Weakly Supervised Learning of a Classifier for Unusual Event Detection

Mark Jäger, Christian Knoll and Fred A. Hamprecht*

Abstract

In this paper, we present an automatic classification framework combining appearance based features and Hidden Markov Models (HMM) to detect unusual events in image sequences. One characteristic of the classification task is that anomalies are rare. This reflects the situation in the quality control of industrial processes, where error events are scarce by nature. As an additional restriction, class labels are only available for the complete image sequence, since frame-wise manual scanning of the recorded sequences for anomalies is too expensive and should therefore be avoided.

The proposed framework reduces the feature space dimension of the image sequences by employing subspace methods and encodes characteristic temporal dynamics using continuous Hidden Markov Models (CHMMs). The applied learning procedure is as follows: 1) A generative model for the regular sequences is trained (one-class learning). 2) The regular sequence model (RSM) is used to locate potentially unusual segments within error sequences by means of a change detection algorithm (outlier detection). 3) 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. Experimental results are presented for image sequences recorded from industrial laser welding processes. We demonstrate that the learning procedure can significantly reduce the user interaction and that sequences with error events can be found with a small false positive rate. It has also been shown that a modeling of the temporal dynamics is necessary to reach these low error rates.

Index Terms

State-space models, Weak labels, One-class learning, Outlier detection, Time series classification

I. Introduction

In many event detection applications, the aberrations of interest occur only rarely. This lack of positive examples is, for instance, a characteristic of efficient industrial processes, and complicates the training of automated systems that rely on statistical learning. In this paper, we consider the task of finding error events within image sequences recorded from an industrial manufacturing process. In this laser welding application, error events can be distinguished from regular frames in terms of their spatial appearance and temporal dynamics. Class labels are only available for the complete sequence (sequence labels) and specify the quality of the produced part (either error free or erroneous). It would be overly time consuming to obtain labels for each individual frame within a sequence (frame labels) due to the large amount of data (in the present application, an image sequence comprises up to 4000 frames). However, it is deemed acceptable to obtain frame labels in very few cases in which the error event cannot be clearly located within a sequence.

As others [1]–[7] have done before, we propose to use Hidden Markov Models (HMMs) to capture the dynamics of the process and ultimately discriminate between regular and erroneous sequences. The focus of this contribution is to find ways that allow to train these HMMs with the smallest possible amount of user interaction. In particular, we propose a strategy that allows to use a strongly imbalanced and only weakly labeled training set. That is, there are only few examples of the error class, and there is only one label for the entire sequence, even if it is only a few frames that account for an error event.

Since the raw data is too high-dimensional to allow for an efficient learning, the recorded images are projected to a low-dimensional feature space. The resulting feature vector can be considered as a multivariate time series over the complete sequence. The temporal dynamics of the process are captured using continuous Hidden Markov Models (CHMMs) with Gaussian Mixture Models (GMMs). Hidden Markov Models (HMMs) are widely used probabilistic models for the analysis of time sequences, especially in the area of speech recognition, gesture recognition and bioinformatics, and efficient algorithms exist for their implementation [8]. The task considered in this paper is closely related to the problem of unusual event detection in video sequences, where HMMs showed promising results [9]–[14]. In particular, it has been found that HMMs are effective for unusual event detection, if the normal activity exhibits little variability [15]. This assumption is rarely fulfilled for video sequences of natural scenes which tend to exhibit high variability even in the normal state, but it is valid for industrial quality control applications. By definition, industrial processes must be highly repeatable, and therefore regular variations in an industrial process are relatively small.

In the setting considered here, error events (positive events) are rare and only examples of regular sequences (negative examples) are available in the beginning. The feature subspace is therefore computed from regular sequences only and the resulting features of the available negative examples are used to train a regular sequence model (RSM) using CHMMs. In the next step, the RSM is used to identify unusual segments within error sequences and the detected outlier segments are used to expand the RSM model with additional states and to thus obtain an error sequence model (ESM).

Related ideas for HMMs with GMMs have been considered for speech recognition [16] and unusual event detection in video sequences [9]. In [9], 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 event with the lowest likelihood under the regular model is used to add an additional state to the regular model. In contrast in our approach, the expansion of the regular model is based on a change detection algorithm on the posterior frame probabilities in order to ensure independence from absolute likelihood values. In addition, we use a discriminative model selection criterion which only adds states to the error sequence model if they better help to discriminate between a regular sequence and an error sequence.

Several authors have investigated the benefits of HMMs for the detection of error states in industrial manufacturing processes [1]–[7]. In [1], [2] it is shown that HMMs can improve the precision of detecting known and unforeseen error states in process control applications. [3], [5] consider the task of finding sudden changes in tool wear for machining processes like drilling or milling with HMMs. In [6] principal component analysis and HMMs are combined for on-line fault detection in industrial processes, and the main contribution of [7] is to use trained HMMs not only for an automatic diagnosis of machining processes but also for their prognostics. In [4], an algorithm for detecting changes in the transition probability matrix is presented to identify faults. To our knowledge, learning strategies for weakly labeled data for the detection of unusual events in industrial processes have not been considered previously.

The paper is organized as follows: Section II describes the proposed framework with its dimension reduction, incremental model building and model selection. Results of unusual event detection in image sequences from industrial laser welding processes are presented in section III, followed by a discussion in section IV. Conclusions are offered in section V.

