Combining Where and What in Change Detection for Unsupervised Foreground Learning in Surveillance

Ivan Huerta[a], Marco Pedersoli[b], Jordi Gonzàlez[c], Albert Sanfeliu[d]

[a]DPDCE, University IUAV of Venice, Santa Croce 1957, 30135, Venice, Italy (e-mail:huertacasado@iuav.it).
[b]PSIVISICS, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium (e-mail:marco.pedersoli@esat.kuleuven.be).
[c]Dept. Computer Science & Computer Vision Center, Edifici O, Campus Univ. Autònoma de Barcelona, 08193 Bellaterra, Spain (e-mail: jordi.gonzalez@uab.cat).
[d]Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Parc Tecnològic de Barcelona, Llorens i Artigas 4-6, 08028 Barcelona, Spain (e-mail: sanfeliu@iri.upc.edu).

Abstract

Change detection is the most important task for video surveillance analytics such as foreground and anomaly detection. Current foreground detectors learn models from annotated images since the goal is to generate a robust foreground model able to detect changes in all possible scenarios. Unfortunately, manual labelling is very expensive. Most advanced supervised learning techniques based on generic object detection datasets currently exhibit very poor performance when applied to surveillance datasets because of the unconstrained nature of such environments in terms of types and appearances of objects. In this paper, we take advantage of change detection for training multiple foreground detectors in an unsupervised manner. We use statistical learning techniques which exploit the use of latent parameters for selecting the best foreground model parameters for a given scenario. In essence, the main novelty of our proposed approach is to combine the *where* (motion segmentation) and *what* (learning procedure) in change detection in

an unsupervised way for improving the specificity and generalization power of foreground detectors at the same time. We propose a framework based on latent Support Vector Machines that, given a noisy initialization based on motion cues, learns the correct position, aspect ratio, and appearance of all moving objects in a particular scene. Specificity is achieved by learning the particular change detections of a given scenario, and generalization is guaranteed since our method can be applied to any possible scene and foreground object, as demonstrated in the experimental results outperforming the state-of-the-art.
Keywords: Object detection, unsupervised learning, motion segmentation, latent variables, support vector machine, multiple appearance models, video surveillance

1. Introduction

Change detection is a fundamental task for scene understanding in the surveillance domain. In the literature, *motion segmentation* [1, 2, 3] has been used for detecting *where* motion is present in a scene. Although motion does not represent all the information in a scene, detecting moving objects is very useful because motion is usually highly correlated with the interesting objects of the scene, such as humans, animals and vehicles (see Fig. 1). However motion segmentation has many drawbacks since, instead of learning foreground objects, it computes a background model as a reference for performing change detection. This has been proven not robust enough for surveillance scenarios, where the usual changes in lighting, viewpoint and weather conditions are uncontrolled.

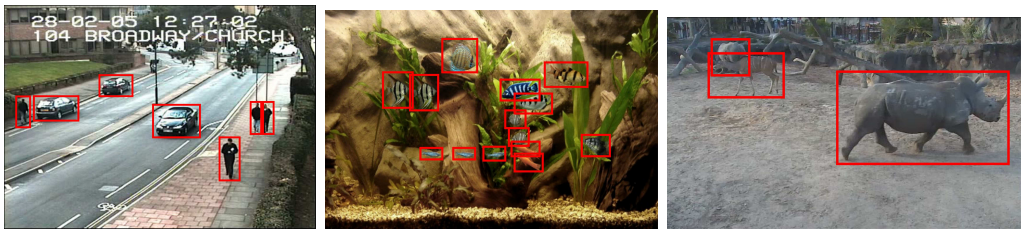


Figure 1: The approach is able to build multi-appearance detectors for unknown and uncontrolled sequences in an unsupervised manner where no pre-trained detectors are available.

Instead, most recent approaches use *object class detection* techniques [4, 5] to learn *what* objects are present in the scene by modelling the highly variable appearance of foreground objects. In this case, instead of modelling the background of a scene, a complex statistical model of those foreground objects which are expected to appear in the scene is learnt. Although learning object categories overcomes the typical problems of motion segmentation [6] (such as illumination changes, camera calibration, weather conditions, and background in motion), object learning is still an open problem due to the enormous variability of the appearances that foreground objects exhibit in the surveillance domain. Also, existing approaches typically requires an extensive collection of positive samples, i.e. annotations of foreground objects, which in the surveillance domain implies an expensive manual labelling process for each possible scene and deployed camera.

In this paper we propose a novel unsupervised methodology which overcomes the limitations of motion segmentation and appearance learning by combining the holistic knowledge obtained from change detection by using these two complementary strategies. On the one hand, motion segmentation provides an initial estimation of the foreground appearance, i.e. statistically

consistent motion changes are considered as objects of interest. These initial hypotheses are then clustered into different appearances to generate the set of foreground object models to be trained. On the other hand, in contrast to current state-of-the-art, our approach does not make any a-priori assumption about the type of foreground object which is expected to appear in the scene: learning the foreground appearance and position based on the clustering step described before is achieved by means of an optimization procedure based on latent variables. Thus we are able to train a specific foreground detector based on the motion segmented in each particular scenario.

The contributions of our method are: (i) substituting the costly manual-labelling task with the use of motion and unsupervised learning for change detection, and (ii) using a discriminative optimization technique based on latent variables able to build accurate multi-class detectors even in the case of noisy and missing motion segmentation. To the best of our knowledge, no method has been proposed to train multiple foreground objects from motion cues in an unsupervised way. To better show the adaptability, generality and robustness of our proposed approach, we have considered different video sequences with no assumptions about the type of foreground object to be detected.

This paper is structured as follows: the next section reviews the works most related to our research while highlighting the advantages of this proposal with respect to the state-of-the-art. Section 3 presents an overview of the methodology used, discusses several critical steps like initialization and the detector used, and describes the multiple appearance learning framework in terms of an optimization problem. The feasibility of the proposed

approach is demonstrated in the experimental results in section 4, while the final discussion and an overview of future avenues of research is presented in section 5.

2. Related Work

Recently there has been a significant interest in semi-supervised and unsupervised learning for object detection, exploiting both labelled and unlabelled data. There are a number of representative approaches that assume different levels of supervision when training object detectors or classifiers.

Among the semi-supervised methods we can find some that use the information from labelled and unlabelled data for co-training manner, such as [7, 8]. Levin et al. [7] use a quantity of labelled data to train two different detectors. Then they use the known relationship between prediction confidence and margin to retrain an improved classifier. However, when the correlation between the two types of inputs is relatively high, co-training does not really improve the detector performance. Javed et al. [8] also used co-training to improve the performance of an initial classifier by selecting new training examples based on PCA. Background subtraction is also used in order to prune stationary-objects in the image. However, the base classifier, which is based on one dimension of a learned PCA model, is relatively weak. Nair and Clark [9] in their approach proposed an on-line detector trained based on an automatic labeller. However, in contrast to ours, this approach needs a manually pre-defined aspect ratio for the automatic labeller. In [10], Wu and Nevatia presented an unsupervised on-line learning approach to improve the performance of boosted object detectors trained from a small labelled

training set by using a large amount of unlabelled data.

