Perspective Multiscale Detection and Tracking of Persons

Marcos Nieto[1], Juan Diego Ortega[1], Andoni Cortes[1], and Seán Gaines

Vicomtech-IK4, Paseo Mikeletegi 57, San Sebastian, Spain
mnieto@vicomtech.org

Abstract. The efficient detection and tracking of persons in videos has widrespread applications, specially in CCTV systems for surveillance or forensics applications. In this paper we present a new method for people detection and tracking based on the knowledge of the perspective information of the scene. It allows alleviating two main drawbacks of existing methods: (i) high or even excessive computational cost associated to multiscale detection-by-classification methods; and (ii) the inherent difficulty of the CCTV, in which predominate partial and full occlusions as well as very high intra-class variability. During the detection stage, we propose to use the homograhy of the dominant plane to compute the expected sizes of persons at different positions of the image and thus dramatically reduce the number of evaluation of the multiscale sliding window detection scheme. To achieve robustness against false positives and negatives, we have used a combination of full and upper-body detectors, as well as a Data Association Filter (DAF) inspired in the well-known Rao-Blackwellization-based particle filters (RBPF). Our experiments demonstrate the benefit of using the proposed perspective multiscale approach, compared to conventional sliding window approaches, and also that this perspective information can lead to useful mixes of full-body and upper-body detectors.

Keywords: Object Detection, Machine Learning, Person Detection, Person Tracking, Homography, Camera Calibration

1 Introduction

People detection and tracking in video sequences using computer vision methods has become a hot topic in the related scientific community due to its potential in CCTV applications like surveillance or forensics. Significant progresses have been made, specially in the object detection-by-classification approaches [16], object tracking using appearance [8, 2], and also to extract semantic information from the sequence [14, 11].

Detection-by-classification is the most promising family of techniques, using the sliding window technique [15], which consists on exhaustively scanning the whole image searching for objects at different scales or levels. Although this methodology is adequate for general problems, it is too much exhaustive for

CCTV applications. On the one hand they may require low computational cost (to analyze many video files in large installations), but on the other hand are typically static enough to use useful prior information of the scene.

Using contextual information is a way enhance such approaches [4]. We propose to exploit the perspective information of the scene to determine the maximum and minimum expected size of persons at different locations in the images and use them to reduce the number of levels to be used. In the context of CCTV systems, it is broadly accepted an initial set-up or installation stage in which prior information can be retrieved using an appropriate GUI. We have observed that the generation of the perspective information with a GUI takes only about 1-2 minutes and might allow for significant speedups (in our experiments from 30% to 80% depending on the perspective), which can result on more video sequences processed with the same computer or less time to process a given video file, and also better results in terms of false positives using the same detectors.

Most related works focus on the detection of full-body [12], upper-body [16, 10, 13], or heads [2], according to the type of targer application. We propose to use both type of detectors and combine them using the perspective. On the one hand, full-body detections are really distinctives when the person is seen completely in the image. On the other hand, upper-body detections are useful in scenarios in which partial occlusions happen.

However, using two detectors imply more computational load, and also the necessity to handle more false positives. The use of the perspective information help us to control these two problems. Particularly, we combine these two type of detections so that (i) each upper-body detection generates a full-body estimation; and (ii) the location of detections is projected into the plane to filter out false positives by checking if its size is between the expected minimum and maximum sizes of persons.

To complete our contributions, we apply a tracking approach based on the Rao-Blackwellization Data Association Particle Filter (RBDAPF) [3] that provides the required temporal coherence to detections by linking detections through time and generating predictions according to object appearances.

The results presented at the end of the paper demonstrate the benefits of using the proposed approach, specially the usage of the perspective of the scene, plus the combination of detectors in surveillance sequences (for this purpose we have used the available dataset from Oxford Active Vision group [2]).

2 Approach overview

Figure 1 illustrates the modular architecture of the proposed approach. Details of each module are given in the next sections. The first step is the generation of the perspective information, that can be done offline, and it is only done once. This information is encoded as the projection matrix P, which is composed by the camera calibration matrix K and the relative pose of the camera, R and $\mathbf{t}$ with respect to a coordinate system placed in the dominant plane of the scene.

