Automatic Learning of Background Semantics in Generic Surveilled Scenes

Carles Fernández, Jordi Gonzàlez, and Xavier Roca

Dept. Ciències de la Computació & Computer Vision Center,
Edifici O, Campus UAB, 08193 Bellaterra, Barcelona, Spain
{carles.fernandez,poal,xavier.roca}@cvc.uab.es

Abstract. Advanced surveillance systems for behavior recognition in outdoor traffic scenes depend strongly on the particular configuration of the scenario. Scene-independent trajectory analysis techniques statistically infer semantics in locations where motion occurs, and such inferences are typically limited to abnormality. Thus, it is interesting to design contributions that automatically categorize more specific semantic regions. State-of-the-art approaches for unsupervised scene labeling exploit trajectory data to segment areas like sources, sinks, or waiting zones. Our method, in addition, incorporates scene-independent knowledge to assign more meaningful labels like crosswalks, sidewalks, or parking spaces. First, a spatiotemporal scene model is obtained from trajectory analysis. Subsequently, a so-called GI-MRF inference process reinforces spatial coherence, and incorporates taxonomy-guided smoothness constraints. Our method achieves automatic and effective labeling of conceptual regions in urban scenarios, and is robust to tracking errors. Experimental validation on 5 surveillance databases has been conducted to assess the generality and accuracy of the segmentations. The resulting scene models are used for model-based behavior analysis.

1 Introduction

The automatic analysis of human behaviors in large amounts of video surveillance footage is becoming critical as the number of cameras installed in public areas increases. This demand has generated novel techniques for the analysis of large collections of archives containing recordings from different outdoor scenarios during long periods. As a result, events of interest are detected, and alarms can be raised online according to predefined criteria [1]. Complementary, events extracted from image sequences can be used for annotation purposes when becoming concepts to index surveillance databases.

There is a clear trade-off between the semantic richness of video events and the robustness of their recognition. The richness of the conceptual knowledge extracted from surveillance sequences greatly determines the limitations of eventual user queries. Ideally, indexing would be based on high-level concepts determined by rich and complete ontologies [2]. However, as events become more specific, their recognition in surveillance data also becomes more challenging.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part II, LNCS 6312, pp. 678–692, 2010.

Important steps forward have been taken in the computer vision domain, where interesting approaches appeared related to the automatic interpretation of human activities within scenes. In surveillance data obtained from static cameras in outdoor scenes, human activities are commonly represented by trajectories of points extracted using detection and/or tracking processes.

On the one hand, different statistical properties of these observed trajectories are computed in order to assess their normal or abnormal nature. There are several strategies to cluster and merge trajectories, like spatial extension [3], Hierarchical Fuzzy C-Means [4], Hierarchical clusters [5], GMMs [6], or splines [7], among others. Subsequently, by analyzing the regions where motion is observed, characteristic regions like roads, walking paths, or entry/exit points can be learned [8]. Statistical techniques have been also used to model semantic regions based on activity correlation [9]. These robust bottom-up processes are scene-independent, and abnormal behaviors like violent struggling can be detected and annotated, e.g., by observing erratic trajectories with high speed variations.

On the other hand, deterministic models provided beforehand by an expert have been also applied successfully in the surveillance domain, such as Situation Graph Trees [10,11], Petri Nets [12], or Symbolic Networks [13], for example. These models can represent complex behaviors (such as *'danger_of_runover'* or *'car_overtaking'*) while performing reasoning based on more simple, but robustly detected, events (e.g., *'turning_left'* or *'accelerating'*), for example those ones extracted using the aforementioned bottom-up processes. Hence, high-level reasoning processes can generate key-words and concepts associated to stronger semantics that can be searched for.

Towards this end, reasoning on events requires of conceptual scene models that semantically represent the background of the surveilled scene. The semantics of the regions in which an agent is found at each time step are used to infer behaviors, such as *'crossing_the_street'* or *'waiting_at_the_crosswalk'*. Unfortunately, each particular scene requires of its own conceptual scene model. Therefore, there is a need for automatic and generic learning procedures able to infer the semantics of thousands of surveillance scenes.

In this paper we present a novel technique for automatic learning of conceptual scene models using domain knowledge, which can be successfully used for further reasoning and annotation of generic surveillance sequences. In essence, we learn spatiotemporal patterns of moving objects to infer the semantic labels for background regions where motion has been observed, such as pedestrian crossings, sidewalks, roads, or parking areas. The derivation of site labels is formulated as a MAP-MRF inference in terms of a pairwise Markov network, whose graph configuration is factored into a joint probability guided by taxonomical knowledge. Finally, we have applied the SGT formalism to demonstrate the applicability of our approach, although other deterministic behavior models that require of conceptual scene models can be used instead.

