Eurographics Italian Chapter Conference (2010)
E. Puppo, A. Brogni, and L. De Floriani (Editors)

An adaptive video surveillance architecture for behavior analysis

L. Zini and N. Noceti and F. Odone

DISI - Dipartimento di Informatica e Scienze dell'Informazione
Università degli Studi di Genova
{Zini,Noceti,Odone}@disi.unige.it

Abstract

Adaptivity to scene changes is a main requirement for video analysis. The interpretation of video streams can be dealt by triggering different techniques depending on the scene properties. We present a work-on-progress for the design of a video surveillance architecture where different tasks in the context of behavior analysis are addressed, depending on the crowd level. A coarse estimation of the scene occupancy allows us to focus on single person or groups, adopting appropriate strategies to model the dynamic information. This paper focuses in particular on the crowd estimation problem: we propose a solution to detect and localize groups of people, able to provide an estimate of the number of people in the scene.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [Image Processing and Computer Vision]: Scene analysis—Motion I.4.9 [Image Processing and Computer Vision]: Applications—

1. Introduction

One of the main goal of the current research in video surveillance is the design of methods able to *automatically* cope with variable scene complexities and environment conditions. In this work we refer specifically to the problem of detecting and modeling behavioral patterns of different complexities. The availability in our reference application of *long-time observations* calls for solutions to be *adaptive* and able to exploit knowledge coming from *previously seen scenarios*.

It has been experienced in the last decades how classical problems of behavior analysis can be better dealt by coupling traditional computer vision techniques with statistical learning from examples [GSRL98, HTWM04, HXF*06, PCV00, RR05, SG00]. If, on one side, the computer vision literature provides nowadays benchmark techniques for video analysis, statistical learning methods, on the other side, represent effective tools when an higher-level of generalization is needed.

In this paper we present a *work-on-progress* on the development of an adaptive video surveillance pipeline to model common behaviors by learning frequent patterns of activities from huge sets of unlabeled data, with a very limited a-priori information included into the pipeline.

To properly deal with the rich variety of possible scene conditions, the adaptability of the system against time should consider different aspects, from physical variations of the layout, to illumination changes occurring during daytime, also different level of occupation of the observed scene. This paper focuses on the latter, since the adopted techniques are dynamically selected depending on the scene complexity:

- When the occupancy of the scene is low then it makes sense to consider the dynamics of single objects – people in our case – or small groups (*people behavior analysis*);
- Instead, if the scene is densely occupied the global motion of the crowd should be taken into account (*crowd behavior analysis*).

For what concerns people behavior analysis, we proposed and validated a pipeline to extract and model the dynamics of single subjects (or small groups) based on clustering temporal series (for more details see [NSO10, Noc10]). A low-level analysis allows us to obtain, at each time instant, static descriptions of interesting targets that are correlated over time to obtain a representation of the dynamic evolution. We assume we are monitoring possibly complex scenarios from a distance where the "action of interest" is the

DOI: 10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2010/057-064

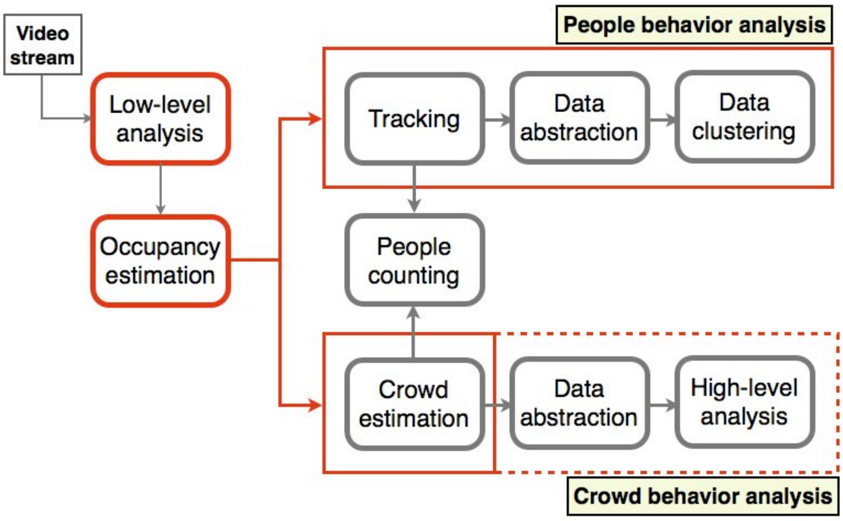

Figure 1: *The pipeline of our adaptive video surveillance architecture: depending on the estimated occupancy of the scene, the dynamics of single person (and small groups) or the global crowd motion are modeled with learning from examples.*

trajectory of a moving object as a whole, and no information is available or needed on the motion of object's parts. We thus explicitly refer to data that may be modeled as trajectories of instantaneous observations. A final higher-level analysis detects frequent patterns of activities by analysing the internal structures of a trajectories collection observed during an appropriate amount of time.

