Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-15855-3

Concept drift adaptation in video surveillance: a systematic review

Vinicius P. M. Goncalves[1] · **Lourival P. Silva**[1] · **Fatima L. S. Nunes**[1] · **João E. Ferreira**[2] · **Luciano V. Araújo**[1]

Received: 22 June 2022 / Revised: 19 February 2023 / Accepted: 15 May 2023

Abstract

The world we live in is dynamic by nature. Frequently, the environment changes in ways we cannot predict. In machine learning, the phenomenon that occurs when a model has its prediction effectiveness degraded due to unforeseen changes is known as concept drift. Applications of smart video surveillance tend to suffer from concept drift due to changes in illumination, weather, and scene structure. This work differs from previous ones as it brings focus to the problem of concept drift from a surveillance video perspective which presents additional challenges compared to other sources of data, such as high dimensionality, spatial and temporal relations between data, and real-time constraints. The approaches and algorithms used to cope with concept drift are compared and discussed. We also present datasets and metrics used to evaluate the effectiveness of the algorithms.

As contributions, we present a new classification of concept drift adaptation methods, delineate the characteristics and limitations of techniques that deal with concept drift, and analyze practical aspects, such as real-time processing and memory constraints. Moreover, we conclude that informed concept drift adaptation methods have been employed 90% less than continuous adaptation ones.

Research directions include using established concept drift detection techniques applied to surveillance video data, exploring datasets for concept drift in surveillance, strategies to deal with the high dimensionality and volume of surveillance video data when adapting existing models, and the creation of frameworks to manage drift adaptation while applying computer vision tasks.

Keywords Concept drift · Change detection · Surveillance · Video · Adaptation

✉ Vinicius P. M. Goncalves
viniciuspires.go@usp.br

[1] School of Arts, Sciences and Humanities, University of São Paulo, Rua Arlindo Béttio 1000, São Paulo, SP 03828-000, Brazil

[2] Institute of Mathematics and Statistics, University of São Paulo, Rua Do Matão 1010, São Paulo, SP 05508-090, Brazil

Published online: 20 June 2023

1 Introduction

Video surveillance systems have become essential tools to assist in the task of monitoring public and private spaces and identifying possible threats, therefore, protecting people and assets. In order to alleviate the burden of the surveillance personnel and to reduce human errors, smart video surveillance approaches aim to automatize tasks that involve the observation of actions, behaviors, or events that present any risk [109]. The advantages of this strategy over conventional video surveillance systems include providing the ability to prevent incidents by analyzing suspicious behaviors, enabling analytical video capabilities that can be used for forensics and content-based retrieval, and identifying objects and actions of interest to monitor [54]. Recent advances in smart video surveillance are being made through the employment of machine learning models.

Traditionally, machine learning models are trained using a static set of data that represents or tries to generalize, all future examples presented to these models later. However, real-world environments are non-stationary, i.e., continually changing and evolving. Therefore, past data tends to, over time, not be able to describe the current context, and prediction models that have been trained using that data have their performance degraded. This phenomenon is described as concept drift [137], and, in recent years, techniques to deal with this event have been developed. One possible approach is to incorporate new information in streams so that, as soon as new data becomes available, it is used to update the model [55, 154]. Another approach is to detect when the drift occurs and then perform adaptation by retraining the model or switching to another one [27, 148].

Dealing with concept drift in predictive models used in surveillance video streams presents a set of challenges that are specific to this kind of application due to factors and constraints such as: (a) ideally, surveillance systems run endlessly, which implies large volumes of data being generated constantly; (b) surveillance cameras, especially those installed in outdoor environments, are in non-controlled environments where illumination and the characteristics of the scenario can change gradually or drastically; (c) although smart surveillance can serve as a post-analysis tool, to enable authorities or security personnel to take action, real-time computation is a concern.

Several comprehensive reviews and surveys of concept drift adaptation have been made in recent years [8, 28, 29, 41, 42, 68, 93]. However, none of them approached the additional challenges that surveillance video data imposes. Noticeably, in [41] the authors categorize the types of concept drifts and present the main drift detection techniques, proposing a general drift adaptation taxonomy. This work is a comprehensive introduction to the problem of concept drift, thus, it does not provide details about any particular machine learning problem. In [8], the focus is placed on exploring feature drift, which occurs when a subset of features becomes irrelevant to the learned concept. In [93], in addition to an extensive literature overview of concept drift, the authors present real-world and synthetic datasets used to evaluate drift detection methods. However, no video dataset was presented, and also no analysis of what features and algorithms are being used to deal with this type of data. In [28] and [42], the authors also present concept drift datasets, but do not present any video dataset. More specifically, in [28] the authors mention some concept drift applications such as forecasting, recommendation systems, and energy demand prediction. A short consideration is made about the challenges that high-dimensional unstructured data presents (e.g., images), but they do not elaborate on how to deal with them. In [42], the authors propose a new unsupervised drift detection algorithm and compare its performance with other popular ones, evaluating it on low-dimensional datasets.

None of these reviews, and to the best of our knowledge no other previous work, focus on surveillance video applications or on how to deal with complex video data in the presence of concept drift, but rather center attention on datasets in which the dimensions, i.e., number of features, are relatively low in comparison with video data.

Dealing with concept drift in surveillance video streams introduces additional complexities. In this context, a fast and continuous flow produces a mutable, large volume of data. Data characteristics are constantly changing due to illumination, weather, complex interactions, and scene changes; especially in non-controlled settings, as is the case in outdoor environments. Moreover, this type of data also presents challenges imposed by the high number of dimensions, scales, and spatio-temporal relations between video frames.

The goal of this systematic review is to comprehensively analyze the state of the art of recent approaches to deal with concept drift in the context of surveillance video streams. This systematic review aims to answer the following research questions:

- What are the existing methods and techniques to deal with concept drift in surveillance videos?
- What feature descriptors and machine learning algorithms are used by such methods and techniques?
- Which methods and techniques can be used in real-time?
- What are the datasets and evaluation metrics used?

The contribution of this work is a comprehensive analysis of aspects involving concept drift adaptation in surveillance contexts, including: (a) concept drift adaptation and the proposal of a new classification to describe the adaptation process; (b) the relation between learning strategies and computer vision tasks; (c) features and machine learning techniques used in surveillance contexts; (d) challenges imposed by real-time processing (e.g., computing capacity, data volume); (e) the datasets employed; and (f) the metrics used to evaluate the effectiveness of the proposed approaches.

In Section 2, we present the theoretical background and introduce concepts related to concept drift adaptation and learning settings. Section 3 describes the method used to conduct the systematic review. In Section 4, concept drift adaptation methods are outlined. In Section 5, the relation between learning settings and computer vision tasks is presented. In Section 6, we present the features and machine learning algorithms employed by the works. In Section 7, the characteristics of the employed real-time approaches are described. In Section 8, we report the datasets and metrics used to evaluate computer vision algorithms in the context of surveillance. In Section 9, we make our considerations about the results. Lastly, the overall conclusion and future research directions are given in Section 10.

2 Theoretical background

In this section, the main concepts around concept drift adaptation in surveillance will be exposed.

2.1 Concept drift

In a machine learning context, learning from examples, or acquiring concept, is what makes feasible the generation of a mathematical model that can make predictions or classifications based on feature points presented to it earlier. Concept drift [137] is a phenomenon that

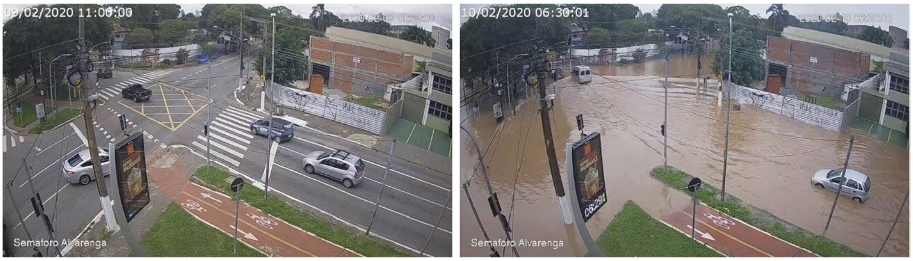


Fig. 1 Example of concept drift in surveillance. The image to the left corresponds to the initial concept. Concerning the image to the right, due to severe rains, most of the area is flooded, which changed the characteristics of the data. Source: Office of Information Technology of USP (STI-USP)

occurs when the context changes in a way that the learned concept does not hold any longer. In other words, the real world presents contexts that are hidden from the model [159].

An example of concept drift in surveillance systems can be illustrated with a model created to detect anomalies in images from a video camera. Suppose that there were no images of rainy or snowy days in the dataset during the training phase. However, at the inference phase, images from outdoor cameras facing diverse weather conditions may cause the model to wrongly classify a rainy scene as an anomaly when it is actually a new context (Fig. 1).

A formal definition of concept drift, as given in [41], between the times t_0 e t_1 is presented in Eq. 1.

$$\exists X : p_{t_0}(X, y) \neq p_{t_1}(X, y) \tag{1}$$

where p_{t_0} represents the joint-distribution at the time t_0, between the set of input variables X and the output variable y.

According to [41], there are two categories of concept drift:

- Real concept drift: occurs when the data distribution $p(y|X)$ changes so that the prediction capacity is affected. This category of drift can happen with or without changes in $p(X)$.
- Virtual concept drift: occurs when the data distribution $p(X)$ changes without altering $p(y|X)$. It is also the case when $p(y|X)$ is not available for inference.

2.1.1 Concept drift detection

Detecting concept drift can be achieved by using change detection algorithms. Usually, the metrics monitored are the classification error or the accuracy returned by a machine learning model. The methods to detect drift can be roughly divided into sequential analysis, statistical process control, and distribution-based [41].

Sequential analysis methods continuously use the most recent observations to evaluate if the mean of the input data significantly deviates from an allowed value. CUSUM [113] and Page-Hinkley [113] are algorithms that belong to this category.

Statistical process control algorithms keep track of the statistical properties of the data distribution. The probability of error is also evaluated by considering a prediction and its true label. This type of method also defines distinct levels of change, such as warning and drift levels. Examples of this type of method are DDM [40] and EDDM [5].

Distribution-based methods aim to compare two data distribution windows (i.e., subsets of the data over time). These distributions include a reference window and a window containing the most recent data. Both windows are then compared using statistical tests to infer whether a change was introduced or not. ADWIN [13] and VFDTc [39] are methods that lie in this category.

2.2 Learning setting

Based on the learning objective, the learning setting can be classified into supervised, unsupervised, and semi-supervised.

Supervised learning It is employed when, for every set of input variables, there is one or multiple known target variables. According to [41], real concept drift can only be verified in a supervised learning setting because it is possible to measure the discrepancy between a prediction and its ground truth.

Unsupervised learning As opposed to supervised learning, in this type of learning setting, the examples do not need to have an annotated target variable, i.e., ground truth. The input variables themselves are used to model the output. Labeling videos is a time-consuming task since, usually, each second of a video produces 25 to 30 frames. Therefore, unsupervised learning proves itself to be an advantageous approach. However, the accuracy of unsupervised algorithms tends to be relatively lower than the supervised ones, given the fact that knowing the target variable beforehand provides a clearer objective [19], [64].

Semi-supervised learning A combination of supervised and unsupervised learning, semi-supervised learning requires the annotation of a subset of the training examples, which reduces the workload associated with the labeling process. An example of a problem that can potentially be addressed with this setting is anomaly detection. In this task, only video segments identified as normal are annotated and used to train a model that learns what normal video clips are like. During inference time, it is possible to tell how much a video clip deviates from the normal and then classify it as normal or abnormal.

2.3 Knowledge acquisition

In order to cope with concept drift, machine learning models need to be updated with new concepts, i.e., newly acquired knowledge must be added to the existing model. Three different strategies of knowledge acquisition are found in the literature: batch, incremental, and active learning.

Batch learning A naive approach to adding new knowledge to a model is to train the model from scratch. This approach is known as batch learning [35] and requires all training examples to be present before the training process starts. The time taken to re-train the whole model is a factor that can make the adaptation process a time-consuming step and, therefore, not ideal for surveillance scenarios where timely actions are needed.

Incremental/Online learning This type of learning allows continuous integration of knowledge into an existing model. It naturally fits within non-stationary environments where the context is constantly changing. In the literature, incremental and online learning are at times defined separately [44], but also used interchangeably [114, 159]. In this review, the terms incremental and online will refer to algorithms that can gradually add new information to an existing model.

Active learning Active learning [138] comprehends strategies to receive annotations given by an oracle (e.g., human) by selecting only the most relevant instances, i.e., the ones classified with the most uncertainty. In this way, the model would benefit from having newly annotated examples that can be used for training and adaptation to new contexts, and the oracle would have to label fewer instances, thus, saving time and effort.

2.4 Computer vision tasks in surveillance

Computer vision [10] involves techniques to analyze and interpret images. Video streams are a rich source of analysis where several computer vision tasks can be performed. In the context of video surveillance, these tasks aim to reveal potential risks to protect people and assets.

Anomaly detection [122, 134, 161] is the task of telling apart abnormal events from normal ones in a dataset. In videos, the detection of anomalies can provide information on where the anomalies are (spatial information, i.e., coordinates inside a frame) and also when the anomalies happen (temporal information) without necessarily indicating their spatial location.