Exploiting tracking information, Kalal et al. [11, 12] present a tracker based on a continuously refined detector. The structure of the data is exploited by positive and negative constraints that restrict the labelling of the unlabelled data. These constrains provide a feedback about the performance of the classifier which is iteratively improved in a bootstrapping fashion. Other approaches such as [13, 14, 15] also use tracking to improve the object detector then used for extracting positive and negative examples from the current frame. Babenko et al. [13] use multiple instance learning (MIL), Zhang et al. [14] use sparse representation, and Lu et al. [15] use weighted multiple instance learning (WMIL). However, these tracking-by-detection approaches are trained with the aim of tracking a single object given an initial bounding box, while in our case, foreground detectors are trained to detect at the same time multiple and different object categories in an unsupervised way and without any specific initialization. Also, in our approach we do not use tracking because visual trackers [16] can introduce more noise to the detection results if the tracker is not reliable enough for random motions.

Ali et al. [17] present a method that learns objects of a single category from sparsely annotated videos using boosting. The boosting procedure together with a convex formulation of the objects flow can iteratively improve the detector using the unannotated data considering the constraints generated from the video trajectories. The main limitations of this method are the lack of dealing with multiple object classes, which is quite common in unconstrained scenarios, the sequentiality of the training images, and the need for some object annotations, although sparse.

Methods which train class object detectors in a weakly supervised manner [18] or using random ferns [19] have a very different goal than our approach. Their objective is to improve generic class detectors. Instead, our goal is to train the best object detectors for a specific scenario. More recently, [20, 21, 22] improve generic offline trained detectors using specific scenarios. However, they need pre-trained detectors to be initialized. In contrast Hoai et al. [23] use weakly labelled data to build better object detectors.

The advantages of our approach with respect to all the aforementioned approaches are that our model is trained based on totally unlabelled data and does not require pre-trained detectors. Likewise, there are other methods which also present a fully unsupervised approach. Celik et al. [24] propose training a detector of the most dominant object class (the most repeated class) in the observed scene that is able to select useful training samples in an autonomous manner. Other techniques for training object detectors without the necessity of hand-label examples are presented in [25, 26] where a virtual scenario or a 3d model are used to train a pedestrian detector. These approaches rely on the strong assumption that only one target [24] or a predefined target [25, 26] can be present in the scene. In contrast, we rely on a global optimization procedure which allows our system to handle an unknown number of objects and unconstrained categories of targets.

An approach also based on motion cues for the detection of interesting objects is [27]. In that work, the input received from the motion segmentation is considered the ground truth and a clustering procedure is used to separate the examples for each detector. A further refinement of the clusters is effectuated in order to avoid wrong clusters assignment. However, unlike

our method, the selection of the training examples and the subsequent object model training are done in two separate and independent procedures. This produces quite poor results, especially when the input data is noisy. This is because the foreground regions are estimated in a bottom-up fashion, without using important information about the final aim, that is distinguishing among foreground objects and background. In contrast, in our method the selection of the positive examples to use for each class, as well as their correct location, are optimized at the same time in a discriminative fashion. Moreover, authors in [27] manually defined the classes which are used to train independents SVM for each class. Once they have defined the possible clusters with some refined examples for each of the appearances, they manually group them in two classes: car and pedestrian. In contrast, our approach uses the data directly by performing a global optimization based on latent variables, thereby being able to train a unique detector which can work with different appearances at the same time.

Summarizing, our proposal is different from the aforementioned approaches because it (i) is fully unsupervised, since there is no need for hand-labelled annotations, (ii) can learn objects never seen before as it does not rely on any a-priori trained detector, and (iii) works with multiple and unlabelled objects.

3. Our Approach

The technique proposed in this paper combines in an unsupervised way *where* to learn (motion segmentation) and *what* (learning procedure) from change detection to improve the specificity and generalization power of trained

foreground detectors at the same time. The *where* will be given by a motion segmentation procedure to subsequently initialize the detectors (section 3.1). In addition, the *what* will be the unsupervised procedure to train the detectors based on the segmented motion, that is what objects do we have? (section 3.2). Consequently, the appearance and position of foreground objects will be learnt by means of an optimization procedure based on latent variables.

An overview of the method is shown in Fig. 2. In the first stage motion cues are used to roughly segment the moving objects. In our experiments this is done by learning a background model and segmenting those regions that have a local motion with respect to the background. Subsequently, based on the statistical distribution of the bounding boxes of the moving regions, the number and appearance of the required detectors are estimated and given as input to the learning procedure. During training, with a global optimization we iteratively and simultaneously learn the correct object location, aspect ratio, and appearance to associate a detector to each moving region.

Since the main purpose of our approach is the detection of the foreground objects in surveillance scenarios instead of the categorization of those detected foreground objects, our approach is not limited to a specific number of categories. That means, different foreground object detectors will be trained based on the variance in aspect ratio of foreground regions instead of based on the nature of the object being learnt. The approach uses the variance in aspect ratio to initialize the foreground detectors. As an example, we can train detectors able to detect pedestrians and cars without explicitly inferring the category the moving objects belong to. For instance, quite

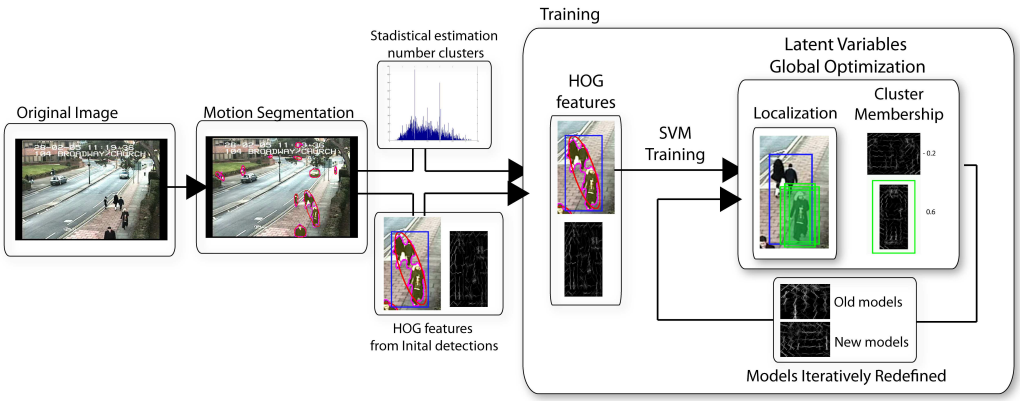


Figure 2: Approach Overview: Firstly, moving objects are roughly segmented using motion cues. This is used as input for the initialization of the learning procedure and the cluster number estimation. Finally, a global optimization iteratively learns the correct object location, aspect ratio, and appearance simultaneously for each of the detectors. See text for more details.

commonly in the surveillance domain, pre-trained detectors are not able to detect a specific category because of occlusions with other objects; however in our scenario the occluded object has an aspect ratio that is different and therefore another detector will be learnt.

In the following sections we give a detailed explanation of the model initialization and the multi-appearance learning.

