Accepted in Advances in Visual Computing, 2015

Spatio-Temporal Fusion for Learning of Regions of Interests over Multiple Video Streams

Samaneh Khoshrou[1,2], Jaime S. Cardoso[1,2], Eric Granger[3], and Luís F. Teixeira[1,2]

[1] INESC TEC, Porto, Portugal,

[2] Faculdade de Engenharia da Universidade do Porto (FEUP), Portugal
skhoshrou,jaime.cardoso@inescporto.pt, lft@fe.up.pt

[3] Lab. d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montreal, Canada
eric.granger@etsmtl.ca

Abstract. Video surveillance systems must process and manage a growing amount of data captured over a network of cameras for various recognition tasks. In order to limit human labour and error, this paper presents a spatial-temporal fusion approach to accurately combine information from Region of Interest (RoI) batches captured in a multi-camera surveillance scenario. In this paper, feature-level and score-level approaches are proposed for spatial-temporal fusion of information to combine information over frames, in a framework based on ensembles of GMM-UBM (Universal Background Models). At the feature-level, features in a batch of multiple frames are combined and fed to the ensemble, whereas at the score-level the outcome of ensemble for individual frames are combined. Results indicate that feature-level fusion provides higher level of accuracy in a very efficient way.

1 Introduction

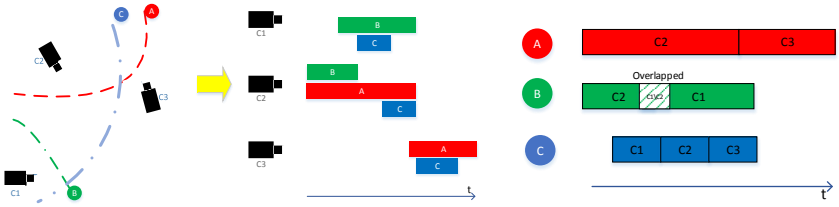
Video surveillance applications, such as activity recognition, are increasingly making use of multiple sensors and modalities. The fusion of multiple diverse sources of information is expected to benefit the system for the recognition of objects, persons, activities and events captured in an array of cameras.

Networks of video cameras are commonly employed to monitor large areas for a variety of applications. A central issue in such networks is the tracking and recognition of individuals of interests across multiple cameras. These individuals must be recognized when leaving the Field of View (FoV) of one camera and re-identified when entering the FoV of another camera. Systems for video-to-video recognition are typically employed for person re-identification (PR). In a FoV, the appearance of an individual may be captured in reference RoIs and representative models may be learned from RoI trajectories. Then, the probe RoI may be matched against the reference model in either live (real-time monitoring) or archived (post-event analysis)[1]. In this paper, we address a PR system over wide network of cameras where no target individual enrolled to the system in advance.

In such environments, where objects move and cross in the FoV of multiple cameras, it is likely to have multiple streams, recorded at different starting points with various lengths, for the same RoI of individuals (see Fig. 1a). The surveillance system must track that person across all cameras whose FoV overlap the person's path.

The final publication is available at link.springer.com http://doi.org/10.1007/978-3-319-27863-6_47

Thus, a suitable outcome for this system could be a time-line graph assigning streams from each camera to an identity for the indicated presence period, as illustrated in Figure 1b. Environmental challenges such as the variation in appearance of individuals due



(a) Typical surveillance scenario (b) An example of time-line graph

Fig. 1: A surveillance scenario including three persons A, B, and C, moving in the scene, crossing the FoV of 3 cameras: c_1, c_2, and c_3 .

to changes in illumination, contrast, positioning of acquisition devices, motion blur as well as occlusion lead to noisy and/or partial RoI captures. These challenges have previously been addressed by a batch divisive strategy in [2], that views a batch of RoI as a unique element to classify, since learning from these batches may reduce noise and fill the gaps caused by dropped-tracks. A batch includes a fixed number of consecutive RoIs (sources of information) of a given stream and a single label is assigned to the batch of same person in time. Each batch of RoI can be learnt using a one-class classifier, and the pool of classifiers generated in one or more FoVs can be combined into an ensemble of classifiers. Fusion of multiple sources into an ensemble have been addressed by three main approaches in the literature: early, mid-level, and late [3]. Early fusion combines the information in the first possible level (so called signal level fusion in image processing), whereas late fusion combines the information as late as possible (decision level fusion) [4,5]. Mid-level fusion is an interesting compromise that combines the information in an intermediate abstraction level [6].

