim_{\Delta t_s \to 0} A'(\Delta t_s)$. In turn the transition probabilities $a_{ij} = P\left(Q(t + \Delta t) = j | Q(t) = i\right)$ for the MMPP can be computed from the generator matrix G by taking the length of the time interval Δt into account:

$$A_k(\Delta t_k) = \exp\left(G \Delta t_k\right) \qquad (7.4)$$

The generator G defines a N_Q-state kinetic scheme [65], which can be described with the following differential equation:

$$\frac{d}{dt} P_i(t) = \sum_{j \neq i} \left(g_{ji} P_j(t) - g_{ij} P_i(t)\right), \qquad (7.5)$$

where $P_i(t)$ is the probability of being in state i at time t.

The matrix exponential can be computed (iff G is diagonalizable, which is not guaranteed) with

$$\exp\left(G \Delta t\right) = B \cdot \exp(D \Delta t) \cdot B^{-1}, \qquad (7.6)$$

where D is a diagonal matrix of the eigenvalues and B the matrix of corresponding eigenvectors of G. The diagonalization in eq. (7.6) is in general not a stable numerical technique. A stable and efficient algorithm for the computation of the matrix exponential is presented in [94][2]. The algorithm employs a Padé approximation with scaling and squaring:

$$\exp\left(G \Delta t\right) = \left[R_{qq}\left(\frac{G \Delta t}{2^\kappa}\right)\right]^{2^\kappa}, \qquad (7.7)$$

where $R_{qq}(z)$ is the (q, q) Padé approximation to e^z and κ is chosen such that $\frac{G \Delta t}{2^\kappa}$ has a norm of order one. Since the inter photon times Δt are observed in the experiment, the transition probability matrix A_k can be calculated for each time instant k.

It is assumed that photon emissions from a single state occur uniformly over time (homogeneous Poisson process), i.e. the number of expected events in an interval is proportional to the length of the interval, regardless of the interval location [3]. We assume that the rates of the Poisson processes in each state are constant within one sequence, i.e. the molecule does not move under the

---

[2]The algorithm presented in [94] is implemented in Matlab with the function *expm*.

[3]An extension to include time varying event rates would be a nonhomogenous Poisson process.

detector and does not undergo other modifications. As explained above the inter photon times Δt are exponentially distributed in each state Q_k and the emission probability in state Q_k is given by:

$$b_k(\Delta t_k) = \lambda_m \exp\left(-\lambda_m \Delta t_k\right). \tag{7.8}$$

7.4.1 Parameter Estimation

Eqs. (7.4) and (7.8) show that the MMPP can be interpreted as a HMM with time-dependent transition rates and time-dependent exponentially distributed emission probabilities.Thus, for ML parameter estimation with the EM algorithm the standard algorithms presented in section 2.3 can be used with a slight modifications in the M-step. As noted earlier the disadvantage of the EM algorithm is that it converges only to a local optimum and can have a slow rate of convergence. In order to avoid local optima, the EM has to be run with a number of different initial guesses. Here the initial guess for the parameters can e.g. be found by using the results from the classical binning approach. The EM algorithm is closely related to the MCMC technique used in [4, 43] for the parameter estimation. The MCMC samples from the conditional distributions, whereas the EM maximizes over them [32]. The advantage of the MCMC is that it provides posterior parameter distributions instead of point estimates as the EM. But MCMC has a higher computational complexity and since the recorded data sets consist in the order of 10^6 data points and the number of model parameters vary from 4 to 12, the EM was selected based on computational considerations. The expected time duration D_i for which a molecule is in state i for $[t_1, t_K]$ is given by:

$$D_i = \sum_{k=1}^{K-1} \gamma_{i,k} \cdot o_k. \tag{7.9}$$

In [72, 91] the update formulas for MMPPs in the **M**-step of the EM are derived. A new parameter estimate $\hat{\Psi} = \{\hat{G}, \hat{\Lambda}\}$ can be obtained with:

$$\hat{\lambda}_i = \frac{\sum_{k=1}^{K} \gamma_i(k)}{D_i} \tag{7.10}$$

and

$$\hat{q}_{ij} = \frac{\sum_{k=1}^{K-1} \xi_{ij,k}}{D_i}. \tag{7.11}$$

The update formulas for the EM show that the data points contribute differently to the parameter estimates. The higher the probability that an observation belongs to a certain state i or triggers a state transition from state i to state j, the higher is its impact on the particular parameter estimate.

7.5 Experimental Results

In the following the complex formation and dissociation of Cu^{2+}-ions with the bidentate ligand 2,2'-bipyridine-4,4'dicarboxic acid (*dcbpy*) in aqueous media recorded by SMFS is considered. Upon copper addition to the *dcbpy*, the fluorescence signal starts fluctuating between two intensity levels which can be associated with the unbound (high photon intensity) and bound states (low photon intensity) of copper [42]. Thus the simplest model to describe the data is a two-state ON and OFF model. In the ON-state (OFF-state) the photon arrival rate is high (low), thus the time interval between photon arrivals Δt tends to be shorter (longer).

7.5.1 Experimental Details

A probe that reports the binding of Cu^{2+} ions to *dcbpy* by fluorescence quenching of tetramethlyrhodamine (TMR) to study the complex formation of copper(II) ions is established [18, 19]. The probe (Fig. 7.3), made of two complementary and modified DNA oligonucleotides that carry the ligand *dcbpy* (1) and the dye TMR as well as a biotin-linker (2) was immobilized on a BSA coated glass surface by binding to strepavidin labeled BSA. Upon addition of $CuSO_4$ to the sample, a strong decrease in the fluorescence emission of TMR is observed. Together with systematic studies on the quenching of the TMR-*dcbpy* conjugate these findings suggest that TMR is quenched strongly by the $Cu[dcbpy]^{2+}$ complex while dynamic quenching by Cu^{2+} ions can be neglected. Thus for time-resolved single molecule studies the observation of a highly fluorescent state corresponding to the free probe and a weakly fluorescent state upon coordination of Cu^{2+} to the ligand *dcbpy* is expected.