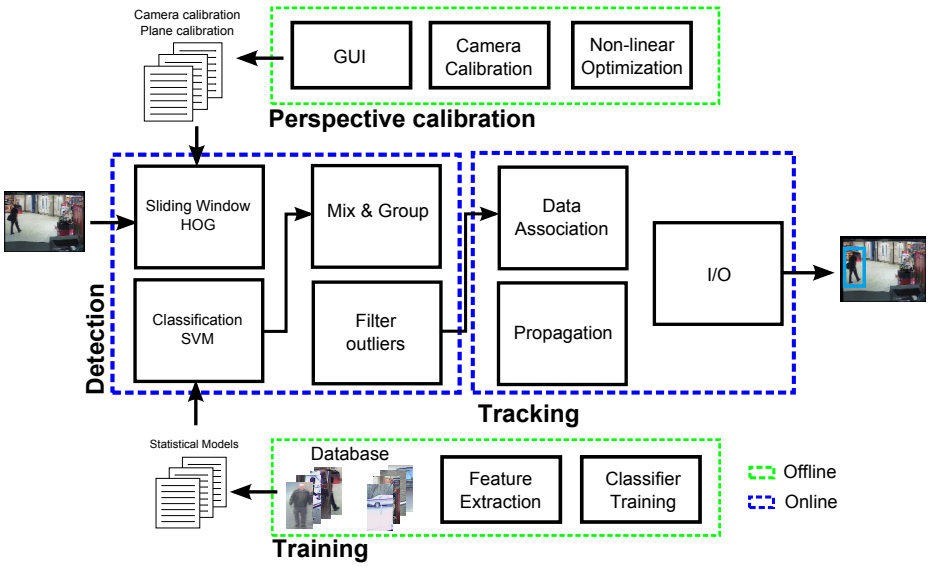

Fig. 1. Perspective multi-scale using detection-by-classification.

The full-body and upper-body detectors load the respective SVM models, and the multiscale sliding window parameters are set according to the perspective. The detector then detects candidate regions of the images likely containing full-bodies and upper-bodies. These regions are mixed (upper-bodies can be upgraded to full-bodies using approximate human dimensions), and filtered (many false positives are removed applying the perspective restriction which determines the expected sizes of human beings at different positions in the image). Also, for long sequences, conventional background substraction methods could be applied, and only those regions which contain a certain amount of foreground pixels are considered as valid detections (in this paper we have not included this module because we wanted to focus on the detectors alone).

As a result, a set of detections is obtained and fed to the tracker, which associates the detections with the tracks. Entering and exiting persons or tracks are handled by the tracker, which creates a new track when detections not associated to existing tracks show time coherence (e.g. appear consecutively during a number of frames), and deletes an existing track when it is not associated to detections during a certain amount of frames.

3 Perspective Multiscale Detection

The calibration of the camera and the computation of its relative pose with respect to the ground plane offers valuable information for the detection of persons in images under the hypothesis that there is a dominant ground plane in the scene and persons are on it. In this work we propose to formalize the ex-

ploitation of the perspective of the scene by means of computing the projection matrix and defining a multi-scale detection approach according to it.

3.1 Perspective Multiscale

Figure 2 illustrates the difference between the typical use of multiscale sliding window detectors and our proposed approach. The simplest way to proceed is to run a multi-scale scanning of the image evaluating each image patch with the classifier in order to determine the presence of objects in the image. Starting from the smallest size, which is determined by the window size parameter of the SVM model (e.g. 64×128 pixels for instance), L copies of the images are created, down-scaled by a factor that is typically 1.05 or 1.1 in the hope that this exhaustive scan will likely find small, medium and large objects; we have called this method *brute-force multiscale*. The total amount of evaluations of image patches against the SVM classifier is given by $N_e = \sum_{i=1}^{L} \left(\frac{W_i}{s} - \left(\frac{w}{s} - 1\right)\right)\left(\frac{H_i}{s} - \left(\frac{h}{s} - 1\right)\right)$, where $W_i \times H_i$ is the size of the image in pixels at each level, s is the window stride, and $w \times h$ is the size of the model. For instance, for a 1920×1080 image with a 64×128 model and $s = 16$, and $L = 10$, then $N_e = 144880$.