Score-level fusion is the most popular way of fusion. A quantitative similarity measure disseminates valuable information about the input, and yet it is still easy to process compared to sensor-level or feature-level data. However the score space is subject to considerable flexibilities, e.g different normalization methods may lead to different decision boundaries. Furthermore, small number of scores in a batch might easily overfit the data [7]. On the other hand, feature-level fusion schemes derive the most abstract form of original multiple feature set by eliminating redundant information. The advantages of this scheme are the use of only one learning stage to combine the information (instead of running individual learning stage for every single feature set) for rapid decisions.

In this paper, two feature-level abstraction schemes that represent the entire batch with a single descriptor are proposed. These descriptors are obtained by combining features of individual frames in different ways. To the best of our knowledge, this work is the first attempt to explore spatial-temporal fusion schemes for RoI batches captured

from video streams generated in a multi-camera surveillance scenario. We compare their performances with two score-level fusion schemes.

Next section 2 provides an overview of the framework. Section 3 briefly reviews the employment of fusion schemes and introduces algorithms. Section 4 discusses the experimental methodology. In Section 5, we experimentally compare the effectiveness of different levels of fusion on several real-world videos.

2 Background on The NEVIL.ubm Approach

A surveillance system should track and recognize the object from the first moment it is captured by a camera and across all cameras whose fields of view overlap the path. In this section, the Never Ending Visual Information Learning with UBM (NEVIL.ubm) framework is briefly presented. NEVIL.ubm [8] is designed for learning in non-stationary environments in which no labelled data is available but the learning algorithm is able to interactively query the user to label the desired outputs at carefully chosen data points.

The system receives multiple visual streams, generated by a typical tracking algorithm, which analyses sequential video frames and tracks RoIs over time. For each RoI the features corresponding to some pre-selected object representation (e.g. bag of words) are extracted ($\upsilon[l] \;\; l = 1,...,B$). A batch $\upsilon_t^{m_t}$ is a temporal sequence of frames $\upsilon_{t,f}^{m_t}$, where f runs over 1 to the batch size B. Initially, the composite model is initial-

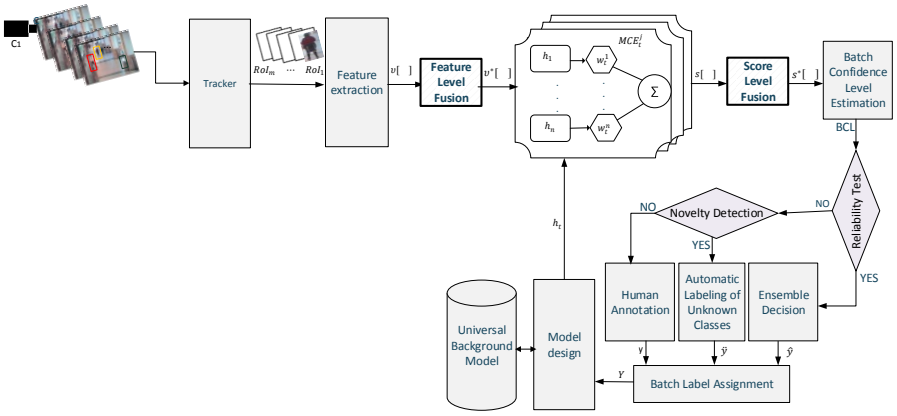


Fig. 2: Block diagram of NEVIL.ubm. (The diagram shows both possible level of fusion applied in the framework)

ized to yield the same probability to every class (uniform prior). When the features of batches of RoIs $\upsilon_{t,f}^{m_i}$ in time slot t become available, the framework starts computing the scores $\mathscr{S}(\upsilon_t^{m_i}|C_k, H_{t-1})$ for each batch $\upsilon_t^{m_i}$ in the time slot. The scores are obtained from the likelihood ratio test of the batch data obtained by the individual class model C_k and the UBM.

The composite model H_t is an ensemble of Micro-classifiers ensembles ($MCE_t^j, j = 1,...,k$). Each MCE_t^j includes classifiers that are incrementally trained (with no access

to previous data) on incoming batches of j_{th} class at t, h_t^j. The individual models h_t^j are combined using a weighted majority voting, where the weights are dynamically updated with respect to the classifiers' time of design.

