Applied Artificial Intelligence
Vol. 00, No. 00, Month-Month 200x, 1–18

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, outliers and illumination changes occur. In order to address the problems induced by the dependency on the camera's viewpoint, multiple cameras are used in an endeavour to exploit redundancies. We initially examine the effectiveness of various information stream fusion approaches based on hidden Markov models, including Student's t-endowed models for tolerance to outliers. Following, 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 improving 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 behaviour recognition. In this paper, we focus on monitoring visually complex environments, such as the production line of an automobile manufacturer. Computer vision and machine learning algorithms attempting to effect activity recognition in complicated environments are confronted with visibility problems, occlusions, outliers, and, in some cases, low intraclass and high interclass similarity of the observed activity classes. Industrial environments pose additional difficulties ranging from background clutter, frequent illumination changes, 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 object based methods cannot cope with the aforementioned challenges. Testing a tracker and a popular person detector both led to failure in our industrial dataset. In particular, the tracker was based on standard particle filtering and the employed features were the color histogram and the edges of the blobs corresponding to the human figure (for details see our previous work (Makris et al., 2011)); the experiments showed that the tracker was losing the target very often. We

*School of Electrical and Computer Engineering, National Technical University of Athens, Greece. E-mail: thanosv@mail.ntua.gr
†Department of Computer Science and Engineering, University of Texas at Arlington, USA.

also tested the HOG person detector (Dalal and Triggs, 2005), which achieved a maximum accuracy of 56.42% in some of the least challenging sequences of our dataset. However, despite the visually complex environment, the observed activities in the production line remain structured to a certain extent, thus making it reasonable to expect that they can be modelled using machine learning methods.

In this context, the need arises to bypass the error-prone detection and tracking algorithms (Doulamis, 2010) by relying on appropriate holistic features for scene representation. Moreover, 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 time series classifiers with outlier tolerant characteristics can increase robustness. Finally, exploiting an expert user's feedback on a small part of the video sequences through a relevance feedback inspired approach can minimise classification error.

Considering the above, 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 based on outlier tolerant hidden Markov models (HMMs) and multicamera fusion; this method allows interaction with the user, who may provide relevance feedback in part of the data. The proposed approach is based on 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 activity recognition as well as relevance feedback is discussed in Section 2. Section 3 focuses on robust multi-camera HMM based activity modelling. 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.

2 Related work

Event detection as well as human action and activity recognition have been the focus of interest of the computer vision community 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 (Hung and Hsieh, 2008), retrieving actions in movies (Laptev and Perez, 2007), 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 image feature extraction methods are used, such as global scene 3D motion (Padoy et al., 2009), object trajectories (Johnson and Hogg, 1996) or other object based approaches (Fusier et al., 2007) 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 (HMMs) (e.g. (Ivanov and Bobick, 2000; Padoy et al., 2009)), due to the fact that they can efficiently model stochastic time series at various time scales. An alternative approach to the HMM for the analysis of complex dynamical systems is the Echo State Networks (ESNs) (see, e.g., (Jaeger et al., 2007)). ESNs have been recently used for industrial activity recognition in workflows using part of the same dataset that we are using (Veres et al., 2010). A limitation of ESNs is that all significant variations of activity order in a given workflow have to be learnt to provide good classification results. As will be shown in the experimental section through comparisons, our approach outperforms ESN based methods. Other approaches for industrial activity recognition have also been proposed, involving sensors and wearable computing, e.g. (Stiefmeier et al., 2008). A recent comprehensive literature review regarding action and activity recognition can be found in (Poppe, 2010).

As far as multiple cameras are concerned, the work that investigates fusion of time series resulting from holistic image representation is 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; Lao, 2009). However, camera calibration or homography estimation is required and in most cases there is still dependency on tracking or on extraction of foreground objects and their position, which can be easily corrupted by illumination changes and occlusions. Later in the paper, several fusion schemes using HMMs are discussed and their applicability to our scenario is scrutinised.

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 user's interaction (Doulamis and Doulamis, 2006). Originally, it has been developed in traditional information retrieval systems (Rocchio, 1971), but it has been now extended to other applications, such as surveillance systems (Oerlemans et al., 2007; Zhang et al., 2010). Relevance feedback is actually an online learning strategy which reweights important parameters of a procedure in order to improve its performance. Reweighting 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, 2005). In this direction, neural network models have been introduced as non-linear function approximation systems (Doulamis et al., 2000). A comprehensive review regarding algorithms of relevance feedback in image retrieval has been provided in (Zhou and Huang, 2003). In this paper, the authors lay emphasis on comparing different techniques of relevance feedback with respect to the type of training data, the adopted organisation strategies, the similarity metrics used, the implemented learning strategies, and the effect of negative samples in the training performance. However, such approaches have been applied mostly in information retrieval systems instead of event recognition or surveillance applications. In information retrieval systems, a query object (image) is compared against a set of stored objects (images), and the time dimension is not present, while activity recognition is accomplished by taking into consideration the "time variation" of the features of several image frames.

