Improving multi-camera activity recognition by employing neural network based readjustment

Athanasios S. Voulodimos∗, Nikolaos D. Doulamis∗, Dimitrios I. Kosmopoulos†, Theodora A. Varvarigou∗

(Received December 2010)

In this paper we propose a method to enhance activity recognition in complex environments, where problems like occlusions, noise and illumination changes are present. In order to address the problems induced by the dependency on the camera’s viewpoint, multiple cameras are commonly used in an endeavour to exploit redundancies. We initially examine the effectiveness of various information stream fusion approaches based on Hidden Markov Models, including enhanced Student-t models for tolerance to outliers. Then we introduce a neural network based readjustment mechanism that fits these fusion schemes and aims at dynamically correcting erroneous classification results for image sequences, thus enhancing the overall recognition rates. The proposed approaches are evaluated under complex real life activity recognition scenarios and the acquired results are compared and discussed.

1 Introduction

The field of event recognition and human activity modelling has been the focal point of researchers from various communities. The main reason that justifies this trend lies in the wide variety of applications linked with event detection and behavior recognition: virtual reality, automated video indexing, human-computer interaction, medical monitoring, assistive living, surveillance and smart monitoring are the most prominent examples. In this paper we focus on monitoring visually complex environments, such as the production line of industrial plants. Computer vision and machine learning algorithms attempting to effect activity recognition in complicated environments are confronted with significant challenges, such as serious visibility problems, occlusions, outliers, and, in some cases, low intraclass and high interclass similarity of the observed activity/event classes. Industrial environments, which constitute the use case in this paper, pose additional difficulties ranging from frequent illumination changes, machinery operation and welding flare to camera shaking and target deformations. Figure 1 depicts typical key frames from the complex industrial environment of our use case, highlighting the challenges posed. Typical activity/behaviour understanding

∗School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Zografou, Greece. E-mail addresses: thanosv@mail.ntua.gr; ndoulam@cs.ntua.gr; dora@telecom.ntua.gr.
†Institute of Informatics and Telecommunications, NCSR Demokritos, 15310 Aghia Paraskevi, Greece. E-mail address: dkosmo@iit.demokritos.gr.
methods are mostly unsuccessful because they tend to rely on object detection and motion tracking, which are adversely affected by the aforementioned challenges to a great extent, thus usually resulting in failure.

In this context, certain issues arise, whose scrutinisation can lead to appropriate methods for activity recognition. Bypassing the error-prone detection and tracking algorithms can be attained by relying on suitable holistic features for scene representation. Exploiting the wider scene coverage provided by multiple viewpoints (which are often available in monitoring applications) may conduce to occlusion solving; on the other hand, endowing models with noise and outlier tolerant characteristics can increase robustness. However, all these may not be enough and an expert user’s feedback may be required to readjust the recognition framework’s results in the direction of overall misclassification error minimisation. This feedback is ideally given on a small part of the video footage.

Taking all of the above into consideration, our work contributes to the solution of activity recognition by proposing an approach for further improving the supplied results after holistic scene representation, robust classification and multicamera fusion; this method allows interaction with the user, who may provide relevance feedback in a part of the data. The proposed approach is based on a classifier like a neural network and early, as well as late fusion feedback schemes are investigated.

The remainder of this paper is structured as follows: Related work regarding event and activity recognition as well as relevance feedback is discussed in Section 2. Section 3 focuses on HMM based activity modelling, and describes the fusion frameworks and their applicability for multi-camera activity recognition, as well as the Student-t observation model. In Section 4 we analyse the neural network based rectification mechanism, which readjusts the classification probabilities provided by the HMM, and we introduce a novel "fusion" approach. The experimental validation is detailed in Section 5, while results
are reported and discussed in Section 6. Finally, Section 7 concludes the paper with a summary of the findings.