II. Incremental Learning for Unusual Event Detection in Image Sequences

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]. 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. 1; the implementation details follow in the next sections.

1) Subspace Computation & Dimension Reduction: The individual images of the recorded sequences can be considered as high-dimensional feature vectors. The number of features is much larger than the number of observations, making an implicit or explicit dimension reduction necessary. We have opted for the latter and have employed principal component analysis (PCA) to find an appropriate linear subspace [17], [18] . Due to the lack of positive examples, the PCA subspace is computed using regular sequences only. The result of the dimension reduction step is a time series of low dimensional feature vectors containing the relevant information of each image from the regular sequences.

2) HMM Training: Next, a model for the multivariate time series is trained. HMMs (with Mixtures of Gaussians as a model for the probability density of the emission distributions) are employed to represent the temporal dynamics of the underlying process and to achieve invariance to varying sequence length and position of the anomalies within a sequence. HMMs with GMMs belong to the class of generative models. A generative learning

[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.

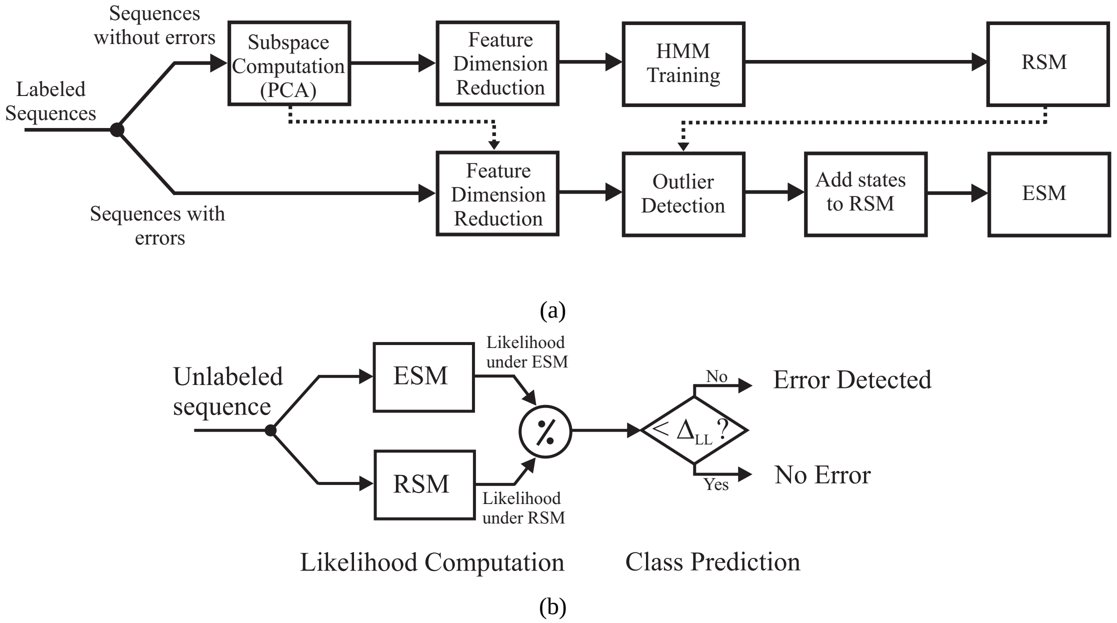


Fig. 1. Schematic overview of the training procedure of the classification system (a) and its application to new sequences (b). In the training stage, a linear subspace is computed from the sequences without errors using principal component analysis (PCA) and the recorded images are projected into this subspace to reduce the feature dimension. Using these features a Hidden Markov Model (HMM) for the regular sequences (RSM) is trained. The RSM 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) by adding additional states. The likelihood ratio of the data given the ESM and RSM is used for the classification decision (b).

approach is opted, since only sequence labels but no frame labels are available in the current application [19]. The discriminative classification of complete sequences would be possible, but this results in a very high-dimensional classification problem and invariance to the sequence length has to be achieved in pre-processing steps. A direct estimation of the class boundary, as in discriminative methods, is not possible on a single frame basis without labels or prior assumptions. HMMs allow to detect unusual frames in terms of their exceptionally low likelihood given the trained model.

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 maximum likelihood (ML), and the model with the optimal number of parameters is referred to as regular sequence model (RSM). It represents the class of sequences without errors.

3) Model Expansion: Sequence labels are available for the training of an error sequence model (ESM) which is obtained by expanding the RSM with additional states. The adaptation works 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 the 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.

4) Classification: For unlabeled sequences, the classification decision is based on the log-likelihood ratio of the data between the ESM and RSM (see Fig. 1(b)). If the log-likelihood ratio is above an empirical threshold Δ_{LL}, which is determined from the training data set, the sequence is marked as erroneous.

In the following, the implementation of the incremental learning approach is specified in more detail.

A. Dimension Reduction

PCA is a traditional technique for dimension reduction. It seeks a projection that best approximates the data in a least-square sense. For the subspace computation, only samples from the majority class are used. The transformation decomposes a feature space into a principal subspace F and an orthogonal complementary space $\overline{F}$. The residual error ϵ, also called **D**istance **F**rom **F**eature **S**pace (DFFS), is the Euclidean distance of a point in feature space from the subspace F. Both the components in the F-Space (corresponding to the directions which describe the major variations in the data) as well as the residual error carry information that can be used for classification [18]. The DFFS signal increases for images that are far from the trained subspace, and is therefore a measure of novelty.