The very final aim of our current work is to integrate into this existing pipeline a module to cope with highly crowded scenes (see Fig. 1). The first point is to establish how to trigger the different analysis according to the scene requirements. Then, the presence of crowd will influence the specific techniques adopted during the modeling phase: when the number of people is high, tracking fails to produce accurate descriptions and it is thus advisable to focus attention on the global motion of crowd rather than of single subjects. After a brief introduction of the architecture we are developing, this paper focuses specifically on the crowd estimation module, able to localize the people in the scene and provide an estimate of their number. The remainder of the pipeline (enclosed with a dotted rectangle in Fig. 1) is object of current and future work.

As for crowd estimation, we start off from the method originally proposed in [KRJ*08], that, at each time instant, relies on an analysis of the image motion segmentation and exploits camera calibration. We propose some variations to the original pipeline that improve the estimates and allow for real-time performance.

We evaluated the pipeline on two experimental scenarios, characterized by rather different complexities. In our evaluations, we show how the pipeline is robust and able to adapt to such very different conditions.

The remainder of the paper is organized as follows. Sec. 2 briefly reports the relevant related works for what concerns people and crowd behavior analysis. In Sec. 3 we go into the details of our architecture. We provide just a sketch of the single person-based analysis, which is not the focus of this paper. We also present the video surveillance scenario where we mainly lead the experiments, enhancing how the camera calibration problem has been addressed. The Sec. 4 is the core of the paper, since it presents the details of the crowd estimation module: we start from the original paper and then present the variations we introduces. The discussion on the experimental evaluation concludes the section. The final section (Sec. 5) closes the paper with discussions on the future developments.

2. Related work

The study and the understanding of human activities from videos has been widely addressed in the last decades (see for instance [SG00, GSRL98, PCV00, PMF08, AC08]), particularly with the availability of an enormous amount of installed video surveillance cameras. A huge amount of video data are daily acquired, becoming more and more difficult to be handled by human operators. This justifies the growing need for computational methods to be adopted for the design of intelligent system.

Learning from examples is a rather conventional way to deal with data complexity. In the literature, approaches based on both *supervised* [PCV00, BKS07] and *unsupervised* settings can be found. The existing architecture we discussed in [NSO10, Noc10] is more related to the second approach.

A good starting point for an overview of existing approaches for unsupervised behavior analysis is a recent special issue

[SI008]. Among the first contributions we mention the influential work by Stauffer and his co-workers [SG00]. More recently [HXF*06] proposed a pipeline based on k-means, while in [PMF08] normal behaviors are associated to one class only, learned with a one-class SVM. For a reading more specifically focused on temporal series clustering a complete survey is [Lia05].

An overview of the literature concerning the analysis of crowd motion and behaviors shows lack of approaches, especially in the case of unsupervised settings. An important ingredient of crowd-centered methods is optical flow. In [RMAS04], as an examples, the authors consider the problem of detection crowd from a moving camera. They look for characteristic patterns of a spatio-temporal representation based on optical flow. The analysis of crowd flows is the core of [AS07], where Lagrangian Particle Dynamics is used to segment high-density (thousands of people) crowds. [GBB09, SBTM08] propose approaches based on tracking. The first one considers a HOG-based tracker to recognize crowd events with respect to a set of pre-defined models learnt from the data. The experimental analysis shows the appropriateness of the choice. In the second work the authors apply a KLT tracker to build crowd motion vectors.

The very recent and interesting work in [MOS09] introduces the concept of social force into a crowd analysis pipeline. They detect and localize abnormal crowd behaviors using again optical flow. In [GCR09] social behaviors are considered, using an approach built upon state-of-art algorithms for pedestrian detection and multi-object tracking.