Activity recognition and localization [59, 128, 153] are tasks that learn predetermined activities from the input data. In surveillance videos, examples of activities are car crashes, fire, robbery, violence, and trespassing. This task can be employed when the objective is clear, e.g., for a camera installed on a highway, it is possible to have a model specialized in car crashes.

Image classification [73, 77, 92] can also be used in videos. It aims to process each video frame and then classify them with respect to a target variable. An example of image classification in a surveillance context is gun detection.

Object detection [24, 124, 162] is a task where in addition to knowing what an object is, it is also relevant to find the location of that object in the image. The location of an object is usually given by a set of coordinates that represent a bounding box around that object.

Re-identification (Re-ID) [53, 63, 69] is a task that aims to match objects in different frames and, in this way, re-identifying specific objects. It can be used in biometric systems and also to identify suspects in video surveillance feeds.

Table 1 Inclusion (I) and exclusion (E) criteria adopted in the Systematic Review

Criterion	Description
I1	Studies that address concept drift adaptation in the context of surveillance videos.
I2	Studies that propose solutions to concept drift through techniques such as continuous learning, active learning, online learning, and adaptive learning.
E1	Studies that only perform image processing techniques without employing any machine learning model. i.e., motion detection, and border detection.
E2	Studies that address drift in tracking.
E3	Studies that do not propose any solution for the concept drift issue.
E4	Studies not published or not available in scientific databases or libraries.
E5	Studies not available for the researcher conducting the systematic review.
E6	Studies in which the method is not described.
E7	Studies in which the research goal is not clear.
E8	Studies that do not present evaluation metrics.
E9	Studies published before the year 2015.

3 Research method

The systematic review was conducted based on [67], which defines the phases of planning, conduction, and report.

The search string was defined as: (video* OR camera* OR visual OR feed) AND ("incremental" OR "continual learning" OR "active learning" OR "continuous learning" OR "online learning" OR "on-line learning" OR "adaptive learning" OR "drift" OR "concept shift" OR "dataset shift" OR "covariate shift" OR "non-stationary" OR "nonstationary") AND ("surveillance" OR "analytics" OR "security") AND NOT (tracking OR education OR classroom OR student).

The search was performed in the scientific databases: IEEE Explore,[1] ACM Digital Library[2] and SCOPUS.[3] In order to select the studies, the inclusion and exclusion criteria were defined as presented in Table 1.

To be included, a study must fulfill all the inclusion criteria and cannot fulfill any of the exclusion criteria.

Note that, for the exclusion criterion E1, only machine learning approaches were considered because concept drift itself is a term defined in a machine learning context [137, 159]. As for the exclusion criterion E2, the term drift, in an object tracking context, has a meaning that differs from drift in concept drift. Drift in object tracking, as defined by [156], occurs when the tracking fails because a significant part of the tracked object is no longer in the updated template. Concept drift is more general and does not focus on single objects in an image, but rather on global characteristics that can change and impact a model's performance.

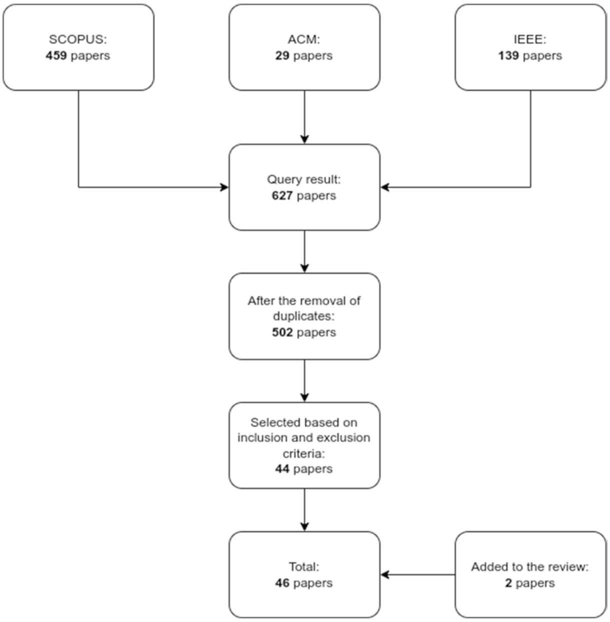
As shown in Fig. 2, a total of 627 papers were found during the searching phase. After removing 125 duplicates, 44 studies were selected based on the inclusion and exclusion criteria defined earlier.

[1] https://ieeexplore.ieee.org/Xplore

[2] https://dl.acm.org

[3] https://www.scopus.com

Fig. 2 Systematic review process



Two studies were included manually due to their relevance to this review. Both studies do not explicitly mention surveillance, but they propose relevant methods to deal with concept drift in videos. The first one is [27], where the authors propose a Convolutional Neural Network (CNN) [78] that can be re-trained only at specific layers that were potentially affected by concept drift. The other one is [81], where the authors propose a Support Vector Machine (SVM) [15] model that can learn incrementally but also forget about irrelevant information. The overall information extracted from the papers is shown in Table 2.

4 Concept drift adaptation

Concept drift adaptation methods can be divided into the ones that can be adapted in a continuous way and the ones that can be adapted in an informed manner. Based on the approaches to deal with drift analyzed in the papers of this review, we propose a new classification of how concept drift adaptation takes place, illustrated in Fig. 3. This classification is different from previous ones [41, 93] as it brings the dimension of active learning and draws the relationship between adaptation types and knowledge acquisition strategies.

Continuous adaptation methods use the newly and continually arriving input data to gradually update the model. In contrast, informed methods keep track of concept drifts, and when one occurs, that information is used to update the model.

Informed adaptation methods can have multiple strategies with respect to how machine learning models are updated once concept drift is detected. To generalize, as shown in Fig. 3, we further divided the adaptation of informed methods into two categories. The first is model selection/ensemble, which means that, in the presence of concept drift, either a new model will be generated or selected from a pool of existing models, or an ensemble strategy will be used with the existing or newly produced models. The second category is

Table 2 General information of the papers included in the systematic review

Author	Adaptation	Adaptation category	Learning setting	Task	Feature	Feature type	Model	Model category	Evaluation metric
Khoshrou et al [71]	Continuous	Incremental with active learning	Supervised	Object detection, Re-ID	Bag of visual terms, PCA	Handcrafted	Ensemble	Ensemble	Accuracy
Lin et al [81]	Continuous	Incremental with passive learning	Semi-supervised	Anomaly detection	HOF	Handcrafted	SVM	SVM	AUC-ROC, EER
Pagano et al [112]	Informed	Ensemble/ Model selection	Supervised	Re-ID	Haar Cascades	Handcrafted	PFAM	Neural network	Precision, Recall, F- measure, AUC- ROC
Alcantara et al [2]	Continuous	Incremental with passive learning	Supervised	Activity recognition	CMS	Handcrafted	kNN	Distance based	Accuracy
Bastani et al [11]	Continuous	Incremental with passive learning	Unsupervised	Anomaly detection	Tracking	Handcrafted	Monte-Carlo	Probabilistic	AUC-ROC, Precision, Recall
Chen et al [20]	Continuous	Incremental with passive learning	Unsupervised	Anomaly detection	3D Gradient, HOF	Handcrafted	Sparse Combination	Sparse Coding	Accuracy
Kharabe and Raghu [70]	Continuous	Incremental with passive learning	Supervised	Re-ID	SIFT	Handcrafted	SVM	SVM	Recognition Rate
Lin et al [80]	Continuous	Incremental with passive learning	Semi- supervised	Anomaly detection	AMHOF	Handcrafted	Weighted Clustering algorithm	Clustering	AUC- ROC, Detection rate
Nguyen [108]	Continuous	Incremental with passive learning	Supervised	Object detection	Haar Cascades	Handcrafted	Decision tree	Decision tree	Precision, Recall, F-measure, PR Curve

Table 2 (continued)

Author	Adaptation	Adaptation category	Learning setting	Task	Feature	Feature type	Model	Model category	Evaluation metric
Bakliwal et al [6]	Continuous	Incremental with active learning	Supervised	Object detection	N/A	Handcrafted	N/A	N/A	AP
Cao et al [17]	Continuous	Incremental with passive learning	Supervised	Re-ID	Eigenvectors	Handcrafted	KISSME	Distance based	Matching rate
Nawaratne et al [106]	Continuous	Incremental with passive learning	Semi- supervised	Anomaly detection	Bounding Box	Handcrafted	Growing Self-Organizing Map (GSOM)	Clustering	Accuracy
Wang et al [155]	Continuous	Incremental with passive learning	Supervised	Re-ID	Conv Filter, MImSF	Handcrafted	Incremental SVM	SVM	Accuracy
Huang et al [61]	Continuous	Incremental with passive learning	Supervised	Re-ID	LOMO	Handcrafted	Null Foley-Sammon Transform (NFST)	Distance based	Recognition Rate
Lv et al [94]	Continuous	Incremental with passive learning	Unsupervised	Re-ID	CNN, Spatiotemporal	Learned, Handcrafted	Bayesian Fusion	Probabilistic	Accuracy
Shin et al [139]	Continuous	Incremental with active learning	Supervised	Object detection	CNN	Learned	CNN	Neural network	AP
Teng et al [149]	Continuous	Incremental with active learning	Supervised	Image classification, Object detection	CNN	Learned	CNN	Neural network	AP
Disabato and Roveri [27]	Informed	Model re-training	Supervised	Image classification	CNN	Learned	CNN	Neural network	Accuracy

Table 2 (continued)

Author	Adaptation	Adaptation category	Learning setting	Task	Feature	Feature type	Model	Model category	Evaluation metric
Torres et al [151]	Continuous	Incremental with passive learning	Supervised	Activity recognition	Optical Flow	Handcrafted	Adaboost	Ensemble	AUCROC
Nallaperuma et al [104]	Informed	Model re-training	Unsupervised	Anomaly detection	Sensors, Social Networks, Videos	Handcrafted	Growing Self-Organizing Map (GSOM)	Clustering	Accuracy
Soomro et al [143]	Continuous	Incremental with passive learning	Supervised	Activity recognition	Superpixel HSI, iDTF, HOG, HOF, MBH, Traj	Handcrafted	SVM	SVM	AUCROC
Sugianto et al [147]	Continuous	Incremental with passive learning	Supervised	Re-ID	CNN	Learned	CNN	Neural network	AP
Ullah et al [154]	Continuous	Incremental with passive learning	Supervised	Activity recognition	CNN	Learned	SVM	SVM	Precision, Recall, Fmeasure, Accuracy
Ali and Bouguila [3]	Continuous	Incremental with passive learning	Supervised	Activity recognition	HOF, MBH	Handcrafted	Hidden Markov models (HMM)	Probabilistic	Accuracy
Campo et al [16]	Continuous	Incremental with passive learning	Unsupervised	Anomaly detection	CNN	Learned	Markov Jump Particle Filter	Probabilistic	Accuracy
Doshi and Yilmaz [30]	Continuous	Incremental with passive learning	Unsupervised	Anomaly detection	Bounding Box, Optical Flow	Handcrafted	kNN	Distance based	AUCROC
Grimmeisen and Theissler [51]	Continuous	Incremental with active learning	Supervised	Image classification	Input Vector	Handcrafted	Random Forest	Decision tree	Accuracy

Table 2 (continued)

Author	Adaptation	Adaptation category	Learning setting	Task	Feature	Feature type	Model	Model category	Evaluation metric
Hasan et al [55]	Continuous	Incremental with active learning	Supervised	Activity recognition	STIP, C3D	Learned, Handcrafted	SVM	SVM	Accuracy
Kumari and Saini [75]	Continuous	Incremental with passive learning	Unsupervised	Anomaly detection	Spatiotemporal	Handcrafted	Gaussian Multivariate Mixture (GMM)	Probabilistic	AUCROC, EER
Li et al [83]	Continuous	Incremental with passive learning	Supervised	Activity recognition	Sensors	Handcrafted	Hoeffding tree (HT), Swarm Decision Table (SDT)	Decision tree	Accuracy, AUCROC
Nawaratne et al [105]	Continuous	Incremental with active learning	Semisupervised	Anomaly detection	CNN	Learned	LSTM	Neural network	AUCROC, EER
Raj et al [121]	Continuous	Incremental with passive learning	Unsupervised	Re-ID	CNN	Learned	kMeans	Clustering	AP
Suprem et al [148]	Informed	Model retraining/selection	Supervised	Object detection	GAN	Learned	Density bands	Clustering	AP
Gonzalez and Prevost [47]	Continuous	Incremental with passive learning	Semisupervised	Image classification	Action Units	Handcrafted	Nearest Class Mean Forest (NCMF)	Decision tree	Accuracy
Joy and Vijayakumar [66]	Continuous	Incremental with passive learning	Supervised	Object detection	CNN	Learned	Fast RCNN	Neural network	mAP, Precision, Recall, Fmeasure
Kim et al [72]	Continuous	Incremental with passive learning	Semisupervised	Anomaly detection	CNN	Learned	CNN	Neural network	Accuracy

Table 2 (continued)

Author	Adaptation	Adaptation category	Learning setting	Task	Feature	Feature type	Model	Model category	Evaluation metric
Lopez-Lopez et al [86]	Continuous	Incremental with passive learning	Unsupervised	Re-ID	CNN	Learned	SVM Ensemble	SVM	Accuracy
Lopez-Lopez et al [87]	Continuous	Incremental with passive learning	Unsupervised	Re-ID	CNN	Learned	SVM Ensemble	SVM	AUCROC
Pillai and Sen [118]	Continuous	Incremental with passive learning	Semi supervised	Anomaly detection	Optical Flow	Handcrafted	Recursive Neural Network	Neural network	AUCROC, EER
Anoopa et al [4]	Continuous	Incremental with active learning	Supervised	Anomaly detection	Optical Flow, CNN	Learned, Handcrafted	LSTM	Neural network	AUCROC, EER
Cao et al [18]	Continuous	Incremental with active learning	Semi supervised	Object detection	CNN	Learned	CNN	Neural network	AP
Doshi and Yilmaz [32]	Continuous	Incremental with passive learning	Semi supervised	Anomaly detection	Bounding Box	Learned	RNN	Neural network	AUCROC, Average Precision Delay
Doshi and Yilmaz [31]	Continuous	Incremental with passive learning	Semi supervised	Anomaly detection	Optical Flow, Poses, CNN, Embeddings	Learned, Handcrafted	NN	Neural network	AUCROC, Average Precision Delay
Kwon and Kim [76]	Continuous	Incremental with passive learning	Supervised	Object detection	CNN	Learned	CNN	Neural network	mAP, Precision, Recall
Nguyen-Meidine et al [107]	Continuous	Incremental with passive learning	Supervised	Object detection	CNN	Learned	CNN	Neural network	mAP
Ren et al [126]	Continuous	Incremental with passive learning	Supervised	Object detection	Haar Cascades, CNN	Learned, Handcrafted	Nearest Mean of Exemplars	Distance based	Precision, Recall, Accuracy

model retraining, which can happen either globally, by replacing the existing model, or locally, by adapting only parts of the current model.