PCA approximates the data in terms of a single multivariate Gaussian distribution, hence only first and second order statistical dependencies of the pixels can be considered. For complex objects such as faces it is often not possible to capture the important information for recognition or discrimination with a single covariance matrix and therefore extensions of PCA have been investigated [20]. Independent Component Analysis (ICA) is one possible method to take into account higher order pixel dependencies. 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 melt pools which are simple objects compared to faces.

B. Hidden Markov Model

HMMs are one of the most popular methods in statistics and machine learning for modeling sequences and are used extensively in applications such as speech or gesture recognition. At time instance k the HMM exists in one of a finite set of states $Q_k = j$ with $1 \leq j \leq N_Q$. 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. Stochastic transitions between states are governed by a transition probability matrix A. Each state Q_k that could be visited at time instance k could emit a single observation o_k according to a probability distribution that is specific to that state. The probability of a particular, continuous observation o_k in state $Q_k = j$ is given by $b_j(o_k) = P(o_k|Q_k = j)$ and is modeled with a GMM with parameters $\lambda_j = \{w_{j,m}, \mu_{j,m}, \Sigma_{j,m}\}$:

$$b_j(o_k) = \sum_{m=1}^{M} w_{j,m} P(o_k|\mu_{j,m}, \Sigma_{j,m}), \tag{1}$$

where M is the total number of mixture components, $w_{j,m}$ is the weight of the m^{th} mixture component of state j, and $P(o_k|\mu_{j,m}, \Sigma_{j,m})$ specifies a multivariate normal distribution with mean vector $\mu_{j,m}$ and covariance matrix $\Sigma_{j,m}$. The complete set of HMM parameters for a particular model is summarized by $\Phi = \{\pi, \lambda, A\}$ where π is the initial state distribution at time $k = 0$. The parameters $\widehat{\Phi}$ of the HMM are estimated from observed data with expectation-maximization (EM) algorithm. Starting from an initial guess, the EM algorithm is an iterative procedure to find the maximum-likelihood (ML) estimate of the unknown parameters Φ [8], [21].

C. Training of the Regular Sequence Model

Instead of immediately training a separate HMM with parameters Φ_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 Φ_R for the regular sequence model (RSM) can be estimated with high precision. The feature vector $o_k = [y_k, \epsilon_k]$ at time k consists of the principal components (PC) y_k and the residual error ϵ_k. The HMM parameters Φ_R are estimated with the EM algorithm and the number of different states N_Q is optimized using the Bayesian Information Criterion (BIC) [22]:

$$BIC(\widehat{\Phi}) = \log P(O|\widehat{\Phi}) - \frac{K_P}{2} \log K_D \tag{2}$$

where $O = \{o_1, o_2, \ldots, o_K\}$, K_P are the number of free model parameters and K_D is the size of the data set. The first term in eq. (2) 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 covariance matrices $\Sigma_{j,m}$ are assumed to be diagnoal and all state transitions are allowed. The BIC is estimated using a 5-fold cross validation from the training data.

D. Expansion of the RSM to the ESM by adding additional states

A crucial step in the incremental learning procedure is to select the outlier data which is used to find a first estimate of the parameters for the additional error states. The selection is based on a temporal change detection algorithm: Each weakly labeled sequence is partitioned into segments of constant size; for each segment s the log-likelihood L_s is approximated using the forward probability α, i.e. the probability of observing the partial sequence $O_{1:k} = \{o_1, ..., o_k\}$ and ending up in state $Q_k = j$ at time k:

$$\alpha_{j,k} = P\left(O_{1:k}, Q_k = j|\Phi\right) \tag{3}$$

The log-likelihood L_s for segment s is approximated with:

$$L_s = \sum_{k \in s} \sum_{j=1}^{N_Q} \alpha_{j,k}, \tag{4}$$

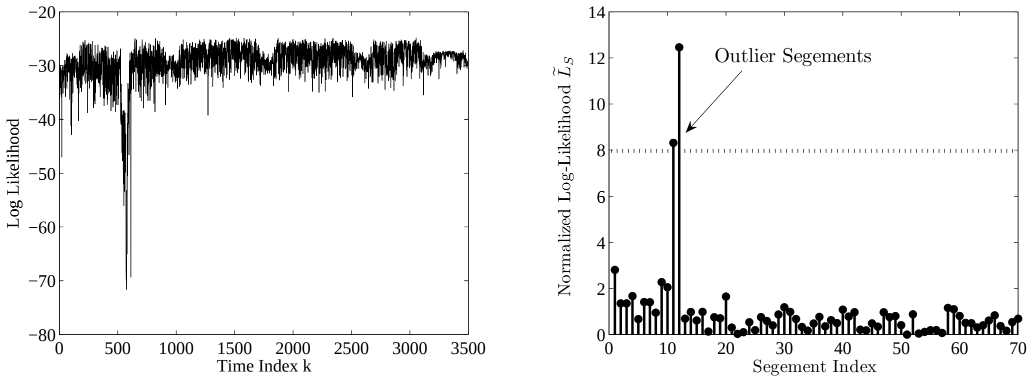
where N_Q is the number of different states of the RSM. A robust temporal change detection is used to flag those segments within a sequence that show an abnormal change in the log-likelihood:

$$\widetilde{L}_s = \frac{|L_s - \mathrm{med}_p(L_p)|}{\mathrm{med}_q\,|L_q - \mathrm{med}_p(L_p)|}, \tag{5}$$

where med is the median operator and p and q are segment indices ranging over the number of segments s In general the unusual event constitutes only part of the sequence and therefore manifests itself through a significant change in the log-likelihood (see e.g. Fig.2(a) e. Outlier segments are found with an empirical fixed threshold T_{L_s}; if $\widetilde{L}_s > T_{L_s}$, then segment s is identified as incompatible with the RSM (see Fig.2(b)).