3.1. Initialization

In our framework the learning procedure is based on latent SVM [5]. We consider object position, and appearance as latent variables. In this way, the latent variables can assume the value that is most discriminative in order to distinguish moving foreground objects from background. However, the optimization problem is not convex due to the latent variables. This means

that the yielded solution is local and an optimal solution requires a proper initialization of the latent variables. For initialization, the detected moving regions are considered as the initial candidates for learning appearances as well as the shape of the detectors that will model those foreground regions.

Motion The estimation of the objects location is provided by bottom-up information. The key idea is that motion segmentation substitutes the tedious hand-labelling task. Specifically, in our approach we use a background subtraction technique to obtain a rough initial estimation of the presence of one or more objects in a certain location of the image.

In order to obtain the moving foreground objects we have employed [28]. It uses a hybrid architecture which exploits the benefits of fusing a chromatic-invariant cone model for colour segmentation, an invariant gradient model which fuses magnitude and orientation for edge segmentation, and intensity cues together with temporal difference. Furthermore, taking advantage of these cues it also detects and removes shadows [1]. An example of the motion segmentation results obtained from CLEAR06 database can be seen in the Fig. 3.(a).

Even though many of the problems of motion segmentation are solved by the approach presented in [28], the detection of moving objects in complex environments is still far from being completely solved [35] since noise and other segmentation errors occur frequently. However our system is robust to such errors thanks to the refinement of the global discriminative optimization, as described next.

[1] In fact, any motion segmentation algorithm such as those presented in [29, 30, 31, 32, 33, 34] could be used instead to obtain the moving regions to be learnt by the detectors

(a) (b)

Figure 3: a) Motion segmentation results from CLEAR06 sequence. b) histogram of bounding box ratios computed from the objects segmented in the CLEAR06 sequence.

Detectors Initialization In order to detect objects of different shapes and sizes, an initial analysis of the objects that most frequently appear in the scene is necessary. In particular we estimate the detector's size and appearance. We evaluate the most distinctive appearances of all objects that appear in the scene, and tailor a set of detectors to best reproduce this distribution. In practice, we obtain the optimal trade-off between representing all the appearances of the objects in the scene and getting enough samples. The initial object clustering could contain clusters with a reduced set of samples. A model trained with that reduced set of samples would in general produce a poor detector. In order to obtain a trade-off between representing all the appearances of the objects in the scene and obtaining good detectors, those clusters with too few samples will be discarted. For doing so we extract a smoothed histogram of the distribution of the bounding boxes aspect ratios obtained from motion segmentation. We take the local maximum of the histograms as the aspect ratio of our detectors. We also split each aspect ratio

to left-right facing samples. To do that we randomly flip a sample and check if the global variance of the HOG [4] features on the samples is smaller than before. In that case we maintain the change. We continue that procedure until no more flips are applied. An example of the bounding box aspect ratios histogram obtained from CLEAR06 sequence can be seen in Fig. 3.(b).

We are interested in estimating the sizes of the objects that appear in the scene to obtain the best trade-off between a high resolution representation of the object (more discriminative) and the risk of not detecting small objects (more robust). For this we set for each appearance a detector with a size that allows it to detect 90% of the samples in the training set.

Some regions are erroneously segmented as belonging to an object. However, in our approach these false positives are statistically considered as outliers given the whole segmented sequence. In the case that the number of different appearances are erroneously considered due to a failure in segmentation or in clustering, these problems do not modify considerably the detection results as later discussed in the experimental results.

3.2. Multi-Appearance Learning

The strength of our approach relies on the learning procedure. Instead of dividing the learning procedure in two separate tasks, *clustering* and *appearance learning*, we propose to learn both tasks in a single, global optimization procedure. In essence, cluster assignment as well as the accurate object position estimation are represented as so-called *latent variables* which can be jointly estimated during training using the latent SVM algorithm as proposed in [5].

In our case, the assignment of the latent variables is based on two joint

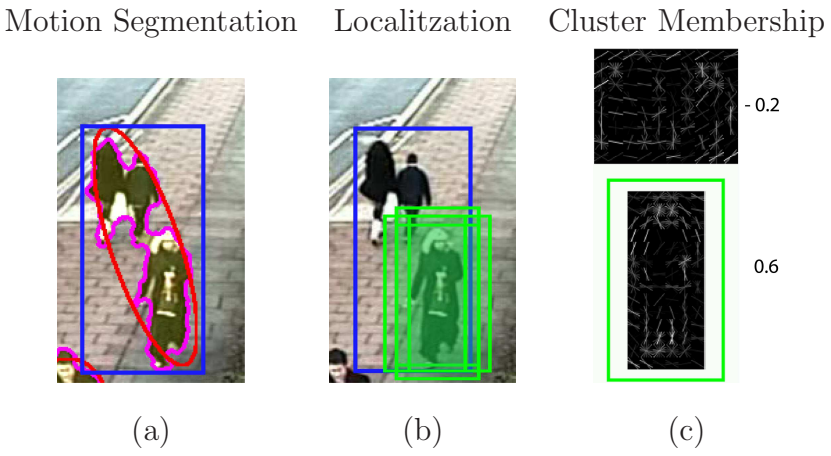


Figure 4: Example of assignment of latent variables. The assignment of the latent variables is effectuated based on two rules: (i) the overlap with the motion segmentation, and (ii) the scoring function of the latent SVM. (a) Motion Segmentation results. (b) Localization: for each object model, the object location is chosen based on the location that maximizes the detection score. (c) Cluster membership: as both object models have enough overlap with the segmentation, the model is chosen based on the maximum score it can obtain. Note that the assignment of two latent variables is effectuated jointly.

rules: (i) the overlapping intersection area between the ground truth and the detected bounding boxes obtained from motion segmentation, and (ii) the scoring function given by the latent SVM, see Fig. 4. Indeed this procedure works well since both tasks are highly interconnected: the object appearance is used to compute a better estimation of the cluster that belongs to each foreground object and its localization, as well as when the foreground objects are well separated into different aspect ratio clusters, object appearances can be better learned by the detectors.

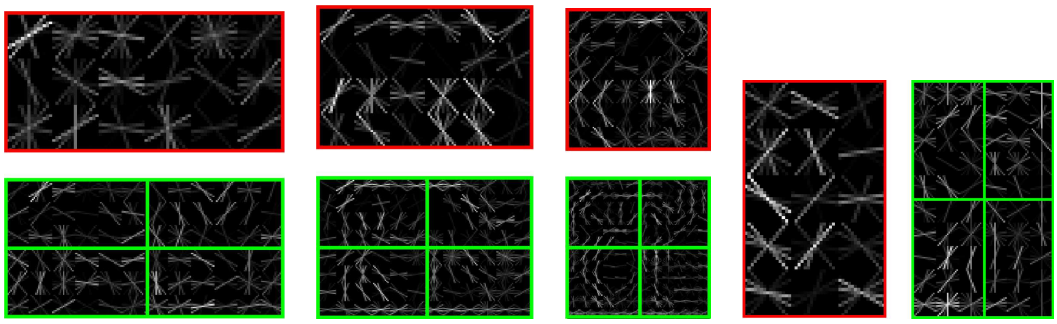
Unfortunately, in our problem the estimation of the object appearance, the cluster membership and the object position cannot be estimated at the

same time because they are mutually dependent. This implies, in contrast to normal SVM, that the corresponding energy function is not convex and its optimization should be performed in an iterative way composed of two parts: a convex optimization of the object model using the current estimation of the latent variables, in addition to a concave optimization of the object model corresponding to a new estimation of the latent variables which minimize the energy function. These two iterative steps are detailed next.