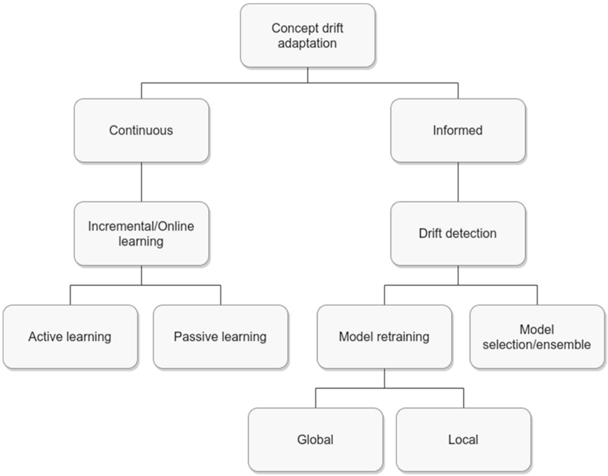
Continuous adaptation methods rely on incremental learning to seamlessly add new instances to the already existing model (Fig. 3). These new instances can be acquired using active learning, which selects only the most informative instances to be labeled by a human oracle. Or without a specific strategy for acquiring knowledge, which can be called passive learning. In Table 2, we present how the studies in this review are classified concerning their adaptation method.

Informed Adaptation From the 46 papers in the review, only 9% employed informed adaptation. In [112], the authors use a drift detection method to train new models when a drift is identified. In [148], the drift is detected by comparing the similarity between known data points and newly added ones. Upon drift detection, an algorithm selects a new model from a pool of existing models or trains a new one from scratch. That new model provides better accuracy than the one previously used. In [104], drift detection is used as an alert mechanism applied to road traffic. It also “forgets” old concepts, a technique called *decremental learning* by the authors, which is able to drop concepts that are no longer relevant. In [27], the authors use an adapted version of CUSUM as the drift detection method. When drift is detected, only the affected layers of the CNN are retrained, while the rest is left untouched.

Continuous Adaptation Continuous adaptation methods correspond to 91% of the analyzed papers. From these works, the combination of incremental and active learning is used in 21% of the papers, whereas incremental learning with passive learning is used in 79% of the studies.

In [149] and [6], the authors propose ways to automatize and simplify the labeling process to make active learning feasible in surveillance. A visual-interactive labeling strategy is proposed by [51], where model-based suggestions and visual cues are combined to ease the labeling process for users. In [55], contextual information is obtained from the newly arriving data to improve the selection of informative instances for posterior human

Fig. 3 Classification of concept drift adaptation



labeling. Thus, once the instances are appropriately labeled, the new examples incrementally are added to the model. Similarly to [51], in [55], the authors use an algorithm to select the most informative examples for user annotation. In [139], the authors use the output of object detection algorithms to select objects detected with low confidence to be double-checked (labeled) by humans. Similarly to [139] in [18], unlabeled images are supplied to an object detection algorithm, then the detections with the smallest confidences considering a threshold are used to retrain the model. The authors call this approach semi-supervised active learning and it is also possible to receive feedback from an oracle.

In [105], an active learning strategy called human-in-the-loop is employed, where human feedback is required whenever an anomaly is detected. If the detection is identified as a false-positive, this information is then used to incrementally update the model, advising that this example should be treated as normal. In [4], although an active learning approach is proposed, there is no specific details of how it is handled.

The remaining studies employ an incremental approach, as seen in Table 2. The algorithms used by each one of the studies are presented in Section 6. The machine learning models employed are capable of incrementally aggregating new unseen information to an existing model.

5 Learning settings and computer vision tasks

Concerning the learning settings used by the studies analyzed in this review, 56% (26 studies) are supervised, 22% (10 studies) are unsupervised, and 22% (10 studies) are semi-supervised (Table 2). The relation between computer vision tasks and learning settings is shown in Fig. 4, where a larger circle represents a greater number of studies using a learning setting to perform the corresponding computer vision task.

It is possible to see that the activity recognition task is employed exclusively in a supervised learning context. Although more commonly applied in a supervised setting, image

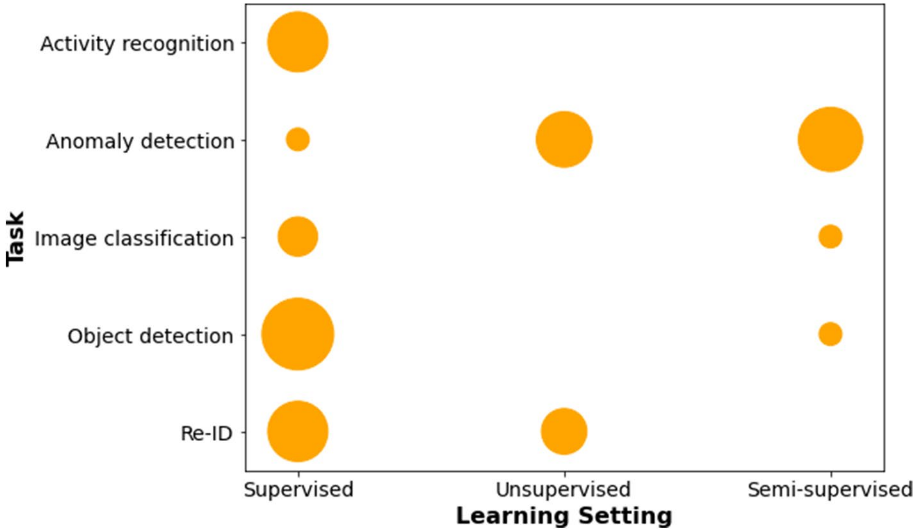


Fig. 4 Computer vision tasks versus learning settings

classification, object detection, and re-ID are also used in settings where not all labels are available. In re-ID tasks, the unsupervised setting can be achieved by measuring similarities between one or more frames.

In object detection, even though some techniques are regarded by the authors as semi-supervised or unsupervised approaches, labeled examples are still needed, so we classified them as supervised approaches. The difference lies in how these labels are acquired, which can be by using another pre-trained object detection model [139, 148] or by employing an algorithm to automatically extract these labels, such as regions that are moving between frames [108].

Among the studies of this review, anomaly detection was only performed in a supervised setup in [4] and was used in a semi-supervised setting roughly as much as it was used in an unsupervised one. Although this task can be performed in a supervised learning setting, it is unusual since knowing all possible anomalies beforehand is not feasible. Also, the definition of anomaly itself is ambiguous. In other words, it is impossible to know beforehand all events that comprehend abnormal and normal activities in every context [30], e.g., a person running in a public square is usually not an anomaly, whereas a person running inside a bank would commonly be.

6 Features and models

Representing videos and images through features is a research topic on its own. Visual features like HOG, Haar Cascades, SIFT, Optical Flow, and 3D spatial gradients have been used extensively for the last few decades [140]. They are considered handcrafted features because they rely on previous knowledge and pre-assumptions about the input data. Recently, CNN models have been used with the reasoning that features can be learned rather than handcrafted, achieving, in this way, better representations than features designed in a manual fashion.

Considering studies that use CNN features as cases where the attributes are learned (or self-learned), analyzing the studies in this review (Table 2), before 2018, there were no papers that employed learned features. In [106], the authors extract bounding boxes using a CNN but do not use the CNN feature vectors. However, after 2018, 67% of the papers (22 out of 33) used learned features.

In recent years, we have also seen approaches that combine handcrafted and learned features to achieve more robust feature representations. Some studies use features provided by other algorithms or systems. For instance, in [30], the authors use the output bounding boxes coordinates from YOLO ([125] along with Optical Flow features.

From Table 2, 52% of the studies use handcrafted features, 37% use self-learned ones, and 11% employ both feature types simultaneously. From the studies using self-learned features, 59% also employ CNNs as the machine learning model, while the other 41% use CNNs as a feature extractor only.

Regarding machine learning models, many distinct algorithms and architectures have been used. Although each one presents its peculiarities, the models were divided into general categories. Figure 5 shows the number of times each category has been used and Fig. 6 shows the distribution of learned and handcrafted feature types among the model categories.

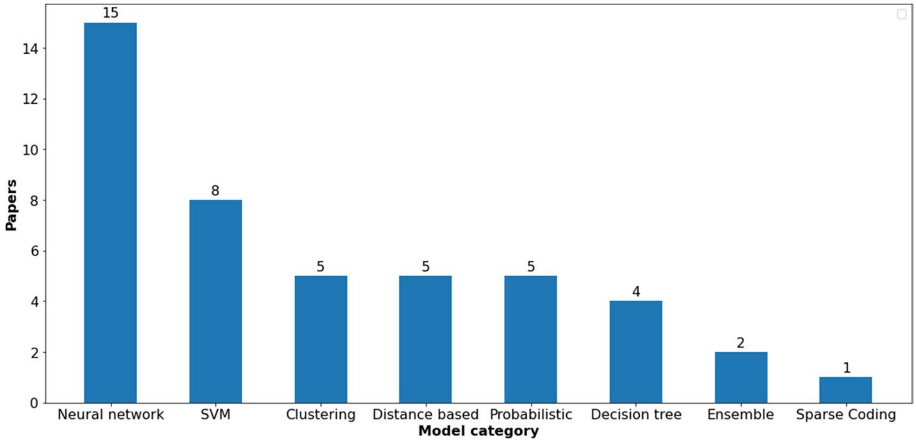


Fig. 5 Use of model categories among the papers

Neural Networks Neural Networks [99] are used in 33% of the papers in this review. This technique uses nodes that are interconnected by weight vectors to learn a prediction function based on the training data. Among papers that employed this machine learning algorithm, 90% used CNNs, a type of neural network primarily created to be used when the inputs are images.

A common issue that arises when training neural networks continuously is a phenomenon known as *catastrophic forgetting* [98], which can be roughly described as the tendency that a neural network has to lose old information learned previously as new examples are presented to the model. In [149], a CNN is used only to improve labeling effort, but not in classification itself. In [139], active learning is employed to reduce training time and improve the quality of object detection. In [105], the authors employ a spatio-temporal autoencoder using a Long Short-Term Memory (LSTM) [58] neural network with convolutional layers. Similarly to [139], active learning is used in order to improve classification accuracy.

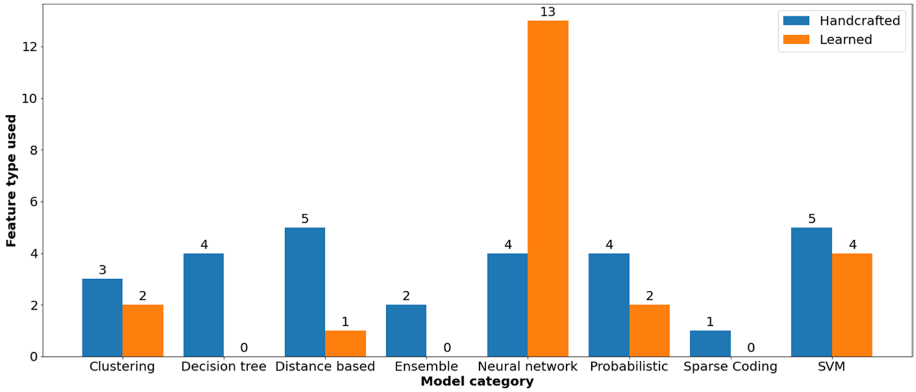


Fig. 6 Feature types used in each one of the model categories. Some models employ both types

In [147], the authors evaluated incremental training in a CNN using four different approaches: (a) Fine-Tuning: retraining only a subset of layers of the network,(b) Knowledge distillation: transferring knowledge from a large network to a smaller one; (c) Learning without Forgetting (LwF): extending from knowledge distillation, it combines cross-entropy loss to learn new tasks and distillation loss function to keep old ones; and (d) Joint-Training: retraining the network from scratch as new examples are added to the existing training set. As discussed by the authors, Joint-Training produced the best accuracy on old data, avoiding the catastrophic forgetting issue. However, as pointed out by the authors, Joint-Training takes, in general, 1.3 times longer to train than the other strategies. Therefore, LwF presents itself as a middle-ground between accuracy and training time, since performance in old tasks is preserved 11% more than Fine-Tuning and it does not require training from scratch like Join-Training.

