Robust human behavior modeling from multiple cameras

Dimitrios I. Kosmopoulos
Institute of Informatics and Telecommunications
N.C.S.R. Demokritos
Aghia Paraskevi 15310, Greece
dkosmo@iit.demokritos.gr

Athanasios S. Voulodimos, and Theodora A. Varvarigou
School of Electrical and Computer Engineering
National Technical University of Athens
Zografou 15773, Greece
{*thanos, dora*}*@telecom.ntua.gr*

***Abstract*—In this work, we propose a framework for classifying structured human behavior in complex real environments, where problems such as frequent illumination changes and heavy occlusions are expected. Since target recognition and tracking can be very challenging, we bypass these problems by employing an approach similar to Motion History Images for feature extraction. Furthermore, to tackle outliers residing within the training data, which might affect severely the training algorithm of models with Gaussian observation likelihoods, we scrutinize the effectiveness of the multivariate Student-t distribution as the observation likelihood of the employed Hidden Markov Models. Additionally, the problem of visibility and occlusions is addressed by providing various extensions of the framework for multiple cameras, both at the feature and at the state level. Finally, we evaluate the performance of the examined approaches under real-life visual behavior understanding scenarios and we compare and discuss the obtained results.**

***Keywords*-behavior modeling; recognition; Hidden Markov Models; fusion;**

I. Introduction

The field of behavior detection and understanding in video has been the focus of the interest of many researchers, mainly because of the vital importance of the related applications in virtual reality, human-computer interaction and smart monitoring. Smart monitoring is applicable to large-scale enterprises like industrial plants, which have a clear need for automated supervision services to guarantee safety, security, and quality by enforcing adherence to predefined procedures. Here, we focus on monitoring the production line of an automobile manufacturer, which is a rather structured process in contrast to monitoring stations or airports, thus being more realistic to believe that it can be modeled using computer vision and machine learning. The identified deviations from this process possibly indicate security and safety related events and will be automatically highlighted. The complexity of detection and tracking of moving objects under occlusions in a typical industrial environment requires more than a single camera and features that will not result from an error-prone tracker. Furthermore, the high diversity and complexity of the behaviors that need to be monitored require new learning methods that will be able to fuse information from multiple streams.

Based on these observations, our work contributes in the following ways:

- A holistic approach for action representation in every video frame, which bypasses the problem of object tracking, that is the proposition of new holistic features.
- A behavior recognition framework based on the aforementioned features, which is extended to solve the multicamera problem in an endeavor to alleviate visibility problems and occlusions.
- A means to enhance robustness to outliers, that is the use of the multivariate Student-t distribution as the observation likelihood of the Hidden Markov Models (HMMs).

The rest of this paper is organized as follows: In Section II we analyze the fusion frameworks and their applicability for multi-camera behavior recognition, as well as the Student-t observation model, which aims to enhance robustness to outliers. In Section III we explain our holistic approach for action representation in each frame and describe the feature extraction method. The experimental results are given in Section IV. Finally, Section V concludes the paper.

II. Robust 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 modeled by a camera-specific HMM.

The HMM framework entails a Markov chain comprising a number of, say, N states, with each state being coupled with an observation emission distribution. The EM (or Baum-Welch) algorithm is very popular for training HMMs under a maximum-likelihood framework. 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.

Among existing approaches *Feature fusion* is the simplest; it assumes that the observation streams are synchronous. This synchronicity is a valid assumption for cameras that have overlapping fields of view and support synchronization. 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:

$$\boldsymbol{o}_t = \{\boldsymbol{o}_{ct}\}_{c=1}^{C} \tag{1}$$

Then, the observation emission probability of the state $s_t = i$ of the fused model, when considered as a k-component mixture model, yields:

$$P(\boldsymbol{o}_t|s_t = i) = \sum_{k=1}^{K} w_{ik} P(\boldsymbol{o}_t|\theta_{ik}) \tag{2}$$

where w_{ik} denotes the weights of the mixtures and θ_{ik} the parameters of the kth component density of the ith state (e.g. mean and covariance matrix of a Gaussian pdf). Both training and testing are performed in the typical way using the obtained concatenated vectors.

In the *state-synchronous multistream HMM* [1] the streams are assumed to be synchronized. Each stream is modeled 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 [1]:

$$P(\boldsymbol{o}_t|s_t = i) = \prod_{c=1..C} [\sum_{k=1}^{K} w_{ik} P(\boldsymbol{o}_{ct}|\theta_{ik})]^{r_c} \tag{3}$$

The weight r_c is associated with the reliability of the information carried by the cth stream.