2 Related work

Event detection as well as human action and activity recognition have been the focus of interest of the computer vision, machine learning and multimedia communities for years. A variety of methods has addressed these problems, including semilatent topic models (Wang and Mori, 2009), spatial-temporal context (Hu et al., 2010), optical flow and kinematic features (Ali and Shah, 2010), and random trees and Hough transform voting (Yao et al., 2010). Wada and Matsuyama (2000) employ a Non-deterministic Finite Automaton as a sequence analyzer to present an approach for multiobject behaviour recognition based on behaviour driven selective attention. Other works focus on more specific domains, e.g. event detection in sports (Sadlier and O’Connor, 2005; Hung and Hsieh, 2008), retrieving actions in movies (Laptev and Perez, 2007), human gesture recognition (using Dynamic Time Warping (Bobick and Wilson, 1997) and Time Delay Neural Networks (Yang and Ahuja, 1998)), and automatic discovery of activities (Hamid et al., 2007). Models might be previously trained and kept fixed (Wang et al., 2008; Antonakaki et al., 2009) or adapt over time (Breitenstein et al., 2009) to cope with changing conditions. A broad variety of extracted image features are used, such as global scene 3D motion (Padoy et al., 2009) or object trajectories (Johnson and Hogg, 1996), which require accurate detection and tracking. Other machine learning and statistical methods that have been used for activity recognition include clustering (Boiman and Irani, 2005) and density estimation (Johnson and Hogg, 1996). A very popular approach is Hidden Markov Models (HMM) (Ivanov and Bobick, 2000; Lv and Nevatia, 2006), due to the fact that they can eciently model stochastic time series at various time scales. Comprehensive literature reviews regarding action and activity recognition can be found in (Aggarwal and Cai, 1999; Hu et al., 2004).

As far as multiple cameras are concerned, the work that investigates fusion of time series resulting from holistic classification in images is rather limited. Some typical approaches seek to solve the problem of position or posture extraction in 3D or on ground coordinates, see, e.g., (Antonakaki et al., 2009). However, camera parameters are required and in most cases there is still dependency on tracking.

The neural network based rectification framework has been inspired by relevance feedback. Relevance feedback is a common approach for automatically adjusting the response of a system regarding information taken from users’ interaction (Doulamis and Doulamis, 2006). Originally, it has been developed in traditional information retrieval systems (Rocchio, 1971; Doulamis et al., 1999), but it has been now
extending to other applications, such as surveillance systems (Oerlemans et al., 2007; Zhang et al., 2010). Relevance feedback is actually an on-line learning strategy which re-weights important parameters of a procedure in order to improve its performance. Re-weighting strategies can be linear or non-linear relying either on heuristic or optimised methodologies. Linear and heuristic approaches usually adjust the degree of importance of several parameters that are involved in the selection process. On the contrary, non-linear methods adjust the applied method itself using function approximation strategies (Doulamis and Doulamis, 2004). In this direction, neural network models have been introduced as non-linear function approximation systems (Doulamis et al., 2000). However, such approaches have been applied for information retrieval systems instead of event recognition or surveillance applications, and it should be noted that it is not straightforward to move from one type of application to another given the different requirements of the two types of applications.

3 Activity modelling via Hidden Markov Models and multicamera fusion

The main attribute of a machine learning model suitable for event and activity understanding from visual sequences is the capability to extract the "signature" of the event/activity from the captured visual input. A popular approach for sequential data modelling, that succeeds in fulfilling the aforementioned requirement is the Hidden Markov Model (HMM) (see e.g. Rabiner (1989)).

3.1 Using HMMs for activity modelling

An HMM entails a Markov chain comprising a number of, say, $N$ states, with each state being coupled with an observation emission distribution. An HMM defines a set of initial probabilities $\{\pi_k\}_{k=1}^N$ for each state, and a matrix $A$ of transition probabilities between the states; each state is associated with a number of (emitted) observations $o$ (input vectors). Gaussian mixture models are typically used for modeling the observation emission densities of the HMM hidden states.

Typically, HMMs are trained under the maximum-likelihood framework, by means of the expectation-maximisation (EM) algorithm (Rabiner, 1989). The HMM model size, i.e. the number of constituent states and mixture components, can affect model performance and efficiency; for this reason, several criteria have been proposed for the purpose of data-driven HMM model selection, e.g. (Li and Biswas, 1999; Ostendorf and Singer, 1997). However, for systems that are expected to operate in nearly real-time, small models are generally preferable, due to their low number of parameters, hence easier learning, and considerably less computational burden for sequential data classification.
Outliers are expected to appear in model training and test datasets obtained from realistic monitoring applications due to illumination changes, unexpected occlusions, unexpected task variations etc, and may seriously corrupt training results. Here we propose the integration of the Student-\( t \) distribution in our streamwise and fusion models (see subsection 3.2) to address the problem. The heavier tails of the Student-\( t \) distribution compared to the Gaussian may provide higher tolerance to outliers. Recently, it has been shown that the adoption of the multivariate Student-\( t \) distribution in the observation models allows for the efficient handling of outliers in the context of the HMM framework without compromising overall efficiency (Chatzis et al., 2009). As shown therein, the estimation of model parameters, including \( \nu \), can be performed automatically using a model training algorithm, e.g. under the EM algorithm framework.

