Bayesian Multimodal Fusion in Forensic Applications*

Virginia Fernandez Arguedas, Qianni Zhang, and Ebroul Izquierdo

Multimedia and Vision Research Group
School of Electronic Engineering and Computer Science
Queen Mary, University of London
Mile End Road, E1 4NS, London, UK
{virginia.fernandez,qianni.zhang,ebroul.izquierdo}@eecs.qmul.ac.uk

Abstract. The public location of CCTV cameras and their connexion with public safety demand high robustness and reliability from surveillance systems. This paper focuses on the development of a multimodal fusion technique which exploits the benefits of a Bayesian inference scheme to enhance surveillance systems' reliability. Additionally, an automatic object classifier is proposed based on the multimodal fusion technique, addressing semantic indexing and classification for forensic applications. The proposed Bayesian-based Multimodal Fusion technique, and particularly, the proposed object classifier are evaluated against two state-of-the-art automatic object classifiers on the i-LIDS surveillance dataset.

1 Introduction

The recent outbreak of vandalism, accidents and criminal activities, has affected the general public's concern about security. Nowadays, higher safety levels and new security measures are demanded. Monitoring private areas (i.e. shopping malls) and public environments prone to vandalism (i.e. bus stations) has become a crucial task generating a great growth of deployed surveillance systems. The enormous amount of information recorded daily for monitoring purposes is typically controlled by surveillance operators and removed some time later due to storage space limitation. Besides, the lack of pre-processing of the video data increases the complexity of the forensic search and restrains the evolution towards autonomous surveillance systems.

Existing limitations of surveillance video systems demand the development of automatic and smart surveillance solutions to detect and classify objects and events. Despite the huge interest in real-time surveillance applications, accurate object indexing and classification techniques are demanded to improve post-investigations and tackle efficient surveillance object and event storage by using semantic indexing/classification.

Over the past several decades, many different approaches have been proposed to automatically represent objects or concepts in videos. Numerous features

* The research was partially supported by the European Commission under contract FP7-SEC 261743 VideoSense.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 466–475, 2012.
© Springer-Verlag Berlin Heidelberg 2012

analysing visual appearance, motion, shape or temporal evolution have been proposed and selected depending on their performance for diverse applications [1]. Single features or inputs are capable of obtaining high accuracy results and tackle specific problems, i.e. object detection. However, the use of complementary information increases the accuracy of the overall decision making process and enhances the possibilities and capabilities of different systems to perform more sophisticated tasks, i.e. object classification, speaker identification, etc. Multimodal fusion research was motivated for the exploitation of complementary resources/inputs to enhance the system performance; gaining much attention in research areas such as machine learning, pattern recognition and multimedia analysis.

Typically surveillance applications are affected by several restrictions such as low quality images or environmental factors. The public location of surveillance cameras, usually in un-controlled environments, affects not only the quality of the images but also the availability of the image itself. Such constraints limit the range of multimodal fusion techniques, demanding a method capable of dealing with lack of information as well as the presence of uncertainty. The close relationship between surveillance applications and safety demands high robustness and their continuous-working mode.

Two contributions are presented in this paper. Firstly, a Bayesian inference scheme able to fuse several diverse-nature cues is presented, providing a multimodal fusion technique capable of handling the absence of information and the presence of uncertainty. Secondly, a surveillance object classifier exploiting the benefits of the proposed Bayesian multimodal fusion approach is proposed based on the analysis of visual and temporal information.

The remainder of this paper is organised as follows. In Section 2, an exhaustive study of the existing multimodal fusion techniques and their impact on surveillance applications is presented. The proposed Bayesian-based multimodal fusion technique is further detailed in Section 3, while Section 4 presents the proposed surveillance object classifier developed based on the proposed Bayesian inference scheme to enhance the classifier performance in situations of absent information. Experimental results and the performance evaluation of the proposed Bayesian-based object classifier against state-of-the-art object classifiers are presented in Section 5. Whilst, Section 6 draws conclusions and presents the potential future work.

2 Literature Review

The variety of media, features or partial decisions provide a wide range of options to address specific tasks. However, the different characteristics of the modalities involved in any analysis hinder the combination for several reasons including, (i) the particular format acquisition of different media, (ii) the confidence level associated to each data depending on the task under analysis, (iii) the independent protection of each type of data and (iv) the different processing times related to the different type of media streams. Multimodal fusion techniques can

be performed at different levels, tackling such constraints from different angles, distinguishing mainly two, feature and decision level [2].