Time-resolved single molecule studies are carried out on a confocal fluorescence microscope equipped with a piezo-stage (see Fig. 7.4) [90]. The three axis piezo-stage allows for a pixel-wise acquisition of the fluorescence signal. For excitation a frequency doubled cw-NdYAG laser emitting at 532 nm is used (average excitation power $0.6 kW/cm^2$, Coherent, Dieburg, Germany) that was sent through a 1/4-wave plate for circular polarization and focused on the sample by a 100×microscope lens with a numerical aperture of 1.45 (Olympus, Hamburg, Germany). The fluorescence photons emitted by the sample were collected by the same lens and detected by an avalanche photo diode (APD). The pinhole in front of the APD realizes a confocal setup. Signals which are coming from points out of the focal plane or from points which are laterally displaced with respect to the focal volume are not properly imaged onto the APD and hence are strongly suppressed. The photons were recorded individually in the FIFO-mode of the employed single photon counting board (SPC 630, Becker & Hickl, Berlin, Germany) with a custom LabView application. For time-resolved measurements of individual probes, first microscopic images are acquired in order to determine the precise position of the individual molecules (see Fig. 7.5.1). Then the molecules were subsequently positioned in the laser focus one by one for time-resolved recording of their fluorescence emission until
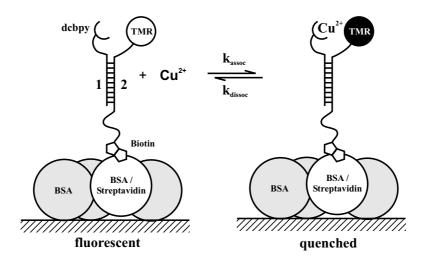
Figure 7.3: The TMR-*dcbpy* conjugate is formed by hybridization of two modified DNA-oligonucleotides. Strand (1) carries the *dcbpy* as metal ion binding site at the 5'-end. The complementary strand (2) is labeled at the 3'-end with TMR and coupled by a hexaethyleneglycol-linker to biotin for immobilization on streptavidin coated surfaces. Upon formation of the $[Cu(dcbpy)^{2+}]$ is strongly quenched.

photo-bleaching occurred. The advantage of the used setup is that a high temporal resolution can be achieved; in particular in the following experiments the photon arrival times t are recorded at 50 ns time-resolution. Recorded traces with a signal/background ratio > 3 were selected for the analysis. A detailed description of the experimental details can be found in [42].

7.5.2 Analysis with the Two-state ON and OFF Model

The SMFS traces recorded with the setup described in section 7.5.1 are analyzed with the MMPP introduced above. Recorded traces with a signal/background ratio > 3 were selected for the analysis [42]. First a two-state model, referred to as model-2, is considered to analyze the dynamics of the complex formation and dissociation of individual Cu^{2+}-ions with *dcbpy*. Model-2 with discrete ON and OFF states (see Fig. 7.2) is the model with the lowest complexity to describe the recorded data and can be linked directly to the following association and dissociation kinetic scheme:

$$\underbrace{M + L}_{ON} \overset{k_{assoc}}{\underset{k_{dissoc}}{\rightleftharpoons}} \underbrace{ML}_{OFF}, \tag{7.12}$$

where M, L and ML stand for the metal (M) Cu^{2+}, the ligand (L) *dcbpy* and the metal ligand complex (ML) $[Cu(dcbpy)^{2+}]$. Hence, the usual kinetic
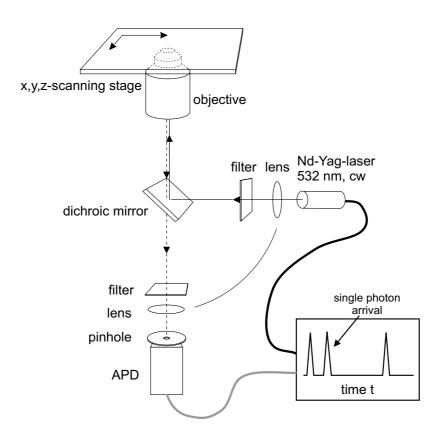
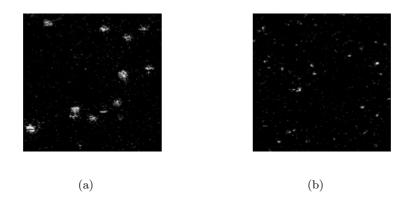Figure 7.4: Setup for a scanning confocal optical microscopy.

Figure 7.5: Microscopic images for the unbound (a) and bound (b) state of copper to the *dcbpy* molecule.

descriptions of the association and dissociation reactions are given by:

$$\frac{dc(L)}{dt} = -k_{assoc}c(M)c(L) + k_{dissoc}c(ML) \qquad (7.13)$$

$$\frac{dc(ML)}{dt} = -k_{dissoc}c(ML) + k_{assoc}c(M)c(L) \qquad (7.14)$$

where $c(M)$, $c(L)$ and $c(ML)$ specify the concentrations of the metal, ligand and ligand metal complex, respectively. For single molecule experiments concentrations loose their validity and have to be replaced with probabilities [51], yielding the following probabilistic description of the kinetic equations (7.13) and (7.14):

$$\frac{dP_{on}(t)}{dt} = -k_{assoc}c(M)P_{on} + k_{dissoc}P_{off} \qquad (7.15)$$

$$\frac{dP_{off}(t)}{dt} = -k_{dissoc}P_{off} + k_{assoc}c(M)P_{on} \qquad (7.16)$$

Comparing equations (7.15) and (7.16) with eq. (7.5) the entries of the generator matrix G

$$G = \begin{pmatrix} -g_{off} & g_{off} \\ g_{on} & -g_{on} \end{pmatrix} \qquad (7.17)$$

can be related to the association and dissociation rates with $g_{on} = k_{assoc} c(Cu^{2+})$ and $g_{off} = k_{dissoc}$, respectively[4].

Fig. 7.6(c) presents examples of the data recorded at $CuSO_4$ concentrations of $2\mu M$ and $6\mu M$ together with the Viterbi path (the most likely sequence of states) computed with a two-state MMPP. Fig. 7.6(b) and Fig. 7.6(d) visualize the corresponding posterior likelihoods (approximated with the forward probability α) for each received photon under the MMPP model. The sign of the a posteriori likelihood indicates if a photon belongs to the ON state (positive value) or OFF state (negative value). The absolute value indicates the reliability of the decision. From visual inspection the decrease of the length of the ON states with increasing $CuSO_4$ concentrations can be seen; the length of the OFF states is expected to be independent from the used $CuSO_4$ concentration. Fig. 7.7(a) demonstrates that even short variations in the photon arrival rate can automatically and accurately be detected with the MMPP. Fig. 7.7(b) shows an excerpt of the sequence in Fig. 7.7(a). The found ON and OFF states agree with a visual examination of the raw data and the shortest state durations consist of approximately 10 (quenched molecule state) to 20 (fluorescent molecule state) photons, only. To analyze the same data set with a binning approach, a binning window of 20 ms is necessary [42]. During 20 ms approximately 30 (quenched molecule state) to 200 (fluorescent molecule state) photons are detected, thus the temporal resolution is improved with the MMPP by a factor of roughly 3 to 10. For comparison, the segmentation on a photon-by-photon basis shown in Fig. 7.7(c) and Fig. 7.7(d) has been performed without

---

[4]Note that the entries of the generator matrix g_{on} and g_{off} are actually the reaction kinetics, normally abbreviated with v_{on} and v_{off}

Table 7.1: Association and dissociation rates k_{assoc} and k_{dissoc} and the complex stability constant $K_s = \frac{k_{assoc}}{k_{dissoc}}$ for the analyzed data, and comparison with previously published results on the same data and the MMPP with different model complexities.