Considering approaches explicitly based on learning, we mention the work in [KN08], based on HMM of spatio-temporal motion patterns, and, for the unsupervised counterpart, [BC06], which presents a data-driven bayesian clustering algorithm to detect individuals on low and medium crowded scenes.

In [SHN09] an evaluation of people tracking, counting and density estimation in crowded environments has been presented. The authors proposed a method coupling a Cluster-Boosted-Tree (CBT) pedestrian detector with a learning-based hierarchical association tracker.

3. The architecture

In this section we provide an overview of the video surveillance architecture we propose, shown in Fig. 1, clarifying the experimental scenario where we lead the evaluation of the method and discussing the structure of the system.

3.1. The experimental scenario

A video surveillance setup (the *Imanalysis suite*, we obtained within a technology transfer program with the company **Imavis srl**, `http://www.imavis.com/`) monitors an indoor open space (one of the main halls of our Department) where a good amount of dynamic events occur during daytime (see Fig. 2). Only people are supposed to be moving in the scene: the monitored environment provides different complexity with respect to the crowd level, which in turns depend on several factors, such as day, temporal interval, period of the academic year (presence of lessons, examinations). The weather conditions strongly affect the scene appearance being the hall illuminated by windows (on the right wall) as well as artificial lights.

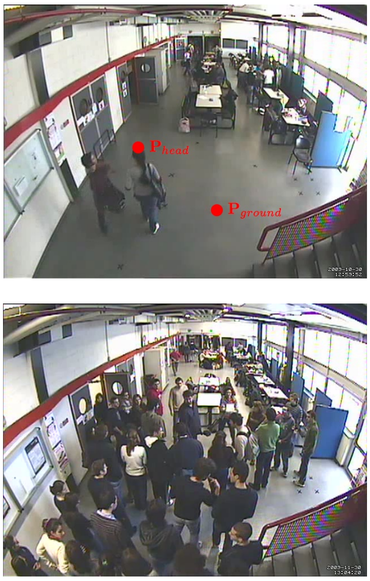

Figure 2: *The real scenario we consider provides an ideal test-bed for our video surveillance pipeline, being characterized by difficult illumination and richness from the standpoint of dynamic content. The frames report examples of low density (above) and crowd (below). Above, examples of points from the planes considered during the system calibration (see Sec. 3.1).*

The physical characteristics of the observed environment and the variety of dynamic events occurring during daytime make the setting of acquisition an ideal test-bed for evaluating the system with respect to the usability in a real video surveillance setup, where computational efficiency and accuracy of the results are important requirements.

Currently the acquisition system is not fully calibrated. Since the implemented crowd analysis module required information on camera calibration, we may simply estimate the homographies that maps the 3D world points into the corresponding image points. For the problem under analysis, in particular, it is important to obtain information on [HZ04]:

- The **ground plane** Π_g
- A **head plane** Π_h

In Fig. 2, above, two examples of 3D world points laying on the ground plane ($\mathbf{P}_{ground}$) and on the head plane ($\mathbf{P}_{head}$) are reported. From an appropriate set of points lying on the two planes we can thus estimate:

- $\mathbf{P} = H_{ground}\mathbf{P}_{ground}$
- $\mathbf{P} = H_{head}\mathbf{P}_{head}$

where $\mathbf{P}$ is expressed in pixel coordinated, $\mathbf{P}_{ground}$ and $\mathbf{P}_{head}$ are in world coordinates. To compute a reliable estimation of H_{head} we assume that all people have a fixed height, h_1.

3.2. The adaptive pipeline

As sketched in Fig. 1, the input video stream is first processed with a low-level analysis: at each time instant, the current frame is segmented with respect to motion information by means of change detection (see an example in Fig. 3, first row, left).

The condition to determine an approximation of people density (or level of occupation) in the scene is based on thresholding the fraction of moving pixels in the binary map at time t. Although such estimation might be unreliable due to the noise when computing the change detection, it is a very simple and immediate way to easily discriminate among low density and high occupation (two examples are in Fig. 2) and consequently activate different paths in the pipeline.

3.2.1. Person-based analysis

When the estimated people density is below a given threshold, the system focuses on the dynamics of single person or small groups. This task is addressed following the *people behavior analysis* pipeline (see Fig. 1, above). Here we just sketch the procedure, we refer the interested reader to [NSO10, Noc10] for further details.