Feature-level multimodal fusion includes all the approaches which combine the available input data before performing the objective task. In this case, the number of features extracted from different modalities must be combined in a unique vector (output) which will be considered as a unique input by the objective task. The main advantages of the feature-level multimodal fusion techniques consist of the need for a unique learning phase for the combined feature vector and the possibility to take advantage of the correlation between multiple features from different modalities [3]. On the other hand, feature-level multimodal fusion presents several disadvantages (i) the difficulty to learn cross-correlation amongst features increases with the number of different media considered, (ii) the feature format should be the same before their fusion and (iii) the synchronisation between features is more complex due to their different modalities [4]. Moreover, Zhang et Izquierdo demonstrated the fundamental need of considering the different nature of the features prior to their fusion, admitting that features "existing" in different feature spaces could not be combined in a linear manner without further consideration [5].

Decision-level multimodal fusion proposes to individually analyse each input, providing local decisions. Those decisions are then combined using a fusion unit to make a fused decision vector that is analysed to obtain a final decision, considering such decisions as the output of the fusion technique. Unlike feature level fusion techniques, decision level multimodal fusion techniques benefit from unique representation despite the use of the multiple media modalities; easing their fusion, the scalability of the system and enabling the use of different and the most suitable techniques to obtain partial solutions. However, the acquisition of partial solutions prevents the considering of the features correlation and is affected by the individual learning process associated to each feature.

In the last few years, multimedia researchers have developed numerous multimodal fusion techniques to perform various multimedia analysis tasks. Some of the most well-known fusion techniques include (i) linear weighted fusion [6,7], (ii) Support Vector Machines (SVM) [8,9], (iii) Bayesian inference [10,11], (iv) Dempster-Shafer theory [12] and (v) Neural Networks [13,14]. In surveillance, automatic object classification is an active research field due to the dependence of the event detection and classification techniques prior to this step.

Typically, surveillance object classifiers are based on binary decisions. For instance, in [15], the authors compute high dimensional features based on edges and use *SVM* to detect human regions. While, Paisitkriangkrai et al. [16] propose a pedestrian detection algorithm based on local feature extraction and SVM classifiers. Within binary classification, several vehicle classification techniques used SVMs to map the detected objects into different categories [17,18]. The former [17] proposes a vehicle classifier where the features to model each object are extracted using Independent Component Analysis while SVM is used to categorise each vehicle into a semantic class. While the latter [18] classifies vehicles in night time traffic using SVMs over their eigenspaces. However, several approaches

propose object classification using a probabilistic framework based on Bayesian Networks, Neural Networks or Hidden Markov Models (HMM) [19,20]. In [19], a multi-class vehicle classification system is presented based on the analysis of rear-side view images. Authors classify the vehicles into four classes including Sedan, Pickup truck, SUV/Minivan and unknown, extracting a set of features to build a feature vector which later is processed by a Hybrid Dynamic Bayesian Network. Zhang et al. addressed the problem of automatic object classification for surveillance videos focusing on traffic monitoring [21]. Their approach consists on the application of Adaboost for feature selection and classification formulation. The classification stage consists of the weighted combination of several weak classifiers. Additionally, generative models like HMM and Graph Models have also been used for object recognition. For instance, in [20], a new detection and recognition method for moving objects is proposed. The authors apply temporal difference for moving object detection, use the discrete wavelet transformation technique to extract the feature vectors and conduct object recognition using HMMs. Finally, in [22], the authors present an empirical performance comparison between several classifiers, such as SVM, Bayesian Network Classifiers or Decision Trees, based on a feature vector built with smoothed discrete cosine transform (DCT) features, 2D moment-based features, horizontal and vertical projection and morphological features, to classify objects in real-world video surveillance scenes.

3 Bayesian-Based Multimodal Fusion

The all-pervasive presence of CCTV cameras, their location in public uncontrolled areas and the strict relationship with safety and security demand a high reliability and continuous work of any surveillance application despite the absence of limited information. Consequently, in this paper, a Bayesian-based multimodal fusion technique is proposed to probabilistically combine diverse-nature cues while addressing the absence of information and the presence of uncertainty by the means of inferring information from previously acquired knowledge.

