Robust workflow recognition using holistic features and outlier-tolerant fused Hidden Markov Models

Athanasios Voulodimos[1], Helmut Grabner[2], Dimitrios Kosmopoulos[3], Luc Van Gool[2,4], and Theodora Varvarigou[1]

[1] School of Electrical & Computer Engineering, National Technical University of Athens, Greece
{thanos, dora}@telecom.ntua.gr
[2] Computer Vision Laboratory, ETH Zurich, Switzerland
{grabner, vangool}@vision.ee.ethz.ch
[3] Institute of Informatics and Telecommunications, N.C.S.R. Demokritos, Greece
dkosmo@iit.demokritos.gr
[4] ESAT-PSI/IBBT, K.U. Leuven, Belgium
luc.vangool@esat.kuleuven.be

Abstract. Monitoring real world environments such as industrial scenes is a challenging task due to heavy occlusions, resemblance of different processes, frequent illumination changes, etc. We propose a robust framework for recognizing workflows in such complex environments, boasting a threefold contribution: Firstly, we employ a novel holistic scene descriptor to efficiently and robustly model complex scenes, thus bypassing the very challenging tasks of target recognition and tracking. Secondly, we handle the problem of limited visibility and occlusions by exploiting redundancies through the use of merged information from multiple cameras. Finally, we use the multivariate Student-t distribution as the observation likelihood of the employed Hidden Markov Models, in order to further enhance robustness. We evaluate the performance of the examined approaches under real-life visual behavior understanding scenarios and we compare and discuss the obtained results.

Key words: robust workflow recognition, Hidden Markov Models, Classifier Grids, multi-camera fusion

1 Introduction

Event understanding in video sequences is a research field rapidly gaining momentum over the last few years. This is mainly due to its fundamental applications in automated video indexing, virtual reality, human-computer interaction, assistive living and smart monitoring. Especially throughout the last years we have seen an increasing need for assisting and extending the capabilities of human operators in remotely monitored large and complex spaces such as public areas, airports, railway stations, parking lots, industrial plants, etc.

Fig. 1. Sequences from the dataset. The relatively low resolution and the several occlusions and self occlusions make very difficult the task of tracking thus necessitating holistic features and a robust model to recognize workflows. The first two rows depict two different tasks that would be difficult to distinguish even for the human eye; the third row shows some example frames of occlusions, outliers, and other challenges faced in this industrial dataset.

Focusing on industrial scenes, the serious visibility problems, the heavy occlusions, along with the high diversity, complexity or sometimes resemblance of the behaviors and events taking place, make workflow recognition extremely challenging. In this paper the case study is an assembly line of an automobile manufacturer, where several different tasks are performed, and a sequence of specific tasks forms a workflow. The goal of recognizing these tasks (classes) and workflows is even more difficult to achieve when taking into consideration the high intraclass and low interclass variance, as shown in Fig. 1. Typical methods tend to fail in such environments, since they rely on object detection and tracking, which are rarely successful under such circumstances. To overcome the aforementioned problems, we propose a robust framework for workflow recognition that contributes to the solution in the three following ways:

- We propose new holistic features, which can be efficiently computed, do not rely on target detection and tracking and can be used to model complex scenes, thus resulting in robust input.
- In addition, we include redundant data by using multiple cameras in order to provide wider scene coverage, solve occlusions and improve accuracy. This is achieved by fusing time series of the above mentioned holistic image features, which is, according to our knowledge, a novel approach.
- Moreover, we scrutinize the effectiveness of the multivariate Student-t distribution, instead of the Gaussian, as the observation likelihood of the employed

Hidden Markov Models (HMMs), so as to solve the problem of outliers and further enhance the robustness of the model.

The rest of this work is organized as follows. In Sec. 2 we briefly survey the related work. Sec. 3,4 and 5 describe details of our approach, with respect to efficiency and robustness respectively. In Sec. 6 we verify our methods experimentally on a real-world dataset from an assembly line of an automobile industry. Finally, Sec. 7 concludes the paper.