A schematic overview of the ESM training is presented in Fig. 3. First the outlier segments are used to estimate the parameters and transition probabilities between the newly added error states. Next, the complete error sequences (including the error free part) are used to estimate the transition probabilities 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. It is therefore possible that unusual segments which could not be found by the conservative outlier detection can now be found by using the EM algorithm [23]. 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 states, and vice versa. The transition probabilities from the states of the RSM to the newly added states are set to small constant values. Since sequences with unusual events are rare, the transition probabilities 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 encode the mean duration of an error event and can be determined directly from the training sequences.

The number of necessary additional error states Q_{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 [24] and the



(a) (b)

Fig. 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) is presented.

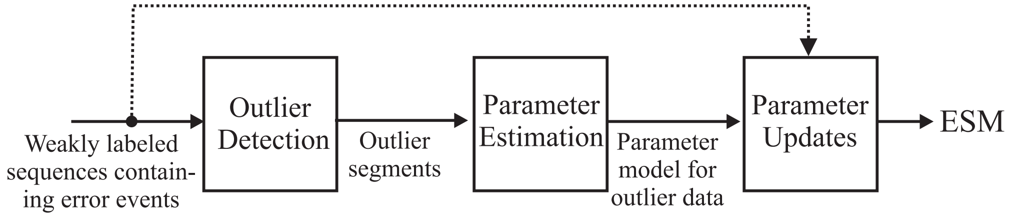

Weakly labeled sequences containing error events
Outlier Detection
Outlier segments
Parameter Estimation
Parameter model for outlier data
Parameter Updates
ESM

Fig. 3. Schematic overview to estimate the parameters of the error sequence model (ESM).

criterion is referred to as the Discriminative Information Criterion (DIC):

$$DIC(\Phi_E) = \frac{1}{N_E}\sum_{u=1}^{N_E} \log P(O_E^u|\Phi_E) - \frac{1}{N_R}\sum_{v=1}^{N_R} \log P(O_R^v|\Phi_E) - \frac{K_{\text{add}}}{2}\log K_D \qquad (6)$$

where K_{add} are the additional parameters for the error sequence model (ESM) and O_E^u/O_R^v is the u^{th}/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 K_{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.

E. Maximum A Posteriori Decoder

Each image sequence belongs to a distinct sequence class ω_i modeled by a HMM with parameters Φ_i. The Maximum A Posteriori (MAP) decoder assigns unlabeled sequences to the class ω_{MAP} with the highest posterior probability:

$$\omega_{MAP}(O) = \arg\max_{\omega_i} \log P(\omega_i|O) \tag{7}$$

In the following, a two class problem is considered: An unlabeled sequence can either be assigned to the regular sequence class ω_R or error sequence class ω_E. Sequences which belong to the error sequence class ω_E contain error events. The two classes can be compared using the posterior log-likelihood ratio:

$$\begin{aligned} \rho(O) &= \log \frac{P(\omega_E|O)}{P(\omega_R|O)} + \ln \left[\frac{C_E}{C_R} \frac{P(\omega_E)}{P(\omega_R)} \right] && (8) \\ &= \log \frac{P(\omega_E|O)}{P(\omega_R|O)} + \Delta_{LL} && (9) \end{aligned}$$

where $\exp\left(\Delta_{LL}\right) = \frac{C_E}{C_R}\frac{P(\omega_E)}{P(\omega_R)}$ is the ratio of model priors weighted with the non-symmetric cost factors C_E and C_R for the sequences containing errors and the error free sequences. If $\rho(O) > 0$, the MAP decoder predicts class $\omega_{MAP} = \omega_E$, otherwise it predicts $\omega_{MAP} = \omega_R$. The weights C_E and C_R are introduced to enable a trade off between the false positive (FP) rate and the false negative (FN) rate. Since $P(\omega_E)/P(\omega_R) << 1$, the MAP decoder would almost always select ω_R in order to a achieve an overall minimum probability of error. But 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 the threshold Δ_{LL} can be interpreted as a measure of the severity of an error event.

III. Experiments

The motivation for the design of the unusual event detection system is the quality inspection of laser welding sequences. Many welds have high quality demands and one possibility to satisfy the quality requirements is to monitor the welding process with high speed cameras. The interaction between the laser radiation and work piece leads to the generation of secondary radiation. This radiation contains information about the stability and the dynamics of the welding process. Therefore many process inspection methods are based on the evaluation of these emissions (see e.g. [25]–[27]). Unusual events in the recorded sequences of the laser welding process correlate with faults on the produced weld seam.