In our approach, since the projection matrix P is known, we can reproject a human model to any position in the scene (see Figure 3) and determine the smallest and largest sizes of it in a region of interest. Therefore we can know the exact scale we have to apply to the multilevel scan procedure to start from the smallest possible detections to the largest. In our experiments we have observed that we can reduce L to 3-5 levels to achieve similar results that the brute-force approach using $L = 10$. Note that the minimum number of scales we propose to use is 2, one corresponds to the smallest person size, and the other to the largest. Any additional scale is an intermediate scale between these two sizes.

The main difference between these alternatives is that the perspective analysis of the scene focuses significantly the effort of the classifier resulting in a much more efficient scan of the image.

3.2 Ground Plane Calibration

The calibration of the scene required to apply the proposed perspective multi-scale approach can be obtained in a single-step process. The user must introduce 4 points in the image that corresponds to a rectangle in the ground plane of the scene, plus the longitudinal and transversal distances between the points.

This information is enough to compute the homography H between the image plane (in pixels) and the ground plane (in metric units) using the DLT (Direct Linear Transform) algorithm [7].

The coordinate system in the ground plane can be selected such that it is defined by $Z = 0$. In such situation, the projection of a point $\mathbf{X} = (X, Y, Z, 1)^\top$ into a image point $\mathbf{x}$ yields:

$$\mathbf{x} = K(R|\mathbf{t})\mathbf{X} = K(\,\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3\ \mathbf{t}\,)\left(\,\mathrm{X}\ \mathrm{Y}\ 0\ 1\,\right)^\top \tag{1}$$

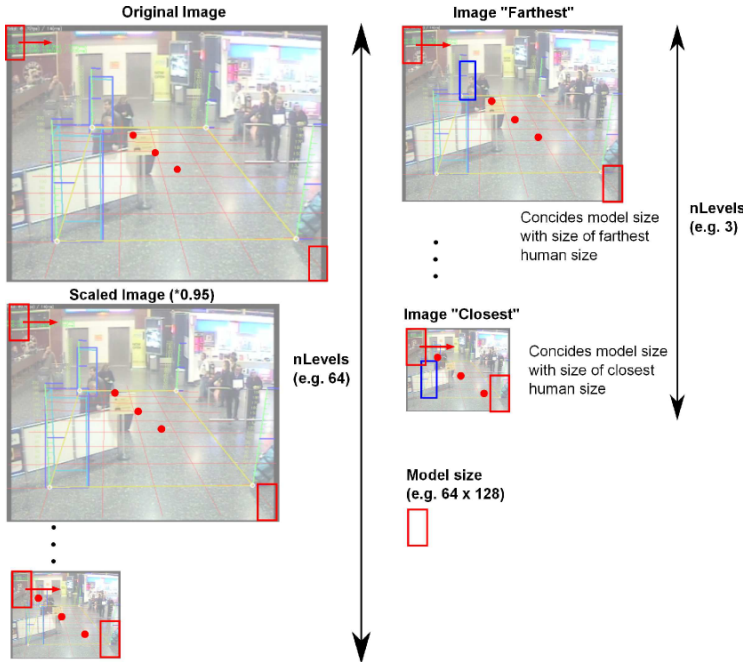

Fig. 2. Sliding window approaches: (left) Brute-force and (right) Perspective multi-scale.

and therefore:

$$\mathbf{x} = K(\,\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{t}\,)\,(\,\mathrm{X}\ \mathrm{Y}\ 1\,)^{\top} \tag{2}$$

which is a 3×3 homography between the image and world plane points:

$$H = K(\,\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{t}\,) \tag{3}$$

As expected, the homography matrix contains all the information about the intrinsics and extrinsics parameters of the projection process. We use the following procedure to estimate the values of K, R and $\mathbf{t}$:

First, the calibration matrix can be assumed to have 1-DoF with the principal point as the center of the image so the only unknown is the focal length that can be computed solving the following expression with SVD (Singular Value Decomposition):

$$\begin{pmatrix} h_{0,0}h_{0,1} + h_{1,0}h_{1,1} \\ h_{0,0}^2 - h_{0,1}^2 + h_{1,0}^2 - h_{1,1}^2 \end{pmatrix} \mathbf{x} = \begin{pmatrix} -h_{2,0}h_{2,1} \\ -h_{2,0}^2 + h_{2,1}^2 \end{pmatrix} \tag{4}$$

as $f = \|x_0\|^{-\frac{1}{2}}$, where x_0 is the first eigenvalue of $\mathbf{x}$.