2 Related work

The field of behavior and workflow recognition has attracted the interest of many researchers. Holistic methods, which define features at the pixel level and try to identify patterns of activity using them directly, can bypass the challenging processes of detection and tracking. Such methods may use pixel or pixel group features such as color, texture or gradient, see e.g. [1] (histograms of spatiotemporal gradients), [2] (spatiotemporal patches). Of particular interest due to efficiency and representation of motion are approaches such as [3], which introduced Motion Energy Images (MEIs) and Motion History Images (MHIs), and [4], where Motion History Volumes are extracted from multiple cameras. Pixel Change History is used in [5] to represent each target separately after frame differencing. What is needed to model complex scenes is a representation that will be able to operate in any adverse condition effected by occlusions, illumination changes or abrupt motion.

As far as multiple cameras are concerned, to our knowledge no previous work has investigated fusion of holistic time series. The works on multicamera behavior recogition that have been reported so far try to solve the problem of position or posture extraction in 3D or on ground coordinates (e.g. [6, 7]). 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 illumniation changes and occlusions.

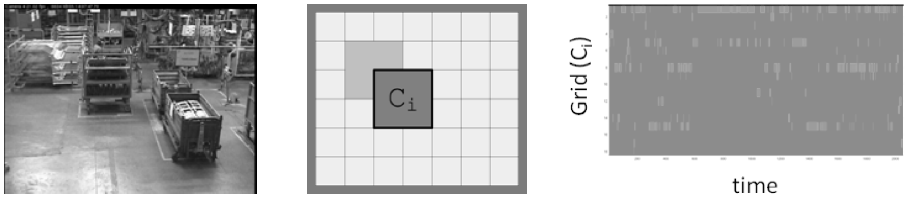
Concerning the classification part, a very popular approach is HMMs ([8], [9], [10]) due to the fact that they can efficiently model stochastic time series at various time scales. Several fusion schemes using HMMs have been presented, which were typically used for fusing heterogeneous feature streams such as audio-visual systems, but can be applied to streams of holistic features from multiple cameras as well. Such examples are the early fusion, the synchronous HMMs [11], the parallel HMMs [12] and the multistream HMMs [13]. The reliability of each stream has been expressed by introducing stream-wise factors in the total likelihood estimation as in the case of parallel, synchronous or multistream HMMs.

3 Robust scene representation

Classifier grids were initially introduced to perform background modeling [14]. In this approach, an input image I_t is spatially (location and scale) sampled with

a fixed highly overlapping grid. For each grid element i, an adaptive classifier C_i is created. These classifiers can now be used in a static camera setting in order to aggregate scene and location specific information. Classifier grids have been successfully used for pedestrian detection, e.g. [15]. Experiments show that very good detection results can be achieved compared with the sliding window technique, which uses a fixed pre-trained classifier which scans the whole image.

In our work, we propose to use the output of the classifier grid as scene descriptor. In other words, the local classifiers can be seen as features which extract "high level" information from each image. Hence, our proposed approach analyses time series, and afterwards all classifier responses are concatenated into one vector. These vectors observed over time t define finally the *grid time matrix*. The principle is depicted in Fig. 2.