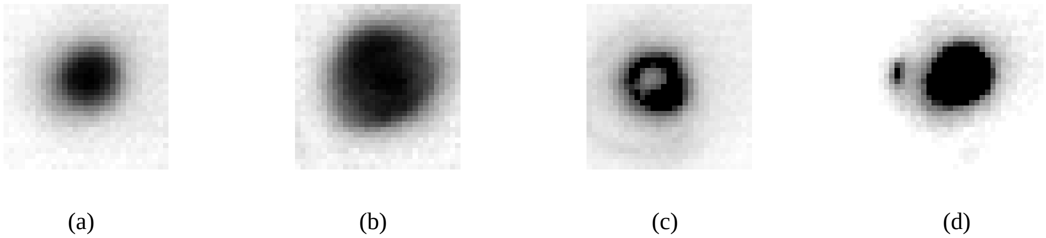
A. Data Description

The experimental data was gathered over a 4 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 $64{\times}64$ pixel. The recorded weld images correspond to a field of view of approximately 0.9×0.9 mm^2 . The welding process was controlled with a temperature sensor to achieve a constant weld seam depth resulting in a very dynamic process, which makes it challenging to distinguish between normal process fluctuations and abnormal error events in the recorded sequences. The manufactured weld seams were visually inspected by experts and matched to the corresponding sequences using an identification number. In total, 99 parts with weld errors were collected and classified in 3 different error classes ω_{E1}, ω_{E2}, and ω_{E3} by visual inspection:

- ω_{E1}: annealing material particles
- ω_{E2}: weld reinforcements / weld break-in [2]
- ω_{E3}: general irregularities

The extent of the error on the manufactured part was also rated between 1 (weak) to 3 (strong) by visual examination. Each sequence was screened to ensure that the position of the fault on the weld seam and the irregularity in the raw image sequence coincide. The screening ensured a fair examination of the algorithms, since it was ensured that the weld seam error is present in the used sequences. In addition, around 1000 sequences from error free welding processes were collected. The sequences were uniformly sampled from the observation period in order to capture the regular process fluctuations. A sample image from a recorded sequence of an error-free weld seam is shown in Fig. 4(a). The disc shaped object is the recorded radiation from the laser induced plasma and is in the following referred to the melt pool. Typical deformations which indicate weld seam faults are shown in Fig. 4(b) to Fig. 4(d).

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



(a) (b) (c) (d)

Fig. 4. Example images of the welding process as recorded by a CMOS camera (inverted colormap). (a) Sample intensity distribution of a regular welding process and (b) - (d) recorded intensity distributions of error events (only a 31×31 region of interest is shown.)

B. Subspace Decomposition

The feature subspace is computed from individual frames of error free welding processes (see Fig. 4(a)). A region of interest (ROI) of size 31×31 pixel is automatically extracted from the recorded images (64×64 pixel). The ROI is centered around the average center of mass of a melt pool determined from all melt pool images belonging to the same welding sequence. . The melt pool can be fully observed in this ROI and, since only part of the recorded image is used, minor translations (± 16 pixel) 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. The melt pool is rotation invariant and no major changes are expected in its scale and brightness unless in case of an error event. Since the observed gray values are directly used for the subspace computation, translation invariance has to be ensured.

The mean image and the first 4 eigenvectors of the computed subspace from the regular sequences are shown in Fig. 5. The eigenvectors describe the principal deviations from the mean image for regular sequences. For the computation of the residual error ϵ the first 20 eigenvectors are used, which cover approximately 98% of the total variance of the recorded images from error free welding sequences. In addition, the first 3 principal components (PC) corresponding to the eigenvectors with the largest eigenvalues are used in the feature vector[3]. The time series of feature vectors are normalized over time for each sequence and for each feature separately, in order to compensate for normal process fluctuations. The mean and variance for normalization are estimated from the data between the first and third quartile, to reduce the effect of outliers which should be detected.

[3] The first 3 principal components cover 55%, 16% and 8% of the total variance of the recorded images, respectively.

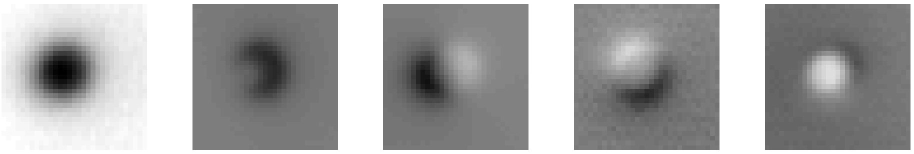


Fig. 5. Mean image followed by the first 4 eigenvectors (“Eigen-MeltPools”) describing the principal deformations of a regular weld sequence.

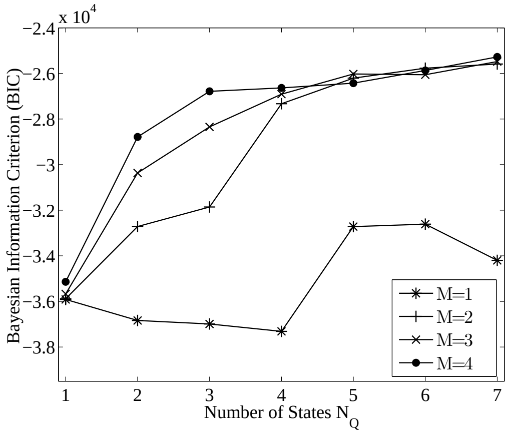


Fig. 6. Topology optimization for the Regular Sequence Model (RSM) with the Bayesian Information Criterion (BIC). The number of states N_Q and the number of mixture elements M are varied.