The prediction output by the composite model MCE_t^j for a given ROI ($\upsilon_{t,f}^{m_i}$) is

$$p(C_k|\upsilon_{t,f}^{m_i}, MCE_t^j) = \sum_{\ell=1}^{t} W_\ell^t h_\ell(C_K|\upsilon_{t,f}^{m_i}) \quad (1)$$

where $h_\ell^j(.)$ is the classifier trained from batches of j_{th} at TS ℓ, W_ℓ^t is the weight assigned to classifier ℓ, adjusted for time t. The weights are updated and normalised at each time slot and chosen to give more credit to more recent knowledge. After combining the decisions of classifiers inside every MC-ensemble, the ensemble will assign a batch to the label of MC-ensemble with highest score ($\mathscr{S}(\upsilon_t^{m_i}|C_k, H_{t-1})$).

Such on-line learning may suffer if labelling errors accumulate, which is inevitable. To help mitigate this issue, the system is designed to interact wisely with a human. Once $\mathscr{S}(\upsilon_t^{m_i}|C_k, H_{t-1})$ is obtained, a batch confidence level (BCL) is estimated. In NEVIL.ubm framework, if the scores associated to all observed classes are significantly low (below a predetermined threshold), it is very likely that this class has not been observed before and it is considered novel and a new label ($\ddot{y}$) is automatically assigned to this batch(es). Having decided that the batch data belongs to an existing class, one needs to decide if the automatic prediction is reliable (the reliability test is positive) and accepted or rather a manual labelling needs to be requested. If BCL is high enough (above a predefined threshold), the predicted label

$$\hat{y} = \arg\max_{C_k} \mathscr{S}(\upsilon_t^{m_i}|C_k, H_{t-1}) \quad (2)$$

is accepted as correct; otherwise the user is requested to label (y) the data batch.

At each time slot, the batches predicted to belong to the same class are used to generate the class model by *tuning the UBM parameters* in a maximum *a posteriori* (MAP) sense. The adaptation process consists in two main estimation steps. First, for each component of the UBM, a set of sufficient statistics is computed from a set of M class specific feature vectors. Each UBM component is then adapted using the newly computed sufficient statistics, and considering diagonal covariance matrices.

Note that the UBM is trained offline, before the deployment of the system. It is designed from a large pool of streams aimed to be representative of the complete set of potentially observable 'objects'.

3 Spatial-Temporal Fusion Schemes Over Frames

Although in many real visual applications different sources of information are available, learning from multiple sources is a less explored area. Early fusion has been applied to define whether the audio signal is consistent with the speaker video file [9]. Pixel-level fusion has shown promising performance in video-based biometric recognition [10] as well as multiple object tracking [11]. Some authors demonstrate [12,13] demonstrated the effectiveness of the decision level fusion strategies on object tracking, video segmentation, and video event detection. Feature-level fusion has gained much importance over the past few years, and various approaches have been introduced in the literature [5,14,15]. Most approaches combined the information of multiple modalities (sensors), while some methods used the complementary descriptors. The former requires

multiple sensors (visible light cameras combined with depth or infra-red camera), and the latter adds more complexity to the system specially in an online application. To the best of our knowledge, the employment of feature-level techniques over frames in a PR scenario has not been addressed before.

Fusion schemes have been successfully used in large-scale recognition systems to address multiple issues confronting these systems such as accuracy, practicality, and efficiency. Inspired by the rationale behind such systems, two fusion schemes to combine the information in a PR system are proposed. Each frame can be considered as an independent source of information and combining such information in different levels could be beneficial for a PR system. The batch score ($\mathscr{S}(\upsilon_t^{m_i}|C_k,H_{t-1})$) can be obtained in two ways: either by combining the scores of individual RoIs in a batch (score-level fusion), or by combining the patterns of M RoIs in a batch (feature-level fusion).