Analysis Method	$k_{assoc}(\times 10^6 M^{-1} s^{-1})$	$k_{dissoc}(s^{-1})$	$K_s(\times 10^6 M^{-1})$
Mono-exp. fit [42]	3.3 ± 0.3	1.2 ± 0.4	2.7 ± 0.9
MMPP Model-2	21.0 ± 2.1	3.6 ± 1.6	5.8 ± 2.6
MMPP Model-3B	6.9 ± 0.6	1.3 ± 0.4	5.3 ± 0.7

using sequence information. The results demonstrate that a segmentation on a photon-by-photon basis without a model using sequence information is not possible.

7.5.3 Determination of Kinetic Rates using the Two-state Model

The EM algorithm is used to determine the dynamics of traces with $CuSO_4$ concentrations of $2\mu M$, $6\mu M$, $10\mu M$ with 3.8, 1.2 and 0.9 million observations, respectively. The rates of the Poisson processes λ_i in each state i are approximately constant within one sequence, but can vary between different sequences since the inter photon times depend on the absolute photon flux which in turn depends on the position of the molecule relative to the focus. The kinetic rates $\hat{q}_{ij}$ are independent of the absolute inter photon times. In order to estimate the kinetic rates with the MMPP over several sequences, the following estimation procedure is used:

- for each sequence, the Poisson rates $\hat{\lambda}_i$ are determined individually using EM (eq. (7.10))

- the individual Poisson rates λ_i for each sequence are kept constant

- the dynamic rates $\hat{g}_{ij}$ are estimated over several sequences using EM (eq. (7.11))

The computed holding times k_{on} and k_{off} for the ON and OFF states are presented in Fig. 7.8(a), respectively. The determined kinetic rate constants for $[Cu(dcbpy)]^{2+}$ are summarized in table 7.1 and compared with previously published results on the same data set. Since the concentration of the TMR *dcbpy* remains constant over time, the holding time k_{on} is expected to depend linearly on the concentration of $CuSO4$ [42]. This agrees with the results presented in Fig. 7.8(a). The kinetic rates presented in table 7.1 for model-2 are significantly higher than those found by binning in [42]; this fact is discussed below.
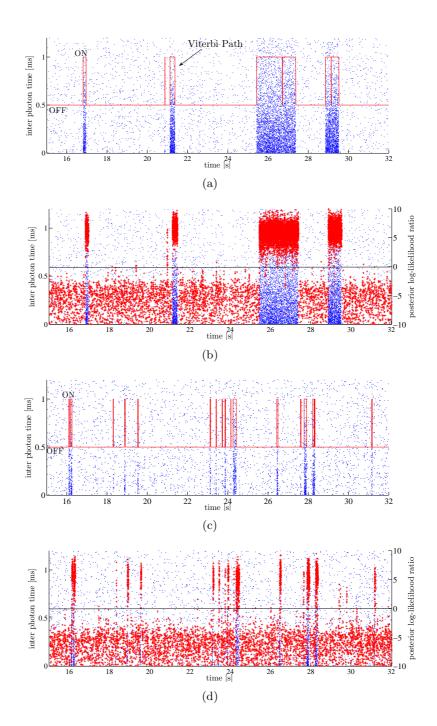
Figure 7.6: SMFS trace for $2\mu M$ $CuSO_4$ in (a) and (b) and $6\mu M$ $CuSO_4$ in (c) and (d). On the x-axis the absolute photon arrival time t is plotted, on the y-axis the time interval Δt between photon arrivals (circels) together with the Viterbi path in (a) and (c) and the posterior log likelihood ratio (crosses) in (b) and (d).
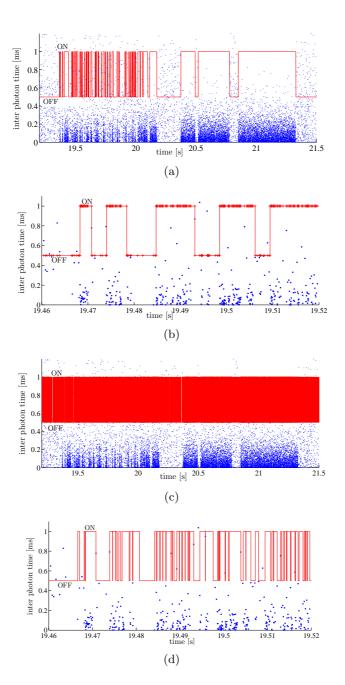
(a)

(b)

(c)

(d)

Figure 7.7: SMFS traces for $CuSO_4$ concentration of 0.1 μM. On the x-axis the absolute photon arrival time t is plotted, on the y-axis the time interval Δt between photon arrivals together with the Viterbi path. (b) and (d) show part of the trace of (a) and (c), respectively. (a) and (b) present the results using the MMMP introduced here and demonstrates that even small variations in the received photon rate can be detected; the marks in the Viterbi path in (b) are photon detections. (c) and (d) present the same traces without using sequence information. It is only assumed that the photon arrival times follow an exponential distribution. The number of state changes increases significantly; in (c) the black bar is caused by rapid state changes. A correct segmentation is not possible.
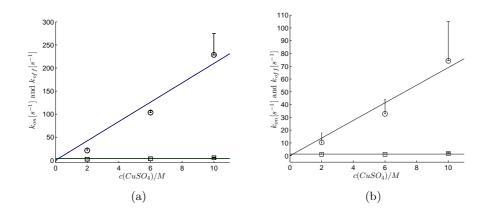
Figure 7.8: Results for the analyzed traces with model-2 (a) and model-3B (b). The on-rate k_{on} (circles) plotted against the used $CuSO_4$ concentrations shows approximately a linear behavior for both models. The off-rate k_{off} (squares) is independent the $CuSO_4$ concentration. To determine the ML parameter estimates the EM was initialized randomly 20 times. The circle/square shows the most likely estimate and the bars visualize the parameter range of the 10 most likely parameter estimates.

7.5.4 Models with increased complexity

The formulas for a general N_Q state model are given in section 7.3. In addition to model-2, which is the model with lowest complexity to describe the data, the following model modifications are considered:

- Model-3A, a model with 3 states, where the OFF- state is split into 2 states, allowing short periods of increased photon arrivals within in the OFF state.

- Model-3B, a model with 3 states, where the ON- state is split into 2 states, allowing short periods of decreased photon arrivals within in the ON state.

- Model-4, a combination of model-3A and model-3B (also considered in [4, 43]).

An overview of the model topologies is shown in Fig. 7.9.