Each connected component in the binary map represent an interesting target that is described with an appropriate set of information at time t, more specifically the target position in the image plane, its velocity expressed in terms of magnitude and direction, and its size. The vectors that statically describe a target at each time instant are then correlated over time with a tracking procedure (Fig. 3, above, rigth). As the system runs, the trajectories are gathered populating a collection of temporal data (the *training set*) that provides a representative sample of what is usually observed in the scene: Fig. 3, below on the left, shows an example.

The final aim of the procedure is to study the internal structure of the training set to detect groups of coherent data, or, in other words, common behaviors. Since in our case this step is based on clustering, a data abstraction phase is required to make the data suitable for a learning framework. We consider string-based representations based on a data partitioning fully data-driven.

The map of Fig. 3, below, right, reports the patterns of activities finally detected by the system: as it can be easily visualized, the patterns reflect very intuitive classes of activities occurring in the hall.

The experimental evaluation of the method has been carried out on two weeks of observations. During the first one, a training set including 1200 dynamic events was gathered (a sampling is shown in Fig. 3, below, left): the phase of acquisition was followed by a simple cleaning procedure of the data to avoid the contamination of the noise (change detection errors, tracking failures) on the models. A test set of 5700 dynamic events has been collected on a second week, without cleaning the data. The training set has been manually annotated with respect to 8 main behavioral patterns; the test set included examples of the 8 known behaviors as well as anomalies (dynamic events) and noisy trajectories. We obtain a percentage of correct events classification of about the 80%, a very good performance if one considers the high complexity of the data.

3.2.2. Crowd-based analysis

As the number of people increases, the tracking fails to compute reliable descriptions of the scene dynamics, because of intersection and occlusion events frequently occurring in the scene, and highly noisy change detection maps. The attention moves instead to the analysis of the global motion of the crowd, requiring the adoption of appropriate techniques. The corresponding plot in Fig. 1 shows that the first step towards this direction is the crowd estimation, in the terms that will be discussed in details in the next section. An interesting side effect of our approach is the capability of providing an estimate of the number of people composing the crowd, task that could not be easily addressed by the direct analysis of the binary map.

The dotted rectangle encloses the modules of the pipeline that will be developed in the near future: once that a first (maybe coarse) crowd estimation has been performed, similarly to what done in the case of people behavior analysis, the system will address the problem of modeling the crowd dynamics. This goal will require to consider appropriate instances for the data abstraction and the high level analysis steps.

4. Crowd estimation

The focus of this section is on the current work on developing the crowd behavior analysis pipeline. In particular, we will provide details on the crowd estimation module and show how an interesting side effect of this initial representation is the capability of estimating the number of people in the crowd.

4.1. Crowd estimation approach

The method we implemented is organized in two different levels of analysis, a *coarse analysis*, that follows the approach proposed in [KRJ*08], and a *real-time refinement* that exploits temporal coherence.

The algorithm is based on the assumptions that only moving people are observed in the scene and the space occupied by