Second, given this initial value of K, we can calibrate the expression of the homography:

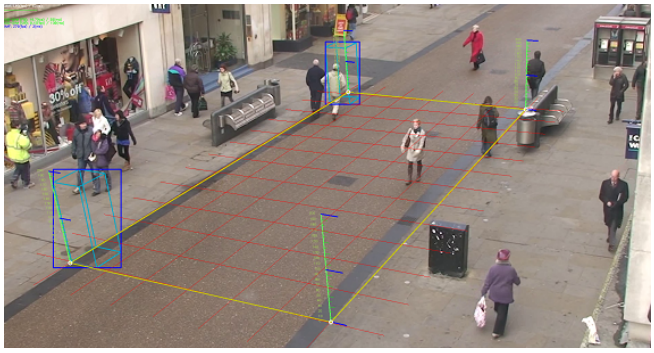

Fig. 3. Example use of the GUI for calibrating a typical CCTV scene. The rulers and the projected boxes help to fit the expected sizes of pedestrians.

$$K^{-1}H = (\,\mathbf{p}_1\ \mathbf{p}_2\ \mathbf{p}_3\,) \tag{5}$$

This way, once we have computed and calibrated the homography we can extract the rotation and traslation from the columns of the resulting matrix. Please note that since these are homogeneous matrices it is necessary to normalize the columns of the matrix in order to get the vectors: $\mathbf{r}_1 = \frac{\mathbf{p}_1}{\|\mathbf{p}_1\|}$, $\mathbf{r}_3 = \frac{\mathbf{p}_2}{\|\mathbf{p}_2\|}$ and $\mathbf{r}_2 = \mathbf{r}_1 \times \mathbf{r}_3$.

Finally, we can use a refinement step that optimizes simultaneously the reprojection error over the set of parameters given by K, R and $\mathbf{t}$. We propose to use the Levenberg-Marquardt non-linear optimization method for which many implementations can be found (e.g. lmfit-3.5, 2013, by Joachim Wuttke, http://apps.jcns.fz-juelich.de/lmfit).

3.3 Model reprojection

The obtained projection matrix $P = K[R|\mathbf{t}]$ can be used to project 3D points into the image. Therefore, we can roughly model a person (more specifically a bounding volume around a person) as a parallelepiped and project it at the closest and farthest point of the defined quadrilateral used for calibration. The projection of a parallelepiped in an image is a convex polygon whose bounding box can be easily computed and used to determine the sizes of the persons that will configure the perspective multiscale approach. Figure 3 shows the two projections of the box model in an image and the corresponding bounding boxes.

4 Detection

Discriminative learning methods have been used in the majority of works referred to object and person detection. Within this kind of methods, variations of SVM and Adaboost algorithms stand out in the literature. The main idea underlying

Fig. 4. Full-body and upper-body detectors can be combined to achieve better resutls: (green) upper-body detections, (blue) full-body detections.

Fig. 5. The perspective can be used to filter out detections that do not represent coherent human sizes.

the training of classifiers is to find a model which could map the input feature vector to a set of output labels. The training stage involves the application of supervised training algorithms to a set of feature vectors extracted from the image database. This database must have positive images (e.g. "person") and negative images (e.g. "non-person").

In this work we have used two detectors, which correspond to full-body person and upper-body person, both using linear SVM and HOG features [5] due to its outstanding capabilities to detect persons in images. Specifically, we have used the Daimler full-body detector [12] available as an SVM file within OpenCV-2.4.5, which corresponds to a dataset of approximately 25k images. For the upper-body we have trained our own detector using a database of approximately 1k positive and negative images taken from TRECVID2013 training dataset, which contain images from 5 different cameras. Although this dataset is not large yet, it has been useful to comprobe that even a simple upper-body detector can help.