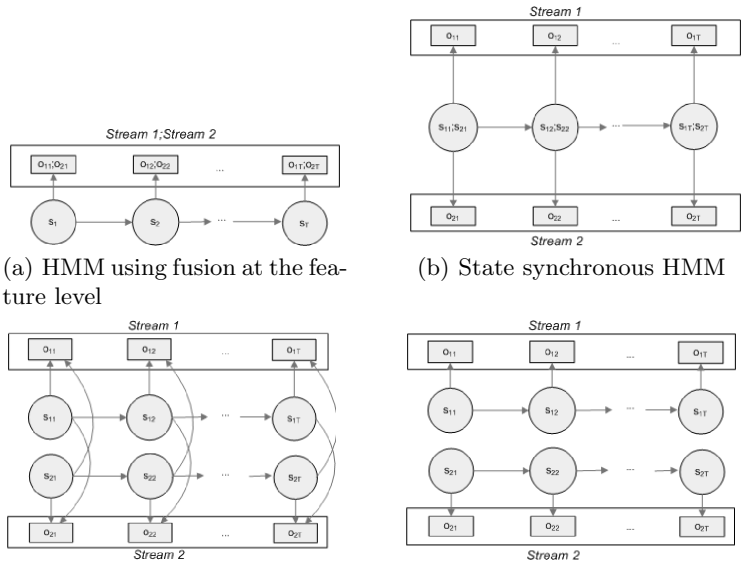
(a) Input image (b) Classifier grid (c) Grid time matrix

Fig. 2. Grid time matrix composition: An input image (a) is analyzed by a highly overlapping grid of classifiers (b). Classifier responses are concatenated over time and used as holistic image description.

4 Multi-view learning

The goal of automatic behavior recognition may be viewed as the recovery of a specific learned behavior (class or visual task) from the sequence of observations O. Each camera frame is associated with one observation vector and the observations from all cameras have to be combined in a fusion framework to exploit complementarity of the different views. The sequence of observations from each camera composes a separate camera-specific information stream, which can be modelled by a camera-specific HMM.

The HMM framework 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 $\boldsymbol{A}$ 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. Given a learned HMM, probability assignment for an observation sequence is performed.



(a) HMM using fusion at the feature level

(b) State synchronous HMM

(c) Multistream fused HMM

(d) Parallel HMM

Fig. 3. Various fusion schemes using the HMM framework for two streams.

In a multicamera setup each sensor stream can be used to generate a stream of observations. The ultimate goal of multicamera fusion is to achieve behavior recognition results better than the results that we could attain by using the information obtained by the individual data streams (stemming from different cameras) independently from each other. We will examine in the following some representative approaches, which can support scalable behavior recognition with several overlapping cameras.

Among existing approaches *Feature fusion* is the simplest; it assumes that the observation streams are synchronous. The related architecture is displayed in Fig. 3(a). For streams from C cameras and respective observations at time t given by $\boldsymbol{o}_{1t}$,..., $\boldsymbol{o}_{Ct}$, the proposed scheme defines the full observation vector as a simple concatenation of the individual observations: $\boldsymbol{o}_t = \{\boldsymbol{o}_{ct}\}_{c=1}^{C}$. Then, the observation emission probability of the state $s_t = i$ of the fused model, when considered as a k-component mixture model, yields:

$$P(\boldsymbol{o}_t|s_t = i) = \sum_{k=1}^{K} w_{ik} P(\boldsymbol{o}_t|\theta_{ik}) \tag{1}$$

where w_{ik} denotes the weights of the mixtures and θ_{ik} the parameters of the kth component density of the ith state.

In the *state-synchronous multistream HMM* (see Fig. 3(b)) the streams are assumed to be synchronized. Each stream is modelled using an individual HMM; the postulated streamwise HMMs share the same state dynamics. 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 [11]:

$$P(\boldsymbol{o}_t|s_t = i) = \prod_{c=1..C} [\sum_{k=1}^{K} w_{ik} P(\boldsymbol{o}_{ct}|\theta_{ik})]^{r_c} \tag{2}$$

The weight r_c is associated with the reliability of the information carried by the c^{th} stream. Another alternative is the *parallel HMM* (see Fig. 3(c)); it assumes that the streams are independent from each other. This HMM model can be applied to cameras that may not be synchronized 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 $\hat{l}$ that maximizes the weighted sum of the classification probabilities from the streamwise HMMs:

$$\hat{l} = \underset{l}{\text{argmax}}([\sum_{c=1}^{C} r_c log P(\boldsymbol{o}_1 ... \boldsymbol{o}_T|\lambda_{cl})]) \tag{3}$$

where λ_{cl} are the parameters of the postulated streamwise HMM of the cth stream that corresponds to the lth class.