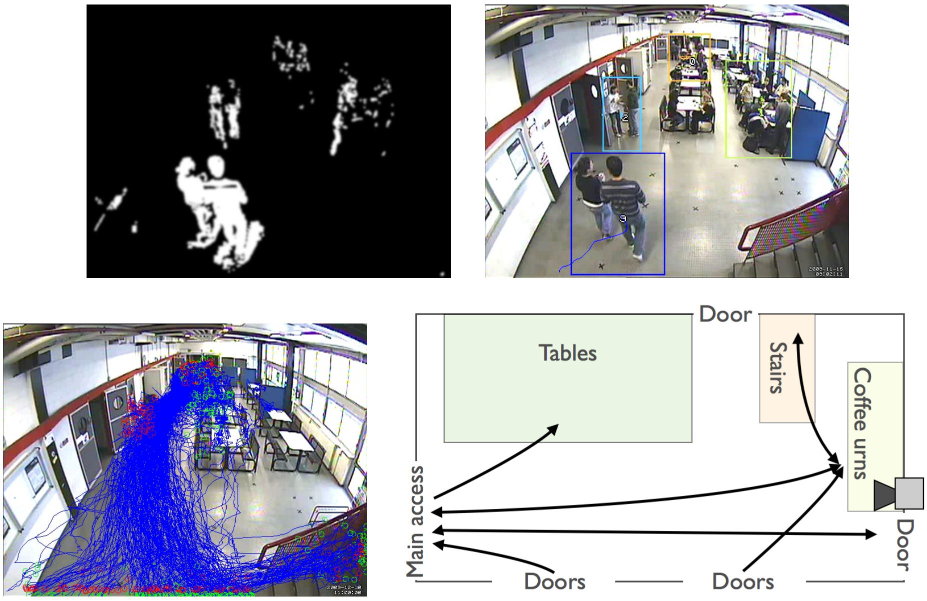

Figure 3: *Above, an example of the resulting binary map after the low-level video analysis: the connected components (left) are first extracted and described with a features vector, then correlated over time (right) by means of tracking to model their dynamic evolution in the scene. Below, left, a collection of temporal data gathered on one week. The final step of higher level analysis, based on clustering trajectories, allows to detect frequent patterns of activity, referred to with arrows on the environment schema (right).*

each person in a crowd is almost constant. In this setting, the problem of estimating the number of pedestrians in a scene can be restated as the problem of estimating the area occupied by them on the ground.

Starting from the binary map resulting from the change detection, we consider each connected component in the map and estimate the corresponding area occupied on the ground, $A(\pi_g)$, using camera calibration. Assuming a substantial homogeneity of the area occupied by each person in the crowd, as already stated, the estimated area $A(\pi_g)$ will be proportional to the number of people in the group.

The area computation is based on projecting the connected component under analysis onto two parallel planes: (1) the first one corresponds to the ground, (2) the second relates to the plane at height h_1 (see Fig. 4). If h_1 is an appropriate candidates of the real average people heights in the group, then the area occupied by each person in the scene is the intersection of his/her projections onto the two fixed planes.

This procedure results in a *coarse* estimation of the area. In [KRJ*08] the shape of the area is the input of a refining step whose objective is to compute height and main axes of a cylinder that projected on the image plane gives the most similar shape to the observed one. The final area occupied on the ground corresponds to the base area of the cylinder and is associated to a statistical confidence estimated from the observations. The number of people in the group N_c is finally approximated by dividing the area by a constant, learnt from the data, that represents the space occupied, on average, by a single person.

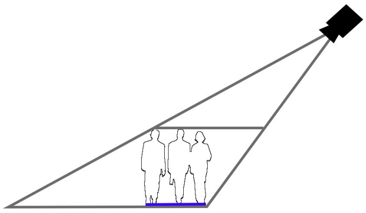

Figure 4: *Given a group in the scene, the corresponding connected component in the binary map is projected onto two parallel planes, corresponding to the ground and an average height h_1. Then the intersection of the projections localizes the area occupied by the group, that will be proportional to the number of people belonging to it.*

We slightly modified this algorithm in two points:

- We skip the cylinder based optimization step, that experimentally showed to be inappropriate for highly crowded scenes and computationally expensive;
- We modified the confidence computation and based it on geometrical aspects.

In the remainder of the section we discuss in details the proposed variations.

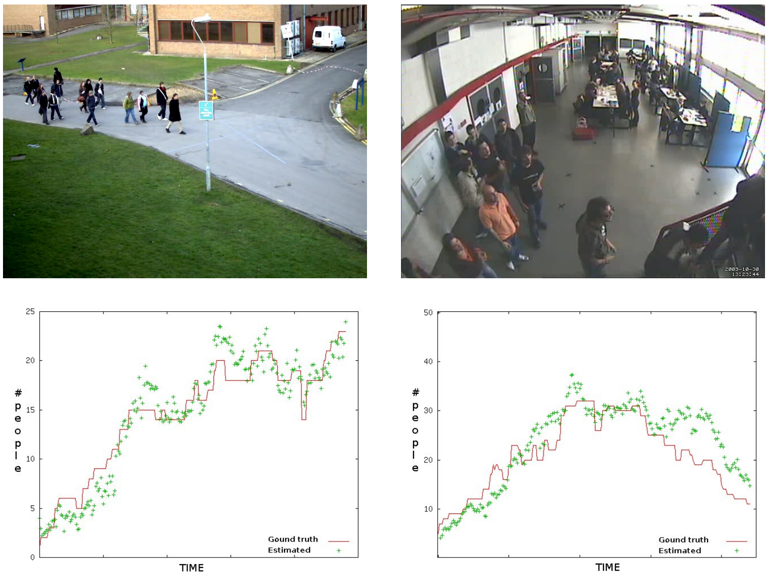

Figure 6: *System feedback for two samples video (from PETS09, left, and from DISI, right). The comparison between estimated number of people and ground truth shows the robustness of our pipeline.*