The full-body detections and extended upper-bodies as full-bodies (see Figure 4) are filtered to group detections which have significant overlap. Also, the detections are filtered according to perspective: a given detection is assumed to correspond to a person in the ground plane, so that it can be reprojected to that plane, and the approximate width and height can be computed. Figure 5 illustrates examples of detections filtered out.

5 Tracking

Tracking is the stage in which intra-frame detections are analyzed through time in order to group them in time and create an inter-frame entity called track. Also, tracking helps to alleviate the problems associated to detection-by-classification methods. Namely, the detections tend to generate noisy, incorrect, missing, and time sparse observations.

First, the tracking methods act as filters, so that noise can be reduced using appropriate models (such as bivariate Gaussian models). Second, tracking methods work on a two-steps fashion: observe and predict. The observation stage reads the observations coming from the detectors and associates them to the existing objects. Incorrect detections are mitigated using association schemes as the one described in the next sections. Also, missing detections can be handled using the prediction step of the filter, in which each object, given its estimated dynamics (e.g. velocity and acceleration) is projected onto the next frame. Therefore, the nature of tracking methods deal well with the drawbacks imposed by detectors and thus a combination of these two types of methods provide a good solution for object detection in video sequences.

5.1 Rao-Blackellized Data Association Particle Filter

The RBDAPF [6] has a special structure that allows to analytically compute the object magnitudes (positions, size, etc.) while the data associations between tracks and detections are approximated by a sampling approach [3]. This method defines a state vector $\mathbf{x}_t = \{\mathbf{x}_t^m, \mathbf{x}_t^a\}$ at time t, where $\mathbf{x}_t^m$ contains the 2D object magnitudes, and $\mathbf{x}_t^a = \{x_k^{a(j)}\}$ encodes the data associations between the tracks and the detections. The j-th data association component relates the j-th detection with a track $x_k^{a(j)} = id0$ or with clutter $x_k^{a(j)} = idC$, where $id0$ is a unique identifier of the track, and idC is the generic identifier for clutter. The mixed and filtered full-body detections are represented by the random variable $\mathbf{z}_t$.

As a Bayesian inference method, the RBDAPF aims to provide an estimate of the posterior density function. The idea of this technique is that solving analytically part of the state vector, and leaving only the non-linear part to the sampling approach gives a more accurate representation of the posterior probability. Intuitively, the variance is smaller because some variables are computed exactly and the non-linear dimensionality is lower that the dimension of the complete state-vector. The Rao-Blackwellization of the posterior density leads to the following expression:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = p(\mathbf{x}_t^m, \mathbf{x}_t^a|\mathbf{z}_{1:t}) = p(\mathbf{x}_t^m|\mathbf{z}_{1:t}, \mathbf{x}_t^a)p(\mathbf{x}_t^a|\mathbf{z}_{1:t}) \tag{6}$$

where $p(\mathbf{x}_t^m|\mathbf{z}_{1:t}, \mathbf{x}_t^a)$ is assumed to be conditionally linear Gaussian, and therefore, with an analytical expression, given by the Kalman filter.

The data association posterior density $p(\mathbf{x}_t^a|\mathbf{z}_{1:t})$ can be expressed as:

$$p(\mathbf{x}_t^a|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_t^a)p(\mathbf{x}_t^a)}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} \tag{7}$$

where $p(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_t^a)$ is the data association likelihood, $p(\mathbf{x}_t^a)$ is the data association prior, and $p(\mathbf{z}_t|\mathbf{z}_{1:t-1})$ the normalization constant.

The data association prior determines the possible associations between tracks and detections, for which several criteria can be applied, such as: (i) each track can be associated only with one or none of the detections; (ii) each detection can be associated only to one track, although several detections can be associated to the clutter object. In this work we are using as likelihood function the Euclidean distance plus a constant clutter model. The data association posterior is approximated using importance sampling [1].