Bayesian Networks enable the robust integration and combination of multiple diverse-nature sources of information applying rules of probability theory. Fusion techniques based on Bayesian Networks benefit from three fundamental advantages. First, the Bayesian inference method allows the combination of multimodal information due to its possibility of adaptation as the information evolves as well as its capability to apply subjective or estimated probabilities when empirical data is absent [2]. Secondly, the hierarchical structure provides flexibility and scalability, facilitating not only the inclusion of additional information, but also enabling the degradation of the a-posteriori probability in case of the absence of a certain cue/s. Finally, Bayesian Networks allow domain knowledge to be embedded in the structure and parameters of the networks, allowing the adjustment of the fusion technique to the domain and scenario's requirements.

The proposed Bayesian-based multimodal fusion technique provides a probabilistic framework capable of combining multimodal cues at the decision-level, unifying the output of several modules to provide a unique output in the

decision-making process. In addition to the advantages provided by the Bayesian Networks, the proposed multimodal fusion technique benefits from (i) the normalised and unique representation of the information despite the multiple media modalities considered within the analysis and (ii) the combination of different nature features considering their own feature space and unique metrics.

The topology applied in the proposed Bayesian-based multimodal fusion technique is shown in Figure 1. The multimodal cues to combine are independent and can be derived from different inputs, i.e. video, metadata, sound. Considering the decision-level fusion as a classification problem, the Bayesian inference scheme can be formulated using the maximum a-posteriori criterion (MAP):

$$D = \underset{i}{argmax}\{P(C_i|F_1, F_2, ..., F_L)\} = \underset{i}{argmax}\{\prod_{j=1}^{L} P(F_j|C_i)P(C_i)\} =$$

$$= \underset{i}{argmax} \begin{pmatrix} \prod_{j=1}^{L} P(F_j|C_1)P(C_1) \\ \prod_{j=1}^{L} P(F_j|C_2)P(C_2) \\ \vdots \\ \prod_{j=1}^{L} P(F_j|C_N)P(C_N) \end{pmatrix}$$

where $P(C_i|F_1, F_2, ..., F_L)$ defines the probability of a concept C_i to be the final decision undertaken by the classifier, D, considering all the individual partial decisions provided by individual classifiers; F_j are the individual classifiers that provide partial decisions to the Bayesian inference scheme; $P(C_i)$ represent the a-priori probability of the i concept; L defines the amount of partial decisions incorporated in the multimodal fusion and N represents the number of concepts involved in the classification problem.

Regarding the conditional probability matrices connecting each partial decision to the network, shown in Figure 1, Bayesian Networks allow specification according to the scenario and application. Consequently, the relationships among the analysed cues can be set manually or learned from training data.

4 Bayesian-Based Object Classifier for Forensic Applications

In order to evaluate the proposed Bayesian-based Multimodal Fusion technique, a surveillance object classifier framework based on the proposed Bayesian inference scheme is presented. The scalable hierarchical structure of the Bayesian-based Multimodal Fusion technique allows the incorporation of various classifiers, enabling a high level of flexibility and adaptation to the scenario under analysis in the form of a-priori probabilities.

Two baseline classifiers provide partial decisions to the proposed decision-level multimodal fusion for the classification of moving objects detected in outdoor surveillance videos monitoring urban scenarios. First, a set of visual features are extracted and combined using a feature-level multimodal fusion technique which preserves the non-linearity of the different feature spaces, as detailed in