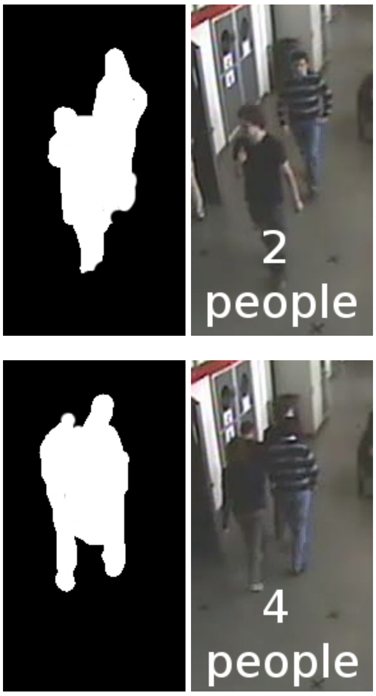

Figure 5: *Different configurations of walking people may generate a very similar change detection maps, introducing instability in the results of the algorithm.*

4.2. Double bound computation

The main source of instability of the method described in the previous section is due to the connected components processed as "input": as shown in Fig. 5 with a simple example, given a map of the change detection is impossible to establish if it is generated by a very compact group of N people or by $M < N$ people that are walking separately.
The algorithm in [KRJ*08] assumes compactness of group, but this is a too restrictive assumption for all settings of acquisition where the angle between camera optical axis and ground plane is very different from 90 degrees. Is this cases, the data ambiguities induced by the mutual position between camera and scene makes results of change detection highly inaccurate and assumption of compactness not reasonable.
This consideration calls for some sort of confidence on the system feedback that may help in evaluating the results. In [KRJ*08], the bounds were built by considering an interval of confidence of size ε centered on the number of people estimated by the coarse analysis, N_c: $N_c \pm \varepsilon$. The constant ε was statistically estimated on the training data. However, as we will show in the experimental analysis (Sec. 4.4), this solution tends to associate to the best confidence a overestimated result, especially in low or medium crowded scenes.
We base instead our estimate on an interval of confidence whose *upper bound* is given by the coarse analysis (N_c) while the *lower bound* computation is based on very practical geometrical considerations. Considering again the connected component coming from the change detection map, our approach is based on looking for the minimum number of people that could generate it. Let us consider the con-

nected component corresponding to a crowd and a pair of people, P_F and P_B, the first located in front of the second. The idea we follow is that the distance between P_F and P_B should be *at least* such that the projection on the image plane of P_B feet encloses the pixels of the head of P_F. The procedure is iteratively repeated until all the connected component pixels have been associated to some person. By counting the resulting number of people, the minimum that could geometrically generate the connected component, we obtain the lower bound on the estimates. The gap between upper and lower bound finally represents the confidence (or uncertainty) on the system feedback.

4.3. Temporal filtering of the results

The final refinement step that we introduced in the algorithm is based on a temporal analysis of the gap between lower and upper bound: at each time instant t, we keep in memory the history of the estimates from the interval $[t-\Delta t, t]$ and select as current feedback the one corresponding to the smaller gap, that is the most stable and thus accurate result.

Considering the temporal evolution is helpful in a real setting where errors might be due to particularly difficult people configurations that cause data ambiguities, and to errors in the change detection (due to, e.g., shadows, difficult illumination, noise).

We experimentally observed that the scene dynamics help in detecting errors coming from the change detection, that are typically temporally limited, and in discriminating between the ambiguous situations exemplified in Fig. 5. In fact, compact groups of people typically generate blobs that remain rather stable over time, as opposite to more spread groups.

Although our solution introduces a slight delay in the system feedback, it significantly improves the performance with respect to the original approach, as shown in the next section.

4.4. Experimental validation

We performed the experimental evaluation of the method considering two rather different scenarios: the environment described in Sec. 3.1 and the benchmark dataset from workshop PETS 2009 (available for download at `http://www.cvg.rdg.ac.uk/PETS2009/a.html#s1`). In what follows, they will be referred to as, respectively, *DISI* and *PETS09*.