The *multistream fused HMM* is another promising method for modeling of multistream data [13] (see Fig. 3(d)) with several desirable features: (i) it is appropriate for both synchronous and asynchronous camera networks; (ii) it has simple and fast training and inference algorithms; (iii) if one of the component HMMs fails, the remaining HMMs can still work properly; and (iv) it retains the crucial information about the interdependencies between the multiple data streams Similar to the case of parallel HMMs, the class that maximizes the weighted sum of the log-likelihoods over the streamwise models is the winner.

5 Robustness to outliers

Outliers are expected to appear in model training and test data sets 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 fusion models, in order to address the problem.

The probability density function (pdf) of a Student-t distribution with mean vector μ, positive definite inner product matrix Σ, and ν degrees of freedom is given by:

$$t\left(x_t; \mu, \Sigma, \nu\right) = \frac{\Gamma\left(\frac{\nu+p}{2}\right) |\Sigma|^{-\frac{1}{2}} (\pi\nu)^{-\frac{p}{2}}}{\Gamma\left(\frac{\nu}{2}\right) \left\{1 + d\left(x_t, \mu; \Sigma\right)/\nu\right\}^{\frac{\nu+p}{2}}} \tag{4}$$

where $\Gamma(.)$ denotes the gamma function and d the Mahalanobis distance. The heavier tails of the Student-t distribution compared to the Gaussian ensure higher tolerance to outliers. The Gaussian distribution is actually a special case of the Student-t for $\nu \to \infty$. 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 [16]. Based on that we propose the following adaptations in the above fusion schemes: For the feature fusion, synchronous, parallel and multistream models we use the Student-t pdf as predictive function for the streamwise models. We use a modified EM training algorithm and solve numerically to obtain ν. For the interstream fusion model we employ a mixture of Student-t functions to increase robustness.

6 Experiments

We experimentally verified the applicability of the described methods. For this purpose, we have acquired very challenging videos from the production line of a major automobile manufacturer[5]. Two synchronized, partially overlapping views are used. Challenges include occlusions, similar colors of the individual people clothing and the background, and real-working conditions, such as shaking cameras and sparks.

Experimental setup. The production cycle on the production line included tasks of picking several parts from racks and placing them on a designated cell some meters away, where welding took place. Each of the above tasks was regarded as a class of behavioral patterns that had to be recognized. A specific sequence of those tasks constitutes a workflow. The information acquired from this procedure can be used for the extraction of production statistics or anomaly detection. The workspace configuration and the cameras' positioning is given in Fig. 4. The behaviors we are aiming to model in the examined application are briefly the following:

1. A worker picks part #1 from rack #1 and places it on the welding cell.
2. Two workers pick part #2a from rack #2 and place it on the welding cell.
3. Two workers pick part #2b from rack #3 and place it on the welding cell.
4. A worker picks parts #3a, #3b from rack #4 and places them on the cell.
5. A worker picks part #4 from rack #1 and places it on the welding cell.
6. Two workers pick part #5 from rack #5 and place it on the welding cell.
7. Welding: two workers grab the welding tools and weld the parts together.

For our experiments, we have used 20 segmented sequences representing full assembly cycles, each one containing each of the seven behaviors/tasks. The total number of frames was approximately 80,000. The videos were shot by two PTZ cameras at an approximate framerate of 25 fps and at a resolution of 704×576. The annotation of these frames has been done manually. For more dependable results, in our experiments we used cross-validation, by repeating the employed training algorithms several times, where in each repetition all scenarios are considered except for one used for testing (leave-one-out cross-validation).

[5] We are currently investigating legal issues of making the dataset publically available.