C. Regular Sequence Model (RSM)

First the parameters for the RSM were determined. Approximately 40% of the data from regular sequences were used to estimate the parameters, the other 60% were used for test purposes. The obtained BIC values for different model complexities for the training data set are compared in Fig. 6. Additional mixture elements offer no significant increase in the BIC beyond $N_Q = 4$.

For the current application, model complexity was severely penalized to avoid overfitting and high computation times. An additional state or mixture element was added only if it increased the BIC by more than 5%. The optimal number of states was found to be $N_Q = 4$ and the optimal number of mixture components $M = 2$. This low number of Gaussian mixture components ensures that temporal dynamics are modeled with different states instead of different mixture elements.

D. Error Sequence Model (ESM)

The information from the sequence labels was used to extend the RSM to different ESMs. A separate HMM was trained for each error class ω_{Ei}. The segment length for the change detection algorithm was set to 50 frames and the threshold T_{L_s} to indicate outlier segments according to eq. (5) was set to 8 (a conservative choice). This design parameter was chosen manually such that only strongly pronounced error events of the training data set are marked in this initialization step. The ESMs were trained with 80% of the gathered error sequences. The number of additional states, optimized with the DIC (see eq. (6)), varied between 2 and 3. The functional principle of the classification system is shown in Fig. 7, where the log-likelihood of a recorded sequence under the RSM and one ESM are presented for a sequence containing an error event (annealing material particles). The RSM cannot describe the error event, therefore the log-likelihood values drop, whereas the likelihood values for the ESM decrease less. This difference in the likelihood is used for the classification decision. Outside the error event, the likelihood values for the ESM and RSM coincide and the larger the difference in the log-likelihood, the more distinct the error event is.

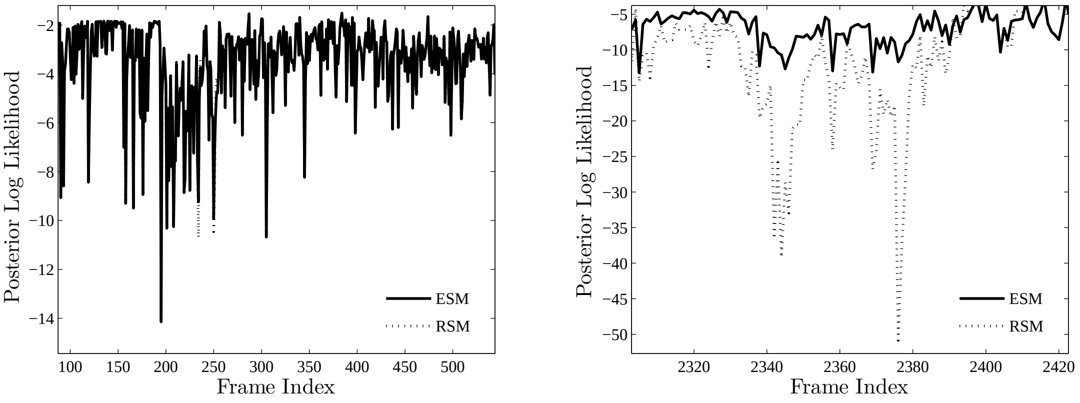
Fig. 8 shows a sequence containing an unusual event from error class ω_{E1} (annealing material particle) along with the computed features (PC and DFFS) and posterior frame log-likelihood ratio:

$$L_F(k) = \log \frac{\frac{1}{|Q_E|}\sum_{j \in Q_E} P(Q_k = j|\Phi_E, O)}{\frac{1}{|Q_R|}\sum_{j \in Q_R} P(Q_k = j|\Phi_R, O)}, \tag{10}$$

where $Q_R/|Q_R|$ and $Q_E/|Q_E|$ are the index/total number of the states from an ESM describing the regular part and erroneous part of a sequence, 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.

E. Classification Results

Fig. 9 presents the receiver operator characteristics (ROCs) for different feature combinations and classification methods for the test data. The ROC curves are computed by varying the parameter Δ_{LL} for the test data set. In industrial applications, it is important to detect all erroneous sequences, therefore the FP rate for a FN rate of 0%, $FP|_{FN=0}$ is an important quality measure for the investigated classification framework. The evaluation procedure and data setss are the same for all used approaches described in the following e



(a)

(b)

Fig. 7. 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.