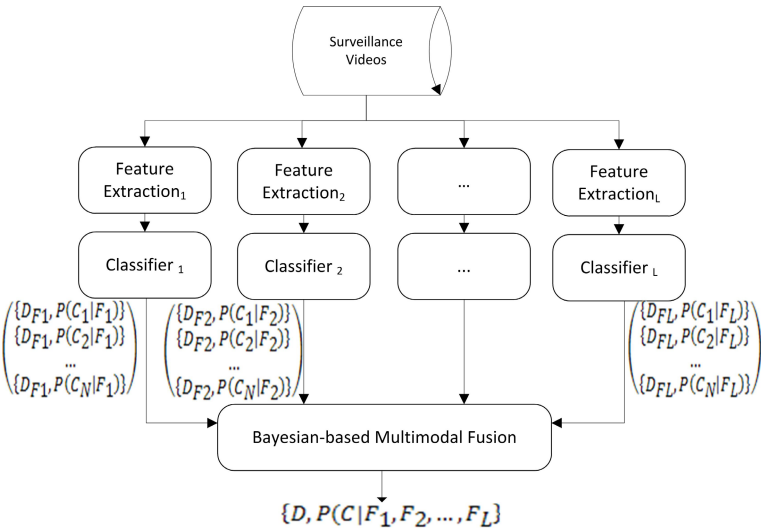

Fig. 1. Bayesian-based object classifier framework

[23]. Secondly, a set of features describing the object's temporal evolution and behaviour are extracted and combined using a behavioural fuzzy classifier, as detailed in [24]. Each of the partial decisions corresponds to an input to the Bayesian inference scheme and represents a classification decision accompanied by a certainty value on the classification. The Bayesian inference scheme is employed to combine the partial decisions considering also the knowledge acquired from the scenario under analysis.

Each individual classifier provides a partial decision coupled together with a conditional probability matrix describing the probability of a detected moving object, or observation O_k, to belong to each of the semantic concepts, $C_i, i = 1, ..., N$, considered within the classification scenario:

$$\begin{pmatrix} P(C_1|F_1) & P(C_1|F_2) \\ P(C_2|F_1) & P(C_2|F_2) \\ \cdots & \cdots \\ P(C_N|F_1) & P(C_N|F_2) \end{pmatrix}$$

where Fj represents each of the individual classifiers whose decisions are fused applying Bayes' probabilistic rules. Each partial decision could perform automatic object classification. However, the integration of several features, derived from different and uncorrelated media, addresses higher robustness, stability, flexibility and adaptation towards the scenario under analysis.

Bayesian Networks enable the continuous work of the multimodal classifier due to their reliability in the presence of missing evidence, either partially or completely. The Bayesian inference scheme allows the system to classify any observation despite the lack of partial decisions, but rigorously decreases the certainty on the classification accordingly.

The proposed Bayesian-based object classifier presents a semantic classification technique based on the fusion of diverse-nature cues to provide semantic indexing of previously detected moving objects. Consequently, semantic indexing would not only facilitate the forensic search and retrieval but would also adapt the search to human understanding. Ultimately the semantic classification provides a step forward towards semantic event indexing.

5 Experimental Results

This section evaluates the performance of the proposed Bayesian-based object classifier for forensic applications. The surveillance dataset and ground truth are further explained and the quantitative evaluation of the experimental results presented. Finally, a comparison between the individual classifiers and the proposed Bayesian-based object classifier is detailed.

5.1 Dataset and Ground Truth

In order to evaluate the performance of the Bayesian-based Multimodal Fusion technique, the proposed Bayesian-based object classifier has been applied to a variety of outdoor video sequences belonging to the i-LIDS dataset[1], provided by the U.K. Home Office. The video sequences recorded under realistic conditions are provided for different scenarios. Our study focuses on urban environments; therefore, the dataset under analysis is Parked Vehicle Detection. Through a careful examination of the dataset, two semantic object categories were noted to be highly repetitive in the sequences, namely Person and Vehicle. The proposed Bayesian-based Object Classifier categorises each moving object detected within the surveillance video as *person*, *vehicle* or *unknown*.

To study the efficiency of the Bayesian-based Multimodal Fusion technique, a ground truth has been manually annotated. A total of 1567 objects were included, 6% were person while 50% were vehicle. Due to the imposed guidelines for the manual annotation, objects presenting certain constraints such as small blob size, partial occlusion of the object over 50% or multiple objects coexisting in a blob, were annotated as *unknown*. The proposed approach automatically classifies objects according to the partial decisions provided by independent classifiers, analysing diverse-nature cues, and inferring information in the presence of (either partially or completely) missing evidence.

5.2 Quantitative Performance Evaluation

To evaluate the performance of the proposed Bayesian-based object classifier and, ultimately, the performance of the Bayesian-based Multimodal Fusion technique, we assumed a tracking algorithm fed the individual classifiers with detected moving objects or observations. Two individual classifiers based on the extraction of visual and temporal information, respectively, categorise each observation into one of the two semantic concepts defined for the scenario. A conditional probability matrix is calculated by each individual classifier and passed

[1] Imagery Library for Intelligent Detection Systems, i-LIDS. `http://ilids.co.uk`

to the Bayesian inference scheme. The Bayesian-based Multimodal Fusion technique combines the different partial decisions to achieve a unique classification considering diverse-nature cues while preserving their individual feature spaces and metrics. The obtained results are shown in Table 1.