Inference In our framework the inference procedure corresponds to the detection of the objects in the scene. This procedure is used in an unconstrained way during testing, where the objects can be found at any location of the image, and in a constrained way during training, where a region of the image is used for training only if a minimum overlap with the motion segmentation is reached. That is, each motion segmentation region represents a sample and then during inference the class of the object as well as its location are estimated. To have an optimal trade-off between speed and accuracy, inference is applied using the coarse-to-fine procedure as proposed in [36]. Notice that this approach, similarly to [5], is also based on parts and therefore can deal with object deformations.

An object model is trained for the detection of the foreground objects in the scene. This model contains the parameters w trained using the latent SVM procedure. It is composed of several components, each one with a different appearance. Also, each appearance is decomposed into several resolution levels. An example of object model with different components and the corresponding parts is shown in Fig. 5.

The multiple resolutions are employed sequentially in a locally greedy



(a) (b) (c) (d)

Figure 5: Object models learned from HoustonZoo_Rino sequence; The model consists of four components (four different appearances), with two levels of resolution. The high resolution is divided into deformable parts.

fashion to find the object model. As the scoring function is locally smooth, the method gets solutions very close to the exact search but in a fraction of the time. To increase the capability of the detector to deal with object deformations, the model is divided into subparts that can move relatively to each other with a certain degree of stiffness that is leaned at training time. For more details see [36].

The scoring function f, for a latent SVM is defined as:

$$f(\mathbf{x}, \mathbf{s}; \mathbf{w}) = \max_{\mathbf{h}} \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{h}, \mathbf{s}) \rangle \tag{1}$$

where each example $\mathbf{x}$ is scored giving a vector of model parameters $\mathbf{w}$, and a region $\mathbf{s}$ represented as a bounding box. Φ is a function that given an image $\mathbf{x}$, the location of the bounding box $\mathbf{s}$ and the set of latent variables $\mathbf{h}$ returns a corresponding feature vector (HOG features in our case).

In our model the latent variable $\mathbf{h}$ represents the position of the detected object in the image, the relative deformation of each object part with respect

to its rigid position, as well as the cluster membership of the model.

To properly train a foreground object detector, the parameters $\mathbf{w}$ of the SVM that minimize the energy function are first computed. As stated before, since this energy function is not convex, a piecewise linear upper bound of the loss is used instead: next we define the resulting energy function and the optimization procedure for such as function.

Energy Definition. We now define the energy function that we want to optimize. Consider a set of input images $\mathcal{X} = \{\mathbf{x}_0, \mathbf{x}_1, ..\mathbf{x}_N\}$ and a set of associated bounding boxes $\mathcal{Y} = \{\mathbf{y}_0, \mathbf{y}_1, ..\mathbf{y}_M\}$ representing the foreground segmentation obtained from motion. That is, we consider the motion segmentation as our ground truth annotations. However, these annotations can have errors that are corrected with the latent localization of the object of interest. As in general we can find more than one bounding box in a single image, we associate each bounding box i with the corresponding image k through the function $l(i) = k$.

We want to find the model parameters $\mathbf{w}$ and the bounding box locations $\mathbf{s} \in \mathcal{S}$, that minimize the following regularized energy function:

$$E(\mathcal{X}, \mathcal{Y}; \mathbf{w}) = \lambda \frac{1}{2} ||\mathbf{w}||^2 + \mu \sum_i \Delta_\tau(\mathbf{y}_i, \mathbf{s}_i) \tag{2}$$

where λ is the trade-off between loss and regularization.

The sum of Eq.(2) represents the loss Δ_τ which punishes detections $\mathbf{s}_i$ that do not overlap[2] with the associated foreground segmentation $\mathbf{y}_i$. The

[2]Here, we considered overlap the intersection area between the ground truth $\mathbf{y}$ and the detected bounding boxes $\mathbf{s}$, normalized by the area of the union of the bounding boxes as defined in Eq.3.

loss is defined as follows:

$$\Delta_\tau(\mathbf{y}, \mathbf{s}) = \begin{cases} 0 & \frac{area(\mathbf{y} \cap \mathbf{s})}{area(\mathbf{y} \cup \mathbf{s})} \quad \geq \tau \\ 1 & otherwise \end{cases} \tag{3}$$

where τ is the overlapping coefficient[3] and *area* is a function that computes the area of a given bounding box. In this way we specify that any detection $\mathbf{s}$ with a sufficient overlap with the foreground segmentation $\mathbf{y}$ would be selected as a positive example, while a detection that falls outside the foreground segmentation or that has too small of an overlap is considered a negative example, and therefore penalized.

Optimization. In order to optimize Eq. (2) we build a piece-wise linear upper bound of the previously defined loss:

$$\Delta'_\tau(\mathbf{y}_i, \mathbf{s}_i, \mathbf{x}_{l(i)}, \mathbf{w}) = \max_{\mathbf{s}_i} \left[f(\mathbf{x}_{l(i)}, \mathbf{s}_i, \mathbf{w}) + \Delta_\tau(\mathbf{y}_i, \mathbf{s}_i) \right] \tag{4}$$

$$- \max_{\mathbf{s} \in \mathcal{S}(\mathbf{y}_i)} \left[f(\mathbf{x}_{l(i)}, \mathbf{s}_i, \mathbf{w}) \right]. \tag{5}$$

The first term of Eq.(5) is the maximization of a linear function and is therefore convex in $\mathbf{w}$, while the second term is the negation of the maximization of a linear function so it is concave.

Now we rewrite Eq.(2) as $E(\mathcal{X}, \mathcal{Y}, \mathbf{w}) = E(\mathcal{X}, \mathcal{Y}, \mathbf{w})_{convex} + E(\mathcal{X}, \mathcal{Y}, \mathbf{w})_{concave}$ where:

$$E(\mathcal{X}, \mathcal{Y}, \mathbf{w})_{convex} = \lambda \frac{1}{2} ||\mathbf{w}||^2 + \sum_i (\max_{s_i} \left[f(\mathbf{x}_{l(i)}, \mathbf{s}_i, \mathbf{w}) + \Delta_\tau(\mathbf{y}_i, \mathbf{s}_i) \right] \tag{6}$$

[3]Empirically τ is set to 0.75, see experimental results

$$E(\mathcal{X}, \mathcal{Y}, \mathbf{w})_{concave} = -\sum_{i} \max_{\mathbf{s}_i \in \mathcal{S}(\mathbf{y}_i)} \left[f(\mathbf{x}_{l(i)}, \mathbf{s}_i, \mathbf{w}) \right] \tag{7}$$

Similarly to [37], the minimization of Eq. (2) can be minimized using the well known CCCP procedure [38]. For the convex optimization of $\mathbf{w}$ in Eq. (6), we use stochastic gradient descent [39] and for the concave part in Eq. 7 we fix $\mathbf{w}$ and optimize over $\mathbf{s}$ which represents the object location as well as the remaining latent variables.