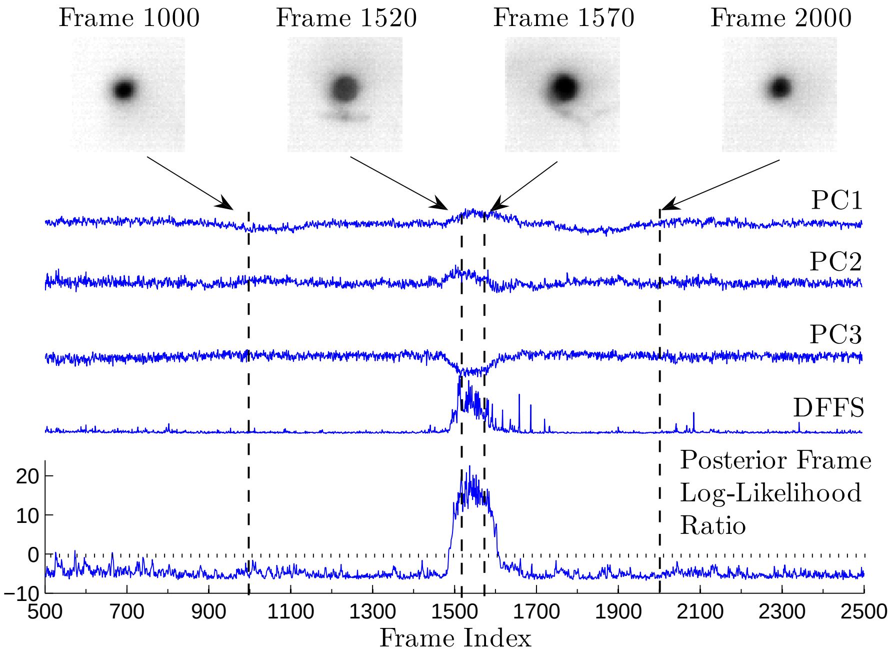
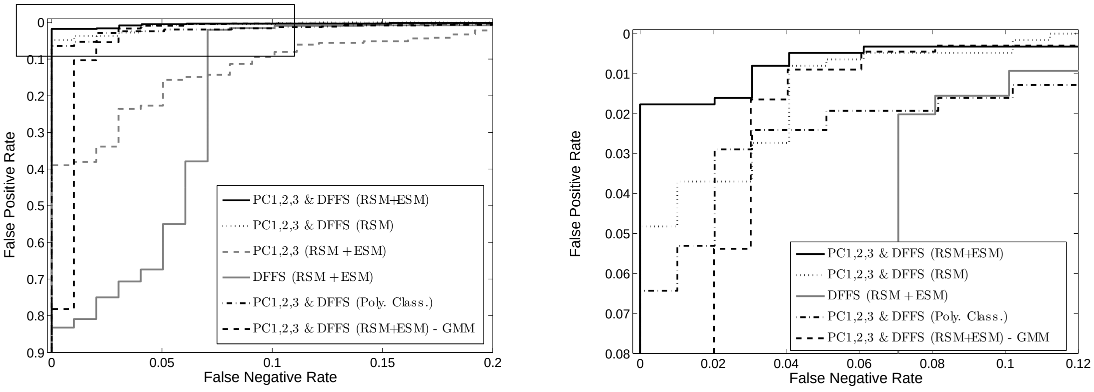


Fig. 8. Part of a sequence containing an error event (annealing material particle) with some recorded images, used features (PC and DFFS) and computed posterior frame log-likelihood ratio. Frames 1000 and 2000 belong to the error free part of the sequence. The images at frames 1520 and 1570 pertain to the error event and show a burning material particle.



(a) (b)

Fig. 9. Comparison of the ROC curves for different classification approaches and feature combinations for the investigated data set. The ROCs are labeled with the used features: principal components (PC) or DFFS (Distance From Feature Space) and whether two-class classification using weakly labeled error sequences (RSM&ESM) or one-class classification using error-free sequences only is employed. Unless otherwise stated, HMMs are used for the classification. For comparison a pure Gaussian Mixture Model (GMM) which cannot model temporal dynamics and a discriminative approach using a Polynomial classifier and time averaging (Poly. Class.) are presented. (b) shows a magnification of the framed part of the ROCs in (a).

1) One-Class vs. Two-Class Classification: The ROCs presented in Fig. 9 show that the classification performance can be improved by using the weak labels. One-class classification using only the RSM (*"PC1,2,3&DFFS (RSM)"*) yields a FP rate $FP|_{FN=0} \approx 4.8\%$ and the area under the ROC (AUC) is 0.998. Two-class classification with the ESMs (*"PC1,2,3&DFFS (RSM+ESM)"*) reduces the FP Rate to $FP|_{FN=0} \approx 1.8\%$ and increases the AUC to 0.999.

A 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 weakly 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 of one and two-class classifications coincide.

2) Feature Selection: The ROCs in Fig. 9 demonstrate that only the combination of features of the principal (PC1,2,3) and residual subspace (DFFS) enable a satisfactory classification performance. The features from the principal subspace detect changes in the overall brightness and translations of the melt pool, whereas the DFFS signal detects deformations which have been not observed in the training data set of regular sequences. The ROC in Fig. 9 for the error detection with the DFFS signal alone (*"DFFS (RSM+ESM)"*)

, clearly demonstrates that this single feature cannot detect all errors. The irregularities in some erroneous sequences are only observable in the principal subspace and therefore the DFFS signal alone cannot recognize them. The same holds for a classification with the features from the principal subspace only s(*"PC1,2,3 (RSM+ESM)"* e.

3) GMM-based Approach: In addition to the HMM approach, the same training method as described in section II for weakly labeled data was used to train a pure Gaussian Mixture Model (GMM) ($N_Q = 1$). In comparison to a HMM, a GMM does not use sequence information to describe the data. The variability of the data is captured by increasing the number of mixtures M, instead of increasing the number of states N_Q and the number of mixtures in a HMM. For the RSM, a GMM with $M = 12$ mixture elements was trained and for each error class between 2 to 3 mixtures were added to obtain the ESMs. The ROC in Fig. 9 shows that the performance(*"PC1,2,3&DFFS (RSM+ESM)* - GMM*"*) is significantly below the performance of the HMM approach for small FN-rates ($FP|_{FN=0} \approx 78.0\%$) and comparable for higher FN rates. Thus it can be concluded that for weakly pronouced error events dynamic information *is* necessary to dinstiguish them from the normal variations of the regular sequences; whereas for strongly pronounced error events, temporal information is not strictly required.