### 3.2 Exploiting redundancies via multicamera fusion

Occlusions are one of the most important challenges that object detection and tracking as well as event and activity recognition algorithms are confronted with. In the cases of complex environments, which are examined in this paper, the vulnerability to occlusions is even more significant, thus highlighting the dependency on the camera viewpoint. Deploying multiple cameras with partly overlapping views and exploiting the redundancies can help solve occlusions and increase robustness. Each camera input provides a different stream of observations. These streams can be combined by means of information fusion techniques, in order to exploit the complementarity of the different views and to attain activity recognition rates better than the ones achieved in the case of independent individual data streams. Here we will examine the most popular HMM fusion approaches, analyse their characteristics and applicability (which will be experimentally verified in subsection 6.1) and propose certain adaptations to increase tolerance to outliers.

In the state-synchronous multistream HMM (Dupont and Luettin, 2000) (Figure 2(a)) the streams are assumed to be synchronised. Each stream is modelled using an individual HMM; the postulated streamwise HMMs share the same state dynamics (identical states, state priors, transition matrices, component priors). Then, the likelihood for one observation is given by the product of the observation likelihood of each stream \( c \) raised to an appropriate positive stream weight \( r_c \) (Dupont and Luettin, 2000):

\[
P(o_t|s_t = i) = \prod_{c=1}^{C} \left( \sum_{k=1}^{K} w_{ikc} P(o_c|\theta_{ikc}) \right)^{r_c}
\]

where \( w_{ikc} \) denotes the weights of the mixtures and \( \theta_{ikc} \) the parameters of the \( k^{th} \) component density of the \( i^{th} \) state of the \( c^{th} \) stream. The weight \( r_c \) is associated with the reliability of the information carried...
by the $\epsilon^{th}$ stream.

Nevertheless, the assumption of synchronised data can be rather confining when attempting event/activity recognition in real world applications. The parallel HMM (Vogler and Metaxas, 1999) (Figure 2(b)) is an alternative that assumes that the streams are independent of each other. A separate HMM for each stream can be therefore trained in the typical way. The parallel HMM can be applied to cameras or other sensors that may not necessarily be synchronised and may operate at different acquisition rates. Similar to the synchronous case, each stream $c$ may have its own weight $r_c$ depending on the reliability of the source. Classification is performed by selecting the class that maximises the weighted sum of the classification probabilities from the streamwise HMMs, i.e. class assignment is conducted by picking the class \( \hat{l} \) with:

\[
\hat{l} = \arg\max_{l} \left( \sum_{c=1}^{C} r_c \log P(o_1...o_T|\lambda_{cl}) \right)
\]

where $\lambda_{cl}$ are the parameters of the postulated streamwise HMM of the $\epsilon^{th}$ stream that corresponds to
the \( l \)th class. As can be inferred by the described architecture a major drawback that plagues the parallel HMM lies in its tendency to neglect any dependencies on the state level between the observation streams.

To this end several architectures attempting to address this issue have been proposed in the literature, such as the coupled HMM (Nefian et al., 2002) and the multistreamed fused HMM (Zeng et al., 2008). Focusing on the latter (Figure 2(c)), the connections between the component stream-wise HMMs of this model are chosen based on a probabilistic fusion model, which is optimal according to the maximum entropy principle and a maximum mutual information criterion for selecting dimension-reduction transforms.

Specifically, if we consider a set of multistream observations \( O = \{ o_t \}_{t=1}^T \) with \( o_t = \{ o_{ct} \}_{c=1}^C \) and \( o^c = \{ o_{ct} \}_{t=1}^T \), the multistream fused HMM models this data based on the fundamental assumption:

\[
P(O) = \frac{1}{C} \sum_{c=1}^{C} P(o^c) \prod_{r \neq c} P(o^r | \hat{s}^c)
\]

where \( \hat{s}^c \) is the estimated hidden sequence of emitting states that corresponds to the \( c \)th stream observations, obtained by means of the Viterbi algorithm, \( P(o^c) \) is the observation probability of the \( c \)th stream-observed sequence, and \( P(o^r | \hat{s}^c) \) is the coupling density of the observations from the \( r \)th stream with respect to the states of the \( c \)th stream model:

\[
P(o^r | \hat{s}^c) = \prod_{t=1}^{T} P(o_{rt} | \hat{s}_{ct})
\]

The probabilities \( P(o_{rt} | \hat{s}_{ct}) \) of the multistream fused HMM can be modelled by means of mixtures of Gaussian densities, similar to the state-conditional likelihoods of the streamwise HMMs. However, in this paper we propose the following adaptation in an endeavour to attain higher tolerance to outliers: The use of Student-\( t \) mixture models instead of Gaussian mixtures can be applied to both the probability models of the streamwise HMM states and the interstream coupling models of the multistream fused HMM to further enhance robustness. Synchronous HMM and parallel HMM will also be adapted by using the Student-\( t \) pdf as predictive function for the streamwise models. We use a modified EM training algorithm and solve numerically to obtain \( \nu \).