4. Experimental Results

In order to show the unconstrained nature of our approach, three different video sequences have been considered. As the approach is generic, we do not assume any prior information about the scene, about the objects that will appear in the sequence, nor about their motion. These sequences correspond to different sources such as a well-known standard database, and publically available web-cam and a synthetically generated video, to show the robustness and generality of the proposed approach.

Databases. In essence, *CLEAR06_PV*[4] dataset shows a real urban scene with multiple people and vehicles at the same time. It is part of a well-known public i-LIDS[5] database previously used in AVSS2007[6] conference. It contains $13,167$ frames for training and $3,929$ frames for testing with more than 236 pedestrian and 357 cars annotated, ground truth from [27].

[4] http://figment.csee.usf.edu/~psoundar/Videos/Surveillance/

[5] https://www.gov.uk/imagery-library-for-intelligent-detection-systems

[6] http://www.eecs.qmul.ac.uk/~andrea/avss2007_ss_challenge.html

FishTank dataset shows fish in an artificially generated fish tank. This is a very challenging dataset due to the multiple occlusions, where fish are constantly splitting and grouping, and the small size of the fish. It contains $1,360$ frames for training and $1,000$ frames for testing. *HoustonZoo_rhino* dataset is directly recorded from an internet web-cam placed in the zoo of Houston [7] that contains rhinos and deer. This challenging dataset contains a lot of camouflage and occlusions in the environment. It has $14,360$ frames for training and $1,860$ for testing[8].

Metrics. For the purpose of comparison we use average precision (AP), which is computed as the average of the detector precision at different values of recall, from 0 to 1. To distinguish between true positive detections and false positive detections we use the VOC overlapping criteria [40]. This is a common metric used for object detection, which evaluates the intersection area between the ground truth and the detection bounding boxes, normalized by the area of the union of the bounding boxes. If it is greater than 0.5 the detection is considered correct, otherwise it is a false detection.

Comparative Analysis. In table 1 we evaluate the AP of our detection algorithm on CLEAR06_PV database which have been previously used in [27]. For a fair comparison, the same training, test, and ground truth (GT) as defined in the [27] have been considered, although the provided GT is not

[7] `http://www.houstonzoo.org/webcam/`

[8] Sequences FishTank and HoustonZoo_Rhino and their hand-segmented GroundTruth are available in `http://www.cvc.uab.es/~ivanhc/ObjDect/huertaDect.html` for the purpose of comparison

method	person	car	both
Pre-trained detector [4]	76	39	-
Celik et al. [27]	58	85	-
Our method w/o latent	64	88	63
Our method w/ latent	77	91	**81.5**

Table 1: Detection Rate at 1 FPPI on CLEAR06 of multiple objects. See text for more details.

.

complete [9]. In this sequence there are mainly two categories of objects: person and car. While in [27] the method learns each object class independently, our approach learns each moving object without even knowing to which category it belongs to, in a single optimization as explained in Sec. 3. In the first row of Table 1 we report the AP of a supervised generic detector [4] pre-trained with an independent set of images of cars and pedestrians. In the same way, in the second row of Table 1 we show the AP obtained for cars and pedestrians with the method proposed in [27] [10].

Our method, does not assume any knowledge about the number and appearance of the different classes that will appear in the scene. As expected, we can not distinguish between cars and pedestrians but we can detect most of them. In order to be able to compare our method with the pre-trained

[9] Annotations of small, partially occluded or partially out of the screen object are missing.

[10] The training and testing methodology as the values for the pre-trained detector [4], and [27] are extracted from [27]

detectors as well as with [27], as in this case the aspect ratio of the objects bounding box is highly correlated with the class (i.e vertical box, pedestrian and horizontal box, car), we manually separate the 4 models generated by our method into one group of 1 cluster containing the car category and another one composed of 3 clusters representing the person category.

Our method clearly outperforms both the pre-trained detectors [4] as well as [27] in both categories. In the third column of Table 1 we also show the global performance of the method without distinguishing between classes. This task cannot be performed by the other methods, as they need to train each class independently.

It is interesting to remark that the AP for the pre-trained detector for car is relatively low. This is because the general detector has not been trained with this specific car view, thereby producing a low recall. One of the problems of a general pre-trained detector vs. a specific object detector (our approach) is that it is not possible to train it for all the specific object appearances. This is the case in this scenario where the pre-trained object detector is trained with the car dataset formed by the frontal and nearly frontal images of cars from the publicly online available ETHZ set. Cars that appear in the current scenario are not from either frontal or lateral views, therefore the pre-trained detector is not able to properly detect the cars. This experiment shows that in surveillance it is not a good strategy to train detectors for specific views of objects, but instead to train detectors for specific scenarios. As can be seen in Table 1, our approach obtains almost perfect detection for cars thus showing the advantages of an appositely-trained detector versus a generic one. Finally, we show the performance of the method without and

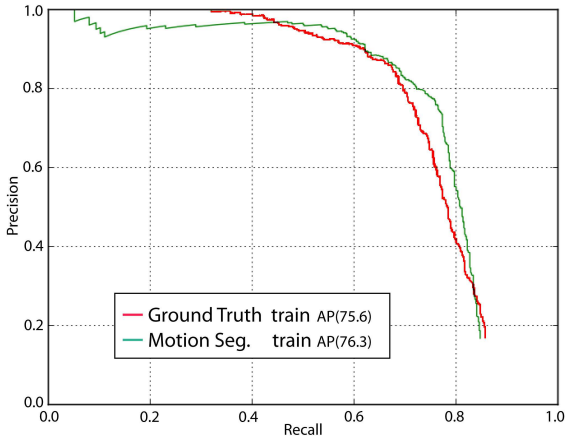


Figure 6: Comparative analysis using our approach trained with Ground Truth, and trained with Motion cues from CLEAR06 sequence.

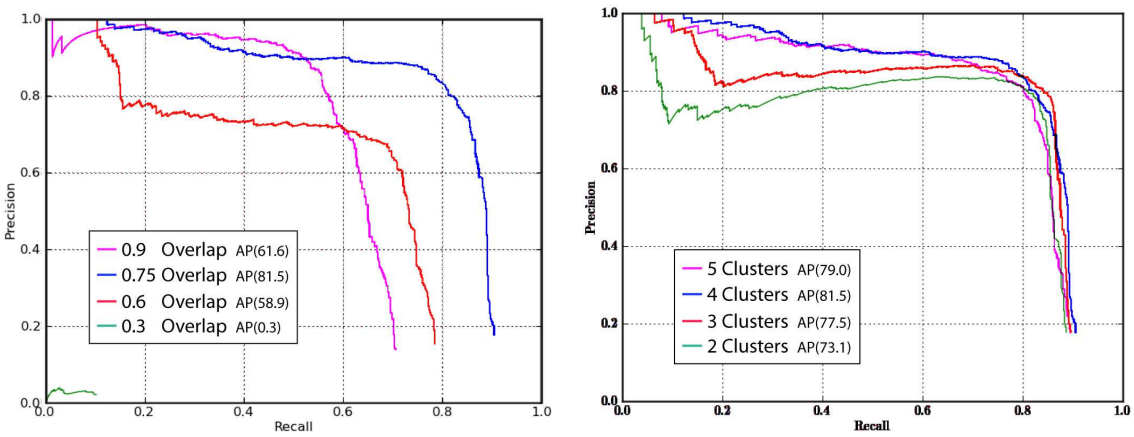
with latent variables. As expected the AP is lower without them.