In [72], the authors combine three networks with distinct purposes (pre-trained prediction, continuous-updating prediction, and weight estimation) to predict future frames in video sequences. In [118], in the context of anomaly detection, a Recursive Neural Network [46] architecture is employed. Then, the estimation is obtained based on autoregression and the moving average of regression errors. In [66], new classes of objects are incrementally aggregated by a Fast R-CNN [45] model. An approach that addresses the concept drift problem indirectly.

In [4], a supervised approach to anomaly detection is presented. The authors perform the preprocessing step of removing the background and then extracting features using both an Optical Flow algorithm and a CNN, which are then input to the LSTM network. In [18], the authors suggest a face mask detection algorithm based on a CNN known as Single Shot Detector (SSD) [82]. The modification made in the original SSD algorithm aimed to improve the accuracy and involved changes including the loss function and the aspect ratio of the network layers.

In [107], the cross-domain adaptation topic (a technique to allow a model trained in one context to be successfully utilized in another context without the need for retraining) is addressed from an object detection perspective. To achieve this, they employ a Domain Transfer Module which consists of a two-layer CNN. Once the model is trained it is able to incrementally add new knowledge using data drawn from the joint representation of previous targets. In [76], the authors deploy a CNN-based object detector with two detection phases: then initial detection performed directly at the camera devices,and a post-processing phase at a server. In order to update the models, the authors claim to have developed a domain adaptation mechanism that can either receive new information from a user feedback or automatically update the domain information related to the spacial and location features.

Among all the works that implemented neural networks, only 20% (3 papers) did not employ CNNs. In [32], the authors put forward a framework to adapt anomaly detection incrementally and continually. The approach involves extracting information from frames, such as bounding boxes, spatial information, and distances as input features to a Recursive Neural Network [152]. In this way, it is argued that the model can be trained continually, thus avoiding the catastrophic forgetting issue. Apparently, a continuation of the previous work, in [31], the authors propose a framework that, in addition to learning continually, is able to implement cross-domain adaptability and few-shot learning (ability to achieve generalization using a relatively small set of representative data). To achieve this, visual information from activities such as object bounding boxes, motion, and poses are transformed into semantic features using the Word2Vec [103] algorithm, i.e., complex activities can be turned into phrases such as “person walking on the sidewalk”. Because this type of feature is more general and simpler than images, it makes cross-domain adaptability and few-shot learning more feasible. Finally, in [112] an incremental learning neural network called Probabilistic Fuzzy ARTMAP (PFAM) [79] is employed. This network provides probabilistic prediction scores based on the categorization of the feature space.

Support vector machines The SVM [15] algorithm is employed in 17% of the analyzed studies. SVM is a classification algorithm that aims to find the decision boundary that presents an optimal margin between classes. In [81], in addition to an incremental approach, the authors also implemented a decremental strategy using sliding windows, making the algorithm capable of removing patterns that are considered obsolete. In [155], an SVM-based algorithm is proposed to make pedestrian recognition adaptive to environmental changes. This method modified the regularization terms to incrementally construct and update its appearance model. In [143], for an action recognition task, the authors used a Structural SVM algorithm applied to short video segments, assuming that the prediction scores of interactions increase over time. In [70, 154], and [55], the SVM algorithms are capable of making incremental updates to the model to prevent them from concept drift. In [87] and [86] the authors propose a face recognition system that uses an ensemble of SVMs that can be self-trained (i.e., the predictions returned by the classifier are used as labels) and in an incremental fashion.

Probabilistic Probabilistic models have been used as classification algorithms in 11% of the analyzed papers. This method builds probability distributions over the training data, so when an example is presented to the model, it can inform the probability one particular example has of belonging to a specific class.

In [11], the authors present a framework to obtain activity patterns from surveillance videos. In this framework, the trajectories classifications and anomaly detections are made using sequential Monte-Carlo techniques. In [94], a combination of handcrafted and learned features is employed to generate a Bayesian fusion model, where the last step is to use a learning-to-rank-based mutual promotion procedure to incrementally update the classifiers based on the newly acquired unlabeled data. In [3], the employed method generates Hidden Markov Models (HMM) [120]. To work incrementally, the likelihood is computed for each incoming video window. When matching a class, a distinct HMM is trained using this data to update an already trained HMM with a weighted average.

In [75], a Gaussian Mixture Model (GMM) is used to detect anomalies in video scenes, where the mixture represents the distribution of abnormal and normal events. The Mahalanobis distance is computed to compare new feature vectors with the mean of the distribution. In [16], features are extracted using variational autoencoder models, and a novelty/anomaly classification is performed using the Markov Jump Particle Filter. Whenever new events are detected, new autoencoder models are deployed.

Clustering Clustering techniques have been used by 11% of the papers. Clustering techniques aim to divide the input data set into groups. Consequently, when a new example is presented, the algorithm is able to predict to which group that instance belongs. This type of algorithm is used in an unsupervised learning setup.

In [106] and [104], the authors use an algorithm named Incremental Knowledge Acquiring and Self-Learning (IKASL) [25], which is based on Growing Self-Organizing Map (GSOM) [1]. The algorithm divides the input set into pathways that can be used in video surveillance tasks. In [80], an algorithm named Online Weighted Clustering is employed in anomaly detection, aiming to model recent events and assign large weights to clusters representing normal events. [121] employ a clustering algorithm to update pose patterns from video cameras dynamically. These extracted patterns seek to improve the labeling process by having more credible images selected for pseudo pose evaluation.

Distance based Used in 11% of the papers in this review, distance-based algorithms use a measure of distance between instances to make inferences. Among these algorithms, k-Nearest Neighbors (kNN) [141] is used in 66% of the papers. KNN is a common choice for adaptive methods because no training is required. This makes the adaptation faster than algorithms such as SVM and neural networks. The inference time, however, is proportional to the number of instances that the model has. In [17], a measure learning algorithm based on statistical probability is employed to incrementally re-identify targets. In [61], in conjunction with Null Folley-Sammon Transform, the Mahalanobis [97] distance is computed.

Addressing object detection, in [126], the authors propose the generation of candidate bounding boxes using a modified Haar Cascades algorithm, these candidate boxes are then used as features along with CNN visual characteristics. These feature vectors serve as input to the posterior classification step using a nearest mean classifier. Through this algorithm, new classes can be added incrementally and there is no need to store all of the training examples.

Other models Among other algorithms, decision tree [119], a tree-based algorithm that derivates decision rules from the training data, was used by 9% of the studies. In [83], the authors use data from sensors to build an incremental model for activity recognition using a swarm decision table. One disadvantage is that people must wear these sensors, which is not always feasible due to constraints such as cost and ease of use. In [108], to perform object detection in videos, the authors automatically label objects based on moving regions and then use these labels to train a decision tree-based model using a co-training strategy for classifier grids. [51] employ a random forest algorithm to incorporate human-assigned labels in an active learning setting. In [47], the authors employ a model called Nearest Class Mean Forest (NCMF) to recognize emotions in images. The NCMF model differs from a random forest, among other characteristics, because only a random subset of available classes is considered in each node.

Ensemble techniques (the combination of predictions of multiple machine learning models) were employed by 4% of the papers. In [71], as new models are generated, old ones are progressively forgotten using a weighting strategy. In [151], an Adaboost algorithm [38] is used to make a global decision by joining a set of weak classifiers. As discussed in [93], ensembles are useful in the case of recurring drift because old models can be simply reused instead of re-trained, which results in a significant saving in computing time.

Lastly, sparse-coding [96], a technique that aims to create sparse linear combinations of basis vectors, has been used in [20] (representing 2% of the papers), where incremental updates are made to existing dictionaries, and the sparse-coding method is used to classify video segments as normal or abnormal.

7 Real-time processing

To prevent damage to people or assets promptly, real-time detection and adaptation are desired capabilities of automated surveillance systems. From all the studies analyzed, 19% (9 papers), reported achieving real-time capacity. To measure the speed of video processing methods, the prevalent Frames per Second (FPS) measure is employed. In Table 3, we summarize the information regarding the studies that employed real-time techniques.

A rate higher than 23 FPS is considered real-time as it is the standard recording rate for cameras and video feeds. If a technique is not able to process frames as they arrive, the processing is delayed.

To make real-time detection possible, we witness a larger use of GPU architectures. In fact, 67% of the papers claiming to use real-time approaches rely on GPU processing [18, 30, 76, 105, 148, 154]. All of these works were published after 2018 and make use of neural networks to obtain the input features.

In [154], considering the steps of feature extraction, dimensionality reduction, and inference, the authors report the speed of 25 FPS for one surveillance video feed. In [30], using pre-trained deep learning models for feature extraction and a kNN model for the testing phase, a speed of 32 FPS is reported. However, it is important to note that kNN needs to compute distances between an incoming example and all of the examples at a specific partition of the training set at inference time. Over time, the number of training examples tends to increase, resulting in performance degradation. The authors also point out that the time for the feature extraction phase can be reduced if a GPU with more computing power is used or if a faster but less accurate version of the deep learning extractors is used.

In [148], whenever concept drift is identified, an object detection model is selected or generated. The baseline model achieves the rate of 24 FPS, while the lightest model architecture yields 144 FPS. Model selection depends on a clustering algorithm, and as stated by the authors, the speed tends to decrease over time as more clusters are created.

In [105], the employed CNN uses eight consecutive frames as input, with 224 × 224 pixels each. The experiment yielded a processing rate of approximately 27 FPS. In [76], the infrastructure takes advantage of cloud computing servers with allocated GPUs. Each camera also has an embedded client that communicates with the server performing object detection. In [18] The neural network proposed, although less accurate than the state-of-the-art model, presents itself as a fast alternative.

We also see less computationally expensive approaches. The remaining 33% of papers rely on regular CPU processing [20, 80, 108]. In [80], a clustering algorithm is used to classify video clips as normal or abnormal, and the reported speed was 30 FPS. In the experiment, the video frames are resized from 158Ã—238 pixels to 120Ã—160,

Table 3 Papers using real-time processing approaches

Author	Feature extraction	Model	Processing	FPS
Lin et al [80]	Optical Flow	Clustering	CPU	30
Chen et al [20]	Gradients, Optical Flow	Sparse Coding	CPU	25
Nguyen [108]	Haar Cascades	Decision tree	CPU	24
Ullah et al [154]	Neural Network	SVM	GPU	25
Doshi and Yilmaz [30]	Neural Network, Optical Flow	Distance based	GPU	32
Suprem et al. [148]	Neural Network	Clustering	GPU	24–144
Nawaratne et al [105]	Neural Network	Neural Network	GPU	27
Cao et al [18]	Neural Network	Neural Network	GPU	42
Kwon and Kim [76]	Neural Network	Neural Network	GPU	55

which requires less computational capacity. Similarly to [30] and [148], the clustering method suffers from speed degradation as new clusters are created. In [20], the spatio-temporal features are generated for every five consecutive frames, and then the dimensions are reduced using PCA. The authors report achieving a processing rate of 25 FPS. In [108], the author states that the implemented system runs at a rate of 24 FPS but does not give details about the type of hardware used.

Lastly, we did not consider the work presented in [149] as being a real-time surveillance method but rather a framework to reduce human effort because the authors employ an already functional real-time application to reduce the annotation to mouse clicks.

8 Datasets and evaluation metrics

8.1 Datasets

In this section, we expose the challenges that handling video data brings, and the characteristics of the datasets used in this review, including the number of frames, resolution, task type, scenes, and annotation type. We also explain why there is still a need for datasets that are proper for concept drift detection.

One of the main characteristics of video surveillance data is that the velocity of data generation is usually high (continuous surveillance video feeds). In addition to that, the number of features is also larger in comparison to other data sources (e.g., tabular data, text, audio). For instance, a video with a resolution as low as 416 × 416 pixels, using the RGB color system, and having a rate of 30 FPS, has 519,168 features (number of pixels) per frame. One minute of that video has 1,800 frames. Furthermore, in [28], the authors mention the challenges faced when dealing with unstructured data in concept drift. In general, video data is multi-dimensional, multi-scale, has spatial relations between frames, and can have an undefined number of labels in each frame.

In [93], several datasets used in concept drift studies are presented. Most of these datasets contain structured data, such as weather and sensor data. There is also text data, which is unstructured, but as explained earlier, video data imposes different challenges, such as spatial relationships between frames and input size. The largest dataset in terms of the number of attributes has 287,034 features and only 10,983 instances.