In a family of models of increasing complexity, the ones with more parameters will always allow for a better fit of the data. To avoid overfitting, it is advisable to verify whether an increase in the number of model parameters is really justified by the improvement of the fit. One popular criterion for model selection is the Bayesian Information Criterion (BIC) [12]:

$$BIC(\Psi) = \log P(y|\widehat{\Psi}) - \frac{K}{2} \log N_P \qquad (7.18)$$
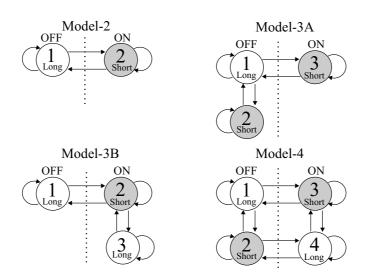
Figure 7.9: Topologies of all models: If two states are not connected, the transition probability between these is constrained to zero during training. The states are marked according to whether the inter photon times Δt are long (low photon arrival rate) or short (high photon arrival rate). For model-2 state 2 is interpreted as the ON-state (fluorescent state) and state 1 as the OFF state (quenched state) of the molecule; for model-3A states 1 and 2 are interpreted as the OFF-states and state 3 as the ON-state; for model-3B state 1 is interpreted as the OFF-state and states 2 and 3 as the ON-states; for model-4 states 1 and 2 are interpreted as the OFF-states and state 3 and 4 as the ON-states. For model-4, the transition rates between state 1 and 3 and between states 2 and 4 are forced to equal values during parameter estimation.
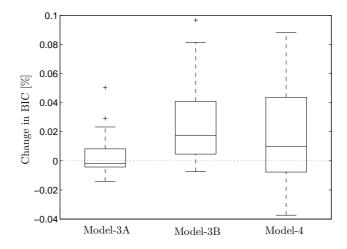
Figure 7.10: The BIC differences (eq. (7.18)) in percent for several sequences from different $CuSO_4$ concentrations ($2\mu M$, $6\mu M$ and $10\mu M$) between model-2 and remaining models are presented as boxplots. In total, 33 traces with approximately $50,000$ to $200,000$ data points each are used.

where N_P is the number of free model parameters and K is the total number of observations. The first term is the likelihood of the data under the model and the second term a penalty for the model complexity. Fig. 7.10 shows the BIC differences over several sequences between model-2 and model-3A, model-3B and model-4, respectively. Note that the BIC differences are very small, indicating a rather flat log likelihood surface around its minimum. According to Occam's razor[5], model-2 should be selected for data analysis purposes. The computed BIC difference indicate that model-3B describes the data slightly better than the other models with increased complexity. Model-3A has lower BICs values than model-3B and shows overall no improvement compared to model-2. Model-4 is a combination of model-3A and model-3B but the BIC does not increase significantly compared to model-3B, indicating that model-3A part of model-4 is not necessary to describe the data. The ON and OFF rates computed with model-3B are presented in Fig. 7.8(b) and the computed kinetic rate constants for $[Cu(dcbpy)]^{2+}$ are summarized in table 7.1. The estimated association and dissociation rates decrease compared to model-2 and are close to the results obtained from the binning approach [42]. An explanation for the better fit of model-3B to the data is given in the next section[6].

---

[5]Occam's razor states that among competing models with similar lack of fit, the one of lowest complexity should be preferred.

[6]Triplet blinking cannot be observed in the recorded photon streams, since the intensity of the received photons is not high enough
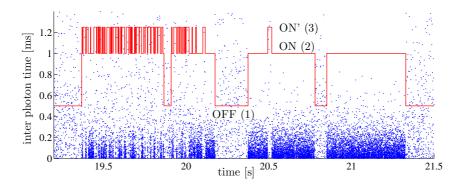
Figure 7.11: Viterbi path segmentation of the trace shown in Fig. 7.7(a) with model-3B. The fluctuations are no longer detected as separate states, but as part of an ON-State with decreased photon arrival times. The additional sub-state is referred to as ON'.

7.6 Discussion

In the sequences, faster kinetics than theoretically expected can be observed over short periods of time. Consider, for example, part of a recorded trace for the $CuSO_4$ concentration of $0.1\mu M$ shown in Fig. 7.7(a). The ON and OFF state durations between 20.3 and $21.5s$ in Fig. 7.7(a) are not in agreement with the rapid fluctuations between 19.2 and $20.3s$. It is statistically unlikely that a simple two-state model can describe these two regimes simultaneously. In contrast, model-3B allows the inter photon arrival time to decrease over a short period of time without triggering a state transition from the ON to the OFF state (see Fig. 7.11). This feature of model-3B allows the description of rapid state fluctuations. Therefore, the ML estimates for the kinetic rates for the ON and the OFF state decrease for model-3B compared to model-2. Since the intermittent fluctuations occur only over a short period of time, the BIC difference to model-2 is small, but the impact on the kinetic rates is significant. Traces recorded from molecules without adding Cu^{2+}-ions did not show these rapid fluctuations. This control experiment rules out the possibility for the fast fluctuations to be a mere experimental artifact.

The MMPP is a generative model and can be used to draw samples and simulate SMFS traces. In Fig. 7.12 realizations generated from model-2, model-3B and model-4 are shown together with the original sequence from which the parameters have been estimated. By visual examination model-3B shows the best correspondence with the original sequence. Model-2 is in poor agreement with the original sequence, which further indicates that a 2-state model is not appropriate to describe the recorded data. Model-2 cannot account for the fact that short fluctuations occur only over a short period of time.
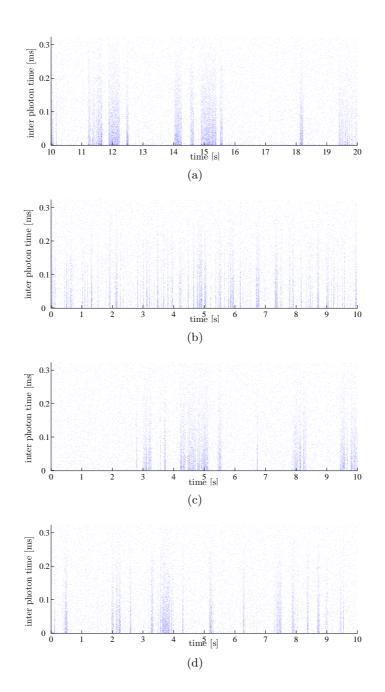
Figure 7.12: Experimental photon trace for a $CuSO_4$ concentration of $0.1\mu M$ (a) and simulated traces for model-2 (b), model-3B (c) and model-4 (d). Visual examination suggests a superiority of model-3B, and rules out model-2.
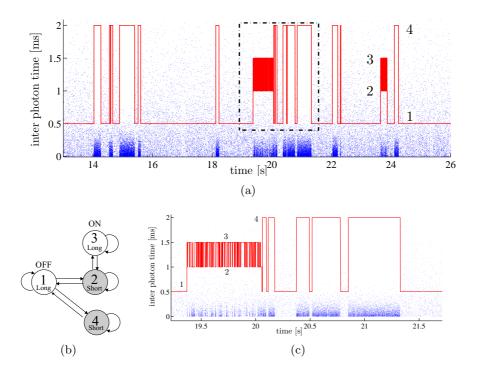
(a)

(b)

(c)

Figure 7.13: Viterbi path segmentation of the trace shown in Fig. 7.7(a) in (a) with a modified four state model presented in (b); (c) shows an excerpt of the sequence in (a). Periods with increased photon arrival rate can be automatically discriminated in a fast fluctuating ON state (state indices 2 and 3) and an ON state with constantly high photon arrival rates (state index 4).