For *PETS09* setting (Fig. 6, above, left) a full camera calibration was provided, allowing for a better accuracy in the results. Also, the mutual position between the camera and the people moving in the scene does not cause data ambiguities. As opposite, in the case of *DISI* dataset (Fig. 6, above, right) the calibration is based on the use of homographies and the acquisitions are characterized by possible high ambiguities on the observations.

Fig. 6, second row, shows the estimated number of people for two videos, one from each dataset, and compares the feedbacks against the ground truth. The plots show, in both cases, the robustness of our estimates.

Table 1: *Comparison of the performances of our method against the results reported in [SHN09]. The values represent the average errors per frame on 3 sequences from PETS09 dataset.*

	Difficulty level	CBTHT	Our method
SEQ1	med.	7.19	4.1
SEQ2	med.	1.37	1.6
SEQ3	high	–	2.25

We first compare on Table 1 the performances of our approach on dataset *PETS09* with the results reported on [SHN09] and obtained combining a learning-based hierarchical association tracker with a Cluster-Boosted-Tree based pedestrian detector. We refer to the method as CBTHT. The values on the table, the average error per frame computed separately on 3 sequences of different complexities, shows that our method performs globally better.

To summarize the results we obtained on the two data sets we consider, we evaluated the estimates with respect to 3 different levels of scene occupation: if N_p is the number of people at a certain time t in the ground truth, we choose 2 thresholds, τ_1 and τ_2, such that (1) $N_p \leq \tau_1$ denotes low occupancy, (2) $\tau_1 < N_p < \tau_2$ defines medium occupancy, while (3) $N_p \geq \tau_2$ represents high occupancy. Because of the difference in the average number of people present in the data sets, we adopt different thresholds for the two settings: $\tau_1 = 8$, $\tau_2 = 25$ for *PETS09*, $\tau_1 = 5$, $\tau_2 = 10$ for *DISI*. Tab. 2 reports the obtained results: it is immediate to note that, although a slight decreasing in the performance on *DISI* due to its higher complexity, the results for the two settings are accurate and comparable.

We finally show how the variations we introduced significantly improve the performance with respect to the original algorithm, in particular for scenarios characterized by low or medium occupancies. In Fig. 7 we report the comparison, performed on a video from *DISI* dataset, between our approach (denoted as "filtered" to enhance the presence of the temporal filtering caused by the final refinement, Sec. 4.3) and the original method. The trend of the ground truth is also reported. It is easy to observe how the original method tends to highly overestimate the correct number of people, as opposite to our approach where the temporal analysis allows to reach a higher robustness. Notice that it is clearly visible the delay, with respect to the ground truth, introduced into the pipeline by the same temporal analysis.

Table 2: *Global evaluation of the system feedback on the two data sets, considering 3 different levels of scene occupation.*

	$GT \leq \tau_1$	$\tau_1 < GT < \tau_2$	$GT \geq \tau_2$
PETS09	97%	85%	98%
DISI	95%	82%	96 %

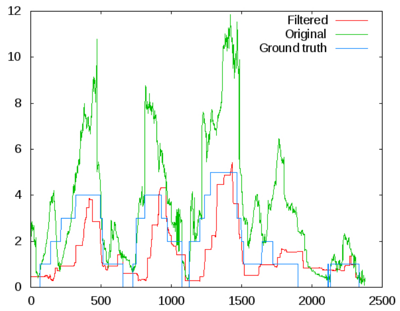

Figure 7: *Comparison between our method (referred to as "filtered") and the algorithm in [KRJ*08] for people counting. It is immediate to observe how the latter tends to overestimate the correct number of people (shown by the ground truth), while our approach, although a small time delay due to the temporal filtering, produced better performances.*

5. Discussion

In this paper we presented a pipeline for behavior analysis, designed to adapt to different scene conditions in terms of occupancy. A condition based on a coarse estimation of the scene occupancy triggers two different pipelines of analysis, centered on people or crowd. If the pipeline of people behavior understanding has been previously presented and evaluated, the one centered on crowd is at an initial stage of development. This paper focused on the current work on a module for crowd detection, whose main side effect is the capability of estimating the number of people in the scene. We started off from the method presented in [KRJ*08] and introduced some variations to improve computational performances and results, as demonstrated in the experimental analysis.

The future work will be devoted to the development of the pipeline towards this direction. We will adopt statistical learning from examples to model the crowd dynamics and finally build general models of its activity. This will require the adoption of appropriate data description (optical flow, space-time features to describe the evolution of the crowd on the video) as well as methods to compare and model the obtained motion descriptions.