3 Activity modelling via Hidden Markov Models and multicamera fusion

On the basis of the activity recognition framework lies the extraction of holistic visual features at the image level, which are further used to associate events and activities with temporal patterns. The extracted information is modelled by employing HMMs, which constitute a popular methodology for sequential data modelling (Rabiner, 1989), while also offering the possibility to exploit redundancies stemming from multiple streams through the utilisation of HMM-based information fusion schemes.

3.1 *Visual observations*

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, using the Pixel Change History (PCH), which is defined as (Xiang and Gong, 2006): $P_{\varsigma,\tau}(x,y,t) = \begin{cases} \min(P_{\varsigma,\tau}(x,y,t-1) + \frac{255}{\varsigma}, 255) & if D(x,y,t) = 1 \\ \max(P_{\varsigma,\tau}(x,y,t-1) - \frac{255}{\tau}, 0) & otherwise \end{cases}$, 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, ς is an accumulation factor and τ is a decay factor. By setting appropriate values to ς and τ we are able to capture pixel-level changes over time.

To represent the resulting PCH images we propose use of Zernike moments. The complex Zernike moments of order p (see, e.g., (Mukundan and Ramakrishnan, 1998)) are defined as: $A_{pq} =$

$\frac{p+1}{\pi}\int_0^1\int_{-\pi}^{\pi} R_{pq}(r)e^{-jq\theta}f(r,\theta)rdrd\theta$ and $R_{pq}(r) = \sum_{s=0}^{\frac{p-q}{2}}(-1)^s\frac{(p-s)!}{s!(\frac{p+q}{2}-s)!(\frac{p-q}{2}-s)!}r^{p-2s}$, where $r = \sqrt{x^2+y^2}$ and $\theta = \tan^{-1}(\mathrm{y/x})$, $-1 < \mathrm{x,y} < 1$, and $p-q = even$, $0 \le q \le p$.

3.2 *Using HMMs for activity modelling*

An HMM entails a Markov chain comprising a number of 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 states; each state is associated with a number of observations $\mathbf{o}$ (input vectors). Gaussian mixture models are typically used for modelling the observation emission densities of the hidden states. Typically, HMMs are trained under the maximum-likelihood framework, by means of the 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. (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. For this we propose the integration of the Student's t-distribution in our models. The probability density function (pdf) of Student's t-distribution with mean vector μ, positive definite inner product matrix Σ, and ν degrees of freedom is given by: $t\left(x_t;\mu,\Sigma,\nu\right) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)|\Sigma|^{-\frac{1}{2}}(\pi\nu)^{-\frac{p}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\{1+d(x_t,\mu;\Sigma)/\nu\}^{\frac{\nu+p}{2}}}$ where $\Gamma(.)$ denotes the gamma function and d the Mahalanobis distance.

Modifying ν enables including outliers in the pdf without corrupting the model. This additional degree of freedom can model heavier tails, which is not possible for the Gaussian, which is a special case of Student's t for $\nu \to \infty$. A detailed presentation on how to learn ν as well as experimental argumentation for the robustness of Student's t-distribution based HMM can be found in (Chatzis et al., 2009).

3.3 *Exploiting redundancies via multicamera fusion*

In the cases of complex environments, which are examined in this paper, the vulnerability to occlusions is 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, to exploit the complementarity of the different views. 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 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): $P(\boldsymbol{o}_t|s_t = i) = \prod_{c=1}^{C} [\sum_{k=1}^{K} w_{ikc} P(\boldsymbol{o}_{ct}|\theta_{ikc})]^{r_c}$, where w_{ikc} denotes the weights of the mixtures and θ_{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 c^{th} stream.

Nevertheless, the assumption of synchronised data can be rather confining when attempting 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 $\hat{l} = \underset{l}{\operatorname{argmax}}([\sum_{c=1}^{C} r_c log P(\boldsymbol{o}_1 ... \boldsymbol{o}_T|\lambda_{cl})])$, where λ_{cl} are the parameters of the postulated streamwise HMM of the c^{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; Brand et al., 1997) and the *multistreamed fused HMM* (Zeng et al., 2008). Brand et al. (1997) couple the current state of one stream with the previous of the other (assuming two streams), while Zeng et al. (2008) couple the current state of one stream to the current of the other, which is a stronger and more intuitive condition and unlike (Brand et al., 1997) does not necessitate approximations which inevitably sacrifice some crucial information. Focusing on multistream fused HMM (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 stateconditional 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's 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's t pdf for the streamwise models.

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 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 in part of the footage, so as to enhance future classification results.