5.2 Data Association Filter

In practice, the use of the RBDAPF imposes handling very carefully entering and exiting objects. The reason is that the data association matrix is sampled using an importance sampling algorithm, and therefore, there are many association hypotheses, and the detections may be associated to different tracks for each sample or hypothesis. Even in the case that the posterior distribution this way defined shows unimodal behaviour (i.e. point-wise estimators can be applied to the set of samples), the different history of associations between observations and tracks at each sample is problematic at the time of considering input and output objects. A track is labeled as exiting the scene if it is not associated to new detections for a period of time (e.g. 5 frames). In that case, the object is considered to have left the scene and removed from the estimation (delete event). On the contrary, observations that are not associated to any existing objects are initially considered as clutter (i.e. erroneous measurements). In our work we create a new clutter object associated to that new observation just in case it receives new observations in the subsequent frames. If this is the case for a number of frames (e.g. 3 frames), the clutter object is upgraded to track, and added to the list of tracks (new event). These events are related to the history of associations and therefore each sample of the RBDAPF filter has its own association history that might lead to different new/delete events. Although filters like RBDAPF can be upgraded to consider samples with different numbers of objects inside it (by means of adding a dimension that spans the number of objects in the scene [9]), the complexity of the filter increases significantly, and the generation of a point-wise estimate from the posterior density function might become a tough task.

We define the Data Association Filter (DAF), which works exactly the same as the RBDAPF but selecting only the best hypothesis (Maximum A Posterior, MAP) during the data association step. Since this is a single sample, the point-wise estimate can be directly retrieved from its components.

6 System Test and Discussion

To evaluate the improvements derived from the use of the proposed perspective multiscale method we have used the TownCentre sequence made available by

the Active Vision Group from Oxford [2] (1920×1080, 4500 frames, with 71460 persons labeled).

We wanted to evaluate the following hypotheses: (i) using the perspective multiscale method the optimum parameters for the detector are found automatically; (ii) our approach reduces significantly the number of SVM evaluations required to achieve similar results than brute force multiscale; (iii) using a combination of full-body and upper-body detectors provide better results; (iv) perspective allows also filtering out numerous false positives; (v) the DAF tracking stage helps to increase the performance of intra-frame detectors thanks to its prediction capabilities.

First, Table 1 shows the performance of the proposed perspective multiscale ($L = 3$ and $L = 5$) at different stages in terms of true positives (TP), false positives (FP), false negatives (FN), and the related Recall (R), Precision (P) and F-measure. In that sense, we have defined a detection to be a TP if the overlap it has with a ground-truth rectangle is larger than half their non-overlapping union area (i.e. overlap is above 50%).

	$L = 3$						$L = 5$					
	TP	FP	FN	R	P	F	TP	FP	FN	R	P	F
FB	21521	358	49926	0.301	0.984	0.461	27725	505	43722	0.388	0.982	0.556
UB	3395	23	68052	0.048	0.993	0.091	3395	23	68052	0.047	0.993	0.091
FBUB	23339	381	48108	0.327	0.984	0.490	29075	528	42372	0.407	0.982	0.575
FBUB*	20099	239	51348	0.281	0.988	0.438	25659	307	45788	0.359	0.988	0.527
DAF	27106	503	44341	0.379	0.982	0.547	32485	642	38962	0.455	0.981	0.621

Table 1. Comparison of the performance of the different combination of detectors (FB: full-body, UB: upper-body, FBUB: both mixed, FBUB*: mixed and filtered according to perspective) and tracker (DAF).

As expected, the usage of UB joint with FB increases the performance of FB alone, even when the UB by itself does not reach good values. After filtering (FBUB*) we can see that many FP have been removed (although TP has slightly decreased, possibly due to the elimination of detections that did not fit exactly to the ground truth). The application of the tracker dramatically enhances these numbers, since a significant number of FN become TP thanks to the prediction capabilities of the filter. Better values are obtained for $L = 5$.

Table 2 compares the proposed scheme with the brute traditional force multiscale method with different values of levels L. We can see that the perspective multiscale gives good values with very few levels. Actually, we have found that the performance is stabilized at 3-5 levels, depending on the sequence, and adding more levels shows no significant improvement while increases the number of operations to carry out. Another remarkable advantage of our approach is that it automatically determines the optimum sizes of the images inside the multiscale pyramid. We deem this feature very practical because otherwise (with brute force

Perspective Multiscale						Brute-force Multiscale				
L	$W \times H$	N_e	R	P	F-measure	L	N_e	R	P	F-measure
3	1920×1080	75929	0.03	0.08	0.05	2	33051	0.15	0.96	0.27
6	1920×1080	117471	0.12	0.18	0.14	3	46226	0.30	0.98	0.46
10	1920×1080	144880	0.47	0.48	0.48	4	60105	0.29	0.96	0.45
20	1920×1080	162968	0.66	0.58	0.62	5	74141	95	263	0.55