Similar to the case of parallel HMMs, the class that maximises the weighted sum of the log-likelihoods over the streamwise models is the winner. Experimental verification of the suitability of the described fusion
Figure 3. Schematic overview of the proposed system. The neural network based rectification mechanism is examined under two different approaches (corresponding to the green and red paths respectively). The first performs stream-wise rectification while the second performs rectification of the fusion result.

schemes for activity recognition, as well as related comparisons and discussion follow in subsection 6.1.

4 A rectification scheme based on a feedforward neural network

In this section we propose a rectification scheme that exploits the expert user’s feedback on the classification (provided by the HMM framework) of part of the footage, so as to readjust, i.e. enhance, future classification results.

Let us denote as $S$ a set that contains the selected samples by the expert user. The set $S = \{(p_i, d_i) \cdots\}$ contains pairs of the form $(p_i, d_i)$, where as $p_i$ we indicate the observation probability vector, generated by the HMM, the elements of which express the probability of the corresponding frame to belong to one of the, say, $M$ available classes. Vector $d_i$ indicates the ideal probabilities for the $i^{th}$ sample. Variable $d_i$ is an indicator vector meaning that all its elements will be zero apart from one which is equal to one. This element indicates the class that this task belongs to. Assuming the existence of a non-linear function able to correct the erroneous classifications of the Hidden Markov Model, we can derive the following equation:

$$d_i = f(p_i)$$ (5)

where $f(\cdot)$ is a vector function indicating the non-linear relationship between $p_i$ and $d_i$.

The main difficulty with the previous equation is the fact that vector function $f(\cdot)$ is actually unknown.
Additionally, the non-linear relationship dynamically changes under different conditions and camera system modification. To address the aforementioned difficulties, we introduce a feedforward neural network model that is able to accurately approximate the unknown vector function \( f(\cdot) \) with a certain degree of accuracy. In this case, equation (5) is written as follows:

\[
d_i = f_w(p_i)
\]  

The main difference between equations (5) and (6) is the introduction of the vector weight \( w \). This means that different parameters (weights) of the network yield different performance of the adaptable classifier. Vector \( w \) includes all the parameters (weights) of the non-linear neural network-based classifier.

To estimate the weights \( w \) we need to apply a training algorithm that actually minimises the mean square error among all data (task sequences) selected from the expert user and the respective output of the network when a particular set of weights is applied. That is,

\[
w = \arg \min_{\text{forall } w} \epsilon = \arg \min_{\text{forall } w} \sum_i (f_w(p_i) - d_i)^2
\]

The backpropagation algorithm (Rumelhart et al., 1988) can provide a solution to this non-linear minimisation problem. In our experiments, we select a small neural network structure of few hidden neurons and one hidden layer. In this case, we try to minimise the number of neural networks parameters, that is the size of weight vector \( w \). It is clear that the samples of the training set \( S \) should be greater than the number of neural network parameters, that is the dimension of the weight vector \( w \). Nevertheless, since the size of the neural network is small few training samples are required. The readjusted probabilities extracted as output of the neural network testing process are used as a basis for enhanced activity recognition by means of selecting the activity yielding the maximum probability in each case. The approach described here is graphically depicted by the green arrow path in Figure 3, which gives a schematic overview of the proposed framework. In the following subsection we introduce a novel approach for integrating the neural network based rectification mechanism into the fusion model.
4.1 Integrating neural network based rectification into the fusion model

In addition to utilising the readjusted likelihoods provided by the neural network as the basis from which to select the winner class for every activity that is to be recognised, we hereby propose an adaptation to the aforementioned parallel HMM fusion scheme, that incorporates the rectified probabilities. The hereby proposed approach corresponds to the red arrow path in Figure 3.

We assume that the probabilities extracted by the individual streamwise HMM frameworks are fed into the rectification mechanism. As a consequence, readjusted probabilities corresponding to the two streamwise models are generated. Let $P_{NN}(o_1...o_T|\lambda_{cl}, n_c)$ be the readjusted probability generated as output from the neural network, where $\lambda_{cl}$ are the parameters of the postulated streamwise HMM of the $c^{th}$ stream that corresponds to the $l^{th}$ class and $n_c$ are the parameters of the neural network that corresponds to the $c^{th}$ stream. In this proposed rectification driven fused HMM (RDFHMM) fusion model class assignment is conducted by picking the class $\hat{l}$ with:

$$\hat{l} = \arg\max_l \left( \sum_{c=1}^{C} r_{cl} \log P_{NN}(o_1...o_T|\lambda_{cl}, n_c) \right)$$ (8)

where $r_{cl}$ is the stream weight factor for the $c^{th}$ stream and the $l^{th}$ class; the stream weight can therefore vary according to the reliability of a stream not only in general terms but also in a class-specific manner, since different camera positions may offer better or worse viewpoints for particular activity classes. The contribution of the proposed non-linear probability readjustment scheme in the improvement of the recognition results is experimentally validated and discussed in subsection 6.2.