Let us denote as S a set that contains the selected samples by the expert user. The set $S = \{\cdots(\mathbf{p}_i, \mathbf{d}_i)\cdots\}$ contains pairs of the form $(\mathbf{p}_i, \mathbf{d}_i)$, where as $\mathbf{p}_i$ 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 $\mathbf{d}_i$ indicates the ideal probabilities for the i^{th} sample. Variable $\mathbf{d}_i$ 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 HMM, we can derive: $\mathbf{d}_i = \underline{f}(\mathbf{p}_i)$ where $\underline{f}(\cdot)$ is an unknown vector function indicating the non-linear relationship between $\mathbf{p}_i$ and $\mathbf{d}_i$. 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 $\underline{f}(\cdot)$ with a certain degree of accuracy. In this case, the previous equation is now written as: $\mathbf{d}_i = \underline{f}_{\mathbf{w}}(\mathbf{p}_i)$. The difference between the two equations is

the introduction of the vector weight $\mathbf{w}$. This means that different parameters (weights) of the network yield different performance of the adaptable classifier. Vector $\mathbf{w}$ includes all the parameters (weights) of the non-linear neural network-based classifier.

To estimate the weights $\mathbf{w}$ 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, $\mathbf{w} = \arg\min_{forall\mathbf{w}} \epsilon = \arg\min_{forall\mathbf{w}} \sum_i (\underline{f}_{\mathbf{w}}(\mathbf{p}_i) - \mathbf{d}_i)^2$.

The backpropagation algorithm 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, i.e., the size of weight vector $\mathbf{w}$. 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 $\mathbf{w}$. 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. Here, the neural network rectifies the "combined" probabilities extracted from the fused HMM. 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, we hereby propose an adaptation to the aforementioned parallel HMM fusion scheme, that incorporates the rectified probabilities. This approach corresponds to the red arrow path in Figure 3, where the neural network rectifies the streamwise probabilities, which are subsequently fused.

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 $P_{NN}(\boldsymbol{o}_1 ... \boldsymbol{o}_T | \lambda_{cl}, n_c)])$ be the readjusted probability generated as output from the neural network, where λ_{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:

$$\hat{l} = \underset{l}{\text{argmax}}([\sum_{c=1}^{C} r_{cl} log P_{NN}(\boldsymbol{o}_1 ... \boldsymbol{o}_T | \lambda_{cl}, n_c)]) \quad (1)$$

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. It should be noted here that it would be possible to include the weight factor r_{cl} in the neural network rectification, i.e., have a "unified" rectification scheme where a neural network would take as input the probabilities of all streamwise HMMs and produce an overall probability vector as output. However, this would raise the complexity of the network, thus requiring a greater number of training samples. We opt for the separate streamwise rectification schemes in the context of RDFHMM, because they involve easier training, fewer training samples required, and lower generalisation error. 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 an automobile manufacturer. The workflow on this line included picking several parts from racks and placing them on a designated welding cell. Each of the above activities/tasks was regarded as a class of behavioural patterns that had to be recognised. Two cameras with partially overlapping views were used. We evaluated the overall efficiency of the proposed system, as well as the framework's different alternative constituent components.

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 seven 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] (Voulodimos et al., 2011) were used for the experiments. Each dataset contains 20 segmented sequences representing 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 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 workers' body silhouettes got overlayed in a random fashion, thus making the motion signatures, i.e., the trajectories of their movement, 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 datasets was done manually. Synchronisation of the employed IP-cameras was approximate by exploiting the server-generated timestamps.

Holistic scene representation. We have used 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). This choice provided a good trade-off between representation quality and real-time performance requirements (higher order moments would require much more computational resources). Limiting the order of moments used was also justified by the fact that the details captured by higher order moments have much higher variability and are more sensitive to noise. For capturing the spatiotemporal variations we have set the parameters at $\varsigma = 10$ and $\tau = 70$, which were defined by the duration of motion that we wanted to capture, and are application specific. 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 approximately 50-60 fps.

[1]The datasets are publicly available on http://www.scovis.eu.

Fused HMM based classification. 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's 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, as well as the weights r_{cl} in the case of RDFHMM, were selected according to the reliability of the individual streams, that is in proportion to the classification accuracy attained by the respective single stream HMM. For each dataset, ten workcycles were used for training of the HMMs and the other ten were used for testing.

Neural network based rectification. In this phase an expert user selected a set of training samples. These samples were 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 was trained so as to adjust the probabilities extracted by the HMM framework to minimise the erroneous classifications. The structure of the feedfoward neural network was selected to be small. In particular, we selected a feedforward neural network with one hidden layer and 15 neurons in this layer. It had 7 input nodes and 7 output nodes (as many as the number of activities). The transfer function was the sigmoid. In these experiments of the second phase, the samples belonging to three workcycles were selected to form the training set, and the remaining were used 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 used recall-precision metrics. Recall corresponds to the number of true positives divided by the total number of positives in the ground truth, whereas precision equals the 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 were averaged across all test sequences per experimental setup.

6.1 *HMM based recognition*

Table 1 shows the obtained results from the HMM based approaches for dataset-1 and dataset-2.