4) Comparison with a Discriminative Classification Approach: The HMM classification system is compared with a two-stage, discriminative classification approach for industrial processes [27]. The classification system in [27] 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 (the most likely sequence of states) 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 s both values have been determined using cross validation. The classification performance is presented in Fig. 9. 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. Note that the HMM

TABLE I

SUMMARY OF FP RATES AT A FN RATE OF 0% (ALL ERRONEOUS SEQUENCES ARE FOUND), AND THE AREA UNDER THE ROCS (AUC) FOR DIFFERENT CLASSIFICATION APPROACHES.

	$FP\|_{FN=0}$	AUC
PC1,2,3 & DFFS (RSM)	4.82%	0.998
PC1,2,3 & DFFS (RSM+ESM)	1.77%	0.999
PC1,2,3 & DFFS (RSM+ESM) - *GMM*	78.0%	0.989
PC1,2,3 & DFFS (Polynomial Classifier)	6.43%	0.994

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. The results for the FP rate $FP|_{FN=0}$ and the AUC for the different classification approaches are summarized in table I.

5) Correlation between Classification Outcome and Error Severity: In Fig. 10 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 the classification system can be seen: the log-likelihood ratio increases with higher error severity. The non-significance of the correlation must be attributed mostly to the sensor system, not 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 corresponding sequence.

IV. DISCUSSION

The observations in section III-E lead to the following practical use of the classification framework for monitoring industrial processes:

- Initialization Stage: 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 Stage: 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

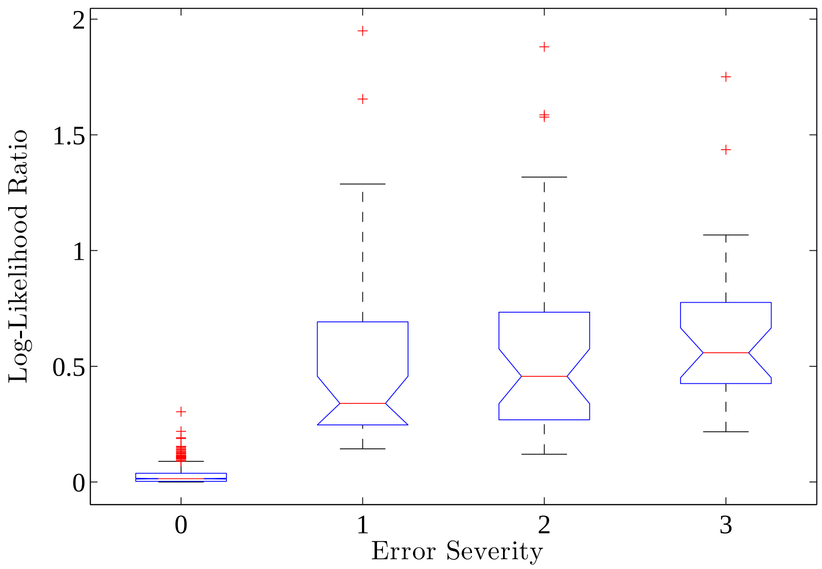


Fig. 10. Box plot of the computed log-likelihood ratio for 1014/35/38/26 parts of error severity 0/1/2/3. By definition the error severity of zero corresponds to error free parts. The observed log-likelihood ratios for each error severity are summarized by the lowest observation, the lower quartile, the median, the upper quartile, and the largest observation (from bottom to top); outliers are marked as additional points in the plot. 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.

procedure, and on the other hand the available information on harmless sequence anomalies that do not jeopardize quality 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 parameter estimates Φ_{new} from the newly collected training data are computed and the model parameters can be updated:

$$\Phi = \xi\Phi_{old} + (1 - \xi)\Phi_{new} \tag{11}$$

where ξ compromises between the new and previous parameter estimates.

This approach combines the benefits of both one- and two-class classification.

V. Conclusions

In an industrial environment, it is imperative that classification systems can be trained with as little user interaction as possible. An automated classification system for the detection of rare events in image sequences has been presented which 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. Starting from a RSM, ESMs are built by using a temporal change detection algorithm to select outlier segments. The usefulness of the classification system has been validated on industrial data from laser welding processes. For the investigated data set, all sequences containing unusual events can be found with a small estimated FP rate of 1.8%.

The use of HMMs allows to take temporal dependencies of the features into account. We have demonstrated that this capability to model the dynamics of a process improves the classification performance compared to a generative approache which does not use temporal information (GMM) and compared to a temporal smoothing of classification scores obtained for individual frames from a discriminative approach (Polynomial Classifier).

REFERENCES

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

[2] ——, “Hidden Markov Models for fault detection in dynamic systems,” *Pattern Recognition*, vol. 27, no. 2, pp. 149–164, January 1994.

[3] 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.

[4] F. LeGland and L. Mevel, “Fault detection in hidden markov models: A local asymptotic approach,” in *Proceedings of the 39th IEEE Conference on Decision and Control*, vol. 5, no. 2000, 2000, pp. 4686 – 4690.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.