5 Experimental validation

We experimentally validated the proposed methods with video sequences obtained from a real assembly line of a major automobile manufacturer. The workflow on this assembly line included tasks of picking several parts from racks and placing them on a designated cell some meters away where welding was performed. Each of the above activities/tasks was regarded as a class of behavioral patterns that had to be recognised. The information acquired from this procedure could be used for the extraction of production statistics or anomaly detection. To tackle the issue of heavy occlusions we used two synchronised cameras with partially overlapping views.

We evaluated the overall efficiency of the proposed system, as well as the framework’s different alternative
Improving multi-camera activity recognition by employing neural network based readjustment

constituent components. We compared recognition rates in the cases of individual streams and fusion methods. We also scrutinised the effectiveness of the use of the outlier-tolerant Student-t distribution as predictive function of the employed HMMs instead of the conventional Gaussian. Finally, we examined the contribution of the neural network based rectification scheme in the improvement of activity recognition results.

5.1 Experimental setup

The workspace configuration and the cameras’ positioning are depicted in Figure 4. According to the manufacturing requirements each workflow consists of the following 7 activities/tasks, which are not necessarily executed sequentially:

Task 1: A part from Rack 1 (upper) is placed on the welding spot by worker(s).
Task 2: A part from Rack 2 is placed on the welding spot by worker(s).
Task 3: A part from Rack 3 is placed on the welding spot by worker(s).
Task 4: Two parts from Rack 4 are placed on the welding spot by worker(s).
Task 5: A part from Rack 1 (lower) is placed on the welding spot by worker(s).
Task 6: A part from Rack 5 is placed on the welding spot by worker(s).
Task 7: Worker(s) grab(s) the welding tools and weld the parts together.

Two datasets\(^1\) were used for the experiments. Each dataset contains 20 segmented sequences representing

---

\(^1\)We are going to make the datasets publicly available. They are currently available for review purposes on http://www.4shared.com/dir/r1zb3XeD/AppliedArtificialIntelligence.html (password:xyz543)
full assembly cycles/workflows. In each workflow all seven activities are performed, but not necessarily in the same order. The total number of frames was approximately 80,000 per camera for each dataset. Challenges of the two datasets include occlusions, visually complex background, similar colours, high intra-class and low inter-class variance. In dataset-1, the assembly process was rather well structured and was performed strictly by two people. Noisy objects were present (other persons or vehicles) but not particularly often. In dataset-2, which was shot several months later, the assembly process was modified, in that a third person was present quite often in the scene, performing tasks in parallel to the tasks executed by the other two workers. Dataset-2 is therefore far more challenging because the silhouettes got overlayed in a random fashion, thus making the motion signatures much more difficult to model. Moreover, variable task durations and overlapping phenomena were far more exacerbated in comparison to dataset-1. The annotation of the data sets was done manually. Synchronisation of the employed IP-cameras was performed by exploiting the server-generated timestamps. Although this provided a good estimate of timing, perfect synchronisation was not ensured, since the cameras were not synchronised at hardware level.

5.2 Holistic scene representation

As is already mentioned, using holistic image based features we obviate the need for successful detection and tracking, which are particularly difficult in complex environments. The features we used are calculated as follows: Firstly we perform background subtraction. We use the foreground regions to represent the multi-scale spatiotemporal changes at pixel level. For this purpose we use a concept proposed by Xiang and Gong (2006), which is similar to Motion History Images, but has better representation capabilities as shown therein. The Pixel Change History (PCH) of a pixel is defined as:

\[
P_{\varsigma,\tau}(x, y, t) = \begin{cases} 
\min(P_{\varsigma,\tau}(x, y, t - 1) + \frac{255}{\varsigma}, 255) & \text{if } D(x, y, t) = 1 \\
\max(P_{\varsigma,\tau}(x, y, t - 1) - \frac{255}{\tau}, 0) & \text{otherwise}
\end{cases}
\]

where \(P_{\varsigma,\tau}(x, y, t)\) is the PCH for a pixel at \((x, y)\), \(D(x, y, t)\) is the binary image indicating the foreground region, \(\varsigma\) is an accumulation factor and \(\tau\) is a decay factor. By setting appropriate values to \(\varsigma\) and \(\tau\) we are able to capture pixel-level changes over time.