The datasets used by the studies in this review are presented in Table 4. Given the fact that none of the datasets have concept drift labels or do not have significant changes in illumination, weather, or camera movements, they are not made specifically for detecting concept drift in videos. Instead, they are intended to be used for particular computer vision tasks (activity recognition, anomaly detection, object detection, image classification, or re-ID). Anomalies (or outliers) cannot be considered concept drifts but rather ephemeral changes. Therefore, the anomaly detection datasets analyzed in this review are not suitable for detecting concept drift. Besides the incidence of occasional variations, the characteristics of the input data do not change in a way that the output variable is affected, resulting in the degradation of the prediction capacity of the models. In [30], the authors use an eight-hour-long YouTube video where it begins to rain at some point, and that rain changes the characteristic of the input data. However, the dataset is not annotated, thus, although the authors present a comprehensive analysis regarding adaptation, it is not clear when the concept drifts starts.

Table 4 Datasets used by studies in this review

Dataset	Tasks	Used by	Frames	Resolution	Information	Viewpoints	Annotations	General description
UCSD Ped [95]	Anomaly Detection	Doshi and Yilmaz [30], Lin et al [81], Lin et al [80], Nawaratne et al [105], Pillai and Sen [118], Anoopa et al [4], Doshi and Yilmaz [31]	18,560	238×158	54 anomalies	2	Spatial, Temporal	Camera overlooking pedestrian walkways
CUHK Avenue [91]	Anomaly Detection	Chen et al [20], Doshi and Yilmaz [30], Doshi and Yilmaz [31]	30,652	640×360	47 anomalies	1	Spatial, Temporal	Front view of a campus avenue
UCF50 [123]	Activity Recognition	Hasan et al [55], Ullah et al [154]	200,000	N/A	50 activities	6676	Bounding boxes	Activities collected from YouTube (e.g., biking, typing, hammering, etc.)
UMN [100]	Anomaly Detection	Lin et al [81], Pillai and Sen [118]	3,855	320×240	11 anomalies	1	Temporal	Anomalies in crowd movement
ShangaiTech [90]	Anomaly Detection	Doshi and Yilmaz [30], Doshi and Yilmaz [31]	317,398	480×856	130 anomalies	13	Spatial, Temporal	Scenes with complex light conditions and camera angles
CAVIAR [37]	Object Detection, Re-ID	Nawaratne et al [105], Khoshrou et al. [71], Nguyen [108]	23,000	384×288	6 activities	2	Bounding boxes	Includes people walking alone, meeting with others, entering and exiting shops, fighting, etc
PASCAL VOC [34]	Object Detection	Shin et al [139], Nguyen-Meidine et al. [107]	2,913	N/A	20 classes	N/A	Bounding boxes, Segmentation masks	Includes objects such as people, vehicles, animals and furniture
PETS2006 [150]	Object Detection, Re-ID	Nguyen [108], Wang et al [155]	308	720×576	1,714 instances	3	Bounding boxes	Videos contain left-luggage scenarios

Table 4 (continued)

Dataset	Tasks	Used by	Frames	Resolution	Information	Viewpoints	Annotations	General description
PETS2009 [36]	Object Detection, Re-ID	Khoshrou et al [71], Joy and Vijayakumar [66]	32,107	768×576	3 classes	8	Bounding boxes, Segmentation masks	Crowd actvities in outdoor envirnoments
CUHK01 [84]	Re-ID	Wang et al [155], Huang et al [61], Lv et al. [94]	3,884	160×60	971 instances	2	Bounding boxes	Two images for every identity from each camera
DukeMTMC [167]	Re-ID	Raj et al [121], Sugianto et al. [147]	17,661	N/A	702 instances	N/A	Bounding boxes, Metadata	Outdoor scenes in a university campus
Market1501 [166]	Re-ID	Lv et al [94], Raj et al [121], Sugianto et al. [147]	500,000	128×64	32,668 instances	6	Bounding boxes	identities captured by six different cameras
VIPeR [50]	Re-ID	Cao et al [17], Huang et al [61], Lv et al [94]	1,264	128×48	632 instances	2	Metadata	Outdoor scenes containing 632 people
50Salads [146]	Activity Recognition	Hasan et al [55]	518,411	640×480	17 activities	1	Temporal, sensor	Contains annotated accelerometer and video data
AVA [52]	Activity Recognition	Hasan et al [55]	385,446	320×400	80 activities	928	Bounding boxes	Actions (e.g., drinking, sleeping, talking) in 15-minute movie clips
CFEE [33]	Activity Recognition	Gonzalez and Prevost [47]	1,610	4000×3000	21 categories	N/A	Action Units	Images database for emotion recognition with 230 human subjects
CK [89]	Activity Recognition	Gonzalez and Prevost [47]	18,000	640×490	7 categories	593	Action Units	Facial expression database composed by video sequences
Collective Activity Dataset [21]	Activity Recognition	Nawaratne et al [106]	2,500	640×480	5 activites	44	Bounding boxes	Activities such as crossing, walking, waiting, talking
HMDB51 [74]	Activity Recognition	Ullah et al [154]	632,655	320×240	51 activities	6849	Spatial, Temporal	Actions such as jumping, kissing and laughing
IOSB [56]	Activity Recognition	Ali and Bouguila [3]	-	800×600	7 activities	2	Temporal	Activities such as throwing, pointing and kicking

Table 4 (continued)

Dataset	Tasks	Used by	Frames	Resolution	Information	Viewpoints	Annotations	General description
IXMAS [157]	Activity Recognition	Alcantara et al [2]	34	390×291	11 activities	5	Human body shapes	Activities such as walking, punching and sitting down in a room
JHMDB [65]	Activity Recognition	Soomro et al [143]	31,838	320×240	21 activities	928	Human body shapes	Clips extracted from HMDB51. Annotations use a 2D articulated human puppet model
KTH [136]	Activity Recognition	Alcantara et al [2]	289,715	160×120	6 activities	3	Temporal	Actions include: walking, jogging, running, boxing, waving and clapping (indoor and outdoor)
MPII Cooking [130]	Activity Recognition	Hasan et al [55]	881,755	1624×1224	65 activities	1	Temporal, Pose	Cooking activities (e.g., squeeze, peel, wash)
MSR-II [164]	Activity Recognition	Soomro et al [143]	40,000	320×240	3 activities	1	Bounding boxes	People hand waving, clapping, and boxing
MuHAVi [142]	Activity Recognition	Alcantara et al [2]	134,085	720×576	17 activities	8	Segmentation masks	Actions performed indoor (e.g., jumping, kicking, waving, etc.)
Okutama-Action [9]	Activity Recognition	Pillai and Sen [118]	77,365	3840×2160	12 activities	43	Bounding boxes	High resolution aerial video dataset
PETS2014 [115]	Activity Recognition	Torres et al [151]	7,300	1280×960	Normal, abnormal and threat events	4	Bounding boxes	Activities recorded around a parked vehicle in a parking lot
TV Human Interaction [116]	Activity Recognition	Soomro et al [143]	30,000	N/A	4 activities	300	Temporal, Pose	Actions such as hand shake, high five, hug and kiss
UCF Aerial Action (of Central Florida, [110]	Activity Recognition	Pillai and Sen [118]	100,000	N/A	9 activities	4	Bounding boxes	Video sequences obtained using a R/C-controlled blimp
UCF Sports [145], [129]	Activity Recognition	Soomro et al [143]	9,450	720×480	10 activities	150	Bounding boxes	Sport activities collected from TV broadcasts

Table 4 (continued)

Dataset	Tasks	Used by	Frames	Resolution	Information	Viewpoints	Annotations	General description
UCF101 [144]	Activity Recognition	Ullah et al [154]	2,430,000	320×240	101 activities	13,320	Bounding boxes	Activities collected from YouTube, extension of UCF50
UCLA Office [117]	Activity Recognition	Hasan et al [55]	72,000	N/A	10 activities	5	Bounding boxes	One- and two-person activities captured indoor
URADL [102]	Activity Recognition	Alcantara et al [2]	73	1280×720	10 activities	1	Temporal	Activities include answering phone, chopping and peeling
UT Interaction [132]	Activity Recognition	Soomro et al [143]	36,000	720×480	6 activities	2	Bounding boxes	Activities such as punching and handshaking
VIRAT [111]	Activity Recognition	Hasan et al [55]	2,400,000	1920×1080	23 activities	17	Bounding boxes	Viewpoints of a university campus
Weizmann [14]	Activity Recognition	Alcantara et al [2]	5,701	180×144	10 activities	81	Bounding boxes	Nine actors performing activities such as bending, jumping and running
Standford Drone [127]	Activity Recognition, Object	Pillai and Sen [118]	929,499	1400×1904	6 classes	6	Bounding boxes	Videos captured from a drone at a university campus
CD2014 [49]	Detection, Tracking	Joy and Vijayakumar [66]	90,000	720×576	6 categories	31	Spatial, Temporal	Scene changes: baseline, dynamic background, camera jitter, shadows, object motion and thermal
NOLA [32]	Anomaly Detection	Doshi and Yilmaz [32]	1,440,000	1280×720	N/A	1	Spatial, Temporal	Segments captured over one week from a single moving camera
ImageNet [131]	Anomaly Detection	Disabato and Roveri [27]	14,000,000	N/A	21 classes	N/A	Bounding boxes, Metadata	Large scale image database
BDD [165]	Image Classification	Suprem et al [148]	120,000,000	1280×720	10 classes	100,000	Bounding boxes, Segmentation masks	Autonomous driving recordings

Table 4 (continued)

Dataset	Tasks	Used by	Frames	Resolution	Information	Viewpoints	Annotations	General description
Cityscape [23]	Object Detection	Nguyen-Meidine et al [107]	25,00	N/A	8 classes	50	Bounding boxes, Segmentation masks	Set of stereo video sequences recorded in streets from 50 different cities
COCO [22]	Object Detection	Joy and Vijayakumar [66]	330,000	N/A	80 classes	N/A	Bounding boxes, Segmentation masks	Large scale image database
KITTI Flow [101]	Object Detection, Tracking	Kim et al [72]	8,379	1392×512	5 classes	400	Segmentation masks	Raw, intended for flow estimation with an application to autonomous driving
KITTI Raw [43]	Object Detection, Tracking	Kim et al [72]	648,000	1392×512	5 classes	N/	Bounding boxes	Traffic data obtained from a moving vehicle. Includes sensor and video data
PDTV [135]	Object Detection, Tracking	Bastani et al [11]	1,000	640×480	51 trajectories and 16 object classes	3	Bounding boxes	Videos of a street crossing with multiple traffic scenarios
Cox Face [62]	Re-ID	Lopez-Lopez et al [86], Lopez-Lopez et al [87]	300,000	1920×1080	1000 instances	3	Temporal	Face recognition with variations in pose, expression, lighting, blur, and face resolution
CUHK03 [85]	Re-ID	Raj et al [121]	13,164	N/A	1360 instances	2	Bounding boxes	Bounding boxes detected from deformable part models
GRID [88]	Re-ID	Lv et al [94]	384,000	320×230	10×10 semantic regions	8	Segmentation masks	Eight uncalibrated and disjoint cameras installed in a underground station
MSMT17 [158]	Re-ID	Raj et al [121]	126,411	N/A	4101 instances	15	Bounding boxes	Twelve outdoor and three indoor viewing angles
PRID2011 [57]	Re-ID	Huang et al [61]	24,541	128×64	200 instances	2	Bounding boxes	Multiple person trajectories recorded from two surveillance cameras

Table 4 (continued)

Dataset	Tasks	Used by	Frames	Resolution	Information	Viewpoints	Annotations	General description
SAVITSOFTBIO [12]	Re-ID	Khoshrou et al [71]	64,472	704×576	150 instances	8	Segmentation masks	Moving targets recorded through a network of eight cameras
Visor [7]	Re-ID	Joy and Vijayakumar [66]	200,000	704×576	200 instances	500	Bounding boxes, Segmentation masks	3D/multi-view surveillance and forensic analysis dataset
Youtube Faces [160]	Re-ID	Lopez-Lopez et al [87]	300,000	N/A	1595 instances	3245	Bounding boxes	Video sequences extracted from Youtube

8.2 Metrics

Although computer vision tasks of different natures have been analyzed, the metrics used to evaluate the outcome of the algorithms can be summarized.

Considering the analyzed studies, 49% (22 papers) have used accuracy or accuracy-based metrics. We considered accuracy-based, the recognition rate and detection rate metrics, which are simply accuracy multiplied by 100.

Precision and recall are informative when classes are imbalanced [133], i.e., anomaly detection, where anomalies represent only a small subset of the data. 15% of the studies use precision and recall, and 60% of these studies also use the F-measure, known as the harmonic mean of precision and recall: a single number that summarizes both metrics.

In addition, 32% of the papers (15 studies) use ROC AUC. The Equal Error Rate (EER) is a metric that can be used along with ROC AUC and that summarizes the trade-off between false positives and false negatives. A lower EER represents a more accurate system. The combined use of ROC AUC and EER represents 10% of the articles reviewed.