Table 2. Results of Perspective Multiscale and Brute-force multiscale

multiscale), the optimum values for L, and *scale* must be found by try-and-error. For instance, Table 2 shows that for these large images, the multiscale approach can not find good performance given the small size of the model unless using a large number of scales. Therefore, the number of levels that best work can only be found launching and evaluating the detector spanning different values for L and scale which can take time and requires the existence of ground truth and evaluation tools.

In the TownCentre video, with the perspective we have computed, the farthest person is represented by a 101×184 bounding box, such that using a 64×128 SVM model, the largest image to be scanned is approximately 1805×1014. Therefore, the reduction of computational load is noteworthy: from 75929 to 46226 (a reduction of 39%). In the case of sequences with lowest perspective, where persons are not so small, (such as those in CAM1 or CAM3 of TRECVID dataset), the gain can reach much largest values, up to $80\% - 90\%$.

7 Conclusions

In this paper we have presented a methodology to apply contextual perspective information of the scene to traditional detection-by-classification schemes that use sliding window scanning. The multiscale procedure typically implies massive amounts of comparisons between windows of the image with a certain model, resulting in variable, a priori unknown, and possibly excessive computational load to achieve good results. Our scheme automatizes the sliding window technique, so that when the perspective information is injected into the solution, the optimum values of the parameters that govern the multiscale approach are found. The experiments carried out show that we can get results comparable to those of traditional (brute force) multiscale with only 3-5 levels, which can lead to computational loads reduction between 30% to 80% depending on the perspective of the scene (more reductions are achieved for images where persons are imaged larger). The addition of a tracking stage based on the Rao-Blacwellization concept helps as well to enhance the detection rates, since the nature of detection-by-classification is often sparse and noisy.

8 Acknowledgements

This work has been partially supported by the European project SAVASA (grant agreement number 285621) under the 7th Marco Framework, and by the program ETORGAI of the Basque Government with the BERRITRANS project.

References

1. Arulampalam, M.S., Maskell, S., Gordon, N.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on Signal Processing 50, 174–188 (2002)
2. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: in Proc. Computer Vision and Pattern Recognition. pp. 3457–3464 (2011)
3. del Blanco, C.R., Jaureguizar, F., Garcia, N.: An advanced bayesian model for the visual tracking of multiple interacting objects. EURASIP Journal on Advances in Signal Processing 130 (2011)
4. Carbonetto, P., de Freitas, N., Barnard, K.: A statistical model for general contextual object recognition. In: In European Conference on Computer Vision (2004)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: In CVPR. pp. 886–893 (2005)
6. Doucet, A., Gordon, N.J., Krishnamurthy, V.: Particle filters for state estimation of jump markov linear systems. IEEE Transactions on Signal Processing 49, 613–624 (1999)
7. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edn. (2004)
8. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. Pattern Analysis and Machine Intelligence 6(1) (2010)
9. Khan, Z., Balch, T., Dellaert, F.: Mcmc-based particle filtering for tracking a variable number of interacting targets. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 2005 (2005)
10. Li, M., Zhang, X., Huang, K.Q., Tan, T.N.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: in Proc. International Conference on Pattern Recognition (2008)
11. Little, S., Jargalsaikhan, I., Clawson, K., Li, H., Nieto, M., Direkoglu, C., O'Connor, N., Smeaton, A., Scotney, B., Wang, H., Liu, J.: An information retrieval approach to identifying infrequent events in surveillance video. In: In ACM International Conference on Multimedia Retrieval. pp. 223–230 (2013)
12. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(11), 1863–1868 (2006)
13. Park, L.J., Moon, J.H.: Exploiting global self similarity for head-shoulder detection. World Academy of Science, Engineering and Technology 0076 (2013)
14. Thonnat, M.: Semantic activity recognition. In: 18th European Conference on Artificial Intelligence. pp. 3–7 (2008)
15. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
16. Zeng, C., Ma, H.: Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In: ICPR. pp. 2069–2072 (2010)