Dataset-1 vs dataset-2. As a first observation, the employed holistic features and HMM based frameworks represented rather well the assembly process. The classification rates attained in dataset-1 were very high, considering the complexity of the environment. The representation capability of PCH based features proved very satisfactory for dataset-1. As expected, success rates in dataset-2 were lower, which can be explained by the far more relaxed structure in the activities performed, the randomly overlayed silhouettes and all the special challenges described above. However, these results were still rather satisfactory for such a difficult dataset, and constituted a good base for the rectification mechanism to follow.

Single stream vs fusion approaches. The results indicated that the individual HMM corresponding to camera 2 (HMM2) tended to yield better recognition rates than HMM1, which can be explained by the generally better viewpoint of the former. The confusion matrices in Figure 5 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 offered a more favourable viewpoint for discerning task 1 from task 5, whereas camera 1 provided a better angle for recognising task 4.

A careful evaluation of the results shown in Table 1 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 outperformed the better of two individual streamwise models in terms of recall and precision by up to 6.2%. This improvement can be put down to the multistream fused model's capability of 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. 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 deteriorated when assuming perfect synchronicity by employing the state-synchronous approach, reflecting the fact that our cameras were indeed not perfectly synchronised.

Gaussian vs Student's t. Using Student's t-distribution instead of the conventional Gaussian as predictive function of the HMMs additionally increased recognition rates to a certain extent (ranging from 1.4% up to 11.4%). The contribution was more apparent in the experimets of dataset-2 (Table 1), where the amount of noise was greater, thus proving the usefulness of Student's t-distribution in enhancing the robustness to outliers in activity recognition from video streams.

6.2 *Neural network based rectification results*

Table 2 contains the results acquired after employing the rectification mechanism. Comparing the measures in Table 2 with the respective results of Table 1, we notice that the proposed rectification scheme provides a substanstial improvement. Recall, precision, and F-measure were all significantly increased compared to the respective experimental setups when no neural network based readjustment was performed. As expected, multistream fused HMM supplemented with the rectification mechanism provided the best results among the approaches that rectify the fused results, since it was also the best performing approach when stand-alone. However, we observe that our proposed RDFHMM, which first readjusts the streamwise probabilities before feeding them into the adapted fusion model, yielded the best results, slightly outperforming MFHMM+RM, with recall rates of up to 95% and 79.8% for datasets 1 and 2 respectively.

Comparing our results with those of (Veres et al., 2010) (the results presented therein concern camera 1 from dataset-1) we observe that the streamwise HMM1 (Student's t) method outperforms the ESN based approach both in terms of recall and precision. The difference in performance increases when considering multistream fusion or rectification. We also experimented with ESN using the features described here, so as to compare the performance of our methods in both datasets. To this end, we used a network of 500 nodes, which was efficient for real time execution and avoided overfitting. It had seven output nodes, each one corresponding to a predicted task. The median of the last 101 estimations was taken to ensure lower jitter in the output. We have used the Matlab toolbox provided by (Jaeger et al., 2007) after parameters' optimisation using trial and error. The F-measures were 80.3% and 82.6% (camera 1 and 2) for dataset-1, and 60.5% and 57.3% (camera 1 and 2) for dataset-2, i.e., comparable to the respective single stream HMM. However, employing the existing HMM based fusion schemes as well as exploiting user feedback through rectification (even better through RDFHMM) can lead to significant improvement of performance.

Figures 6(a) and 6(b) display the % classification error for all experimental setups with and without the rectification mechanism for datasets 1 and 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 7(a) and 7(b). Clearly, rectification significantly enhanced the performance of the proposed framework, especially when implemented in the form of the proposed RDFHMM.

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. The extraction of holistic features to bypass tracking, the employment of Student's t-distribution and multicamera fusion can address the challenges involved. 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 (such as the hidden Markov model and its fused versions) according to a set of data selected by an expert user through an interactive framework. The non-linear rectification is accomplished 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 approaches. In the first, the rectification mechanism readjusts the probabilites stemming from the fused stochastic model and produces the final activity recognition decision; in the second, the rectification mechanism readjusts the streamwise probabilities and feeds its output to the proposed Rectification Driven Fused HMM (RDFHMM), which fuses the readjusted probabilites and extracts the recognised activity.

We have tested the proposed methodology in very challenging datasets from a real production line of an automobile industry. The results illustrate significant improvement when applying the rectification mechanism, while the proposed RDFHMM yields the best recognition rates. Regarding the practical implications of our results, the demonstrated experiment concerns real industrial workflows without any sort of environment engineering. So far no assumptions have been made about occlusions, illumination changes, or workers motion, etc, so the setting is very challenging. The recognition rate is not expected to be perfect using the proposed method under such conditions. For accuracy that approximates 100% we would need to apply some additional constraints in the monitored scene, e.g., controlled illumination, enforced paths to workers, controlled timing for tasks, etc. The application of such constraints is not unusual in production environments and if they are adopted it would be realistic to expect nearly perfect performance, because the repeatability of the tasks would be much higher.