7.6.1 Distinction between Different Kinds of ON states

To show the modeling power of a MMPP approach a modified state topology, presented in Fig. 7.13(b), is considered. It is assumed that the molecule can reside either in the OFF state (state number 1) or in one of two different ON states: one which allows for intermittence of the ON state (state numbers 2 and 3) and one in which the arrival rate of photons is consistently high (state number 4). Fig. 7.13(a) presents a segmentation of a sequence of photon detections with the new four state model structure. With the modified model, periods of intermittence in the photon arrival rates and periods with constantly high arrival rates can be automatically distinguished. Please note that the modified model correctly detects the transition between the fast fluctuation ON state and the ON with constantly high photon arrival rates in Fig. 7.13(c). The automatic segmentation with the MMPP matches the expected segmentation of the raw data by visual examination almost perfectly.

7.7 Conclusions

A MMPP was applied for the quantitative evaluation of the dynamics of the complex formation and dissociation of individual Cu^{2+}-ions with the *dcbpy* molecule, based on SMFS experiments. Like autocorrelation the MMPP has no limitations regarding the time resolution, and SMFS sequences can be analyzed with reduced user interaction, compared to techniques using binning. While autocorrelation does not need an a priori model to characterize dynamic processes, information about dynamic heterogeneities is lost in the process of averaging over the individual transitions occurring in single molecule traces. In contrast, dynamic heterogeneities are preserved in the Viterbi path generated by the MMPP and can be used for further data analysis allowing deeper insights into more complex molecular dynamics. Indeed, the MMPP has accurately detected intermittent fluctuations in the photon arrival intensities, which were not found when analyzing the data with binning [42]. The physical reason for these intermittent fluctuations that appear in parts of a sequences only is yet unknown and the subject of further investigation. The BIC values (eq. (7.18)) of the MMPP models investigated differ little; however the highest values are found for model-3B, a three state model that allows for intermittence of the ON-state. The same model also yields simulations that are perceptually closest to the experimental data.

Additionally, extension to multi-parameter detection schemes, like emission wavelength, fluorescence lifetime or polarization, is more straightforward for MMPP than for autocorrelation [41], thus opening new perspectives on the analysis of data acquired from single molecular events.

8 Conclusions

The goal of this work was to explore the use of state-space models to overcome hitherto unsolved problems in industrial process monitoring and the life sciences. State-space models were successfully employed for unusual event detection in weakly labeled image sequences for manufacturing process control and the estimation of kinetic rates from photon streams recorded from single-molecule spectroscopy experiments. For the different applications sustainable improvements with respect to labeling costs, classification performance and analysis accuracy compared to previously published techniques could be achieved.

A main contribution of this thesis in the field of industrial process control is the investigation of frameworks which can handle weakly labeled datasets. The performance of imaging sensor increases, i.e. higher and higher frames can be achieved and as a result the amount of gathered data grows for industrial monitoring systems. While for a small dataset elaborate labeling is feasible, for larger datasets, e.g. image sequences with a high number of frames, labeling gets tedious and expensive. Human expertise is necessary to improve the performance of a system but the interaction with the classification system must be as simple as possible.The presented work is a step toward the development of user friendly training and operation of classification systems for industrial image processing applications. Using only weakly labeled data for process control applications is an important factor to make classification more efficient and enable operation of the systems by non-data analysis experts.

Weakly labeled data was used for training classification systems for laser welding processes. The first proposed framework combined appearance based features and Hidden Markov Models to detect unusual events in image sequences. Starting from a regular sequence model, error sequence models are built incrementally using a temporal change detection algorithm to select outlier segments and then the regular sequence model is expanded with additional states. The performance of the framework showed promising results on data sets gathered from industrial laser welding processes as well as in an endurance tests performed directly at a production line. All erroneous sequences could be found with a small false positive rate. In particular, the use of *Eigen-MeltPools* have led to a significant improvement in the classification performance. Dimension reduction can be performed automatically from a training data set without the need for hand-crafted object-features. The second framework flags suspicious objects with a change detection algorithm, tracks suspicious objects with a Kalman Filter and bases the classification on trajectory features. The proposed tracking framework is used for the automatic online detection of sputter events. With this new approach for sputter detection , it was possible to

significantly reduce the false positive rate and the labeling costs compared to previously published methods.

In the life sciences a Markov-Modulated Poisson Process (MMPP) was applied for the automatic, quantitative evaluation of the dynamics of single molecules investigated with Single-Molecule Fluorescence Spectroscopy (SMFS). With minor modifications a MMPP can be evaluated with the same algorithms as HMMs. In view of the large amount of data which are acquired with SMFS when detecting individual photons, a simple Expectation-Maximization algorithm was chosen for parameter estimation based on considerations of computational performance. The developed framework was validated for the photon-by-photon analysis of the complex formation and dissociation of individual Cu^{2+}-ions with the *dcbpy* molecule. It was demonstrated that a MMPP can accurately detect fast fluctuations in the photon intensity which were not found in analyzing the data with traditional techniques such as binning. Compared to the binning approach, it could be shown that on the considered data set the temporal resolution could be increased by a factor of 3 to 10. In particular, in the analyzed sequences from the formation and dissociation of Cu^{2+}-ions with *dcbpy*, rapid fluctuations in photon intensity occur occasionally. These fluctuations were first observed when analyzing the raw data with MMPPs, without prior binning. The results clearly demonstrate that advanced analysis techniques for SMFS are necessary for a quantitative analysis of the sequences and that MMPPs can open new perspectives on the analysis of data acquired from single molecular events.