Another alternative is the *parallel HMM* [2]; it assumes that the streams are independent of each other, and, hence we can train one individual HMM for each stream in the typical way. This HMM-type 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 that maximizes the weighted sum of the classification probabilities from the streamwise HMMs, i.e. class assignment is conducted by picking the class $\hat{l}$ with:

$$\hat{l} = \underset{l}{\operatorname{argmax}}([\sum_{c=1}^{C} r_c log P(\boldsymbol{o}_1 ... \boldsymbol{o}_T|\lambda_{cl})]) \tag{4}$$

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 (MFHMM)* is another promising method for modeling of multistream data [3] with several desirable features: a) State transitions do not necessarily happen simultaneously, which makes the method approapriate for both synchronous and asynchronous camera networks; b) it has simple and fast training and inference algorithms; c) if one of the component HMMs fails due to noise or some other reason, the rest of the constituent HMMs can still work properly; and d) it still retains the crucial information about the interdependencies between the multiple data streams, which coupled HMMs tend to neglect. 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.

As far as outliers are concerned, they 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 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:

$$t\left(x_t; \mu, \Sigma, \nu\right) = \frac{\Gamma\left(\frac{\nu+p}{2}\right) |\Sigma|^{-\frac{1}{2}} (\pi\nu)^{-\frac{p}{2}}}{\Gamma\left(\frac{\nu}{2}\right) \left\{1 + d\left(x_t, \mu; \Sigma\right)/\nu\right\}^{\frac{\nu+p}{2}}} \tag{5}$$

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 [4]. 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 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.

III. Feature extraction

The features are calculated as follows: Firstly we perform background modeling. We use the foreground regions to represent the multi-scale spatiotemporal changes at pixel level. For this purpose we use a concept proposed in [5], 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:

$$P_{\varsigma,\tau}(x,y,t) = \begin{cases} \min(P_{\varsigma,\tau}(x,y,t-1) + \frac{255}{\varsigma}, 255) \\ if D(x,y,t) = 1 \\ \max(P_{\varsigma,\tau}(x,y,t-1) - \frac{255}{\tau}, 0) \\ otherwise \end{cases} \quad (6)$$

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. Zernike moments are very attractive because of their noise resiliency, their reduced information redundancy and their reconstruction capability. The complex Zernike moments of order p are defined as: (see for example [6]):

$$A_{pq} = \frac{p+1}{\pi} \int_0^1 \int_{-\pi}^{\pi} R_{pq}(r) e^{-jq\theta} f(r,\theta) r dr d\theta \quad (7)$$

where $r = \sqrt{x^2+y^2}$ and $\theta = \tan^{-1}(\mathrm{y/x})$ and $-1 < \mathrm{x,y} < 1$ and:

$$R_{pq}(r) = \sum_{s=0}^{\frac{p-q}{2}} (-1)^s \frac{(p-s)!}{s!(\frac{p+q}{2}-s)!(\frac{p-q}{2}-s)!} r^{p-2s} \quad (8)$$

where $p-q = even$ and $0 \le q \le p$. The higher the order of moments that we employ, the more detailed the region reconstruction will be, but also the more processing power will be required. Limiting the order of moments used is also justified by the fact that the details captured by higher order moments have much higher variability and are more sensitive to noise.

IV. Experiments and results

We experimentally verified the applicability of the described methods. For this purpose, we have acquired some very challenging videos from the production line of a major automobile manufacturer.

A. 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. The information acquired from this procedure can be used for the extraction of production statistics or anomaly detection. Partial or total occlusions due to the racks made the classification task difficult to effect using a single camera and therefore two synchronized, partially overlapping views are used. The workspace configuration and the cameras' and racks' positioning is given in Fig. 1. The work cycle that we

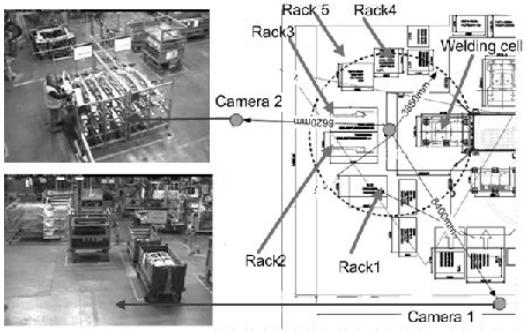

Figure 1. Depiction of workcell

sought to model, despite the noise and the several outliers, (e.g. persons walking into the working cell, vehicles passing by, etc), remains a structured process and is a good candidate to model with holistic methods.

The behaviors we were aiming to model in the examined application are briefly the following:

1) One 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 up parts #3a and #3b from rack #4 and places them on the welding cell.
5) A worker picks up part #4 from rack #1 and places it on the welding cell.
6) Two workers pick up 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 sequences representing full assembly cycles, each one containing at least one of the seven behaviors. The total number of frames was approximately 80,000. The annotation of these frames has been done manually. We have synchronized the cameras using the time stamps embedded by the camera server of our ip cameras. We have used cross validation by training using all scenarios except for one that was used for testing. For capturing the spatiotemporal variations we have set the parameters at ς = 10 and τ = 70. Furthermore, 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 very good scene reconstruction without too high a dimension (31). 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 rate of approximately 50-60 fps.