Commonly used in object detection, Average Precision (AP) or mean Average Precision (mAP), is a metric that takes into account the precision at different recall intervals. It summarises the shape of the Precision-Recall curve. In its respective equation at Table ??, the definition given by [34] is used. AP is defined as the mean precision at a set of eleven equally spaced recall levels $[0, 0.1, 0.2, 0.3, ..., 1]$. The precision at a recall level r is interpolated by taking the maximum precision corresponding to the next recall value greater than the current one:

$$p\mathrm{interp}(r) = \max_{\widetilde{r}:\widetilde{r}\geq r} p(\widetilde{r}) \tag{2}$$

where $p(\widetilde{r})$ is the observed precision at the recall level r. We observe that 18% of all the studies, and 50% of the studies where the task is object detection, used AP.

In Fig. 7, we present the relation between computer vision tasks and the metrics chosen to evaluate them in this review. It is possible to notice that accuracy and AUC-ROC are the most commonly used metrics. Also, the task of object detection presents a strong relation with the AP metric. Some metrics were not used along with some tasks, e.g., AUC-ROC was not used as an evaluation metric in any object detection paper.

9 Discussion

The compilation of the works evaluated in this review allowed us to delineate research potentials, limits, and challenges concerning concept drift adaptation in video-based surveillance regarding four dimensions: adaptation types; features and algorithms; datasets and metrics; and practical aspects. Therefore, our discussion has been structured to present the characteristics of each one of the four defined dimensions.

9.1 Concept drift adaptation

Concerning the first dimension, concept drift adaptation, we explore the implications derived from the choice of adaptation type, adaptation velocity and weighting, and concept drift awareness.

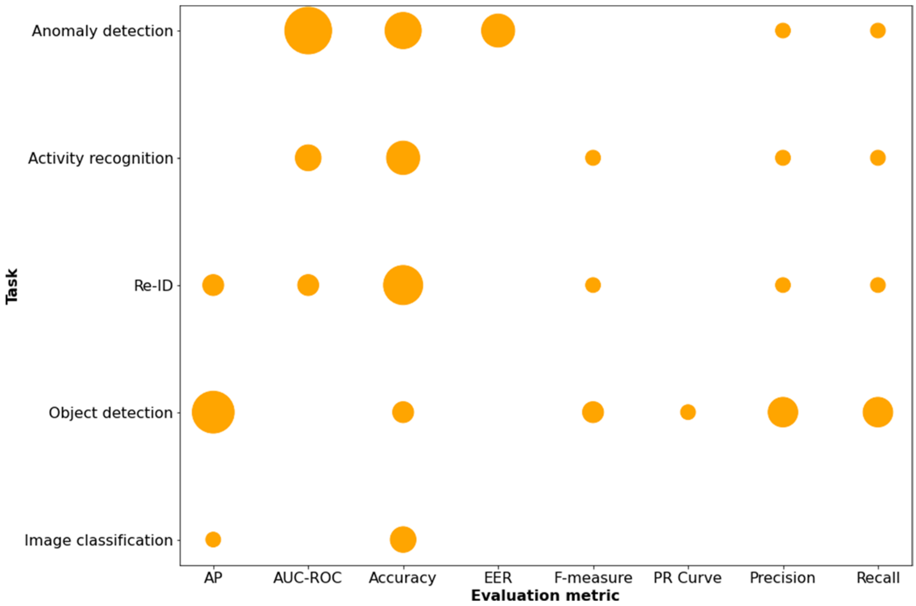


Fig. 7 Relation among tasks and metrics. For simplification purposes, we grouped all accuracy-based metrics (recognition rate, detection rate and matching rate) under the label "Accuracy"

As observed in this systematic review, most concept drift adaptation methods used in surveillance contexts in recent years rely on continuous rather than informed adaptation, which is also pointed out in [41]. While continuous adaptation approaches can constantly aggregate new information into a machine learning model, one disadvantage is that the explicability is sacrificed. In other words, it is not possible to know when a drift occurs.

Furthermore, a strategy on how fast adaptation happens must be defined in incremental approaches, thus, introducing a trade-off between robustness in the presence of noise and adaptation pace. The more weight is given to new examples, the faster adaptation to new concepts happens. However, the model becomes more prone to be affected by noise. On the other hand, when less weight (or no weight) is given to new examples, adaptation occurs slower, but the model becomes more robust to noise. We name this behavior the *robustness-pace trade-off*: a concept introduced in this work.

When all the examples are stored and used for training, besides memory concerns (storage is not unlimited), adaptation to new concepts tends to be slower. In incremental approaches where the data is not stored and single example instances are used to update the model, old concepts are completely forgotten over time, which can be an undesirable feature in contexts where recurring drifts occur.

There are also approaches where specific windows, i.e., sets of sequential training examples, are kept in memory. Larger windows are slower to adapt to new concepts, and smaller windows are faster. In all the incremental approaches, weighting strategies can be applied to make adaptation faster or slower, considering the robustness-pace trade-off explained earlier.

Informed adaptation provides information about when the concept drift happened, a knowledge that, in surveillance contexts, can also be used to trigger alerts to the security personnel, informing them that a context change has occurred. In addition to that, even though space is still a constraint, knowing when a concept drift happens eases the process of getting rid of information that is no longer required. The disadvantages of the informed approach are: a) it is highly dependent on the drift detection algorithm; b) when the model needs to be trained globally, the time spent in adaptation can be long (depending on the learning algorithm and dataset size) and thus, affecting the timing needed in surveillance contexts.

The development of techniques that can circumvent the disadvantages of incremental and informed concept drift adaptation is a topic that deserves more attention. For instance, drift detection could be done at the same time as an incremental training pipeline is running. Thus, it would be possible to know at which moment the concept changed.

9.2 Features and algorithms

In this section, we address the second dimension, features and algorithms, discussing aspects involving feature representation and extraction, handcrafted versus learned features, incremental versus continuous learning, supervised versus unsupervised concept drift detection, and active versus passive learning.

Feature representation and extraction directly impact a model's effectiveness [26],hence, it plays a decisive role in pattern recognition. As presented in this review, in recent years, we have been witnessing a shift from the use of handcrafted features to learned ones and also the combination of both (Section 6). The use of learned features demands, in general, higher computing costs. Consequently, it might not be feasible in cases where more powerful computing architectures (e.g., GPUs) are not available.

Although CNNs have been largely used for feature extraction, 50% of the studies used them exclusively for that purpose and did not employ CNNs as the classifier algorithm. The use of CNNs solely as feature extractors is potentially due to the fact that this type of neural network tends to take a longer time to train than other algorithms, such as SVM [55, 154], clustering algorithms [121], and probabilistic algorithms [94]. Also, as discussed in [30], neural networks suffer from the catastrophic forgetting issue, which causes the performance to degrade over time. To overcome this issue, [27] and [147] suggested approaches to, respectively, retrain only layers affected by drift and employ learning distillation. However, both techniques do not completely solve the forgetting and long training time issues.

As for the other categories of models, even though the training process can be done incrementally and usually faster, and effective storage management strategy needs to be defined to cope with the robustness-pace trade-off. In learning settings where no examples are discarded from memory, the training time increases incrementally as more examples are added. Similarly, for clustering algorithms such as kNN, the inference time grows as new instances are aggregated to the model.

Works that combine learned and handcrafted features do so as a means to explore different characteristics of the input data, thus, improving generalization and accuracy. Besides, another justification for this hybrid approach is reducing computing time. This is the case in [31], where the authors use Optical Flow features along with CNN ones and then transform them into semantic representations that have fewer dimensions than the original data.

Regarding machine learning algorithms used to cope with concept drift, in [28], a review on learning on non-stationary environments, the authors mention that among the continuous adaptation methods, decision trees were one of the most popular algorithms when considering non-ensemble approaches. This differs from the analysis made in recent years, as decision tree based models represent 9% of the total. This is potentially due to the increasing adoption of neural networks approached since the study was published.

Although methods to detect real concept drift rely on annotated instances being evaluated in an existing model [41], the amount of generated data makes the labeling process time-consuming and expensive in surveillance contexts. Thus, detecting drift from a supervised learning approach becomes a less attractive solution. Concept drift detection based on data distribution [93] has been explored as a viable solution [104, 148]. In this case, concept drift is detected by analyzing the dataset itself, which does not require labeled examples.

In [148], the video frames are represented in a lower-dimensional space using a combination of an adversarial autoencoder and a GAN [48]. Then, a clustering algorithm is used to detect concept drift. However, as mentioned earlier, the clustering method tends to get slower and take up more memory space as more instances are aggregated to the model. Similarly, in [104], the authors use a clustering method that brings the same disadvantages.

More research can be done towards approaches that, considering the high dimensionality and high volume of video data, are able to explore the use of established concept drift detection techniques (Section 2.1.1) that can work with less memory and computation constraints. For instance, dimensionality reduction techniques such as PCA or autoencoders can be employed to summarize the data distribution, and then new data points can feed the reconstruction error to a drift detection algorithm such as EDDM or Page-Hinkley.

Active learning aims to enable the acquisition of labels by having an active oracle available while the training process occurs. Nevertheless, acquiring information in this way presents challenges, such as: a) trusting the oracle's labels is not always possible since the quality of these labels can drop over time. An open research problem is to make machine learning models capable of evaluating the quality of these labels [138] b) how to interact with oracles in a way that reduces their effort and optimizes the quality of the labels. General protocols, frameworks, and tools could be developed for this end.

Obtaining labeled information automatically is preferable, but not always achievable. In [108], the authors use background subtraction to extract moving regions and automatically annotate them. Nevertheless, this approach is not extensible to tasks such as activity recognition, where a label is still needed in order to inform which type of activity is taking place.

9.3 Datasets and metrics

The use of datasets and metrics is the third dimension. Regarding this dimension, we discuss the issue of the lack of annotated datasets for concept drift detection, as well as the employment of more distinctive evaluation metrics.

Regarding the datasets used by studies in this review, it was possible to conclude that there is no dataset made specifically to detect drift in surveillance contexts. Beyond the lack of annotations of when drifts occur, the datasets do not present significant changes at the scenes. Illumination, weather, and structural scene changes are some of the phenomena that frequently happen in surveillance contexts and are missing from such datasets. In [93], the authors present several datasets employed in concept drift detection. Although the datasets presented by them do not provide explicit concept drift annotations as well, they do contain drifts that are usual to their respective context (sensor, weather, spam, etc.), a factor not present in the datasets analyzed in this review. Hence, surveillance video datasets where concept drifts occur could be developed and published.

Concerning the metrics employed to assess the quality of the classifiers, accuracy has been the most used one. Despite being straightforward to compute and understand, using accuracy alone is problematic since it does not handle well problems of imbalanced class distribution, with the minority class being less favored than the majority one [60]. Metrics such as ROC AUC and F-measure, for general classification tasks, and average precision, for object detection, are more distinctive and robust measures than accuracy. Thus, more works using metrics that are more distinctive than accuracy can be explored in the subject of video surveillance drift adaptation.

9.4 Practical aspects

Finally, with regard to the fourth and last dimension, we explore practical aspects regarding machine learning and concept drift adaptation in surveillance contexts, including multi-camera environments, real-time processing, data storage, handling sensitive and personal data, and machine learning frameworks for video-based surveillance data.

In large outdoor environments, or even in places with more than one room, a common scenario for video surveillance is to deploy multiple cameras in order to monitor different locations or viewpoints. In this context, considering real-time detection, inference has to be performed in parallel instead of sequentially for each camera. From this practical point of view, real-time approaches using less computational power are preferable. Most of the papers that claim real-time capacity rely on GPU processing, which can potentially use all the GPU units of the computer to perform a single inference. In addition to the cost of having multiple of these powerful machines, configuration, (e.g., deploying new machines) and smart allocation of resources to prevent idle cameras from wasting computational power are other concerns.

Data storage is another issue. Data from surveillance feeds and closed-circuit television (CCTV) are usually generated ceaselessly at a greater velocity than they can be analyzed. Hence, techniques and protocols to process data in distributed ways and also discard it when it is no longer needed could be explored and employed. Additionally, as surveillance video data usually contain sensitive and personal data, data security and privacy are other aspects to consider and are gaining more research attention over the last few years [163]

There is also a need for frameworks and tools not only to analyze surveillance video data but also to manage and cope with concept drift. This could be done by providing a set of concept drift techniques that can be used, compared, and extended along with techniques to perform computer vision tasks (e.g., object detection, activity recognition, anomaly detection) using traditional and deep learning techniques that can be reused and shared.

10 Conclusion

The main contributions of this work are the delineation, the limitations, and research opportunities involving methods, techniques, and strategies to deal with concept drift in surveillance, as well as practical aspects involving computing resources and real-time processing; and the proposal of a new classification of concept drift adaptation method. This classification differs from previous ones as it establishes a relationship between adaptation types and knowledge acquisition strategies; and includes active learning, a relevant technique to acquire new concepts in the presence of drift.

The results show that much more attention has been given to methods that adapt to new concepts in a continuous way rather than in an informed one, and, although blind adaptation in non-stationary environments has advantages, the information on when the concept drift occurred is not available. The continuous adaptation methods include approaches using incremental learning and active learning, while the informed adaptation settings include model selection, model creation, and retraining, which can be done locally or globally.