To represent the resulting PCH images we propose use of Zernike moments. Zernike moments are very attractive because of their noise resiliency, their reduced information redundancy and their reconstruction
capability. For more information on Zernike moments see for example (Mukundan and Ramakrishnan, 1998). In this case, we have used as feature vector the Zernike moments up to sixth order (excluding four angles that were always constant), along with the center of gravity and the area, thus having a good scene reconstruction without too high dimension (31). For capturing the spatiotemporal variations we have set the parameters at $\varsigma = 10$ and $\tau = 70$. Zernike moments have been calculated in rectangular regions of interest of approximately 15,000 pixels in each image, to limit the processing and allow real time feature extraction. The processing was performed at a framerate of approximately 50-60 fps.

5.3 **Fused HMM learning**

The models were trained using the EM algorithm. We used the typical HMM model for the individual streams as well as state-synchronous, parallel and multistream fused HMMs. We have experimented with the Gaussian and the Student-$t$ distribution. All experimental variations were performed on both dataset-1 and dataset-2, thus making a total of 20 different experimental setups. We used three-state HMMs with a single mixture component per state to model each of the seven tasks described above, which is a good trade-off between performance and efficiency. For the mixture model representing the interstream interactions in the context of the multistream fused HMM we used mixture models of two component distributions. Full covariance matrices were employed for the observation models. The stream weights $r_c$ in the fusion models were selected according to the reliability of the individual streams, that is in proportion to the classification accuracy of the respective single stream HMM. For more dependable results, we used cross-validation, by repeating the employed training algorithms several times, where in each repetition all scenarios were considered except for one used for testing (leave-one-out cross-validation).

5.4 **Neural network based rectification**

In this phase an expert user selects a set of training samples. These samples are represented using the respective probability vector, as extracted by the HMM framework, and the targeted correct classification of this task. Following, a feedforward neural network model is trained so as to adjust the probabilities extracted by the HMM framework in order to minimise the erroneous classifications. The structure of the feedforward neural network is selected to be small. In particular, we select a feedforward neural network with one hidden layer and 5 neurons in this layer. It has 7 input nodes and 7 output nodes (as many as the number of activities), thus making a total of 70 weights to be learnt (see Figure 5). The transfer function is the sigmoid.
In these particular experiments of the second phase, we have 140 samples in total (20 workcycles containing 7 tasks each) in each experimental setup. Among all these samples, we assume that half of them (70) are used for training in the second phase while the remaining half (70) for testing.

6 Results

We evaluated the overall efficiency of the proposed system, as well as the framework’s different alternative constituent components. For a quantitative evaluation, we use recall-precision metrics. Recall corresponds to the correct classification rate (number of true positives divided by the total number of positives in the ground truth), whereas precision relates to the trust in a classification (number of true positives divided by number of true and false positives). The F-measure is the harmonic mean of these two measurements. The measurements presented are averaged across all test sequences per experimental setup.

6.1 HMM based recognition

Tables 1 and 2 show the obtained results from the HMM based approaches for dataset-1 and dataset-2 respectively.

Dataset-1 vs dataset-2. As a first observation, the employed holistic features and HMM based frameworks represent rather well the assembly process. The classification rates attained in dataset-1 are very high, considering the complexity of the environment. The representation capability of PCH based features proves very satisfactory for dataset-1. As expected, success rates in dataset-2 are lower, which can be explained by
Table 1. Results obtained from dataset-1 using i) individual HMMs to model information from Stream 1 (HMM1); ii) individual HMMs to model information from Stream 2 (HMM2); iii) state-synchronous HMMs (SYNC); iv) parallel HMMs (PARAL); and v) multistream fused HMMs (MULTI) with a) Gaussian and b) Student-t distribution as observation likelihood.

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM1</td>
<td>Gauss</td>
<td>82.9%</td>
<td>78.3%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>85.0%</td>
<td>80.2%</td>
</tr>
<tr>
<td>HMM2</td>
<td>Gauss</td>
<td>86.3%</td>
<td>82.9%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>88.6%</td>
<td>84.1%</td>
</tr>
<tr>
<td>SYNC</td>
<td>Gauss</td>
<td>71.4%</td>
<td>62.9%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>72.9%</td>
<td>64.4%</td>
</tr>
<tr>
<td>PARAL</td>
<td>Gauss</td>
<td>83.6%</td>
<td>79.0%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>84.3%</td>
<td>80.1%</td>
</tr>
<tr>
<td>MULTI</td>
<td>Gauss</td>
<td>89.3%</td>
<td>86.8%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>92.1%</td>
<td>89.8%</td>
</tr>
</tbody>
</table>

Table 2. Results obtained from dataset-2 using i) individual HMMs to model information from Stream 1 (HMM1); ii) individual HMMs to model information from Stream 2 (HMM2); iii) state-synchronous HMMs (SYNC); iv) parallel HMMs (PARAL); and v) multistream fused HMMs (MULTI) with a) Gaussian and b) Student-t distribution as observation likelihood.