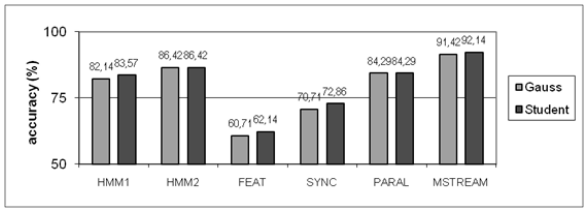

Figure 2. Success rates obtained using (i)individual HMM from stream1 (ii)individual HMM from stream2 (iii)feature-level fusion (iii)state-synchronous HMMs (iv)parallel HMMs and (v)multistream fused HMMs

B. Results

We trained our models using the EM algorithm. We used the typical HMM model for the individual streams as well as various HMM fusion approaches, namely 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.

The obtained results are given in Fig. 2, where the success rates using (i)individual HMM from stream 1, (ii)individual HMM from stream 2, (iii)feature-level fusion, (iii)state-synchronous HMMs, (iv)parallel HMMs, and (v)multistream fused HMMs, are shown. 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 (i.e. concatenation of feature vectors), or state-synchronous HMMs, reflecting the known restrictions of these approaches. Finally, the employment of the Student-t HMM provided some extra accuracy, thus proving its utility in visual behavior recognition applications, where outlier robustness is always of interest. In Fig. 3 we present the confusion matrix for the experiment conducted using the Student-t distribution. Each cell contains three numbers which correspond to the respective number of actual tasks i that were predicted as tasks j by using (i)multistream fusion / (ii)individual HMM from stream 1 / (iii)individual HMM from stream 2. The superiority of the multistream approach over the individual streams is obvious, while the lower performance of HMM 1 when it comes to discerning task 1 from task 5 can be justified by taking into consideration the similar description of the two tasks, as well as the position of camera 1 (Fig. 1), which provides the corresponding stream 1.

Actual \ Predicted	1	2	3	4	5	6	7
1	18/16/17	0/0/0	0/0/0	0/0/0	1/4/1	0/0/0	1/0/2
2	1/1/1	19/19/17	0/0/0	0/0/0	0/0/0	0/0/0	0/0/2
3	0/1/0	0/0/0	19/17/18	0/0/0	0/0/0	0/1/1	1/1/1
4	0/0/0	0/0/0	0/0/1	18/18/15	0/0/0	1/0/1	2/2/3
5	1/8/1	0/1/0	0/0/1	0/0/0	17/11/17	0/0/0	2/0/1
6	0/1/0	0/0/0	0/1/0	0/0/0	0/0/0	19/17/18	1/1/2
7	1/1/0	0/0/0	0/0/0	0/0/1	0/0/0	0/0/0	19/19/19

Figure 3. Confusion Matrix

V. Conclusion

In this work, we have presented a framework for fusion of multiple streams and we have applied it for recognition of visual tasks in an industrial environment using two cameras viewing the work cell from different angles to avoid occlusions. We have bypassed the challenging problem of tracking by having an image-based approach and by considering the foreground pixels. The proposed classification framework is appropriate for visual behavior recognition tasks and can be used to extend existing HMM-based behavior recognition systems to create scalable multicamera systems. Finally, through the complementarity from multiple views and by employing an outlier-tolerant observation model based on the Student-t multivariate distribution, enhancing accuracy is possible.

Acknowledgment

This work is partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 216465.

References

[1] S. Dupont & J. Luettin. Audio-visual speech modeling for continuous speech recognition, IEEE Trans. Multimedia, 2:141-151, 2000.

[2] C. Vogler & D. Metaxas. Parallel hidden Markov models for American sign language recognition. In Proc. ICCV 1999, vol.1, pp.116–122.

[3] Z. Zeng, J. Tu, B. Pianfetti, & T. Huang. Audio-visual affective expression recognition through multistream fused hmm. IEEE Trans. Multimedia, 10(4):570-577, June 2008.

[4] S. P. Chatzis, D. I. Kosmopoulos, & T. A. Varvarigou. Robust sequential data modeling using an outlier tolerant hidden markov model. IEEE Trans. on Pat. Anal. & Mach. Intel., 31(9):1657-1669, 2009.

[5] T. Xiang & S. Gong. Beyond tracking: Modelling activity and understanding behaviour, Int. J. Comput. Vision, vol. 67, no. 1, pp. 21-51, 2006.

[6] R. Mukundan and K. R. Ramakrishnan, Moment Functions in Image Analysis: Theory and Applications. New York: World Scientific, 1998.