This paper is structured as follows: Section 2 formally defines our labeling task in terms of maximization of region compatibility. It comprises two steps: the compatibility with observed evidence is computed from motion features in

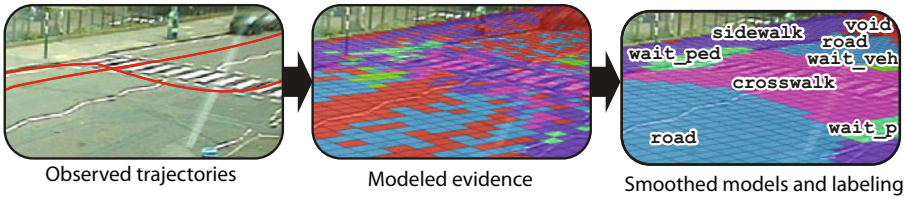


Fig. 1. Scheme of the proposed framework for labeling urban scenarios

Section 3, and inter-region compatibility for smoothness is modeled in Section 4. A preprocessing stage to improve efficiency is explained in Section 5. The method is tested thoroughly for experimental validation in Section 6, and applied to the SGT formalism in Section 7. Finally, we discuss the results and provide some concluding remarks.

2 Background Labeling by Compatibility

The semantic learning of a background model consists of partitioning an arbitrary scenario of the domain into a lattice of regions, and have each region learn a spatiotemporal model. Each model should be estimated based on trajectory properties, and finally assigned an explicit label that categorizes it. Here, we tackle the problem of *semantic region learning* as one of *multiclass semantic segmentation*. Towards this end, efficient techniques have been developed, such as MRF [14] and its variants, like DRF [15], or LCRF [16], or alternatives like Semantic Textons [17]. In our case, the categorization of regions from their statistical models will be posed as a labeling task and formulated as a MAP-MRF inference problem, defined by irregular sites and discrete labels [18].

2.1 Sites and Labels

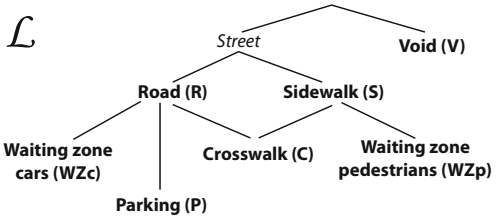
The lattice of irregular regions to be labeled is usually defined either by perceptual groups –out of a segmentation process–, or by clusters of recognized features within the scene [18]. Instead, we aim to define lattices that capture the condition of far-field projectivity, characteristic of scenarios in our domain.

To do so, we compute the scene to ground-plane homography [19], so that each lattice is a set of regions $\mathcal{R}$ obtained as the projection of a rectangular grid from ground-plane to scene. In addition to the sites, a set $\mathcal{L}$ of seven discrete labels defines generic, common, and relevant locations in urban surveillance. Labels are organized taxonomically as shown in Table 1a. A void label (V) is made available for those cases in which none of the labels applies, as in [20].

2.2 Inference

Having defined the set of sites and labels, we next describe the process of assigning a label $l \in \mathcal{L}$ to each region $r \in \mathcal{R}$. The disparity of labels is assumed

Table 1. (a) Taxonomy of locations for urban surveillance. (b) Each location is a vector of the trinary features *ped*=Pedestrian, *veh*=Vehicle, *wai*=Wait, and *stp*=Stop.



(a)

Label from $\mathcal{L}$		ped	veh	wai	stp
C	$f^1 = [$	+1	+1	0	-1]
S	$f^2 = [$	+1	-1	-1	-1]
R	$f^3 = [$	-1	+1	-1	-1]
WZp	$f^4 = [$	+1	-1	+1	-1]
WZc	$f^5 = [$	-1	+1	+1	-1]
P	$f^6 = [$	0	+1	0	+1]
V	$f^7 = [$	-1	-1	-1	-1]

(b)

to be piecewise smooth in the lattice of regions. A series of observation vectors $o = \{x, y, a\}$ constitutes the evidence from the trajectories, where (x, y) is the estimated position of the agents in the image plane –the centroid of the ellipsoid projected to the ground-plane–, and a is a binary parameter stating whether the agent is a vehicle or a pedestrian. The derivation of the site labels $\{l\}$ is formulated as a MAP-MRF inference in terms of a pairwise Markov network, whose graph configuration is factored into the joint probability

$$P(\{l\}, \{o\}) = \frac{1}{Z} \prod_{r \in \mathcal{R}} \phi_r(l_r, o_r) \prod_{\{r,s\} \in \mathcal{N}} \psi_{r,s}(l_r, l_s), \tag{1}$$

where Z is a normalization factor. The *data compatibility* function $\phi_r(l_r, o_r)$ is interpreted as the likelihood of choosing label l for region r, given o observed in r. This function is learned by trajectory analysis as explained in Section 3.