Table 1. Performance evaluation of the proposed Bayesian-based object classifier

Concepts		True Positive	True Negative	False Positive	False negative
Vehicle	Observations	619	31	16	17
	%	97	66	34	3
Person	Observations	31	619	17	16
	%	66	97	3	34

Typically, in surveillance applications, there are two fundamental objectives: (i) to achieve a high true positive rate balanced with a low false negative rate which reveals the capability of the classifier to detect the desired concepts and (ii) to maintain a low false positive rate in order to avoid false alarms within the surveillance application. The results provided by the Bayesian-based object classifier (refer to table 1) reveal, for the semantic concept Vehicle, a high rate of true positive detections coupled with a low false negative rate, 97% and 3%, respectively, while maintaining a moderate rate of false positive detection, 34%. The semantic concept Person presents lower true positive and false negative rate, 66% and 34%, respectively, while scoring a remarkable false positive rate, 3%. The results, both the false positive rates for vehicles and the true positive rate for person, are directly affected by the sparseness of the concept person within the ground truth.

The main objective of this paper was to present a multimodal fusion technique which would allow the integration of various different-nature features independently of which media were they derived from, to benefit from (i) the representability provided by each feature, (ii) their un-correlation in order to cover a bigger spectrum, and (iii) the robustness acquired by the system due to the consideration of multiple partial decisions rather than relying in a single decision. In order to demonstrate the improvement on the performance, the proposed Bayesian-based object classifier is compared with the individual classifiers which provided the partial decisions (refer to table 2).

According to the comparative results shown in Table 2, the proposed Bayesian-based object classifier outperforms both individual classifiers. While independent classifiers, based on visual and temporal features, achieve a true positive rate of 77% and 79% respectively, this is exceeded by the Bayesian object classifier in 20% for the semantic concept Vehicle. However, the improvement undertaken by the proposed fusion approach is smaller, increasing the true positive rate by 2% and 9% for the visual and temporal features classifiers. Similarly, the other rates (true negative, false positive and false negative) present improvements whenever the Bayesian-based object classifier is applied compared to the results provided by the independent classifiers. Finally, detailed analysis reveals that

Table 2. Performance comparison between the proposed classifier and the two intermediate state-of-the-art classifiers

Concepts		True Positive	True Negative	False Positive	False Negative
Vehicle	Visual Features [23](%)	77	64	36	23
	Temporal Features [24](%)	79	57	43	21
	Bayesian (%)	97	66	34	3
Person	Visual Features [23](%)	64	77	23	36
	Temporal Features [24](%)	57	79	21	43
	Bayesian (%)	66	97	3	34

the proposed multimodal fusion enhances the object classification procedure, increasing positive detection while reducing false alarms.

6 Conclusions and Future Work

In this paper, a probabilistic multimodal fusion technique was proposed to integrate diverse-nature cues in surveillance applications. In order to evaluate the fusion technique, a Bayesian-based object classification framework was presented. The main objective was to create a scalable technique which allowed the probabilistic combination of multiple features while preserving their nature. The proposed Bayesian inference scheme addressed the, partial or total, absence of information, by degrading the classification results accordingly. The proposed object classifier combined the decisions provided by two state-of-the-art object classifiers in a probabilistic framework which also considered the scenario a-priori knowledge. The proposed approach outperformed both state-of-the-art classifiers, demonstrating the benefits of combining uncorrelated features to improve the classification results and to enhance the robustness of the classification framework.

Considering the dependence of the event classifiers on the object classification results, in the future, we plan to use the Bayesian-based Multimodal Fusion technique to combine classification results, arisen from various object classifiers, to perform event detection and classification.