As future research, we plan to exploit adaptive neural network models in order to recursively readjust the classification probabilities during the activity execution and to investigate dynamic methods for readjusting the learning process of the involved stochastic models.

References

Ali, S. and M. Shah (2010). Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*(2), 288 –303.

Antonakaki, P., D. Kosmopoulos, and S. Perantonis (2009). Detecting abnormal human behaviour using

multiple cameras. *Signal Processing 89*(9), 1723 – 1738.

Boiman, O. and M. Irani (2005). Detecting irregularities in images and in video. In *Proc. of the 10th IEEE International Conference on Computer Vision (ICCV) 2005*, Volume 1, pp. 462 – 469.

Brand, M., N. Oliver, and A. Pentland (1997). Coupled hidden markov models for complex action recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1997*, pp. 994 –999.

Breitenstein, M., H. Grabner, and L. Van Gool (2009). Hunting nessie - real-time abnormality detection from webcams. In *Proc. of the 12th IEEE International Conference on Computer Vision Workshops (ICCV Workshops) 2009*, pp. 1243 –1250.

Chatzis, S. P., D. I. Kosmopoulos, and T. A. Varvarigou (2009). Robust sequential data modeling using an outlier tolerant hidden markov model. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(9), 1657–1669.

Dalal, N. and B. Triggs (2005). Histograms of oriented gradients for human detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*, pp. 886–893.

Doulamis, A. (2005). Knowledge extraction in stereo video sequences using adaptive neural networks. Intelligent Multimedia Processing with Soft Computing, pp. 235–252. Springer-Verlag.

Doulamis, A. (2010). Dynamic tracking re-adjustment: A method for automatic tracking recovery in complex visual environments. *Multimedia Tools and Applications 50*(1), 49–73.

Doulamis, N. and A. Doulamis (2006). Evaluation of relevance feedback schemes in content-based in retrieval systems. *Signal Processing: Image Communication 21*(4), 334 – 357.

Doulamis, N., A. Doulamis, and S. Kollias (2000). Nonlinear relevance feedback: improving the performance of content-based retrieval systems. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME) 2000*, Volume 1, pp. 331 –334 vol.1.

Dupont, S. and J. Luettin (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia 2*(3), 141–151.

Fusier, F., V. Valentin, F. Bremond, M. Thonnat, M. Borg, D. Thirde, and J. Ferryman (2007). Video understanding for complex activity recognition. *Machine Vision and Applications 18*, 167–188.

Hamid, R., S. Maddi, A. Bobick, and M. Essa (2007). Structure from statistics - unsupervised activity analysis using suffix trees. In *Proc. of the 11th IEEE International Conference on Computer Vision (ICCV) 2007*, pp. 1 –8.

Hu, Q., L. Qin, Q. Huang, S. Jiang, and Q. Tian (2010). Action recognition using spatial-temporal context. In *Proc. of the 20th International Conference on Pattern Recognition (ICPR) 2010*, pp. 1521 –1524.

Hung, M.-H. and C.-H. Hsieh (2008). Event detection of broadcast baseball videos. *IEEE Transactions on Circuits and Systems for Video Technology 18*(12), 1713 –1726.

Ivanov, Y. and A. Bobick (2000). Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(8), 852 –872.

Jaeger, H., W. Maass, and J. Principe (2007). Special issue on echo state networks and liquid state machines. *Neural Networks 20*(3), 287 – 289.

Johnson, N. and D. Hogg (1996). Learning the distribution of object trajectories for event recognition. *Image and Vision Computing 14*(8), 609 – 615.

Lao, W., H. J. d. W. P. (2009). Automatic video-based human motion analyzer for consumer surveillance system. *IEEE Transactions on Consumer Electronics 55*(2), 591–598.

Laptev, I. and P. Perez (2007). Retrieving actions in movies. In *Proc. of the 11th IEEE International Conference on Computer Vision (ICCV) 2007*, pp. 1 –8.

Makris, A., D. Kosmopoulos, S. Perantonis, and S. Theodoridis (2011). A hierarchical feature fusion framework for adaptive visual tracking. *Image and Vision Computing 29*(9), 594 – 606.

Mukundan, R. and K. R. Ramakrishnan (1998). *Moment Functions in Image Analysis: Theory and Applications*. New York: World Scientific.

Nefian, A., L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy (2002). A coupled hmm for audio-visual speech recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2002*, Volume 2, pp. II –II.

Oerlemans, A., J. T. Rijsdam, and M. S. Lew (2007). Real-time object tracking with relevance feedback. In *Proc. of the 6th ACM international conference on Image and Video Retrieval (CIVR) 2007*, pp. 101–104.

Ostendorf, M. and H. Singer (1997). HMM topology design using maximum likelihood successive state splitting. *Computer Speech & Language 11*(1), 17–41.