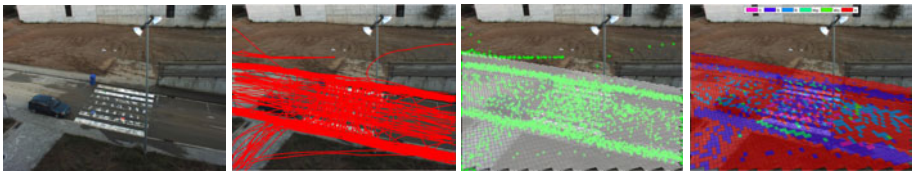
On the other hand, smoothness constraints are encoded into $\psi_{r,s}(l_r, l_s)$, so-called *internal binding*, which models how neighboring regions affect to each other regarding their classes. In this term, the set $\mathcal{N}$ contains all pairs of interacting regions, in our case adjacent 8–connected regions in the projected grids. In our work, $\psi_{r,s}(\cdot)$ is a prior set of constraints directly taken from topological assumptions that are derived from a defined hierarchy of labels depicting domain knowledge, as later explained in Section 4.

Once $\phi_r(\cdot)$ and $\psi_{r,s}(\cdot)$ are defined, a max-product belief propagation (BP) algorithm [20] derives an approximate MAP labeling for Eq. (1).

3 Data Compatibility

We define the function $\phi_r(l_r, o_r)$ as the likelihood of region r to be labeled as l, having observed a series of vectors o_r in the region, and according to a motion-based model that encodes prior domain knowledge.

Challenges arisen by semantic scene –similarly, by document analysis or medical imaging– deal with overlapping classes that are not mutually exclusive. Hence, we characterize scenario regions following the prototype theory, which defines class labels as conjunctions of required (*+1*), forbidden (*-1*), and irrelevant (*0*) features [21]. Here, labels are modeled using 4 features: target (i) is a



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

Fig. 2. Region modeling by trajectory analysis: (a) original image, (b) smoothed trajectories, (c) sampled control points, (d) initial labeling

pedestrian or (ii) a *vehicle*, (iii) is *waiting*, and (iv) has *stopped*, see Table 1b. A series of prototypical feature vectors $\{f^1 \dots f^{|\mathcal{L}|}\}$ results from this step.

Next step consists of online smoothing and sampling data from tracking. To do so, each new complete trajectory is fitted by iteratively increasing a sequence of connected cubic b-splines [7], see Fig. 2b: an adjustment step divides a spline into connected sub-splines more fitted to the trajectory, and a termination step validates a subsequence when its maximum distance to the trajectory is below a 10% of the total length. Once the recursion is done, the global sequence of splines is sampled to generate time-equidistant control points (Fig 2c), each one having an observation $o = \{x, y, a\}$. The position (x, y) is estimated by a multi-target tracker [22], and the target type (a) is identified using a scene-invariant discriminative approach as in [23]. When a new control point is generated, its enclosing region updates an histogram of the 4 features described above. Lastly, each region is assigned an online averaged vector of observed features f_o.

The data compatibility of the observations in region r with label $l \in \mathcal{L}$ is a softmax function of the Hamming distance between the averaged vector of features observed, f_o, and the vector defined for that label, f^l:

$$\phi_r(l_r, o_r) = \frac{\exp(-d_H(f_o, f^l))}{\sum_{m \in \mathcal{L}} \exp(-d_H(f_o, f^m))}. \tag{2}$$

Learned data compatibilities provide an initial rough scene model. This initial labeling omits the inference phase, and simply assigns to each region the label with a highest value of $\phi_r(\cdot)$, see Fig. 2d. Due to the limited coverage of the scene by the control points, there is a massive presence of *Void* labels, in red.

4 Smoothness

The smoothness term $\psi_{r,s}(l_r, l_s)$ specifies inter-region compatibilities, stating how the system privileges or disfavors label l_r at expenses of l_s when r and s are adjacent. In other words, it conditions *a priori* those neighborhoods formed by a certain combination of semantic categories. The goal here is to specify compatibilities that discard unlikely labelings, smooth poorly sampled ones, and preserve detailed information that are scarce but consistent.