Initialization Test. We want to evaluate the effect of substituting the real bounding boxes of the objects of interest (hand annotated ground truth) with the regions obtained by motion segmentation. To do that, we trained one model on CLEAR06 with the ground truth bounding boxes from the original sequence (5700 frames), and another one with the same frames but using the noisy data obtained by the motion segmentation. Surprisingly, the model trained with motion obtains an AP slightly better than using ground truth data, see Fig. 6. This is because the original annotations from [27] are quite conservative in the sense that they discard many examples, like partially occluded and truncated examples. As our learning is based on an iterative refinement of the location and appearance of each example, those difficult examples can also be exploited, as is done when using motion cue. This is the reason why the training effectuated with the initialization based on motion is able to achieve slightly better recall. In contrast, the training

using ground truth obtains better precision at low recall because fewer but better examples are used.

Latent Variables Test. In this experiment we show the effect of varying the amount of freedom assigned to the latent variables. In our problem the space of valid configurations of the latent variables is parametrized by the overlapping coefficient τ defined in section 3.2. For instance setting $\tau = 0.5$ means that only those detections with an overlap higher than 0.5 with the initialization given by motion segmentation can be considered as valid configurations. Fig. 7 (a) shows how the overlap criteria affect the latent variables. When the overlapping is very high (0.9) the space of possible variations of the latent variables is reduced and in the end it is like considering the initialization as ground truth and no latent estimation is computed. In the other side, when the overlap threshold is set to 0.3 the estimated detection can be quite far from the initialization which can produce a training with false positive data. This explains why in this case the AP is so low.

Number of Clusters Test. In Fig. 7 (b) we evaluate the performance of our system changing the number of clusters used during training. As expected, increasing the number of detectors increases the precision of the system and therefore its global performance. This is true up to a certain limit (in this case 4 clusters). After that, more detectors tend to overfit the data. In general we can see that while the overlapping value highly affects the overall performance of the system, the number of used clusters is a relatively steady parameter. This justifies the heuristic explained in sec. 3.1 for the selection of the number of clusters to use. Interestingly, independently of the chosen number of clusters, in all the configurations our proposed approach



(a) (b)

Figure 7: Comparative analysis from CLEAR06 sequence using our approach with (a) different overlapping criteria, and (b) different number of clusters.

Seq	NumFrTr	NumFrTest	NumClus	Ini	Final
CLEAR06_PV	13167	3929	4	63.6	**81.5**
FishTank	1360	1000	3	55.9	**62.3**
HoustonZoo_rhino	14360	1860	4	61.3	**68.6**

Table 2: Performance analysis using different sequences. See text for more details.

.

obtains better results than using pre-trained generic detectors [4] and using the approach presented in [27].

Overall Evaluation. We evaluate our method on two more challenging sequences, where no pre-trained detectors are available. One is a synthetic video of a fish tank. The other is a a video collected from a web-cam placed in the zoo of Houston, HoutonZoo_Rhino. Note that the pre-trained detector cannot be evaluated in these sequences because the generic object detection

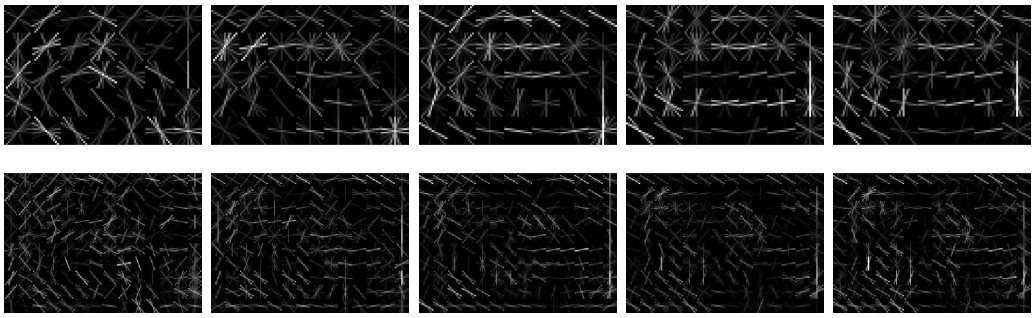


Figure 8: Appearance models over iterations. During the latent variables iterations the appearance model is refined obtaining a better representation of a car for CLEAR06_PV sequence.

datasets such as PASCAL VOC [11], INRIA[12], Daimler[13] does not contains fish or rhinos, thereby showing one of the advantages of our approach in comparison with the ones that need a pre-trained object detector.

Training and testing with frames that are too similar are avoided as follows: for training just 1 out of 10 frames is considered, while for testing 1 out of 20 for CLEAR06_PV and FishTank datasets, 1 out of 15 for the HoustonZoo_Rhino sequence.

The AP performance of our approach, as well as the number of training frames, GT frames for test, and number of clusters employed is shown in Table 2. *Ini* values correspond to the AP for the first estimation of the latent SVM optimization where latent variables have not been correctly estimated yet. *Final* values correspond the final AP once the iterative optimization is finished. In Fig. 8 how one appearance model changes during the iterations

[11] `http://pascallin.ecs.soton.ac.uk/challenges/VOC/`

[12] `http://pascal.inrialpes.fr/data/human/`

[13] `http://www.science.uva.nl/research/isla/downloads/pedestrians/`

of the global optimization procedure for CLEAR06_PV sequence can be seen.

Note that the AP performance obtained in Table 2 (81.5) is different from the one presented in Fig. 6 (76.3) because we use a different amount of training images. In the second case $5,700$, those that come provided with bounding box annotations, and $13,167$ in the first case. This shows that, in fact, when more data is feasible, detection performance can be improved by learning with longer sequences.

Fig. 9 shows the trained models and our detection results for each sequence. Lastly, Fig. 10 shows more detection results for all the sequences, where people, cars, fish, rhinos are correctly detected respectively for each sequence.