Padoy, N., D. Mateus, D. Weinland, M.-O. Berger, and N. Navab (2009). Workflow monitoring based on 3d motion features. In *Proc. of the 12th IEEE International Conference on Computer Vision Workshops (ICCV Workshops) 2009*, pp. 585 –592.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing 28*(6), 976 – 990.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In *The Smart retrieval system -*

experiments in automatic document processing, pp. 313–323. Englewood Cliffs, NJ: Prentice-Hall.

Stiefmeier, T., D. Roggen, G. Troster, G. Ogris, and P. Lukowicz (2008). Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing 7*(2), 42 –50.

Veres, G., H. Grabner, L. Middleton, and L. V. Gool (2010). Automatic workflow monitoring in industrial environments. In *Proc. of the Asian Conference on Computer Vision (ACCV) 2010*, pp. 200–213.

Vogler, C. and D. Metaxas (1999). Parallel hidden markov models for american sign language recognition. In *Proc. of the 7th IEEE International Conference on Computer Vision (ICCV) 1999*, pp. 116–122.

Voulodimos, A., D. Kosmopoulos, G. Vasileiou, E. Sardis, A. Doulamis, V. Anagnostopoulos, C. Lalos, and T. Varvarigou (2011). A dataset for workflow recognition in industrial scenes. In *Proc. of the 18th IEEE International Conference on Image Processing (ICIP) 2011*, pp. 3310–3313.

Wada, T. and T. Matsuyama (2000). Multiobject behavior recognition by event driven selective attention method. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(8), 873 –887.

Wang, X., K. T. Ma, G.-W. Ng, and W. Grimson (2008). Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*, pp. 1 –8.

Wang, Y. and G. Mori (2009). Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(10), 1762 –1774.

Xiang, T. and S. Gong (2006). Beyond tracking: modelling activity and understanding behaviour. *International Journal of Computer Vision 67*, 21–51.

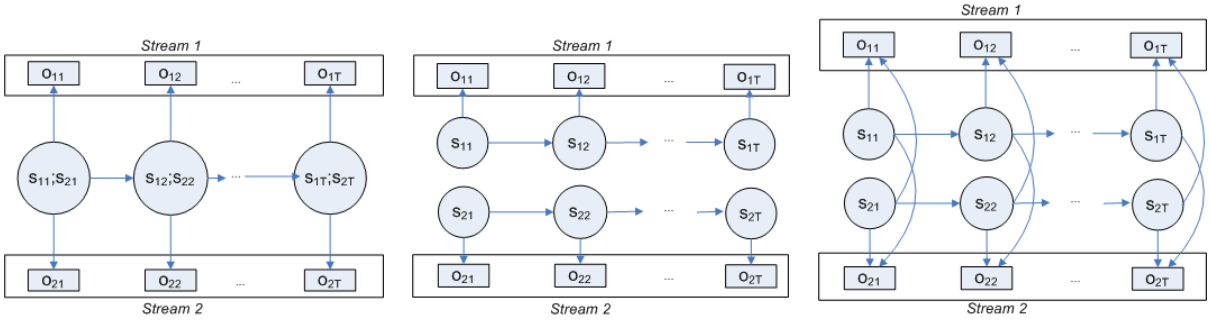
Yao, A., J. Gall, and L. Van Gool (2010). A hough transform-based voting framework for action recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*, pp. 2061 –2068.

Zeng, Z., J. Tu, B. Pianfetti, and T. Huang (2008). Audio-visual affective expression recognition through multistream fused HMM. *IEEE Transactions on Multimedia 10*(4), 570–577.

Zhang, C., W.-B. Chen, X. Chen, L. Yang, and J. Johnstone (2010). A multiple instance learning and relevance feedback framework for retrieving abnormal incidents in surveillance videos. *Journal of Multimedia 5*(4).

Zhou, X. S. and T. S. Huang (2003). Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems 8*(6), 536–544.

Figure 1. Sequences from our industrial environment dataset. Object tracking as well as activity recognition is extremely challenging due to occlusions, low resolution, and high intraclass and low interclass variance. The first two rows depict two different activities that are executed during the production cycle: their resemblance is so high, that they would be difficult to distinguish even for the human eye; the third row shows some example frames of occlusions, outliers, sparks, abnormalities, etc.



(a) State-synchronous HMM (b) Parallel HMM (c) Multistream fused HMM

Figure 2. HMM based fusion approaches for two streams. Symbols s and o stand for states and observations respectively. The first index indicates the stream and the second the time.