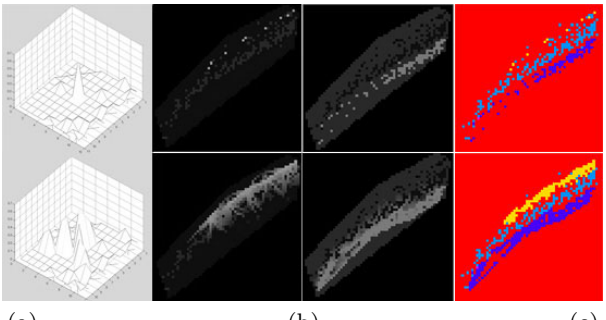
In our case, advantage is taken on the hierarchical organization of $\mathcal{L}$ to constrain discontinuities between labels. $\mathcal{L}$ fixes topological constraints of set inclusion, by establishing relations of particularization as seen in Table 1a; e.g., a *parking* is a concrete segment of *road*, and also constrains the adjacency between different regions. Consequently, compatibilities are fully specified by

$$\psi_{r,s}(l_r, l_s) = \begin{cases} 1 & l_r = l_s \\ \alpha & Adj(l_r, l_s) \\ \beta & \text{otherwise} \end{cases} \tag{3}$$

where $1 > \alpha > \beta > 0$, and $Adj(l_r, l_s)$ states that l_r and l_s are adjacent in the topological map, i.e., have direct links in the taxonomy. For example, P–R, C–R or C–S are adjacent pairs, but WZc–P or R–S are not. This model firstly maintains the identity of the labels, secondly favors dilation and erosion between adjacent regions, and ultimately allows relabeling for region smoothness.

5 Geodesic Interpolation

Having defined compatibilities for observed evidence and sought smoothness, the application of an efficient BP algorithm [20] approximates an optimal labeling via MAP-MRF inference. Nonetheless, In cases of very poor sampling, e.g., when estimating models of parkings, the regions obtained by MAP-MRF inference with the smoothness prior are often still disconnected or not representative, making it difficult to obtain accurate segmentations.



(a) (b) (c)

Fig. 3. Top: non-smoothed marginal probabilities viewed (a) as a discrete mesh and (b) as intensity maps, and (c) initial label assignment (best viewed in color). Bottom: effects of the interpolation.

To solve this problem, a preprocessing stage is used before the inference step to reinforce spatial coherence by interpolating lines in a geodesical manner. The idea is to create linear ridges that connect high-valued and isolated samples in each label's marginal probability map (Fig. 3a), in order to emphasize the

presence of connected structures in them (Fig. 3b). As a result, the subsequent MAP-MRF process is reinforced with these structures and guides more sensible inferences for an eventual labeling, as shown in Fig. 3c.

The algorithm recursively finds non-void assigned categories that are isolated, i.e., have no neighbor with the same category. Regions with the same class assignment are searched within an area of influence –we used 1.5 meters in the calibrated map–, and the two regions are connected with a linear ridge, which modifies the marginal class of the regions on that line only if the original marginal value increases. The class probabilities of each region are finally normalized for each region, and new assignments are applied.

6 Evaluation

The presented framework has been evaluated in 5 urban scenarios with diverse characteristics, obtained from camera recordings. The *Hermes* dataset[1] presents an interurban crosswalk scenario with more pedestrians than vehicles; *Oxford centre*[2] shows an intersection highly populated by both target types; *Devil's Lake*[3] presents moderate agitation but challenges with an intense projectivity; *Kingston–1* contains a partially seen bus stop close to a crosswalk, and *Kingston–2* shows a minor street with perpendicular parking spaces used for long periods of time. These two last scenarios are extracted from the Kingston dataset [24].

Evaluation is carried out using 25 ground truth (GT) images –5 participants per scenario–, consisting of pixel-level maps segmented into the 7 categories of Table 1. Participants were asked to visually identify the semantic regions by observing recorded footage, and partition them accordingly. In order to evaluate discriminant capability, and given that manual labeling is prone to vary across humans, the system will perform well if segmentation errors compare to inter-observer variability. This validation criterion is commonly used in biometrics [25]. To accomplish this, each GT image has been divided into the cells of its corresponding grid, and a modal filter has been applied over each cell, assigning the most repeated pixel label to that region. Finally, each label assignment has been evaluated against the other GTs and averaged for each GT and scenario.

In order to obtain quantitative comparisons, we have computed 3 different accuracy scores over the 5 datasets, evaluating both techniques against the GT assignments. In the evaluation tests, the maximum number of iterations for the MAP-MRF has been limited to 15. The values of α and β are 0.8 and 0.6 respectively, for all experiments.