The relation between computer vision learning tasks and learning settings (supervised, unsupervised, and semi-supervised) used in surveillance was explored, and while tasks such as activity recognition were only performed in a supervised setting, other tasks like anomaly detection were usually done in unsupervised or semi-supervised settings. In such settings, real concept drift cannot be verified because ground truth annotations are not available. Therefore, techniques that explore virtual concept drift must be employed.

Regarding features and machine learning algorithms, even though we witness an increase in the adoption of learned features over handcrafted ones. Traditional methods, such as SVMs and clustering algorithms, tend to be employed more than modern deep learning strategies due to the time taken for adaptation combined with the phenomenon of catastrophic forgetting, usually present in neural networks.

This literature review will help researchers of areas related to machine learning in surveillance to have a comprehensive vision of how the phenomenon of concept drift, in the context of video surveillance, has been handled in recent years, serving as a foundation for other research works. As video surveillance is crucial to improve the security of public and private spaces, the real-world impact of this work is enabling the understanding of alternatives to deal with concept drift, consequently, improving the overall performance of learning methods in this specific scenario, which, as outlined before, presents inherent characteristics and additional complexities over other use cases.

Future research directions include more exploration of informed concept drift adaptation approaches for surveillance, the creation of datasets crafted for non-stationary surveillance environments, the investigation of strategies for data management for continuous surveillance video streams, and the combination of CNNs and traditional approaches.

Acknowledgements The authors would like to thank the Office of Information Technology of USP (STI-USP).

Funding This project is supported by the São Paulo Research Foundation (FAPESP) (grant #2020/06950–4), the Research Dean—PRP-USP (grant #668/2018), the Brazilian National Council of Scientific and Technological Development (CNPq) (grant #309030/2019-6), and the National Institute of Science and Technology in Medicine Assisted by Scientific Computing (INCT-MACC) (grant #157535/2017-7).

Data availability The authors declare that the data supporting the findings of this study are available within the article.

References

1. Alahakoon D, Halgamuge SK, Srinivasan B (2000) Dynamic self-organizing maps with controlled growth for knowledge discovery. IEEE Trans Neural Networks 11(3):601–614
2. Alcantara MF, Moreira TP, Pedrini H (2016) Real-time action recognition using a multilayer descriptor with variable size. J Electron Imaging 25(1):013020
3. Ali S, Bouguila N (2020) Online learning for beta-liouville hidden markov models: Incremental variational learning for video surveillance and action recognition. In 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 10
4. Anoopa S, Salim A, Beevi SN (2022) Advanced video anomaly detection using 2d cnn and stacked lstm with deep active learning-based model. Kuwait J Sci 6
5. Baena-Garcia M, del Campo-Ávila J, Fidalgo R, Bifet A, Gavaldã R, Morales-Bueno R (2017) Early drift detection method. 4th ECML PKDD International Workshop on Knowledge Discovery from Data Streams, 6
6. Bakliwal P, Hegde GM, Jawahar CV (2017) Collaborative contributions for better annotations. In VISIGRAPP 2017 Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications volume 6

7. Baltieri D, Vezzani R, and Cucchiara R (2011) 3dpes: 3d people dataset for surveillance and forensics. In MM'11 Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops JHGBU 2011 Workshop, J-HGBU'11
8. Barddal JP, Gomes HM, Enembreck FC, fahringer BP (2017) A survey on feature drift adaptation: Definition, benchmark, challenges and future directions. J Syst Softw 127:278–294. https://doi.org/10.1016/j.jss.2016.07.005
9. Barekatain M, Marti M, Shih HF, Murray S, Nakayama K, Matsuo Y, Prendinger H (2017) Okutama-action: An aerial view video dataset for concurrent human action detection. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, volume 2017
10. Barrow HG, Tenenbaum JM (1981) Computational vision. Proc IEEE 69(5):572–595
11 Bastani V, Marcenaro L, Regazzoni CS (2016) Online nonparametric bayesian activity mining and analysis from surveillance video. IEEE Trans Image Process 25(5):2089–2102
12. Bialkowski A, Denman S, Sridharan S, Fookes C, Lucey P (2012) A database for person re-identification in multi-camera surveillance networks. In 2012 International Conference on Digital Image Computing Techniques and Applications DICTA
13. Bifet A, Gavaldá R (2007) Learning from time-changing data with adaptive windowing. In Proceedings of the 7th SIAM International Conference on Data Mining
14. Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space-time shapes. In Proceedings of the 7th SIAM International Conference on Data Mining, volume II
15. Bernhard E. Boser, Guyon IM, Vapnik VN (1992) Training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory
16. Campo D, Slavic G, Baydoun M, Marcenaro L, Regazzoni C (2020) Continual learning of predictive models in video sequences via variational autoencoders. In Proceedings International Conference on Image Processing Process ICIP, volume-October
17 Cao W, Han H, Sun XK, Fang ZJ (2017) Target re-identification based on adaptive incremental kiss measure learning. Memetic Comput 9(1):23–30
18. Cao Z, Qin Y, Li Y, Xie Z, Guo J, Jia L (2022) Face detection for rail transit passengers based on single shot detector and active learning. Multimed Tools Appl 8(29):42433–42456
19. Chaovalit P, Zhou L (2005) Movie review mining: a comparison between supervised and unsupervised classification approaches. In Proc 38th Annual Hawaii Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp 112c–112c
20. Chen H, Zhao X, Wang T, Tan M, Sun S (2016) Spatial-temporal context-aware abnormal event detection based on incremental sparse combination learning. In 2016 12th World Con Intell Contr Autom (WCICA). IEEE, 6
21. Choi W, Shahid K, Savarese S (2009) What are they doing?: Collective activity classification using spatio-temporal relationship among people. In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
22. COCO Consortium (2019) Coco detection evaluation metrics
23. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2016-December
24. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In Proceedings- 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005
25. De Silva D, Alahakoon D (2010) Incremental knowledge acquisition and self learning from text. In 2010 International Joint Conference on Neural Networks (IJCNN) 1–8
26. Ding S, Zhu H, Jia W, Chunyang Su (2012) A survey on feature extraction for pattern recognition. Artif Intell Rev 37:3
27. Disabato S, Roveri M (2019) Learning convolutional neural networks in presence of concept drift. In Proceedings of the International Joint Conference on Neural Networks
28 Ditzler G, Roveri M, Alippi C, Polikar R (2015) Learning in nonstationary environments: A survey. IEEE Comput Intell Mag 10(4):12–25
29. Dongre PB, Malik LG (2014) A review on real time data stream classification and adapting to various concept drift scenarios. In 2014 IEEE International Advance Computing Conference (IACC), 533–537
30. Doshi K, Yilmaz Y (2020) Continual learning for anomaly detection in surveillance videos. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops
31. Doshi K, YilmazY (2022) Multi-task learning for video surveillance with limited data. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 6, pp. 3888–3898
32. Doshi K, Yilmaz Y (2022) Rethinking video anomaly detection - a continual learning approach. In Proceedings 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, 1

33 Du S, Tao Y, Martinez AM (2014) Compound facial expressions of emotion. Proc Nat Acad Sci United States Am 111(15):E1454

34 Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338

35. Fang SC, Venkatesh SS (1995) On batch learning in a binary weight setting. In Proceedings of 1995 IEEE International Symposium on Information Theory, pp 170

36. Ferryman J, Shahrokni A 2009. Pets: Dataset and challenge. In, 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance IEEE 12

37. Fisher R, Santos-Victor J, Crowley J (2007) Caviar: Context aware vision using image-based active recognition

38 Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139

39. Gama J, Fernandes R, Rocha R (2006) Decision trees for mining data streams. Intell Data Anal 10(1):23–45

40. Gama J, Medas P, Castillo G, Rodrigues P (2004) Learning with drift detection. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3171

41 Gama J, Zliobaite I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. ACM Comput Surv 46(4):1–37

42. Gözüack A, Can F (2020) Concept learning using one-class classifiers for implicit drift detection in evolving data streams. Artif Intell Rev

43 Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: The kitti dataset. Int J Robot Res 32(11):1231–1237

44. Gepperth A, Hamme B (2016) Incremental learning algorithms and applications. In European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium

45. Girshick R (2015) Fast r-cnn. In 2015 IEEE International Conference on Computer Vision (ICCV), pp 1440–1448

46. Goller C, Kuechler A (1996) Learning task-dependent distributed representations by backpropagation through structure. In IEEE International Conference on Neural Networks Conference Proceedings, 1

47. Gonzalez J, Prevost L (2021) Personalizing emotion recognition using incremental random forests. In 2021 29th European Signal Processing Conference (EUSIPCO), IEEE, 8 pp. 781–7852021

48. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In Advances in Neural Information Processing Systems, 3

49. Goyette N, Jodoin PM, Porikli F, Konrad J, Ishwar P (2012) Changedetection.net: A new change detection benchmark dataset. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops

50. Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, reacquisition, and tracking. 10th International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), 3

51. Grimmeisen B, Theissler A (2020) The machine learning model as a guide: Pointing users to interesting instances for labeling through visual cues. In ACM International Conference Proceeding Series

52. Gu C, Sun C, Ross DA, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, Schmid C, Malik J. (2018) Ava: A video dataset of spatio-temporally localized atomic visual actions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition

53. Hamdoun O, Moutarde F, Stanciulescu B, Steux B (2008) Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In 2008 2nd ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC

54. Hampapur A, Brown L, Connell J, Pankanti S, Senior A, Tian Y (2003) Smart surveillance: applications, technologies and implications. In Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, volume 2, vol.2 pp. 1133–1138

55. Hasan M, Paul S, Mourikis AI, Roy-Chowdhury AK (2020) Context-aware query selection for active learning in event recognition. IEEE Trans Pattern Anal Mach Intell 42(3):554–567

56. Hilsenbeck B, Munch D, Grosselfinger AK, Habner W, Arens M (2017) Action recognition in the longwave infrared and the visible spectrum using hough forests. In Proceedings 2016 IEEE International Symposium on Multimedia, ISM

57. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 6688 LNCS

58 Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

59. Hoogs A, Perera AGA (2008) Video activity recognition in the real world. In Proceedings of the National Conference on Artificial Intelligence, 3

60 Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Proc 5(2):01–11

61. Huang X, Xu J, Guo G (2018) Incremental kernel null foley-sammon transform for person re-identification. In Proceedings International Conference on Pattern Recognition, volume 2018-August

62. Huang Z, Shan S, Wang R, Zhang H, Lao S, Kuerban A, Chen X (2015) A benchmark and comparative study of video-based face recognition on cox face database. IEEE Trans Image Proc 24(12):5967–5981

63. Hu B, Yang C, Shao Y, Yang S (2019) Video-based person re-identification. Nanjing Hangkong Hangtian Daxue Xuebao/Journal of Nanjing University of Aeronautics and Astronautics, 51

64. Ismail MH, Pakhriazad HZ, Shahrin MF (2009) Evaluating supervised and unsupervised techniques for land cover mapping using remote sensing data. Geografia : Malaysian Journal of Society and Space, 01

65. Jhuang H, Gall J, Zuffi S, Schmid C, Black MJ (2013) Towards understanding action recognition. In Proceedings of the IEEE International Conference on Computer Vision

66. Joy F, Vijayakumar V (2021) Multiple object detection in surveillance video with domain adaptive incremental fast rcnn algorithm. Ind J Comput Sci Eng, 12

67. Keele S (2007) Guidelines for performing systematic literature reviews in software engineering: Technical report. EBSE Technical Report EBSE-2007–01

68 Khamassi I, Sayed-Mouchaweh M, Hammami M, Ghadira K (2018) Discussion and review on evolving data streams and concept drift adapting. Evolv Syst 9(1):1–23

69. Khan A, Zhang J, Wang Y (2010) Appearance-based re-identification of people in video. In Proceedings 2010 Digital Image Computing: Techniques and Applications, DICTA

70. Kharabe SR, Raghu B (2016) Matching of video objects taken from different camera views by using multi-feature fusion and evolutionary learning methods. In Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom

71. Khoshrou S, Cardoso JS, Teixeira LF (2015) Learning from evolving video streams in a multi-camera scenario. Mach Learn 100(2–3):609–633

72. Kim W, Tanaka M, Okutomi M, Sasak Y (2021) Adaptive future frame prediction with ensemble network. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 12667 LNCS

73 Krishna MM, Neelima M, Harshali M, Rao MVG (2018) Image classification using deep learning. Int J Eng Technol (UAE) 7(2):614

74. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: A large video database for human motion recognition. In Proceedings of the IEEE International Conference on Computer Vision

75. Kumari P, Saini M (2020) Multivariate adaptive gaussian mixture for scene level anomaly modeling. In 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM). IEEE, 9

76. Kwon B, Kim T (2022) Toward an online continual learning architecture for intrusion detection of video surveillance. IEEE Access 10:89732–89744

77. Lawrence S, Giles CL, Tsoi AC, Back AD (1997) Face recognition: A convolutional neural-network approach. IEEE Trans Neur Netw, 8:98–113

78. Lecun Y, Leon Bottou Y, Bengio PH (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

79. Lim CP Harrison RF (1995) Probabilistic fuzzy artmap: an autonomous neural network architecture for bayesian probability estimation. In 1995 Fourth International Conference on Artificial Neural Networks, pp 148–153