3.1 Feature-Level Fusion

Finding a joint representation for a group of frames is a challenging problem in visual applications. There is a considerable body of research works that addressed this problem by choosing a key frame, which represents the entire batch. As the quality of the batch representation relies heavily on the representative sample and an inappropriate choice may lead to unreliable results, such methods seem impractical for challenging environments. This is the main rationale behind approaches exploiting fusion schemes. In this paper, two feature-level fusion that aggregate descriptors of all the frames in a given batch are proposed. Let $\upsilon_{t,f}^{m_t}$ be the descriptor of $f-th$ frame in a batch, the

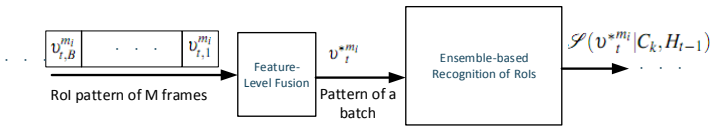


Fig. 3: Block diagram of feature-level fusion (performed before micro-ensembles recognition)

average histogram that combines the information of entire batch in a single histogram defined by:

$$\upsilon^{*m_i}_t = \frac{1}{B}\sum_{f=1}^{B} \upsilon^{m_i}_{t,f}(b) \;\; where \;\; b = 1,...,M \tag{3}$$

Where M is the number of histogram bins.

In our scenario, it is very likely to obtain outlier values for some frames in a batch due to occlusion or miss tracking. The median might be seen as a better indication of central tendency than the arithmetic mean in such cases, since it is less susceptible to the exceptionally large or small values in data. Hence, as an alternative option we consider estimating the descriptor of a given batch by:

$$\upsilon^{*m_i}_t = \mathscr{M}edian \;\; \upsilon^{m_i}_{t,f}(b) \;\; where \;\; b = 1,...,M \tag{4}$$

Given the single representation, a score $\mathscr{S}(\upsilon_t^{m_i}|C_k,H_{t-1})$ is calculated for the batch.

3.2 Score-Level Fusion

The composite model, H_{t-1}, can be used to predict directly $p(\upsilon_{t,f}^{m_i}|C_k,H_{t-1})$ but not $p(\upsilon_t^{m_i}|C_k,H_{t-1})$. The individual scores per frame $\mathscr{S}(\upsilon_{t,j}^{m_i}|C_k,H_{t-1})$ can then be immediately obtained as $\mathscr{S}(\upsilon_{t,j}^{m_i}|C_k,H_{t-1}) = \frac{p(\upsilon_{t,j}^{m_i}|C_k,H_{t-1})}{p(\upsilon_{t,j}^{m_i}|UBM)}$. The batch label prediction can be analysed as a problem of combining information from multiple (B) classification decisions. Considering that, per frame, the composite model produces approximations to the likelihoods/scores for each class, different combination rules can be considered to build the batch prediction from the individual frame predictions. Applying arithmetic

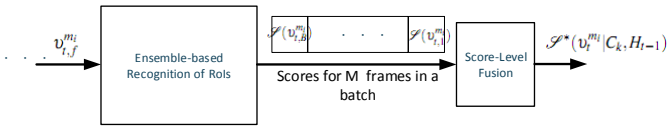


Fig. 4: Block diagram of score-level fusion (performed after micro-ensembles recognition)

mean, the score per batch is obtained as:

$$\mathscr{S}^*(\upsilon_t^{m_i}|C_k,H_{t-1}) = \frac{\sum_{j=1}^{B}\mathscr{S}(\upsilon_{t,j}^{m_i}|C_k,H_{t-1})}{B} \tag{5}$$

As an alternative choice, the median of the scores were also evaluated, since it may be more robust to the outliers. The batch score is defined by:

$$\mathscr{S}^*(\upsilon_t^{m_i}|C_k,H_{t-1}) = \mathscr{M}edian\mathscr{S}(\upsilon_{t,j}^{m_i}|C_k,H_{t-1}) \tag{6}$$

Although other robust statistics could be considered from the individual frame scores, experimentally we will only compare the two options. In the end, NEVIL.ubm assigns each batch to the class maximizing $\mathscr{S}(\upsilon_t^{m_i}|C_k,H_{t-1})$.

4 Experimental Methodology

4.1 Datasets

In order to explore the properties of the proposed framework, we evaluated it on multiple datasets covering various possible scenarios in a multi-camera surveillance system. We conducted our experiments on a number of CAVIAR video clips including: OneLeave ShopReenter1, Enter ExitCrossingPaths1, OneShopOneWait1, OneStop Enter2 and WalkBy Shop1front as well as PETS2009. These sequences present challenging situations with cluttered scenes, high rates of occlusion, different illumination conditions as well as different scales of the person being captured. We employ an automatic tracking approach to track objects in the scene and generate streams of bounding boxes, which define the tracked objects' positions. As the method may fail to perfectly track the targets, a stream often includes frames of distinct objects. A hierarchical bag-of-visterms method is applied to represent the tracked objects, resulting in a descriptor

vector of size 11110 for each frame (refer to [16] for more information). In order to avoid the curse of dimensionality that system may suffer from, PCA is applied to the full set of descriptor features as a pre-processing step. Hence, the number of features in each stream is reduced to 85.