The matricial configuration of the lattice reduces computational effort in both region modeling and label inference. Observations update the region models online as trajectories are complete. Regarding the final inference over regions learned, for a grid size of 75×75 geodesic interpolation takes at most 3 seconds to

[1] `http://www.hermes-project.eu/`
[2] `http://webcam.oii.ox.ac.uk/`
[3] `http://www.gondtc.com/web-cams/main-street-large.htm`

Table 2. Number of correctly tracked (a) pedestrians and (b) vehicles in each scenario, and amount of observation errors due to: (c) agent misclassification, (d) lost or missed tracks, and (e) false detections

	Correct			Erroneous			
Scenario (total tracks)	(a)	(b)	Total	(c)	(d)	(e)	Total
Hermes (161)	103	26	129	13	10	9	32
Oxford centre (180)	87	62	149	20	8	3	31
Devil's Lake (179)	49	98	147	17	10	5	32
Kingston–1 (161)	85	53	138	12	9	2	23
Kingston–2 (87)	35	33	68	7	4	8	19

complete, and the BP algorithm with maximum iterations takes approximately 90 seconds in a Pentium II 3 GHz machine with 2 Gb RAM.

We analyze the consistency of the results by testing over a wide range of grid size values, which is the main parameter involved in the sampling process: given that each control point sampled from a trajectory affects uniquely its enclosing region, the number of cells tesselating the scenario is indicative of the area of influence of tracked objects during region modeling. The dimensions of the projected grid in our experiments range from 40×40 to 150×150. Lower cell resolutions do not capture the details of the scenario, thus not being suitable to model semantic regions. Greater resolutions show performances that are similar to the displayed range, but require substantially more computational resources.

Additionally, the tracked trajectories used as observations incorporate errors. Each error consists of one or more of the following cases: misclassification of agents, lost tracks, and false detections. Table 2 gives numerical information on the agents involved in each scenario and the number and type of erroneous observations. The system has been evaluated with and without the presence of errors, in order to test its robustness.

6.1 Quality Scores

Performances have been evaluated in terms of accuracy. Three scores have been considered: overall accuracy (OA), segmentation accuracy (SA), and weighted segmentation accuracy (WSA). The two former scores are defined by

$$OA = \frac{TP+TN}{TP+FP+TN+FN}, \qquad SA = \frac{TP}{TP+FP+FN},$$

where OA is traditional accuracy, typically overfavored in multiclass contexts given the high value of True Negatives as the number of classes increments. For this reason, SA has been increasingly used to evaluate multiclass segmentations, as in the PASCAL-VOC challenge [4]. Additionally, WSA is defined by

$$WSA = \frac{TP^*}{TP^*+FP^*+FN^*},$$

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

in which an assignment is now considered true positive if the inferred label is either equal to the ground truth, or is its direct generalization in the taxonomy of Table 1a; and negative otherwise, thus modifying the account of errors. For instance, an actual *parking* is here positively labeled as *road*, and a *pedestrian waiting zone* is correctly labeled as *sidewalk*. Note that this score does not necessarily benefit our approach, since our smoothness constraints do not award class generalization. Instead, the goal of this metric is to penalize wrong particularizations. GT evaluation in Fig. 4a shows that WSA takes into account consistency in different GT realizations –unlike SA–, while penalizing differences harder than OA.

6.2 Median Filter

We have compared our method to median filters. They are the most used nonlinear filters to remove impulsive or isolated noise from an image, a typical type of noise found in our problem domain. Median filters preserve sharp edges, which makes them more robust than traditional linear filters and a simple and cheap solution to achieve effective non-linear smoothing. They are commonly used for applications of denoising, image restoration, and interpolation of missing samples, all of which are applicable in our context.

We have compared the performances obtained by a median filter after 15 iterations and by our proposed inference framework, to evaluate the contributions of taxonomy-based constraints to the smoothing task. The filter is applied for each marginal probability map $P(f_r = l), l = 1 \dots |\mathcal{L}|$, maintaining the MRF neighborhood defined. A median-filtered labeling is performed by assigning the most probable label to each region, once the process has converged or exceeded the maximum number of iterations allowed.

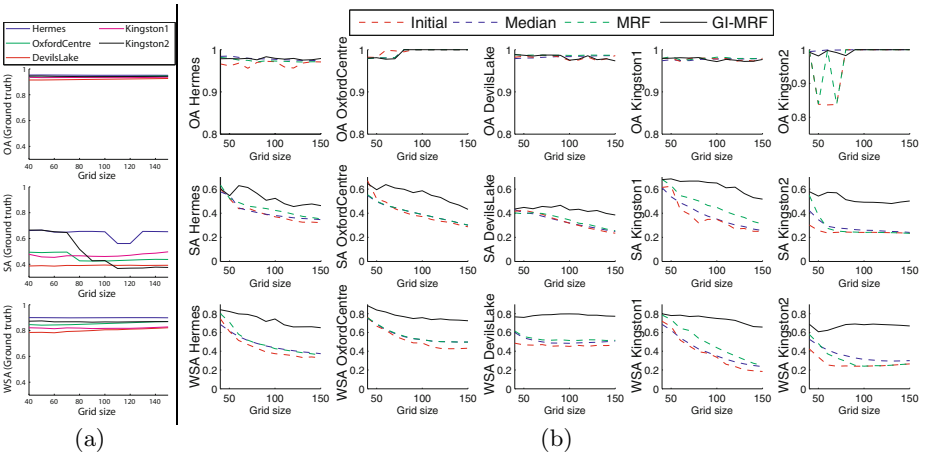


Fig. 4. (a) Evaluation of the inter-observer variability in GT segmentations. (b) Statistical scores for the 5 considered scenarios. More details in the text.