Discussion First, some remarks on the computational complexity and the execution time for a possible real-time application are discussed. Later, a discussion of the limitations of the current approach is presented. In terms of computational complexity, the motion segmentation has a cost that is linear in the number of the pixels in the image. The specific implementation used in the experiments [28] runs at around 3 fps in matlab. However, a faster reimplementation or the use of other algorithm [29, 30, 31] can lead to more than real-time performance. Also, even if the image is at very high resolution, as we need just a rough segmentation of the moving objects to initialize the learning algorithm, real time performance can be easily obtained by subsampling the image. For detection, [5] runs at around 0.1 fps. The coarse-to-fine detector [36] that has been used in the experiments already runs around 10 times faster. Still, there is room for further improvements until real-time performance is achieved, as recently shown in [41, 42, 43]. Finally

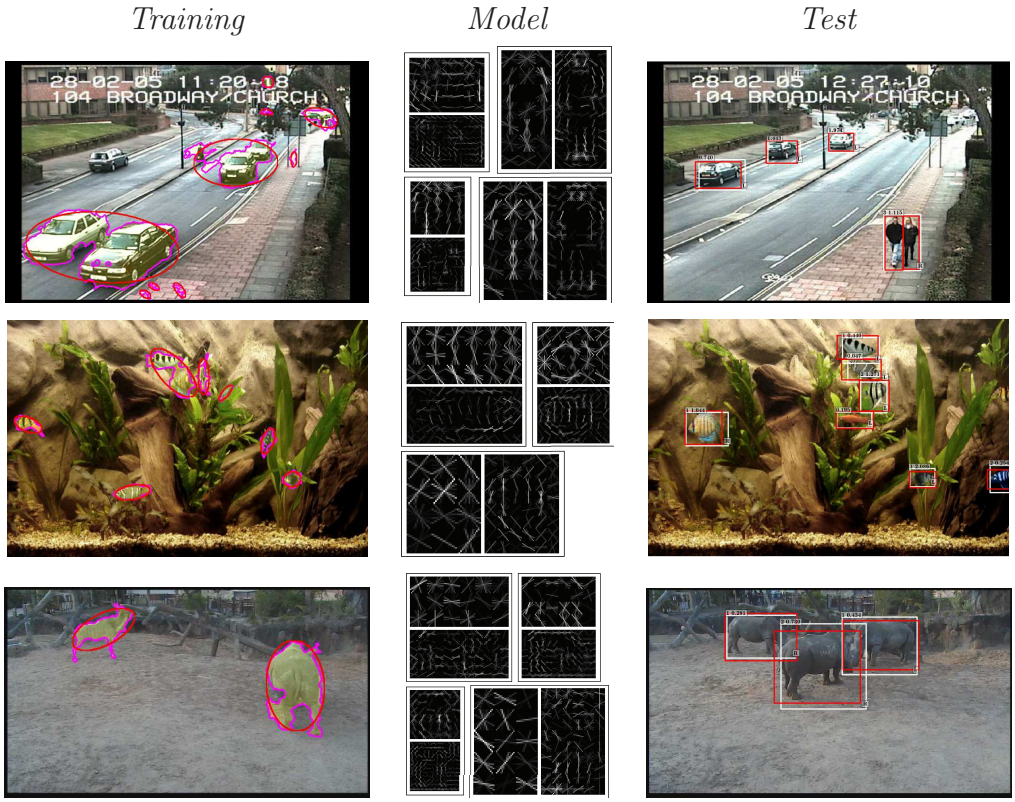


Figure 9: Experimental Results using CLEAR06_PV, FishTank, and HoustonZoo_rhino databases. First column shows one frame from the motion segmentation, the second column shows the learned object models, and the third column shows our detection results. The red bounding boxes are the ground truth annotations while white bounding boxes are our algorithm detections, thereby showing people, cars, fish, rhinos are correctly learned and detected respectively for each sequence.

the last step for a real-time application is a fast on-line training. This is easily achievable with stochastic gradient descent whose computational complexity is independent on the number of samples [39].

Now, some advantages and drawbacks of our approach are proffered. The presented approach has some advantages and drawbacks. The main advan-

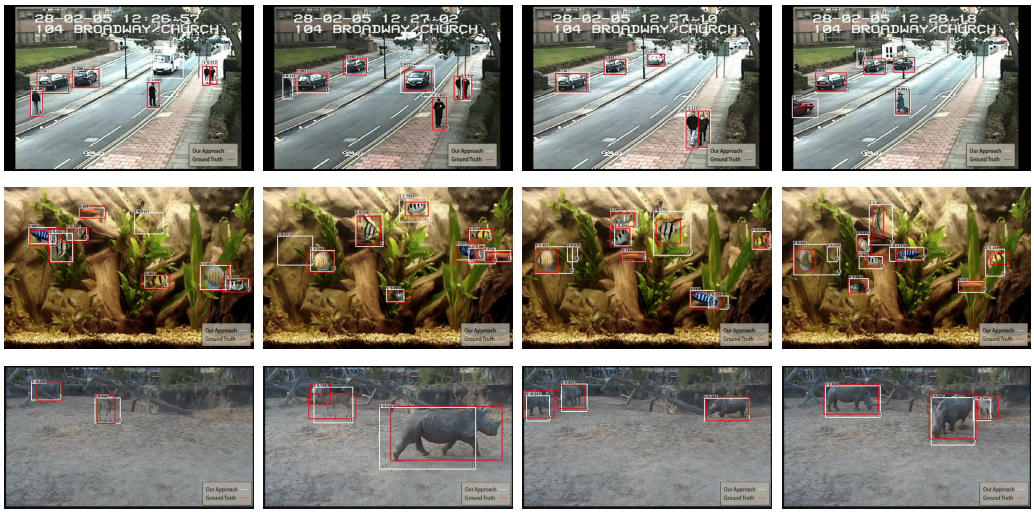


Figure 10: Detection Results using our approach in CLEAR06_PV, FishTank, and HoustonZoo_rhino databases. The red bounding boxes are the ground truth and the white bounding boxes are our detections, thereby showing that people, cars, fish, rhinos are correctly detected.

tage of our approach is that it does not need any type of ground truth annotations of the objects bounding box and does not assume any pre-determined category; it can learn all the objects that appear in the scene in an unsupervised manner. However, in contrast with generic object detectors that are trained for any possible view, our approach cannot learn a specific view of an object that has not appeared in the training of the approach. Although, in certain situations this is a disadvantage, it is also a way to specifically tune the detector to the real content of the scene, avoiding learning views or objects that will never appear and that can be a source of false detections.

5. Conclusions

In this paper we propose a new method for the detection of unknown and multiple moving objects in video sequences. It uses motion cues for an initial estimation of the object location thus avoiding annotation tasks. Subsequently, the system learns an appearance model of multiple clusters using a global discriminative optimization that refines the initial object estimations. Our proposal is unsupervised since there is no need of hand-labelled annotations, works with unknown information since there is no need of any a-priori information of the scene, and is able to deal with multiple appearances while learning multiple foreground regions at the same time.

This work creates an initial framework where multiple lines of future work can be taken. At the moment the iterative learning procedure is off-line, when all the data is already present. A possible extension of the work would be to modify the algorithm in such a way that it is possible to run it on-line.

Currently in the experimental part we have tested the proposed methodology using motion data captured from a static camera using background subtraction. However, it would be possible to extend the procedure to videos obtained from moving cameras. In this case, motion cues could be provided from optical flow computation, but the motion clustering and detector learning steps would be quite similar.