80. Lin H, Deng JD, Woodford BJ, Shahi A (2016) Online weighted clustering for real-time abnormal event detection in video surveillance. In MM 2016 Proceedings of the 2016 ACM Multimedia Conference

81. Lin H, Deng JD, Woodford BJ (2015) Anomaly detection in crowd scenes via online adaptive one-class support vector machines. In Proceedings International Conference on Image Processing, ICIP, volume 2015-December

82. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. Lecture Notes in Computer Science, pp. 21–37

83. Li T, Fong S, Wong KKL, Ying W, Yang XS, Li X (2020) Fusing wearable and remote sensing data streams by fast incremental learning with swarm decision table for human activity recognition. Information Fusion 60:41–64. https://doi.org/10.1016/j.inffus.2020.02.001

84. Li W, Wang X (2013) Locally aligned feature transforms across views. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition

85. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition
86. Lopez-Lopez E, Regueiro CV, Pardo XM, Franco A, Lumini A (2021) Towards a self-sufficient face verification system. Expert Syst Appl, 174
87. Lopez-Lopez E, Regueiro CV, Pardo XM. (2021) An adaptive video-to-video face identification system based on self-training. In 2020 25th International Conference on Pattern Recognition (ICPR), pp 2590–2596. IEEE, 1
88. Loy CC, Xiang T, Gong S (2009) Multi-camera activity correlation analysis. In 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 6
89. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2010
90. Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked rnn framework. In Proceedings of the IEEE International Conference on Computer Vision
91. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE International Conference on Computer Vision
92 Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. Int J Remote Sensing 28(5):823–870
93. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G (2019) Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31
94. Lv J, Chen W, Li Q, Yang C (2018) Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition
95. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 6
96. Mairal J, Bach F, Ponce J, Sapiro G (2009) Online dictionary learning for sparse coding. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09, page 689-696, New York, NY, USA, 2009. Association for Computing Machinery
97. Martos G, Muñoz A, González J (2013) On the generalization of the mahalanobis distance. In José Ruiz-Shulcloper and Gabriella Sanniti di Baja, editors, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 125–132, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg
98. McCloskey M, Cohen NJ (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, Psychology of Learning and Motivation, volume 24, pages 109–165. Academic Press
99. McCulloch WS, Pitts W (1988) *A Logical Calculus of the Ideas Immanent in Nervous Activity*, pp 15-27. MIT Press, Cambridge, MA, USA
100. Mehran R, Oyama A, Shah M (2011) Umn dataset
101. Menze M, Geiger A (2015) Object scene flow for autonomous vehicles. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition
102. Messing R, Pal C, Kautz H (2009) Activity recognition using the velocity histories of tracked keypoints. In Proceedings of the IEEE International Conference on Computer Vision
103. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space
104. Nallaperuma D, Nawaratne R, Bandaragoda T, Adikari A, Nguyen S, Kempitiya T, De Silva D, Alahakoon D, Pothuhera D (2019) Online incremental machine learning platform for big data-driven smart traffic management. IEEE Trans Intell Transport Syst 20(12):4679–4690
105. Nawaratne R, Alahakoon D, De Silva D, Yu X (2020) Spatiotemporal anomaly detection using deep learning for real-time video surveillance. IEEE Transactions on Industrial Informatics 16(1):393–402
106. Nawaratne R, Bandaragoda T, Adikari A, Alahakoon D, De Silva D, Yu X (2017) Incremental knowledge acquisition and self-learning for autonomous video surveillance. In Proceedings IECON 2017 43rd Annual Conference of the IEEE Industrial Electronics Society, 2017-January
107. Nguyen-Meidine LT, Kiran M, Pedersoli M, Dolz J, Blais-Morin LA, Granger E (2022) Incremental multi-target domain adaptation for object detection with efficient domain transfer. Pattern Recognit 129:108771. https://doi.org/10.1016/j.patcog.2022.108771
108. Nguyen DB (2016) Context-based classifier grids learning for object detection in surveillance systems. In Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, volume 165

109. Norman TL (2017) Chapter 6 electronics elements: A detailed discussion - originally from integrated security systems design. thomas norman: Butterworth-heinemann, 2015. updated by the editor, elsevier, 2016. In Lawrence J. Fennelly, editor, Effective Physical Security (Fifth Edition), pages 95 137. Butterworth-Heinemann, fifth edition edition
110. UCF University of Central Florida. (2011) Ucf aerial dataset
111. Oh S, Hoogs A, Perera A, Cuntoor N, Chen CC, Lee JT, Mukherjee S, Aggarwal JK, Lee H, Davis L, Swears E, Wang X, Ji Q, Reddy K, Shah M, Vondrick C, Pirsiavash H, Ramanan D, Yuen J, Torralba A, Song B, Fong A, Chowdhury AR, Desai M (2011) A large-scale benchmark dataset for event recognition in surveillance video. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition
112. Pagano C, Granger E, Sabourin R, Marcialis GL, Roli F (2015) Adaptive classification for person re-identification driven by change detection. In ICPRAM 2015 4th International Conference on Pattern Recognition Applications and Methods, Proceedings, 1
113. Page ES (1954) Continuous inspection schemes. Biometrika, 41
114. Pérez-Sánchez B, Fontenla-Romero O, Guijarro-Berdiñas B (2018) A review of adaptive online learning for artificial neural networks. Artif Intell Rev 49(2):281–299
115. Patino L, Ferryman J (2014) Pets 2014: Dataset and challenge. In 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS
116. Patron-Perez A, Marszalek M, Zisserman A, Reid I (2010) High five: Recognising human interactions in tv shows. In British Machine Vision Conference, BMVC 2010 Proceedings
117. Pei M, Jia Y, Zhu SC (2011) Parsing video events with goal inference and intent prediction. In Proceedings of the IEEE International Conference on Computer Vision
118. Pillai GV, Sen D (2021) Anomaly detection in nonstationary videos using time-recursive differencing network-based prediction. IEEE Geoscience and Remote Sensing Letters
119. Quinlan JR (1986) Induction of decision trees. Machine Learning, 1
120. Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
121. Raj SS, Prasad MVNK, Balakrishnan R (2020) Deep manifold clustering based optimal pseudo pose representation (dmc-oppr) for unsupervised person re-identification. Image and Vision Computing, 101,103-956. https://doi.org/10.1016/j.imavis.2020.103956
122. Ramchandran A, Sangaiah AK (2019) Unsupervised deep learning system for local anomaly event detection in crowded scenes. Multimed Tools Appl 79(47–48):35275–35295
123 Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Appl 24(5):971–981
124. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition
125. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv: 1804.02767
126. Ren S, He Y, Wang X, Guo K, Barra S, Li J (2022) Ciod: an intelligent class-incremental object detection system with nearest mean of exemplars. J Ambient Intell Human Comput, 7
127. Robicquet A, Sadeghian A, Alahi A, Savarese S (2016) Learning social etiquette: Human trajectory understanding in crowded scenes. In European Conference on Computer Vision (ECCV) 549–565
128 Rodriguez-Moreno I, Martinez-Otzeta JM, Sierra B, Rodriguez I, Jauregi E (2019) Video activity recognition: State-of-the-art. Sensors (Switzerland) 19:7
129. Rodriguez MD, Ahmed J, Shah M (2008) Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In 26th IEEE 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, CVPR
130. Rohrbach M, Amin S, Andriluka M, Schiele B (2012) A database for fine grained activity detection of cooking activities. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition
131. Russakovsky O, Deng J, Hao S, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
132. RyooMS, Aggarwal JK. (2009) Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In Proc IEEE International Conference on Computer Vision
133 Saito T, Rehmsmeier M (2009) The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10(3):1–21
134 Saligrama V, Konrad J, Jodoin PM (2010) Video anomaly identification. IEEE Signal Proc Mag 27(5):18–33
135. Saunier N (2010) A public video dataset for road transportation applications. In 93rd Annu. Meeting Transp Res Board, 1–12

136. Schüldt C, Caputo B, Sch C, Barbara L (2017) Recognizing human actions : A local svm approach recognizing human actions. Pattern Recognit, 2004. ICPR 2004. Proc 17th Int Conf, 3
137 Schlimmer JC, Granger RH (1986) Incremental learning from noisy data. Mach Learn 1(3):317–354
138. Settles B (2011) From theories to queries: Active learning in practice. In Isabelle Guyon, Gavin Cawley, Gideon Dror, Vincent Lemaire, and Alexander Statnikov, editors, Active Learning and Experimental Design workshop In conjunction with AISTATS 2010, volume 16 of Proceedings of Machine Learning Research, pages 1–18, Sardinia, Italy, 16 May 2011. JMLR Workshop and Conference Proceedings
139. Shin DK, Ahmed MU, Rhee PK (2018) Incremental deep learning for robust object detection in unknown cluttered environments. IEEE Access 6:61748–61760. https://doi.org/10.1109/ACCESS.2018.2875720
140. Shobha BS, Deepu R (2018) A review on video based vehicle detection, recognition and tracking. In 2018 3rd Int Conf Comput Syst Inform Technol Sustain Solut (CSITSS), pages 183–186
141. Silverman BW, Jones MC (1951) E. fix and j.l. hodges: An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951). Int Stat Review / Revue Int de Statistiq 57:1989
142. Singh S, Velastin SA, Ragheb H (2010) Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In Proc IEEE International Conference on Advanced Video and Signal Based Surveillance, AVS
143 Soomro K, Idrees H, Shah M (2019) Online localization and prediction of actions and interactions. IEEE Trans Pattern Anal Mach Intell 41(2):459–472
144. Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild
145 Soomro K, Zamir AR (2014) Action recognition in realistic sports videos. Adv Comput Vision Pattern Recognit 71:181–208. https://doi.org/10.1007/978-3-319-09396-3_9
146. Stein S, McKenna SJ (2013) Combining embedded accelerometers with computer vision for recognizing food preparation activities. In UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing
147. Sugianto N, Tjondronegoro D, Sorwar G, Chakraborty P, Yuwono EI (2019) Continuous learning without forgetting for person re-identification. In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS
148 Suprem A, Arulraj J, Calton P, Ferreira J (2020) Odin: Automated drift detection and recovery in video analytics. Proc VLDB Endow 13(12):2453–2465
149. Teng E, Falcao JD, Huang R, Iannucci B (2018) Clickbait: Click-based accelerated incremental training of convolutional neural networks. In Proceedings Applied Imagery Pattern Recognition Workshop
150 Thirde D, Li L, Ferryman F (2006) Overview of the pets2006 challenge. In 9th IEEE IEEE Int. Workshop Perform. Eval.Tracking Surveill. (PETS) 47–50
151. Martinez Torres D, Correa HL, Bravo EC (2006) Online learning of contexts for detecting suspicious behaviors in surveillance videos. Image Vis Comput 89:1–26. https://doi.org/10.1007/BFb0053993
152. Tsoi AC (1998) Recurrent neural network architectures: An overview, pages 1–26. Springer Berlin Heidelberg, Berlin, Heidelberg
153 Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008) Machine recognition of human activities: A survey. IEEE Trans Circuits Syst Vid Technol 18(11):1473–1488
154. Ullah A, Muhammad K, Haq IU, Baik SW (2019) Action recognition using optimized deep autoencoder and cnn for surveillance data streams of non-stationary environments. Future Gener Comput Syst 96:386–397. https://doi.org/10.1016/j.future.2019.01.029
155. Wang H, Yan Y, Hua J, Yang Y, Wang X, Li XL, Deller JR, Zhang G, Bao H (2017) Pedestrian recognition in multi-camera networks using multilevel important salient feature and multicategory incremental learning. Pattern Recognit 67:340–352. https://doi.org/10.1016/j.patcog.2017.01.033
156. Wang X, Hu Y, Radwin RG, Lee JD (2018) Frame-sub sampled, drift-resilient long-term video object tracking. In 2018 IEEE International Conference on Multimedia and Expo (ICME), 1–6
157. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Understan 104(2–3):249–257
158. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition
159 Widmer G (1996) Learning in the presence of concept drift and hidden contexts. Mach Learn 23(1):69–101
160. Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition

161 Xiang T, Gong S (2008) Video behavior profiling for anomaly detection. IEEE Trans Patt Anal Mach Intell 30(5):893–908

162. Xiao Y, Tian Z, Jiachen Yu, Zhang Y, Liu S, Shaoyi Du, Lan X (2020) A review of object detection based on deep learning. Multimed Tools Appl 79(33–34):23729–23791

163. Yang P, Xiong N, Ren J (2020) Data security and privacy protection for cloud storage: A survey. IEEE Access 8:131723–131740

164 Yuan J, Liu Z, Ying W (2011) Discriminative video pattern search for efficient action detection. IEEE Trans Patt Anal Mach Intell 33(9):1728–143

165. Yu F, Xian W, Chen Y, Liu F, Liao M, Madhavan V, Darrell T (2018) Bdd100k: A diverse driving video database with scalable annotation tooling. *Arxiv*

166. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2017) Scalable person re-identification : A benchmark scalable person re-identification : A benchmark. The IEEE International Conference on Computer Vision (ICCV)

167. ZhengZ, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.