Table 3. Quantitative OA, SA, and WSA scores for a grid size of 75×75, without and with the presence of erroneous trajectories

		Overall accuracy (OA)				Segmentation accuracy (SA)				Weighted segmentation accuracy (WSA)			
		Initial	Median	MRF	GI–MRF	Initial	Median	MRF	GI–MRF	Initial	Median	MRF	GI–MRF
Only correct	Hermes	0.98	0.96	0.97	0.98	0.40	0.40	0.45	**0.64**	0.50	0.44	0.51	**0.77**
	Oxford Centre	0.98	0.97	0.98	0.98	0.46	0.52	0.58	**0.61**	0.65	0.66	0.75	**0.93**
	Devil's Lake	0.98	0.99	0.99	0.99	0.37	0.39	0.39	**0.44**	0.49	0.46	0.52	**0.78**
	Kingston–1	0.98	0.97	0.98	0.99	0.43	0.37	0.50	**0.66**	0.46	0.44	0.59	**0.76**
	Kingston–2	1.00	0.84	1.00	0.98	0.27	0.24	0.28	**0.56**	0.36	0.24	0.35	**0.69**
	Average	0.98	0.94	0.98	0.98	0.39	0.38	0.44	**0.58**	0.49	0.45	0.54	**0.78**
Correct and erroneous	Hermes	0.98	0.97	0.97	0.98	0.40	0.40	0.45	**0.53**	0.51	0.45	0.52	**0.78**
	Oxford Centre	0.98	0.97	0.98	0.98	0.46	0.53	0.56	**0.57**	0.66	0.68	0.76	**0.94**
	Devil's Lake	0.98	0.99	0.99	0.99	0.37	0.39	0.40	**0.43**	0.50	0.47	0.53	**0.78**
	Kingston–1	0.97	0.98	0.98	0.98	0.43	0.40	0.50	**0.65**	0.46	0.50	0.60	**0.76**
	Kingston–2	1.00	0.84	0.99	0.98	0.28	0.24	0.34	**0.55**	0.38	0.26	0.40	**0.76**
	Average	0.98	0.95	0.98	0.98	0.39	0.39	0.46	**0.55**	0.50	0.47	0.56	**0.80**

6.3 Results

Fig. 4a shows the results of the inter-observer evaluation for the GT, which constitute the baseline of the system's performance. Fig. 4b shows quantitative scores for OA, SA, and WSA in the 5 scenarios. Each plot draws the results of 4 different approaches, applied to the 5 series of GT available. These approaches correspond to: (i) assigning labels using only observed evidence from trajectories, i.e., neglecting smoothness priors (*Initial*); (ii) using a median filter over the initial models (*Median*); (iii) applying MAP-MRF inference (Eq. 1) to the initial models (*MRF*); and (iv) applying a preprocessing step based on geodesic interpolation to the region models (*GI–MRF*).

Results are similar to GT inter-observer variability. Only occasional plot oscillations appear in Kingston2 for the OA measure, due to the non-linear operation of sampling GT images into lattices of a concrete size. Moreover, increasing the cell resolution progressively lowers the quality of the initial models, as well as the accuracy on posterior labelings. Nonetheless, it is shown that interpolation grants a performance almost invariant to the grid size used. This is emphasized in case of poor sampling, e.g, parking spaces.

Table 3 shows numerical results for a grid of 75×75 cells, with and without noisy trajectories. As seen in this table, OA is excessively favored due to the high number of true negatives produced in a multiclass context, thus suggesting SA and WSA as more convenient to compare the different techniques. Particularly, WSA should be interpreted as the precaution to avoid wrong particularizations. With these metrics, experiments using geodesic interpolation and smoothness constraints practically always achieve the maximum score, whereas a median filter fails dramatically as the grid resolution increments, or in case of

ill-convergence, e.g., it fails to preserve parking regions in *Kingston–2*. Additionally, it is seen that even by incorporating erroneous trajectories to the datasets, letting them be about a 20% of the total, the accuracy values remain stable.