A Abbreviations

AUC	Area Under Receiver Operator Characteristic (ROC)
BIC	Bayesian Information Criterion
CML	Conditional Maximum Likelihood
CMOS	Complementary Metal Oxide Semiconductor
CHMM	Continious Hidden Markov Model
dcbpy	bidentate ligand 2,2'-bipyridine-4,4'- dicarboxic acid
DIC	Discriminative Information Criterion
DFFS	Distance From Feature Space
EM	Expectation-Maximization
ESM	Error Sequence Model
FCS	Fluorescence Correlation Spectroscopy
fps	frames per second
FRET	Fluorescence Resonance Energy Transfer
FN	False Negative
FP	False Positive
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
KFM	Kalman Filter Model
LDA	Linear Discriminant Analysis
MCE	Minimum Classification Error
MCMC	Markov Chain Monte Carlo
MHT	Multiple Hypothesis Tracking
ML	Maximum Likelihood
MMI	Maximum Mutual Information
MMPP	Markov Modulated Poisson Process
PC	Principal Component
PCA	Principal Component Analysis
PLS	Partial Least Squares
ROC	Receiver Operator Characteristic
ROI	Region Of Interest
RSM	Regular Sequence Model
SMFS	Single Moelcule Fluorescence Spectroscopy

A Abbreviations

SMS   Single Molecule Specroscopy
SVM   Support Vector Machine
TISC   Two Stage Image Sequence Classification
TMR   tetramethlyrhodamine
TN   True Negative
TP   True Positve
VCA   Video Content Analysis

Bibliography

[1] C. Alippi, P. Braione, V. Piuri, and F. Scotti, "A methodological approach to multisensor classification for innovative laser material processing units," in *Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference (IMTC'01)*, vol. 3, May 2001, pp. 1762 – 1767.

[2] J. Alon, S. Sclaro, G. Kollios, and V. Pavlovic, "Discovering clusters in motion time-series data," in *IEEE Computer Vision and Pattern Recognition Conference (CVPR'03)*, vol. 1, Madison, USA, 2003, pp. I–375 – I–381. [Online]. Available: citeseer.ist.psu.edu/alon03discovering.html

[3] E. Andrade, S. Blunsden, and R. Fisher, "Characterisation of optical flow anomalies in pedestrian traffic," in *The IET International Symposium on Imaging for Crime Detection and Prevention (ICDP'05)*, London, UK, 2005.

[4] M. Andrec, R. Levy, and D. Talaga, "Direct determination of kinectic rates from single-molecule photon arrival trajectories using Hidden Markov Models," *J. Phys. Chem. A*, vol. 107, pp. 7454 – 7464, 2003.

[5] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, January 2001.

[6] L. Atlas, M. Ostendorf, and G. D. Bernard, "Hidden Markov Models for monitoring machining tool-wear," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, Istanbul, Turkey, 2000, pp. 3887 – 3890.

[7] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Transactions on Information Theory*, vol. 20, pp. 284–287, March 1974.

[8] P. Baruah and R. Chinnam, "HMMs for diagnosis and prognostics in machining processes," *International Journal of Production Research*, vol. 43, no. 6, pp. 1275 – 1293, March 2005.

[9] J. F. Beausang, C. Zurla, C. Manzo, D. Dunlap, L. Finzi, and P. Nelson, "DNA looping analyzed using diffusive Hidden Markov Model," *Biophys. J. BioFAST*, 2007.

Bibliography

[10] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[11] Y. Bengio, "Markovian models for sequential data," *Neural Computing Surveys*, vol. 2, pp. 129–162, 1999.

[12] A. Biem, "A model selection criterion for classification: Application to HMM topology optimization," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Edinburgh, UK, 2003, pp. 104 – 108.

[13] ——, "Minimum classification error training for online handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1041–1051, 2006.

[14] J. A. Blimes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models," International Computer Science Institute, Berkely, CA, 94704 USA, Tech. Rep. TR-91-021, April 1998.

[15] A. Bollig, S. Mann, R. Beck, and S. Kaierle, "Einsatz optischer Technologien zur Regelung von Laserstrahlschweißprozeßen," *Automatisierungstechnik*, vol. 53, pp. 513 – 521, October 2005.

[16] M. Brocke, "Statistical image sequence processing for temporal change detection," in *Pattern Recognition*, ser. Lecture Notes in Computer Science, L. V. Gool, Ed., no. 2449, 24th Annual meeting of the German Association for Pattern Recognition. Springer, Heidelberg, 2002, pp. 215–223.

[17] ——, "Statistische Ereignisdetektion in Bildfolgen," Ph.D. dissertation, University of Heidelberg, 2002. [Online]. Available: http://www.ub.uni-heidelberg.de/archiv/3065

[18] J. Brunner and R. Krämer, "Copper(II)-quenched oligonucleotide probes for fluorescent DNA sensing," *J. Am. Chem. Soc.*, vol. 126, pp. 13 626–13 627, 2004.

[19] J. Brunner, A. Mokhir, and R. Krämer, "DNA-templated metal catalysis," *J. Am. Chem. Soc.*, vol. 125, pp. 12 410–12 411, 2003.

[20] S. Burk, "Optimierung und Verifizierung eines high-speed kamera systems zur automatisierten defekterkennung bei Laserschweißprozessen in der Großserienfertigung," Master's thesis, University of Siegen, 2006.

[21] T. Burzykowski, J. Szubiakowski, and T. Ryden, "Analysis of photon count data from single-molecule fluorescence experiments," *Chemical Physics*, vol. 288, pp. 291 – 307, 2003.

[22] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models.* Springer, New-York, 2005.

[23] M. T. Chan, A. Hoogs, J. Schmiederer, and M. Petersen, "Detecting rare events in video using semantic primitives with HMM," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, 2004, pp. 150 – 154.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* Toronto: John Wiley & Sons, Inc., 2001.

[25] *Heurisko, Digital Image Processing Software*, Fa. Aeon Verlag & Studio Walter H. Dorn, Hanau, Germany. [Online]. Available: http://www.heurisko.de

[26] T. Fawcett, "ROC graphs: Notes and pratical considerations for researchers," HP Laboratories, Tech. Rep., 2004.

[27] D. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, "Bayesian filtering for location estimation," *IEEE Pervasive Computing*, vol. 2, pp. 24–33, July-September 2003.

[28] A. E. Gamal and H. Eltoukhy, "CMOS image sensors," *IEEE Circuits and Devices Magazine*, vol. 21, pp. 6 – 20, 2005.

[29] T. Gevers and A. Smeulders, "Content-based image retrieval: An overview," in *Emerging Topics in Computer Vision*, G. Medioni and S. Kang, Eds. Prentice Hall, 2004.

[30] G. Grimmet and D. Stirzaker, *Probabilty and Random Processes.* Oxford University Press, 2001.

[31] S. Hader and F. A. Hamprecht, "Two-stage classification with automatic feature selection for an industrial application," in *Classification - the ubiquitous challenge, Proceedings of the 28th Annual Conference of the German Classification Society*, B. Michaelis and G. Krell, Eds. Springer, Heidelberg, 2004, pp. 136 – 144.

[32] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* Springer-Verlag Heidelberg, 2001.

[33] J. Hornegger, D. Paulus, and H. Niemann, "Probabilistic modeling in computer vision," in *Handbook of Computer Vision and Applications*, B. Jähne, H. Haußecker, and P. Geißler, Eds. Academic Press, San Diego, 1999, vol. 2.