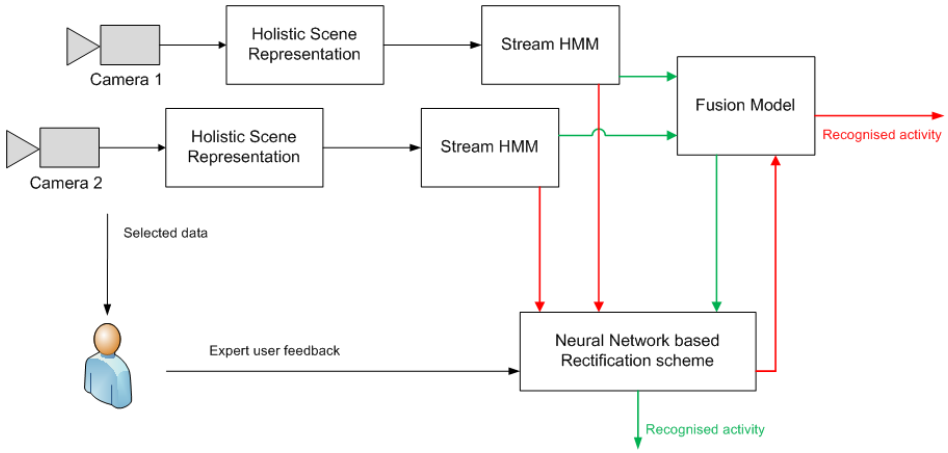


Figure 3. Schematic overview: The neural network based rectification mechanism is examined under two different approaches (corresponding to the green and red paths respectively). The green approach rectifies the fused result produced by the fused HMM, while the red one performs streamwise rectification and in the sequel the rectified streams are fused (RDFHMM).

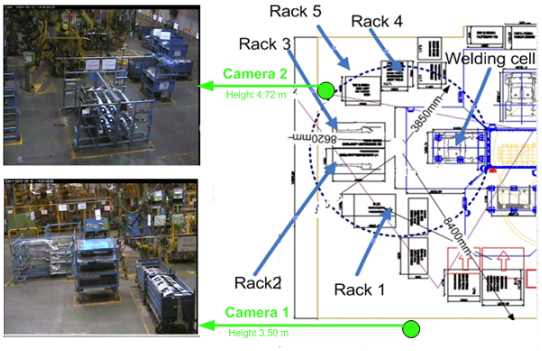
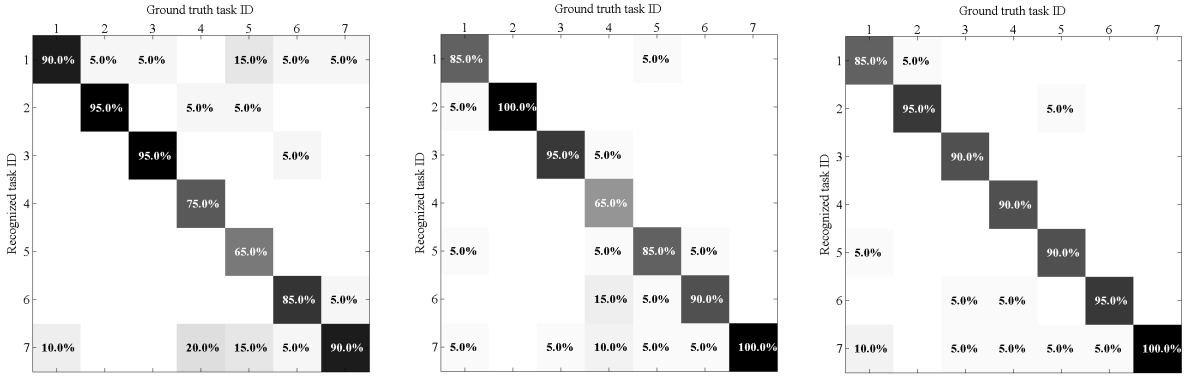
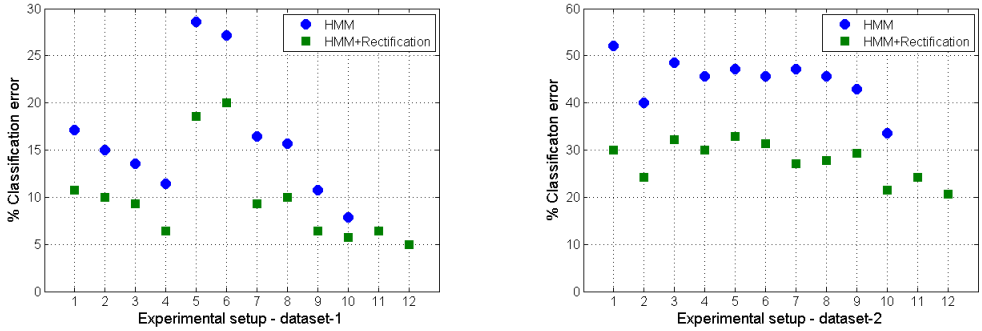


Figure 4. Depiction of workcell along with the position of the cameras and racks #1-5.