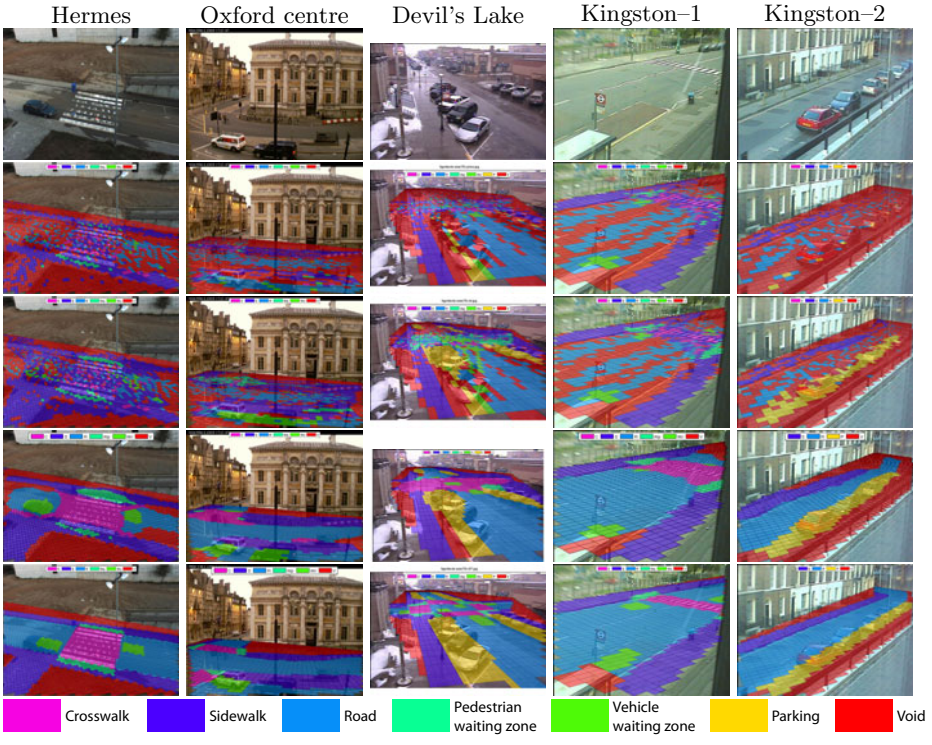


Fig. 5. Labeling step results for a 75×75 grid: First row shows original image, second row the initial labeling only from observations, third row the labeling with geodesic interpolation, fourth row the inference labeling using both interpolation and smoothness constraints, and the bottom row shows the GT. Best viewed in color.

Fig 5 depicts qualitative step results of the labeling process for a grid size of 75×75. For visualization purposes, results are shown within a ROI. The depicted results represent the activity of the tracked objects, rather than the visual appearance of the scenario. Instead, appearance is commonly used to guide manual labelings. We also identify an edge-effect of *Void* regions, given that control points near the edges often lack of precedent or consecutive samples to update their regions. This happens especially for vehicles, due to their higher speed and poorer sampling. Finally, cases of intense projectivity –e.g., *Devil's Lake*–, make it more difficult for the models to emphasize the presence of connected regions, thus provoking generalized smoothing.

7 Application

Finally, the conceptual scene model have been used to exploit model-based behavior analysis. This has been achieved using the Situation Graph Tree (SGT) [10] shown in Fig. 6(a,b), although any symbolic approach requiring conceptual scene models could be used instead, like Petri Nets or Symbolic Networks. We choose

Fig. 6. SGT used to interpret behaviors of (a) vehicles and (b) pedestrians. (c,d) Selected frames from each interval. Semantic predicates are generated deterministically using (e) the learned region maps and (f) their corresponding GT maps.

SGTs because they reason about the events observed in the learned semantic regions, and can annotate situations of interest and traffic behaviors.

The scenario-independent SGT used generates conceptual descriptions when certain conditions happen, such as vehicles entering sidewalks or pedestrians entering roads. In addition, basic interpretations are formulated; e.g., if a vehicle stops in front of a crosswalk where a pedestrian is found, it is *giving_way* to this person; and if a vehicle stops in a parking, it has *parked* there. In essence, basic conceptual predicates are inferred by a *fuzzy metric-temporal reasoner*, we refer the reader to [10,11] for implementation details.

Fig. 6 shows predicates generated in *OxfordCentre* and *DevilsLake* at different time intervals. Most frequently, the generated predicates differ only at the beginnings or endings of the temporal intervals; this is due to slight variations among region boundaries. In Fig 6c, two predicates from the left column are not found in the right one, since a *WZc* zone has not been identified in the GT model. Nevertheless, alarms and simple interpretations are correctly generated.

8 Conclusions

We have shown an effective motion-based method to automatically label semantic zones. The method has been applied to different urban scenarios using the same behavioral models. Our approach enhances state-of-the-art on background labeling by using taxonomical knowledge to guide consistent inferences during labeling. It is scene-independent, viewpoint-invariant and of reduced computational cost, for it does not require to compute costly image descriptors.