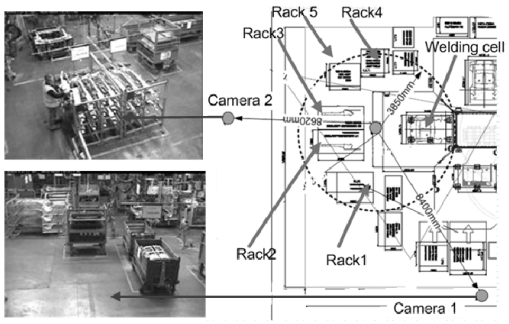


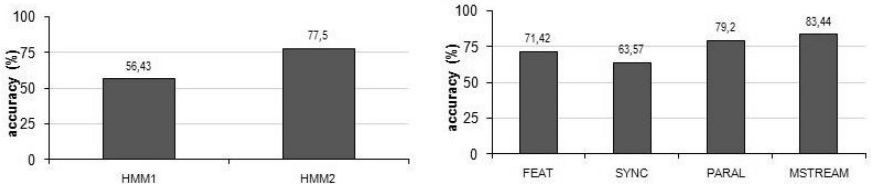
Fig. 4. Depiction of workcell along with the position of the cameras and racks #1-5.

Representation and feature extraction. We created a classifier regular grid with overlap 0.5 (50%). Each frame was eventually represented by a 42-dimensional feature vector. For learning and adapting the classifiers we have used a simple motion based heuristic. Each local classifier learns a simple background model [17]. As classification function, the amount of moving pixels, i.e. the differnce between the current image and the background model, is used. For each stream corresponding to a different viewpoint we have selected a region of interest, to which the classifier grids have been applied, as the activity taking place in the remaining area of the frame is noise.

Learning. We trained our models using the EM algorithm. We used the typical HMM model for the individual streams as well as feature fusion, synchronous, parallel and multistream HMMs. We experimented with the Gaussian observation model as well as with the multivariate Student-t model. 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 HMM we use mixture models of two component distributions.

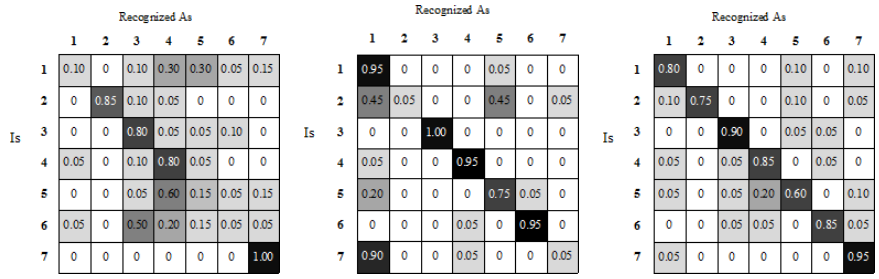
Results. The obtained results of the experiments are shown in Fig. 5. It becomes obvious that the sequences of our features and the respective HMMs represent quite well the assembly process. Information fusion seems to provide significant added value when implemented in the form of the multistream fused HMM, and about similar accuracy when using parallel HMMs. However, the accuracy deteriorates significantly when using simple feature level fusion or state-synchronous HMMs, reflecting the known restrictions of these approaches.

The confusion matrices in Fig. 6 show the percentage of successful and unsuccessful task recognitions averaged across all classes (tasks). A look at the matrices would justify the complementarity between the two camera streams



(a) Single camera (b) Fusion

Fig. 5. Success rates obtained using Student-t distribution and (i) individual HMM for camera 1 (HMM1); (ii) individual HMM for camera 2 (HMM2); (iii) feature-level fusion (FEAT); (iii) state-synchronous HMMs (SYNC); (iv) parallel HMMs (PARAL) and (v) multistream fused HMMs (MSTREAM)



(a) Camera 1 (HMM1) (b) Camera 2 (HMM2) (c) Multistream fusion

Fig. 6. Confusion matrices for individual tasks.