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM1</td>
<td>Gauss</td>
<td>47.9%</td>
<td>42.8%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>60.0%</td>
<td>53.8%</td>
</tr>
<tr>
<td>HMM2</td>
<td>Gauss</td>
<td>51.4%</td>
<td>46.9%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>54.3%</td>
<td>47.5%</td>
</tr>
<tr>
<td>SYNC</td>
<td>Gauss</td>
<td>52.9%</td>
<td>42.5%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>54.3%</td>
<td>43.9%</td>
</tr>
<tr>
<td>PARAL</td>
<td>Gauss</td>
<td>52.9%</td>
<td>43.9%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>54.3%</td>
<td>46.3%</td>
</tr>
<tr>
<td>MULTI</td>
<td>Gauss</td>
<td>57.1%</td>
<td>49.9%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>64.4%</td>
<td>58.7%</td>
</tr>
</tbody>
</table>

the far more relaxed structure in the activities performed, the randomly overlayed silhouettes and all the special above described challenges. However, these results are still rather satisfactory for such a difficult dataset, and constitute a good base for the rectification mechanism to follow.

**Single stream vs fusion approaches.** The results indicate that the individual HMM corresponding to camera 2 (HMM2) tends to yield better recognition rates than the one corresponding to camera 1 (HMM1), which can be explained by the generally better viewpoint of the former. The confusion matrices in Figure 6 display the impact of the complementarity of the views on the results as well as the successful exploitation of this fact in the case of multistream fused HMM. For example, camera 2 offers a more favourable viewpoint for discerning task 1 from task 5, whereas camera 1 provides a better angle for recognising task 4.

A careful evaluation of the results leads to the conclusion that information fusion provides significant added value when implemented in the form of multistream fused HMM. In all experimental setups the multistream fused approach outperforms the better of two individual streamwise models in terms of recall.
and precision by at least 3% (and up to 6.4%). This improvement can be put down to the fact that the multistream fused model succeeded in capturing the state interdependencies, without assuming strict synchronicity. The parallel HMM approach provided slightly inferior or slightly superior success rates (depending on the experimental setup) in comparison to the best individual streamwise model. As was mentioned above, this approach considers the streams to be totally asynchronous and is thus unable to make use of state interdependencies. On the other hand, recall and precision rates deteriorate when assuming perfect synchronicity by employing the state-synchronous approach, reflecting the fact that our cameras were indeed not perfectly synchronised (i.e. they were not synchronised at hardware level).

**Gaussian vs Student-t.** Using Student-t distribution instead of the conventional Gaussian as predictive function of the HMMs additionally increases recognition rates to a certain extent (ranging from approximately 2% up to 10%). The contribution is more apparent in the experiments of dataset-2 (Table 2), where the amount of noise is greater, thus proving the usefulness of Student-t distribution in enhancing the robustness to outliers in event and activity recognition from video streams.

### 6.2 Neural network based rectification results

Tables 3 and 4 contain the results acquired after employing the rectification mechanism described in subsection 5.4 to readjust the classification probabilities provided by the HMM based approaches. Comparing the measures in Tables 3 and 4 with the respective results of Tables 1 and 2, we notice that the proposed rectification scheme provides a substantial improvement in the recognition rates. Recall, precision, and F-measure are all significantly increased comparing to the respective experimental setups when no neural network based readjustment was performed. As expected, multistream fused HMM supplemented with
Table 3. Results obtained from dataset-1 after applying the rectification mechanism (RM) using i) individual HMMs to model information from Stream 1 (HMM1); ii) individual HMMs to model information from Stream 2 (HMM2); iii) state-synchronous HMMs (SYNC); iv) parallel HMMs (PARAL); v) multistream fused HMMs (MULTI); and vi) rectification driven fused HMM (RDFHMM) with a) Gaussian and b) Student-t distribution as observation likelihood.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM1+RM</td>
<td>Gauss</td>
<td>89.3%</td>
<td>83.7%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>90.0%</td>
<td>86.5%</td>
</tr>
<tr>
<td>HMM2+RM</td>
<td>Gauss</td>
<td>90.7%</td>
<td>86.2%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>93.6%</td>
<td>90.9%</td>
</tr>
<tr>
<td>SYNC+RM</td>
<td>Gauss</td>
<td>81.4%</td>
<td>77.4%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>80.0%</td>
<td>73.8%</td>
</tr>
<tr>
<td>PARAL+RM</td>
<td>Gauss</td>
<td>90.7%</td>
<td>87.2%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>90.0%</td>
<td>87.7%</td>
</tr>
<tr>
<td>MULTI+RM</td>
<td>Gauss</td>
<td>93.6%</td>
<td>91.0%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>94.3%</td>
<td>91.8%</td>
</tr>
<tr>
<td>RDFHMM</td>
<td>Gauss</td>
<td>93.6%</td>
<td>91.0%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>95.0%</td>
<td>93.2%</td>
</tr>
</tbody>
</table>