[34] S. Humbert, "Partikel-Verfolgung beim Laser-Schweißen mit dem Kalman-Filter," Master's thesis, University of Heidelberg, May 2005.

[35] M. Inoue and N. Ueda, "Exploitation of unlabeled sequences in Hidden Markov Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1570–1581, 2003.

Bibliography

[36] T. Jaakkola and D. Hausser, "Exploiting generative models in discriminative classifiers," in *Proceesings of the 1998 conference on Advanaces in neural information processing systems II*, S. Solla and D. Cohn, Eds., vol. 11. MIT Press, Cambridge, MA, USA, 1999, pp. 487–493.

[37] A. K. Jain, *Fundamentals of Digital Image Processing*, T. Kailath, Ed. Prentica Hall, 1989.

[38] C. Joo, S. McKinney, M. Nakamura, I. Rasnik, S. Myong, and T. Ha, "Real-time observation of RecA filament dynamics with single monomer resolution," *Cell*, vol. 126, pp. 515 – 527, 2006.

[39] H. Junno, P. Laurinen, E. Haapalainen, L. Tovinen, and J. Röning, "Resistance spot welding process indentification using an extended knn method," in *Proceedings of the IEEE International Symposium on Industrial Electronics*, Dubrovnik, Croatia, June 2005, pp. 7–12.

[40] S. Kapadia, "Discriminative training of Hidden Markov Models," Ph.D. dissertation, Cambridge University, 1998.

[41] P. Kapusta, M. Wahl, A. Benda, M. Hof, and J. Enderlein, "Fluorescence lifetime correlation spectroscopy," *Journal of Fluorescence*, vol. 17, pp. 43 – 48, 2007.

[42] A. Kiel, A. Mokhir, R. Krämer, and D.-P. Herten, "Direct monitoring of formation and dissociation of individual metal complexes by single molecule fluroscence spectoscropy," *Angewandte Chemie*, vol. 46, no. 18, pp. 3363–3366, April 2007.

[43] S. Kou, X. S. Xie, and J. S. Liu, "Bayesian analysis of single-molecule experimental data," *Applied Statistics*, vol. 53, pp. 469 – 506, 2005.

[44] C. Kratzsch, "Realisierung eines kamerabasierten Prozessüberwachungssystems am Beispiel des Laserstrahlschweißens," Ph.D. dissertation, RWTH Aachen, 2002.

[45] A. Krogh, M. Brown, I. S. Mian, K. Sjlander, and D. Haussler, "Hidden Markov Models in computational biology applications to protein modeling," *Journal of Molecular Biology*, vol. 235, pp. 1501–1531, 1994.

[46] H. Kurihata, T. Takahashi, I. Ide, Y. Mekada, H. Murase, Y. Tamatsu, and T. Miyahara, "Rainy weather recognition from in-vehicle camera images for driver assistance," in *Proceedings IEEE Intelligent Vehicles Symposium*, 2005, pp. 205 – 210.

[47] J. A. Lasserre, C. M. Bishop, and T. P. Minka, "Principled hybrids of generative and discriminative models," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 87–94.

[48] D. Li, A. Biem, and J. Subrahmonia, "HMM topology optimization for handwriting recgonition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP"01)*, vol. 3, 2001, pp. 1521 – 1524.

[49] S. Z. Li and J. Lu, "Face detection, alignment, and recognition," in *Emerging Topics in Computer Vision*, G. Medioni and S. B. Kang, Eds. Prentice Hall, 2004.

[50] M. Lippitz, F. Kulzer, and M. Orrit, "Statistical evaluation of single nano-object fluorescence," *Chem Pys Chem*, vol. 6, pp. 770 – 789, 2005.

[51] H. P. Lu, L. Xun, and X. S. Xie, "Single-molecule enzymatic dynamics," *Science*, vol. 282, pp. 1877 – 1882, 1998.

[52] E. McDermott, T. J. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 203 – 223, January 2007.

[53] K. McHale, A. Berglund, and H. Mabuchi, "Bayesian estimation for species identification in single-molecule fluorescence micropscopy," *Biophysical Journal*, vol. 86, pp. 3409 – 3422, 2004.

[54] S. A. McKinney, C. Joo, and T. Ha, "Analysis of single-molecule FRET trajectories using Hidden Markov modeling," *Biophysical Journal*, vol. 91, pp. 1941 – 1951, 2006.

[55] B. H. Menze, B. M. Kelm, and F. A. Hamprecht, "From eigenspots to fisherspots - latent spaces in the nonlinear detection of spot patterns in a highly variable background," *Advances in Data Analysis, Studie sin Classification, Data Analysis and Knowledfe Organization*, vol. 33, to appear 2007. [Online]. Available: http://klimt.iwr.uni-heidelberg.de/mip/publications/menze_06_eigenspots.pdf

[56] D. C. Minnen and C. P. Wren, "Finding temporal patterns by data decomposition," in *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04)*, Seoul, Korea, 2004, pp. 608 – 613.

[57] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696 – 710, 1997.

[58] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing.* Upper Saddle River, NJ: Prentice Hall, 1999.

[59] J. Müller-Borhanian, C. Deininger, F. Dausiger, and H. Huegel, "Spatially resolved on-line monitoring during laser beam welding of steel and aluminum," in *Proceedings of 23rd International Congress on Applications of Lasers and Electro-Optics (ICALEO'04)*, 2004.

Bibliography

[60] K. P. Murphy, "Dynamic bayesian networks: Representation, inference and leraning," Ph.D. dissertation, University of California, Berkeley, 2002.

[61] D. Naso, B. Turchiano, and P. Pantaleo, "A fuzzy-logic based optical sensor for online weld defect-detection," *IEE Transactions on Industrial Informatics*, vol. 1, pp. 259–273, November 2005.

[62] K. Nigam, A. MacCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 31, pp. 103 – 134, 2000.

[63] J. Petereit, P. Abels, S. Kaierle, C. Kratzsch, and E. Kreutz, "Failure recognition and online process control in laser beam welding," in *Proceedings of 21st International Congress on Applications of Lasers and Electro-Optics (ICALEO)*, 2002, pp. 91–98.

[64] *User Manuals for Photonfocus Products*, Photonfocus AG, Lachen, Switzerland. [Online]. Available: http://www.photonfocus.com/html/eng/support/userManuals.php

[65] F. Qin, A. Auerbach, and F. Sachs, "Hidden Markov Modeling for single channel kinetics with filtering and correlated noise," *Biophysical Journal*, vol. 79, pp. 1928 – 1944, October 2000.

[66] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257 – 286, 1989.

[67] J. Rehs, "Appearance based features for the evaluation of image sequences of a camera system monitoring a laser weld process," Master's thesis, TU Ilmenau, 2007.

[68] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19 – 41, 2000.

[69] C. Rosenberg, "Semi-supervised training of models for appearance-based statistical object detection methods," Ph.D. dissertation, Carnegie Mellon University, May 2004.