References

[AC08] ANJUM N., CAVALLARO A.: Multifeature object trajectory clustering for video analysis. *IEEE Trans. on CSVT 18*, 11 (2008). 2

[AS07] ALI S., SHAH M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR* (2007), IEEE Computer Society, pp. 1–6. 3

[BC06] BROSTOW G., CIPOLLA R.: Unsupervised bayesian detection of independent motion in crowds. IEEE Computer Society. 3

[BKS07] BASHIR F., KHOKHAR A., SCHONFELD D.: Object trajectory-based activity classification and recognition using hidden markov model. *IEEE Trans. on IP 16* (2007). 2

[GBB09] GARATE C., BILINSKY P., BREMOND F.: Crowd event recognition using hog tracker. In *IEEE PETS-Winter* (2009), pp. 1–6. 3

[GCR09] GE W., COLLINS R., RUBACK B.: Automatically detecting the small group structure of a crowd. In *WACV* (2009), pp. 1–8. 3

[GSRL98] GRIMSON W. E., STAUFFER C., ROMANO R., LEE L.: Using adaptive tracking to classify and monitor activities in a site. In *CVPR* (1998), pp. 22–29. 1, 2

[HTWM04] HU W., TAN T. N., WANG L., MAYBANK S. J.: A survey on visual surveillance of object motion and behaviors. *IEEE Tran. on Systems, Man and Cybernetics 34*, 3 (2004), 334–352. 1

[HXF*06] HU W., XIAO X., FU Z., XIE D., TAN T., MAYBANK S.: A system for learning statistical motion patterns. *IEEE Trans on PAMI 28*, 9 (2006). 1, 3

[HZ04] HARTLEY R., ZISSERMAN A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, 2004. 3

[KN08] KRATZ L., NISHINO K.: Spatio-Temporal Motion Pattern Modeling of Extremely Crowded Scenes. In *MLVMA-ECCV* (2008). 3

[KRJ*08] KILAMBI P., RIBNICK E., JOSHI A. J., MASOUD O., PAPANIKOLOPOULOS N.: Estimating pedestrian counts in groups. *CVIU 110*, 1 (2008), 43–59. 2, 4, 5, 6, 8

[Lia05] LIAO T. W.: Clustering of time series data: a survey. *Pattern Recognition 38*, 11 (2005). 3

[MOS09] MEHRAN R., OYAMA A., SHAH M.: Abnormal crowd behavior detection using social force model. 3

[Noc10] NOCETI N.: Learning to classify visual dynamic cues. *PhD thesis, DISI, Universitá di Genova* (2010). 1, 2, 4

[NSO10] NOCETI N., SANTORO M., ODONE F.: Learning behavioral patterns for video surveillance. *MLVMA* (2010). 1, 2, 4

[PCV00] PITTORE M., CAMPANI M., VERRI A.: Learning to recognize visual dynamic events from examples. *IJCV* (2000). 1, 2

[PMF08] PICIARELLI C., MICHELONI C., FORESTI G. L.: Trajectory-based anomalous event detection. *IEEE Trans on Circuits and Systems for Video Technology 18*, 11 (2008). 2, 3

[RMAS04] REISMAN P., MANO O., AVIDAN S., SHASHUA A.: Crowd detection in video sequences. In *IEEE Intelligent Vehicles Symposium* (2004), Citeseer, pp. 66–71. 3

[RR05] ROBERTSON N., REID I.: Behaviour understanding in video: a combined method. In *IEEE Proc. on ICCV* (2005), vol. 1. 1

[SBTM08] SAXENA S., BRÉMOND F., THONNAT M., MA R.: Crowd behavior recognition for video surveillance. In *ACIVS* (2008), Springer, pp. 970–981. 3

[SG00] STAUFFER C., GRIMSON E.: Learning patterns of activity using real-time tracking. *IEEE Trans. on TPAMI 22*, 8 (2000). 1, 2, 3

[SHN09] SHARMA P. K., HUANG C., NEVATIA R.: Evaluation of people tracking, counting, and density estimation in crowded environments. In *PETS* (2009), pp. 39–46. 3, 7

[SI008] Special issue on event analysis in videos. *IEEE Trans on Circuits and Systems for Video Technology 18*, 11 (2008). 3