due to the different viewpoints. Camera 1 performs well for task number 2 and 7 while camera 2 performs better for the rest. For example, camera 1's viewpoint is such, that discerning task 1 from task 5 is extremely difficult - even for a human - hence the low success rates in these particular tasks; on the contrary, camera 2's viewpoint is much better for viewing tasks 1 and 5 and therefore allows for a significantly higher performance, which can be confirmed by noticing the confusion matrices. This complementarity of the two streams results in the improvement of the accuracy by the streams' fusion when the latter is implemented as a multistream fused HMM. Finally, the employment of the Student-t distribution as observation likelihood of the employed HMM provides additional improvement from 81.43% (Gaussian) to 83.44% (Student-t) in recognition rates.

7 Conclusion

It has been shown that a fused holistic scene representation, which uses a grid time matrix, is very well suited for monitoring and classifying well structured processes such as the production tasks in an assembly line. Using the proposed holistic features to bypass the challenging tasks of detection and tracking, which are usually unsuccessful in such environments, leads to a rather satisfactory

representation. Furthermore, exploiting redundancies by fusing time series from multiple cameras using the multistream fused HMMs results in higher recognition rates than those achieved when employing one single camera. Finally, employing an outlier-tolerant observation model based on the Student-t multivariate distribution instead of the Gaussian further enhances accuracy and robustness.

References

1. Zelnik-Manor, L.: Statistical analysis of dynamic actions. IEEE Trans. Pattern Anal. Mach. Intell. **28**(9) (2006) 1530–1535 Member-Irani, Michal.
2. Laptev, I., Pe'rez, P.: Retrieving actions in movies. In: Proc. Int. Conf. Comp. Vis.(ICCV'07), Rio de Janeiro, Brazil (October 2007) 1–8
3. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3) (2001) 257–267
4. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Comput. Vis. Image Underst. **104**(2) (2006) 249–257
5. Xiang, T., Gong, S.: Beyond tracking: modelling activity and understanding behaviour. International Journal of Computer Vision **67** (2006) 21–51
6. Antonakaki, P., Kosmopoulos, D., Perantonis, S.: Detecting abnormal human behaviour using multiple cameras. Signal Processing **89**(9) (2009) 1723 – 1738
7. Lao, W., H.J.d.W.P.: Automatic video-based human motion analyzer for consumer surveillance system. IEEE Trans. on Consumer Electronics **55**(2) (2009) 591–598
8. Bregler, C., Malik, J.: Learning appearance based models: Mixtures of second moment experts. In Mozer, M.C., Jordan, M.I., Petsche, T., eds.: Advances in Neural Information Processing Systems. Volume 9., The MIT Press (1997) 845
9. Ivanov, Y.A., Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8) (2000) 852–872
10. Bashir, F.I., Qu, W., Khokhar, A.A., Schonfeld, D.: Hmm-based motion recognition system using segmented pca. In: ICIP (3). (2005) 1288–1291
11. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. Multimedia, IEEE Transactions on **2**(3) (Sep 2000) 141–151
12. Vogler, C., Metaxas, D.: Parallel hidden markov models for american sign language recognition. (1999) 116–122
13. Zeng, Z., Tu, J., Pianfetti, B., Huang, T.: Audiovisual affective expression recognition through multistream fused hmm. IEEE Trans. Mult. **10**(4) (2008) 570–577
14. Grabner, H., Bischof, H.: On-line boosting and vision. In: Proc. CVPR. Volume 1. (2006) 260–267
15. Stalder, S., Grabner, H., van Gool, L.: Exploring context to learn scene specific object detectors. In: Proc. PETS. (2009)
16. Chatzis, S., Kosmopoulos, D., Varvarigou, T.: Robust sequential data modeling using an outlier tolerant hidden markov model. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(9) (sept. 2009) 1657 –1669
17. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Proc. CVPR. Number 2 (1999) 246–252