[70] C. Rosenberg, M. Henert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *IEEE Workshop on Applications of Computer Vision (WCAV'05), Breckenridge, Colorado*, 2005. [Online]. Available: http://www.cs.cmu.edu/~chuck/pubpg/wacv-2005/wacv2005.pdf

[71] A.-V. Rosti, "Linear gaussian model for speech recognition," Ph.D. dissertation, University of Cambridge, 2004.

[72] T. Ryden, "An EM algorithm for parameter estimation in Markov Modulated Poisson Processes," *Computational Statistics and Data Analysis*, vol. 21, pp. 431 – 447, 1996.

[73] G. K. Schenter, H. P. Lu, and X. S. Xie, "Statistical analyses and theoretical models of single-molecule enzymatic dynamics," *J. Phys. Chem. A.*, vol. 103, pp. 10 477 – 10 488, 1999.

[74] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[75] R. Schwarte, H. Heinol, B. Buxbaum, T. Ringbeck, Z. Xu, and K. Hartmann, *Principles of Three-dimensional Imaging Techniques*, B. Jähne, H. Haußecker, and P. Geißler, Eds.   Academic Press, San Diego, 1999.

[76] S. L. Scott, "A Bayesian paradigm for designing network intrusion systems," *Computational Statistics and Data Analysis*, vol. 45, no. 1, pp. 69 – 83, 2004.

[77] A. Selinger and R. Nelson, "Minimally supervised acquisition of 3D recognition models from cluttered images," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, vol. 1, 2001, pp. 213–220.

[78] F. Sha and L. K. Saul, "Large margin Hidden Markov Models for automatic speech recognition," in *Processdings of Neural Information Processing Systems (NIPS'07)*, B. Schölkopf, J. Platt, and T. Hofmann, Eds., vol. 19.   MIT Press, Cambridge, MA, 2007.

[79] J. Shao and Y. Yan, "Review of techniques for on-line monitoring and inspection of laser welding," *Journal of Physics: Conference Series*, vol. 15, pp. 101–107, 2005. [Online]. Available: http://stacks.iop.org/1742-6596/15/101

[80] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A*, vol. 4, pp. 519–524, 1987.

[81] N. Smith and M. Gales, "Using SVMs to classify variable length speech patterns," Cambridge University, Tech. Rep. CUED/F-INFENG/TR.412, 2002. [Online]. Available: http://mi.eng.cam.ac.uk/~mjfg/tr412.ps.gz

[82] P. Smyth, "Markov monitoring with unkown states," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 1600 – 1612, 1994.

[83] J. Snoek, J. Hoey, L. Stewart, and R. S. Zemel, "Automated detection of unusual events on stairs," in *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, 2006, pp. 5–5.

Bibliography

[84] N. Stache, H. Zimmer, J. Gedicke, B. Regaard, A. Olowinsky, A. Knepper, and T. Aach, "Approaches for high-speed melt pool detection in laser welding applications," pp. 217–224, November 22-24 2006.

[85] W. M. Steen, *Laser material processing*, 3rd ed. London: Springer, 2003.

[86] P. Strohm, "Defekterkennung und Klassifikation von Laserschweißprozessen anhand von Multi-Kamera-Bildsequenzen," Master's thesis, Technical University Ilmenau, 2005.

[87] R. Sun, "Introduction to sequence learning," in *Sequence Learning - Paradigms, Algorithms, and Applications*, R. Sun and C. L. Giles, Eds. Springer, Heidelberg, 2001, vol. 1828.

[88] D. S. Talaga, "Information theoretical approach to single molecule experimental design and interpretation," *Physical Chemistry A*, vol. 110, pp. 9743 – 9757, 2006.

[89] I. The MathWorks, *Matlab Help*, Natick, MA, United States. [Online]. Available: http://www.mathworks.com/access/helpdesk/help/helpdesk.shtml

[90] P. Tinnefeld, D.-P. Herten, and M. Sauer, "Photophysical dynamics of single dye molecules studied by spectrally-resolved fluorescence lifetime imaging microscopy (SFLIM)," *J. Phys. Chem. A.*, vol. 105, pp. 7989–8003, 2001.

[91] W. Turin, "Fitting probabilistic automata via the EM algorithm," *Communications in Statistics – Stochastic Models*, vol. 12, pp. 405 – 424, 1996.

[92] M. Turk and A. P. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71 – 86, March 1991.

[93] I. Ulusoy and C. Bishop, "Generative versus discriminative methods for object recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE Computer Society, Washington, DC, USA, 2005, pp. 258–265.

[94] C. van Loan, "Computing integrals involving the matrix exponential," *IEEE Transactions on Automatic Control*, vol. 3, pp. 395 – 404, 1978.

[95] R. Veaux, L. Ungar, and J. Vinson, "Statistical approaches to fault analysis in multivariate process control," in *Proceedings of the American Control Conference*, 1995, pp. 1274–1276. [Online]. Available: citeseer.ist.psu.edu/231120.html

[96] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis Part i: Quantitative model-based methods," *Computers and Chemical Engineering*, vol. 27, pp. 293–311, 2003.

[97] M. Wahl, I. Gregir, M. Patting, and J. Enderlein, "Fast calculation of fluorescence correlation data with asynchronous time-correlated single-photon counting," *Optics Express*, vol. 11, no. 26, pp. 3583 – 3591, 2003.

[98] L. Wang, M. G. Mehrabi, and E. Kannatey-Asibu, "Hidden markov model-based tool wear monitoring in turning," *Journal of Manufacturing Science and Engineering*, vol. 124, no. 3, pp. 651 – 658, 2002.

[99] M. Yu and M. Zjou, "A model reduction method for traffic described by mmpp with unkown rate limit," *IEEE Communications Letters*, vol. 10, pp. 302 – 304, 2006.

[100] D. Zhang, "Probabilistic graphical models for human interaction analysis," Ph.D. dissertation, Swiss Federal Institute of Technology, Lausanne, 2006. [Online]. Available: http://www.idiap.ch/publications/zhang-rr-06-78.bib.abs.html

[101] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted HMMs for unusual event detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE Computer Society, Washington, DC, USA, 2005, pp. 611–618.

[102] H. Zhong, J. Shi, and M. Viontai, "Detecting unusual acitivity in video," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, vol. 2, 2004, pp. 819–826.

[103] H. Zhou and D. Kimber, "Unusual event detection via multi-camera video mining," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*. Hong Kong, China: IEEE Computer Society, Washington, DC, USA, 2006, pp. 1161–1166.

[104] S. Zhou, J. Zhang, and S. Wang, "Fault diagnosis in industrial processes using principal component analysis and Hidden Markov Model," in *Proceeding of the 2004 American Control Conference*, vol. 6, Boston, MA, United States, 2004, pp. 5680–5685.