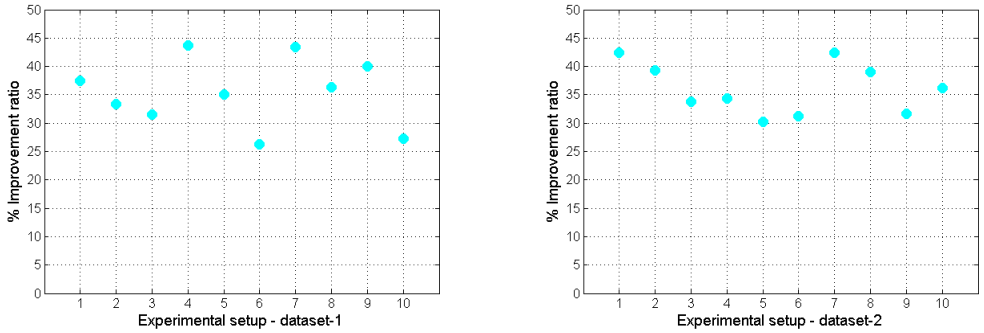
(a) Camera 1 HMM (b) Camera 2 HMM (c) Mulistream fused HMM

Figure 5. Confusion matrices from dataset-1 for a) individual HMM for camera 1, b) individual HMM for camera 2 and c) mulistream fused HMM, using Student's t-distribution.



(a) Dataset-1 (b) Dataset-2

Figure 6. Classification error % with and without the rectification mechanism for all experimental setups: 1.HMM1-Gauss, 2.HMM1-Student-t, 3.HMM2-Gauss, 4.HMM2-Student-t, 5.SYNC-Gauss, 6.SYNC-Student-t, 7.PARAL-Gauss, 8.PARAL-Student-t, 9.MULTI-Gauss, 10.MULTI-Student-t, 11.RDFHMM-Gauss, 12.RDFHMM-Student-t (11 & 12 have no corresponding non-rectified setup).



(a) Dataset-1 (b) Dataset-2

Figure 7. Improvement ratio % in terms of error decrease for all experimental setups: 1.HMM1-Gauss, 2.HMM1-Student-t, 3.HMM2-Gauss, 4.HMM2-Student-t, 5.SYNC-Gauss, 6.SYNC-Student-t, 7.PARAL-Gauss, 8.PARAL-Student-t, 9.MULTI-Gauss, 10.MULTI-Student-t, (above mentioned 11 & 12 have no corresponding non-rectified setup therefore no improvement ratio can be calculated).

Table 1. Results obtained from dataset-1 and 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's t-distribution as observation likelihood.

		dataset-1			*dataset-2*		
		Recall	Precision	F-measure	Recall	Precision	F-measure
HMM1	Gauss	82.4%	78.1%	80.2%	51.8%	49.8%	50.8%
	Student-t	85.0%	80.2%	82.5%	63.2%	58.6%	60.8%
HMM2	Gauss	86.4%	82.7%	84.5%	53.7%	49.9%	51.7%
	Student-t	88.6%	84.1%	86.3%	56.3%	53.4%	54.8%
SYNC	Gauss	70.3%	62.1%	65.9%	53.9%	49.5%	51.6%
	Student-t	73.3%	64.1%	68.4%	57.1%	53.9%	55.5%
PARAL	Gauss	83.3%	78.9%	81.0%	54.9%	49.9%	52.3%
	Student-t	84.5%	80.8%	82.6%	59.3%	54.3%	56.7%
MULTI	Gauss	89.1%	86.5%	87.8%	59.1%	52.5%	55.6%
	Student-t	92.1%	89.8%	90.9%	67.6%	59.6%	63.3%

Table 2. Results obtained from dataset-1 and 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's t-distribution as observation likelihood.

		dataset-1			*dataset-2*		
		Recall	Precision	F-measure	Recall	Precision	F-measure
HMM1+RM	Gauss	89.6%	83.9%	86.7%	72.0%	62.4%	66.9%
	Student-t	90.2%	86.8%	88.5%	77.4%	66.7%	71.7%
HMM2+RM	Gauss	91.0%	86.4%	88.6%	69.5%	61.3%	65.1%
	Student-t	93.5%	90.7%	92.1%	73.2%	62.3%	67.3%
SYNC+RM	Gauss	80.7%	77.1%	78.9%	70.4%	65.2%	67.7%
	Student-t	80.1%	74.8%	77.4%	72.5%	66.4%	69.3%
PARAL+RM	Gauss	90.5%	87.2%	88.8%	73.9%	67.7%	70.7%
	Student-t	90.3%	87.6%	88.9%	74.3%	66.3%	70.1%
MULTI+RM	Gauss	93.7%	91.2%	92.4%	73.1%	68.5%	70.7%
	Student-t	94.2%	91.8%	92.9%	78.9%	72.3%	75.4%
RDFHMM	Gauss	93.8%	91.3%	92.5%	75.7%	70.4%	72.9%
	Student-t	95.0%	93.2%	94.1%	79.8%	77.3%	78.5%