4.2 Confidence Measure

Various criteria have been introduced as uncertainty measures in literature for a probabilistic framework.

Most confident measure (MC): Perhaps the simplest and most commonly used criterion relies on the probability of the most confident class, defining the confidence level as

$$\max_{C_k} \mathscr{S}(C_k|\upsilon_t^{m_i}, H_{t-1}) \tag{7}$$

Modified margin measure (MM): MC only considers information about the most probable label. Thus, it effectively "throws away" information about the remaining label distribution [17].To correct this, an option is to adopt a margin confidence measure based on the first and second most probable class labels under the model. We evaluate experimentally the BCL base on the *ratio* of the first and second most probable class labels:

$$\mathscr{S}(C^*|\upsilon_t^{m_i}, H_{t-1})/\mathscr{S}(C_*|\upsilon_t^{m_i}, H_{t-1}), \tag{8}$$

where C^* and C_* are the first and second most probable class labels, respectively.

4.3 Evaluation Criteria

Active learning aims to achieve high accuracy using as little annotation effort as possible. Thus, a trade-off between accuracy and proportion of labelled data can be considered as one of the most informative measures.

Accuracy In a classical classification problem the disparity between real and predicted labels explains how accurately the system works. However, in our scenario the labels do not carry any semantic meaning (it is not a person recognition problem). The same person should have the same label in different batches, whichever the label. As such, when evaluating the performance of our framework we are just comparing the partition of the set of batches as defined by the reference labelling with the partition obtained by the NEVIL labelling. We adopted a generic partition-distance method for assessing set partitions, initially proposed for assessing spatial segmentations of images and videos [18]. Thus, the accuracy of the system is formulated as:

$$Accuracy = \frac{N - Cost}{N} \tag{9}$$

where N denotes the total number of batches, and $Cost$ refers to the cost, yielded by the assignment problem.

Annotation Assume MLB and TB denote the manually labelled batches and all the batches available during a period (includes one or more time slots), respectively. The *Annotation Effort* is formulated as:

$$Annotation\ effort = \frac{\#MLB}{\#TB} \tag{10}$$

It is expected that the accuracy increases with the increase of the annotation effort.

Confidence Measure	Combination Rule	Reenter1	Reenter2	front	Paths1	Enter2	Wait1	Enter1	PETS09
MC	Median	0.96(1)	0.96(2)	0.91(4)	0.87(2)	0.96(1)	0.90(4)	0.90(1)	0.85(2)
	Mean	0.94(3)	**0.97(1)**	**0.96(1)**	**0.88(1)**	**0.96(1)**	**0.91(3)**	**0.90(1)**	**0.86(1)**
MM	Median	0.95(2)	0.94(3)	0.93(3)	0.86(3)	0.96(1)	0.93(1)	0.88(2)	0.79(3)
	Mean	0.96(1)	0.96(2)	0.95(2)	0.87(2)	0.96(1)	0.92(2)	0.90(1)	0.73(4)

Table 1: ALC of fusion at feature-level on videos. The rank of each setting in a given dataset is presented next to the ALC between parentheses. Highlighted row indicates the optimal design. Values in bold indicate better performance than score-level fusion for optimal setting.

Area under the learning curve (ALC) [19] is a standard metric in active learning research that combines *accuracy* and *annotation effort* into a single measurement. ALC, which provides an average of accuracy over various budget levels, seems to be a more informative metric. Herein, the learning curve is the set of accuracy plotted as a function of their respective annotation effort, a, $Accuracy = f(a)$. The ALC is obtained by:

$$ALC = \int_0^1 f(a)da \tag{11}$$

5 Results

Table 1 shows the ALC performance of the proposed fusion techniques using all datasets along with the mean of ALC rank averaged over all the experiments (the std of the results is always below ± 0.01). The table shows that settings in where sum rule have been applied for combining the information occupy the two top spots for both feature-level and score-level fusion. The results indicate that the most confident class as batch confidence measure selects more informative batches than modified margin, as settings employing the former have better mean rank. Based on the average rank, we conclude that the arithmetic mean as fusion rule and the most confident as selection criterion presents the optimal design. Comparing the ALC of identical designs of two fusion schemes (highlighted rows in table 1 and 2) for every dataset, we observe that for 6 out of 8 datasets feature-level fusion attains better performance (higher ALC) than score-level fusion.