References

1. Fernandez Arguedas, V., Zhang, Q., Chandramouli, K., Izquierdo, E.: Vision Based Semantic Analysis of Surveillance Videos. In: Anagnostopoulos, I.E., Bieliková, M., Mylonas, P., Tsapatsoulis, N. (eds.) Semantic Hyper/Multi-media Adaptation. SCI, vol. 418, pp. 83–126. Springer, Heidelberg (2012)
2. Atrey, P., Hossain, M., El Saddik, A., Kankanhalli, M.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems 16, 345–379 (2010)
3. Snoek, C., Worring, M., Smeulders, A.: Early versus late fusion in semantic video analysis. In: ACM Multimedia (2005)

4. Wu, Z., Cai, L., Meng, H.: Multi-level Fusion of Audio and Visual Features for Speaker Identification. In: Zhang, D., Jain, A.K. (eds.) ICB 2005. LNCS, vol. 3832, pp. 493–499. Springer, Heidelberg (2005)
5. Zhang, Q., Izquierdo, E.: Combining low-level features for semantic inference in image retrieval. EURASIP Journal on Advances in Signal Processing 12 (2007)
6. Jaffre, G., Pinquier, J.: Autdio/video fusion: a preprocessing step for multimodal person identification. In: MMUA (2006)
7. Kankanhalli, M., Wang, J., Jain, R.: Experiential sampling in multimedia systems. IEEE Transactions on Multimedia 8, 937–946 (2006)
8. Nirmala, D., Paul, B., Vaidehi, V.: A novel multimodal image fusion method using shift invariant discrete wavelet transform and support vector machines. In: ICRTIT, pp. 932–937 (2011)
9. Arsic, D., Schuller, B., Rigoll, G.: Suspicious behavior detection in public transport by fusion of low-level video descriptors. In: ICME, pp. 2018–2021 (2007)
10. Bahlmann, C., Zhu, Y., Ramesh, V., Pellkofer, M., Koehler, T.: A system for traffic sign detection, tracking, and recognition using color, shape, and motion information. In: IEEE Intelligent Vehicles Symposium, pp. 255–260. IEEE (2005)
11. Meuter, M., Nunn, C., Görmer, S., Müller-Schneiders, S., Kummert, A.: A decision fusion and reasoning module for a traffic sign recognition system. IEEE Transactions on Intelligent Transportation Systems, 1–9 (2011)
12. Klausner, A., Tengg, A., Rinner, B.: Vehicle classification on multi-sensor smart cameras using feature-and decision-fusion. In: ICDSC, pp. 67–74. IEEE (2007)
13. Xiao, J., Wang, X.: Study on traffic flow prediction using rbf neural network. In: ICMLC, vol. 5, pp. 2672–2675 (2004)
14. Ozkurt, C., Camci, F.: Automatic traffic density estimation and vehicle classification for traffic surveillance systems using neural networks. Mathematical and Computational Applications 14, 187 (2010)
15. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, pp. 886–893. IEEE (2005)
16. Paisitkriangkrai, S., Shen, C., Zhang, J.: Performance evaluation of local features in human classification and detection. IET Computer Vision 2, 236–246 (2008)
17. Chen, X., Zhang, C.: Vehicle Classification from Traffic Surveillance Videos at a Finer Granularity. In: Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., Chia, L.-T. (eds.) MMM 2007. LNCS, vol. 4351, pp. 772–781. Springer, Heidelberg (2006)
18. Thi, T., Robert, K., Lu, S., Zhang, J.: Vehicle classification at nighttime using eigenspaces and support vector machine. In: ICISP, vol. 2, pp. 422–426. IEEE (2008)
19. Kafai, M., Bhanu, B.: Dynamic bayesian networks for vehicle classification in video. IEEE Transactions on Industrial Informatics, 1 (2012)
20. Cho, W., Kim, S., Ahn, G.: Detection and recognition of moving objects using the temporal difference method and the hidden markov model. In: CSAE, vol. 4, pp. 119–123 (2011)
21. Zhang, Z., Li, M., Huang, K., Tan, T.: Boosting local feature descriptors for automatic objects classification in traffic scene surveillance. In: ICPR, pp. 1–4 (2008)
22. Gurwicz, Y., Yehezkel, R., Lachover, B.: Multiclass object classification for real-time video surveillance systems. Pattern Recognition Letters (2011)
23. Fernandez Arguedas, V., Zhang, Q., Chandramouli, K., Izquierdo, E.: Multi-feature fusion for surveillance video indexing. In: WIAMIS. IEEE (2011)
24. Fernandez Arguedas, V., Izquierdo, E.: Object classification based on behaviour patterns. In: ICDP (2011)