Initial region models are learned from trajectory features, and updated as new trajectories are available. Smoothness is taken into account using a MAP-MRF inference, whose parameters are conditioned by prior taxonomical domain knowledge. The framework is scenario-independent: it has been applied to 5 datasets showing different conditions of projectivity, region content and configuration, and agent activity. Step results are shown at every stage of the process, to capture the particular contributions of each task. The method has been compared to a median filter, showing its better performance on the 3 scores tested.

Our work makes it possible to use predefined behavior models in generic surveillance scenes. By automatically learning the conceptual scene model behind lots of outdoor scenes, we can evaluate existing deterministic models (SGT, Petri Nets, Symbolic Networks) in terms of generalization or scaling criteria. Further steps include improving the accuracy of inter-region boundaries and extending the system to indoor scenarios. Such environments incorporate more complex semantics on agent actions and interactions, so deterministic behavior models using domain knowledge can be used to extract key concepts for annotation.

Acknowledgements

This work has been supported by the Spanish Research Programs Consolider-Ingenio 2010:MIPRCV (CSD200700018) and Avanza I+D ViCoMo

(TSI-020400-2009-133); and by the Spanish projects TIN2009-14501-C02-01 and TIN2009-14501-C02-02. We acknowledge the valuable collaboration of Dr. Pau Baiget.

References

1. Robertson, N., Reid, I.: A general method for human activity recognition in video. CVIU 104, 232–248 (2006)
2. Ballan, L., Bertini, M., Serra, G., Del Bimbo, A.: Video annotation and retrieval using ontologies and rule learning. IEEE Multimedia (2010)
3. Makris, D., Ellis, T.: Learning semantic scene models from observing activity in visual surveillance. IEEE TSCM, Part B 35, 397–408 (2005)
4. Hu, W., Xiao, X., Fu, Z., Xie, D.: A system for learning statistical motion patterns. PAMI 28, 1450–1464 (2006)
5. Piciarelli, C., Foresti, G.L.: On-line trajectory clustering for anomalous events detection. PRL 27, 1835–1842 (2006)
6. Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: CVPR, Anchorage, USA (2008)
7. Baiget, P., Sommerlade, E., Reid, I., Gonzàlez, J.: Finding prototypes to estimate trajectory development in outdoor scenarios. In: 1st THEMIS, Leeds, UK (2008)
8. Wang, X., Tieu, K., Grimson, E.: Learning semantic scene models by trajectory analysis. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 110–123. Springer, Heidelberg (2006)
9. Li, J., Gong, S., Xiang, T.: Scene segmentation for behaviour correlation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 383–395. Springer, Heidelberg (2008)
10. Nagel, H.H., Gerber, R.: Representation of occurrences for road vehicle traffic. AI-Magazine 172, 351–391 (2008)
11. Gonzàlez, J., Rowe, D., Varona, J., Xavier Roca, F.: Understanding dynamic scenes based on human sequence evaluation. IVC 27, 1433–1444 (2009)
12. Albanese, M., Chellappa, R., Moscato, V., Picariello, A., Subrahmanian, V.S., Turaga, P., Udrea, O.: A constrained probabilistic petri net framework for human activity detection in video. IEEE TOM 10, 982–996 (2008)
13. Fusier, F., Valentin, V., Bremond, F., Thonnat, M., Borg, M., Thirde, D., Ferryman, J.: Video understanding for complex activity recognition. MVA 18, 167–188 (2007)
14. Kumar, M., Torr, P., Zisserman, A.: Obj. Cut. In: CVPR (2005)
15. Kumar, S., Hebert, M.: Discriminative fields for modeling spatial dependencies in natural images. In: Advances in Neural Information Processing Systems, vol. 16 (2004)
16. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: CVPR, pp. 37–44 (2006)
17. Shotton, J., Johnson, M., Cipolla, R., Center, T., Kawasaki, J.: Semantic texton forests for image categorization and segmentation. In: CVPR (2008)
18. Li, S.: Markov random field modeling in image analysis. Springer, Heidelberg (2001)
19. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge Univ. Press, Cambridge (2003)
20. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. IJCV 70, 41–54 (2006)

21. Croft, W., Cruse, D.: Cognitive linguistics. Cambridge Univ. Press, Cambridge (2004)
22. Rowe, D., Gonzàlez, J., Pedersoli, M., Villanueva, J.: On tracking inside groups. Machine Vision and Applications 21, 113–127 (2010)
23. Bose, B., Grimson, E.: Improving object classification in far-field video. In: CVPR (2004)
24. Black, J., Makris, D., Ellis, T.: Hierarchical database for a multi-camera surveillance system. Pattern Analysis and Applications 7, 430–446 (2004)
25. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33, 159–174 (1977)