Figure 5 presents the results of optimal design (arithmetic mean as fusion rule and the most confident as selection criteria) for two fusion levels on all video clips. Since ALC measures the average performance over various budget levels, it does not give detailed information for every single budget level. We chose the point obtained by labelling 20% of batches for a more detailed analysis. Given that budget while employing mid-level fusion, we obtain 100% accuracy for four scenarios (OneLeaveShopReenter2, OneLeaveShopReenter1, OneStopEnter2, and WalkByShop1front). For more complex scenarios, such as OneStopMoveEnter1 (in where 42 streams from 14 classes are available) 88% of batches are correctly classified, showing an improvement over score-level

		Datasets							
Confidence Measure	Combination Rule	Reenter1	Reenter2	front	Paths1	Enter2	Wait1	Enter1	PETS09
MC	Median	0.96(1)	0.93(3)	0.93(2)	0.86(2)	0.95(2)	0.88(4)	0.87(3)	0.79(2)
	Mean	**0.96(1)**	**0.97(1)**	0.90(3)	0.87(1)	0.95(2)	0.90(3)	0.90(1)	0.85(1)
MM	Median	0.96(1)	0.91(4)	0.95(1)	0.87(1)	0.93(3)	0.91(2)	0.87(3)	0.71(4)
	Mean	0.96(1)	0.95(2)	0.93(2)	0.85(3)	0.96(1)	0.92(1)	0.89(2)	0.75(3)

Table 2: ALC of fusion at score-level on videos. The rank of each setting in a given dataset is presented next to the ALC between parentheses. Highlighted row indicates the optimal design. Values in bold indicate better performance than score-level fusion for optimal setting.

fusion results (80% accuracy). The results indicate the better performance of feature-level over score-level fusion.

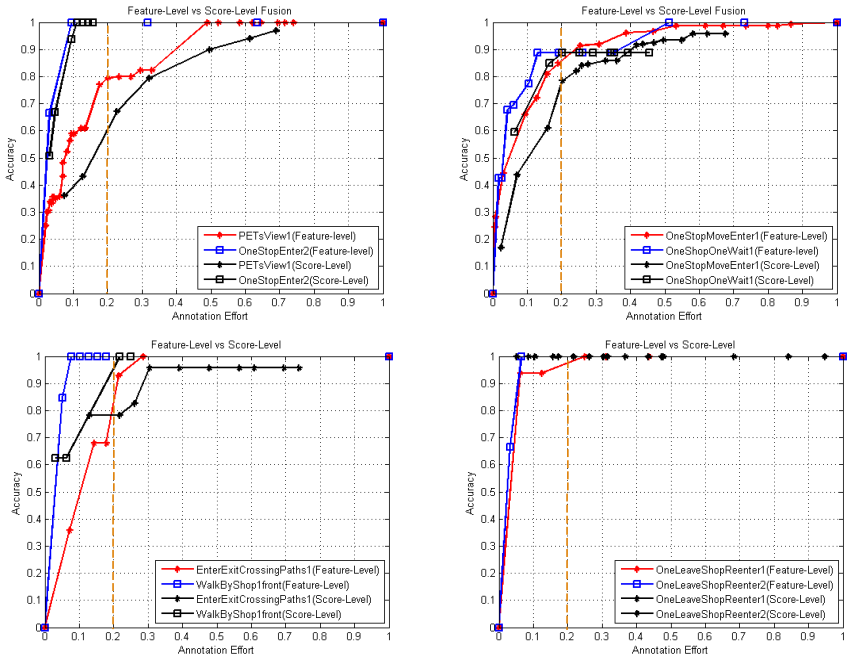


Fig. 5: ALC vs annotation effort for feature-level with score-level fusion on the various videos. – – – highlights 20% budget.

Complexity Improving the accuracy is not the only advantage of feature-level fusion. In real-time learning, when massive amount of information is available, efficiency is equally important. In contrary to score-level fusion, where an independent recognition process is applied to every single RoI (of M RoI in a batch) and then the results are mathematically combined, feature-level fusion employs a single learning stage on the joint representation of a batch of M frames. Thus, the time and complexity of the framework decrease dramatically. Since the framework was developed in MATLAB without any efficiency concerns, a straightforward assessment of the time efficiency is not adequate. Nevertheless our experiments shows that combining the information at feature-level is able to process the streams almost twice as fast as score-level fusion, for a framerate of 25fps (running in an Intel Core i7 at 3.2GHz).