Acknowledgements

This work was partially funded by the DPI2013-42458-P and TIN2012-39051.

References

[1] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, PR 36 (3) (2003) 585–601.

[2] T. B. Moeslund, A. Hilton, V. Kruger, A survey of advances in vision-based human motion capture and analysis, CVIU 104 (2006) 90–126.

[3] Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, C. Rosenberger, Comparative study of background subtraction algorithms, Journal of Electronic Imaging 19 (3) (2010) 1–12.

[4] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. CVPR, 2005.

[5] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, PAMI 32 (9) (2010) 1627–1645.

[6] K. Toyama, J.Krumm, B.Brumitt, B.Meyers, Wallflower: Principles and practice of background maintenance, in: Proc. ICCV'99, Vol. 1, Kerkyra, Greece, 1999, pp. 255–261.

[7] A. Levin, P. Viola, Y. Freund, Unsupervised improvement of visual detectors using cotraining, in: CVPR03, Vol. 1, Nice, France, 2003, pp. 626–633.

[8] S. A. O. Javed, M. Shah, Online detection and classification of moving objects using progressively improving detectors, in: IEEE CVPR'05, 2005, pp. 696–701.

[9] V. Nair, J. Clark, An unsupervised, online learning framework for moving object detection, in: IEEE CVPR'04, 2004, pp. 317–324.

[10] B. Wu, R. Nevatia, Improving part based object detection by unsupervised, online boosting, in: IEEE CVPR'07, 2007, pp. 1–8.

[11] Z. Kalal, J. Matas, K. Mikolajczyk, P-n learning: Bootstrapping binary classifiers by structural constraints, in: IEEE CVPR'10, 2010, pp. 1–8.

[12] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE TPAMI 34 (7) (2012) 1409–1422.

[13] B. Babenko, M. Yang, S. Belongie, Visual tracking with online multiple instance learning, in: CVPR09, Vol. 1, Miami,FL, 2009, pp. 983–990.

[14] K. Zhang, H. Song, Real-time visual tracking via online weighted multiple instance learning, PR 46 (1) (2013) 397–411.

[15] X. Lu, Y. Yuan, P. Yan, Robust visual tracking with discriminative sparse learning, PR 46 (7) (2013) 1762–1771.

[16] A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, M. Shah, Visual tracking: An experimental survey, IEEE TPAMI.

[17] K. Ali, D. Hasler, F. Fleuret, Flowboost - appearance learning from sparsely annotated video, in: IEEE CVPR'11, 2011, pp. 1433–1440.

[18] A. Prest, C. Leistner, J. Civera, C. Schmid, V. Ferrari, Learning object class detectors from weakly annotated video, in: CVPR12, Providence, RI, 2012, pp. 3282–3289.

[19] M. Villamizar, J. Andrade-Cetto, A. Sanfeliu, F. Moreno-Noguer, Bootstrapping boosted random ferns for discriminative and efficient object classification, PR 45 (9) (2012) 3141–3153.

[20] X. Wang, G. Hua, T. Han, Detection by detections: Non-parametric detector adaptation for a video, in: CVPR12, Providence, RI, 2012, pp. 350–357.

[21] Y. Yang, G. Shu, M. Shah, Semi-supervised learning of feature hierarchies for object detection in a video, in: CVPR13, Portland, OR, 2013, pp. 1650–1657.

[22] G. Shu, A. Dehghan, M. Shah, Improving an object detector and extracting regions using superpixels, in: CVPR13, Portland, OR, 2013, pp. 3721–3727.

[23] M. Hoai, L. Torresani, F. D. la Torre, C. Rother, Learning discriminative localization from weakly labeled data, PR 47 (3) (2014) 1523–1534.

[24] H. Celik, A. Hanjalic, E. A. Hendriks, A framework for unsupervised training of object detectors from unlabeled surveillance video, JAISE 3 (3) (2011) 213–235.

[25] J. Marín, D. Vázquez, D. Gerónimo, A. López, Learning appearance in virtual scenarios for pedestrian detection, in: IEEE CVPR'10, 2010, pp. 137–144.

[26] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormaehlen, B. Schiele, Learning people detection models from few training samples, in: CVPR, Colorado Springs, USA, 2011, pp. 1473–1480.

[27] H. Celik, A. Hanjalic, E. A. Hendriks, Unsupervised and simultaneous training of multiple object detectors from unlabeled surveillance video, CVIU 113 (10) (2009) 1076–1094.

[28] I. Huerta, A. Amato, F. X. Roca, J. Gonzàlez, Exploiting multiple cues in motion segmentation based on background subtraction, Neurocomputing 100 (6) (2013) 183–196.

[29] A. Prati, I. Mikic, M. Trivedi, R. Cucchiara, Detecting moving shadows: Algorithms and evaluation, IEEE TPAMI 25 (7) (2003) 918–923.

[30] Z. Zivkovic, F. Heijden, Efficient adaptive density estimation per image pixel for the task of background subtraction, PRL 27 (7) (2006) 773–780.

[31] P. Jodoin, M. Mignotte, J. Konrad, Statistical background subtraction using spatial cues, Circuits and Systems for Video Technology 17 (12) (2007) 1758–1763.

[32] Y. Chen, C. Chen, C. Huang, Y. Hung, Efficient hierarchical method for background subtraction, PR 40 (10) (2007) 2706–2715.

[33] A. Leone, C. Distante, Shadow detection for moving objects based on texture analysis, PR 40 (4) (2007) 1222–1233.

[34] I. Huerta, M. Holte, T. Moeslund, J. Gonzàlez, Detection and removal of chromatic moving shadows in surveillance scenarios, in: ICCV2009, Kyoto, Japan, 2009.

[35] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, P. Ishwar, Changede-

tection.net: A new change detection benchmark dataset, in: CDW12 at CVPR12, Providence, RI, 2012, pp. 1–8.

[36] M. Pedersoli, A. Vedaldi, J. Gonzàlez, A coarse-to-fine approach for fast deformable object detection, in: CVPR, 2011.

[37] A. Vedaldi, A. Zisserman, Structured output regression for detection with partial occulsion, in: Proc. NIPS, 2009.

[38] A. Yuille, A. Rangarajan, A. L. Yuille, The concave-convex procedure (cccp, in: Advances in Neural Information Processing Systems 14, MIT Press, 2002.

[39] Y. Singer, N. Srebro, Pegasos: Primal estimated sub-gradient solver for svm, in: In ICML, 2007, pp. 807–814.

[40] M. Everingham, A. Zisserman, C. Williams, L. V. Gool, The pascal visual obiect classes challenge 2007 (voc20067) results, Tech. rep. (2007).

[41] M. Pedersoli, J. Gonzalez, X. Hu, X. Roca, Toward real-time pedestrian detection based on a deformable template model, IEEE Intelligent Transportation Systems 15 (1) (2014) 355–364.

[42] J. Yan, Z. Lei, L. Wen, S. Z. Li, The fastest deformable part model for object detection, in: CVPR, 2014.

[43] M. Sadeghi, D. Forsyth, 30hz object detection with dpm v5, in: ECCV, 2014.