Table 4. Results obtained from dataset-2 after applying the rectification mechanism (RM) using i) individual HMMs to model information from Stream 1 (HMM1); ii) individual HMMs to model information from Stream 2 (HMM2); iii) state-synchronous HMMs (SYNC); iv) parallel HMMs (PARAL); v) multistream fused HMMs (MULTI); and vi) rectification driven fused HMM (RDFHMM) with a) Gaussian and b) Student-t distribution as observation likelihood.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM1+RM</td>
<td>Gauss</td>
<td>70.0%</td>
<td>62.4%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>75.7%</td>
<td>66.7%</td>
</tr>
<tr>
<td>HMM2+RM</td>
<td>Gauss</td>
<td>67.9%</td>
<td>59.3%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>70.0%</td>
<td>61.3%</td>
</tr>
<tr>
<td>SYNC+RM</td>
<td>Gauss</td>
<td>67.1%</td>
<td>62.2%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>68.6%</td>
<td>61.4%</td>
</tr>
<tr>
<td>PARAL+RM</td>
<td>Gauss</td>
<td>72.9%</td>
<td>67.9%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>72.1%</td>
<td>64.5%</td>
</tr>
<tr>
<td>MULTI+RM</td>
<td>Gauss</td>
<td>70.7%</td>
<td>66.5%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>78.6%</td>
<td>70.0%</td>
</tr>
<tr>
<td>RDFHMM</td>
<td>Gauss</td>
<td>75.7%</td>
<td>67.4%</td>
</tr>
<tr>
<td></td>
<td>Student-t</td>
<td>79.3%</td>
<td>72.3%</td>
</tr>
</tbody>
</table>

The rectification mechanism provides the best results among the approaches that rectify the fused results, since it is also the best performing approach when stand-alone. However, we observe that our proposed rectification driven fused HMM (RDFHMM), which readjusts the stream-wise probabilities before feeding them into the adapted fusion model, yields the best results, slightly outperforming MFHMM+RM (by up to 5% in recall). Figures 7(a) and 7(a) display the % classification error for all experimental setups with and without the rectification mechanism for dataset-1 and dataset-2 respectively. The improvement ratio (in terms of % error decrease) in relation to the sole use of the HMM based approaches is depicted in Figures 8(a) and 8(b). It is therefore clear that the described neural network based approach significantly enhances the performance of the proposed activity recognition framework, especially when implemented...
7 Conclusion

In this work we have presented a framework for activity recognition in complex environments, such as the production line of an industrial plant, which although visually complicated remains a structured process. As we showed, the extraction of holistic features to bypass tracking, the employment of Student-\(t\) distribution and multicamera fusion may improve results. However, all these together may be further improved by a rectification mechanism. Inspired by relevance feedback, this mechanism is based on a non-linear
classification scheme that aims to re-adjust the probabilities of the stochastic models (like the Hidden Markov Model and its fused versions) according to a set of relevant / irrelevant data selected by an expert user through an interactive framework. The non-linear rectification is accomplished in this paper using a feedforward neural network model that takes as input the classification probabilities of the stochastic models and generates as output the adjusted probabilities. We differentiate between two different approaches. In the first, the rectification mechanism readjusts the probabilities stemming from the fused stochastic model and produces the final activity recognition decision, whereas in the second, the rectification mechanism readjusts the stream-wise probabilities and feeds its output to the proposed Rectification Driven Fused HMM (RDFHMM), which performs a type of fusion of the readjusted probabilities and extracts the recognised activity.

We have tested the proposed methodology in very challenging datasets from real-life production line of an automobile industry. The results illustrate significant improvement when applying the rectification mechanism, i.e. up to 15% in terms of outperformance in recall and on average 35% in terms of error decrease. The proposed RDFHMM yields the best recognition rates. As future research, we plan to exploit adaptable neural network models in order to recursively re-adjust the classification probabilities during the activity execution and to investigate dynamic methods for readjusting the learning process of the involved stochastic models.

References


REFERENCES