6 Conclusions

In this paper, two spatio-temporal fusion strategies to combine the patterns of RoIs in various streams captured in a multi-camera surveillance system are presented. We experimentally investigated the impact of feature-level and score-level fusion on the performance of the PR system. Experiments indicate the potential of feature-level fusion for on-line applications, as they attained the best performance with much lower

time complexity. For future work, we plan to exploit descriptors that are specifically designed to represent video shots.

References

1. Dewan, M.A.A., Granger, E., Marcialis, G.L., Sabourin, R., Roli, F.: Adaptive appearance model tracking for still-to-video face recognition. Pattern Recognition (2015)
2. Khoshrou, S., Cardoso, J.S., F.Teixeira, L.: Active learning of video streams in a multi-camera scenario. In: 22nd International Conference on Pattern Recognition. (2014)
3. Dietrich, C., Palm, G., Schwenker, F.: Decision templates for the classification of bioacoustic time series. Information Fusion **4** (2003) 101–109
4. Jiang, B., Martínez, B., Valstar, M.F., Pantic, M.: Decision level fusion of domain specific regions for facial action recognition. In: 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014. (2014) 1776–1781
5. Abouelenien, M., Wan, Y., Saudagar, A.: Feature and decision level fusion for action recognition. (2012) 1–7
6. Schels, M., Glodek, M., Meudt, S., Scherer, S., Schmidt, M., Layher, G., Tschechne, S., Brosch, T., Hrabal, D., Walter, S., Traue, H.C., Palm, G., Schwenker, F., Rojc, M., Campbell, N.: Multi-Modal Classifier-Fusion for the Recognition of Emotions. In: Converbal Synchrony in Human-Machine Interaction. CRC Press (2013) 73–97
7. Tao, Q., Veldhuis, R.: Hybrid fusion for biometrics: Combining score-level and decision-level fusion. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Workshop on Biometrics, Los Alamitos, IEEE Computer Society Press (2008) 1–6
8. Khoshrou, S., Cardoso, J.S., Teixeira, L.F.: Learning from evolving video streams in a multi-camera scenario. Machine Learning **100** (2015) 609–633
9. Fisher, J.W., Darrell, T.: Signal level fusion for multimodal perceptual user interface. In: workshop on Perceptive user interfaces. (2001) 1–7
10. Colores-Vargas, J.M., García-Vázquez, M.S., Ramírez-Acosta, A.A., Pérez-Meana, H., Nakano-Miyatake, M.: Video images fusion to improve iris recognition accuracy in unconstrained environments. In: 5th Mexican Conference on Pattern Recognition. (2013) 114–125
11. Cvejic, N., Nikolov, S., Knowles, H., Loza, A., Achim, A., Bull, D., Canagarajah, C.: The effect of pixel-level fusion on object tracking in multi-sensor surveillance video. In: ICVPR. (2007) 1–7
12. Krishna Mohan, C., Dhananjaya, N., Yegnanarayana, B.: Video shot segmentation using late fusion technique. In: ICMLA. (2008) 267–270
13. Kamishima, Y., Inoue, N., Shinoda, K. EURASIP J. Image and Video Processing (2013)
14. Sharma, V., Davis, J.W.: Feature-level fusion for object segmentation using mutual information. In: Augmented Vision Perception in Infrared. Springer (2009) 295–320
15. Chen, C., Jafari, R., Kehtarnavaz, N.: Improving human action recognition using fusion of depth camera and inertial sensors. IEEE T. Human-Machine Systems **45** (2015) 51–61
16. Teixeira, L.F., Corte-Real, L.: Video object matching across multiple independent views using local descriptors and adaptive learning. Pattern Recognition Letters **30** (2009) 157–167
17. Settles, B.: Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison (2009)
18. Cardoso, J.S., Corte-Real, L.: Toward a generic evaluation of image segmentation. IEEE Transactions on Image Processing **14** (2005) 1773–1782
19. Cawley, G.C.: Baseline methods for active learning. In: Active Learning and Experimental Design@ AISTATS. (2011) 47–57