Spatio-Temporal Clustering of Probabilistic Region Trajectories

Fabio Galasso [1] Masahiro Iwasaki [2] Kunio Nobori [2] Roberto Cipolla [1]
[1] Department of Engineering, University of Cambridge, United Kingdom
[2] Advanced Technology Research Laboratory, Panasonic Corporation, Japan

Abstract

We propose a novel model for the spatio-temporal clustering of trajectories based on motion, which applies to challenging street-view video sequences of pedestrians captured by a mobile camera. A key contribution of our work is the introduction of novel probabilistic region trajectories*, motivated by the non-repeatability of segmentation of frames in a video sequence. Hierarchical image segments are obtained by using a state-of-the-art hierarchical segmentation algorithm, and connected from adjacent frames in a directed acyclic graph. The region trajectories and measures of confidence are extracted from this graph using a dynamic programming-based optimisation. Our second main contribution is a* Bayesian framework *with a twofold goal: to learn the optimal, in a maximum likelihood sense, Random Forests classifier of motion patterns based on video features, and construct a unique graph from region trajectories of different frames, lengths and hierarchical levels. Finally, we demonstrate the use of* Isomap for effective spatio-temporal clustering *of the region trajectories of pedestrians. We support our claims with experimental results on new and existing challenging video sequences.*

1. Introduction

Fig. 1 illustrates a basic task attracting increasing attention in the Vision community: the detection, segmentation and tracking of moving pedestrians. The topic is regarded for purposes such as surveillance, automatic navigation, or as the base for action recognition and video database indexing. Many factors concur to make it extremely challenging: camera motion, non-rigid deformations, perspective and scale changes, occlusion, illumination changes *etc*.

Recent results in image segmentation have provided important tools for clustering the image pixels based on appearance. However, it is widely agreed that segmentation is ambiguous in most cases, *e.g.* a dark coat is often more similar to the background than to the cloths of the same person. Furthermore, the "right" segmentation needs usually defining a task, *e.g.* a vision system may be employed to detect pieces of clothing, human silhouettes, or a moving crowd.

Contemporary people detectors achieve this goal by learning the appearance and shape from single frames, often using images annotated with bounding boxes. While working commercial solutions are already available, proving that repeatable results may be obtained with these algorithms, one can anticipate the need for larger databases for including all the possible appearances of the world pedestrians.

At the junction of unsupervised image segmentation and appearance learning-based methods, we present a new approach to spatio-temporal clustering based on learnt motion patterns. Motivated by the non-repeatability of segmentation of frames in a video sequence, in Sec. 3 we introduce novel *probabilistic region trajectories*. Based on a hierarchical segmentation algorithm [1] and on dynamic programming (DP) optimisation, the probabilistic region trajectories aim to track the most probable image segments across the frames and the hierarchical levels of the video sequence.

A *Bayesian framework* is introduced in Sec. 4 to address our novel purposes: *defining a task* for the spatio-temporal clustering, and learning complex motion patterns from novel and more descriptive low-level *video features*, as opposed to traditional frame-based features. There, we also illustrate the effective use of the Isomap technique.

Our model, generally applicable to any motion pattern, is validated in Sec. 5 on the difficult scenario of pedestrians in crowds, in new and existing outdoor video sequences. We conclude the paper and discuss future prospects in Sec. 6.

2. Related Work

The spatio-temporal clustering of region trajectories is closely related to the problems of detection, segmentation and tracking. Especially for the case of pedestrians, a number of techniques have been recently proposed in these areas, based on learning the appearance of body parts ([10, 22, 14]), particle filters ([9, 12]), and feature point trajectories ([4, 17, 18, 7, 8, 5]). The presented work proceeds in a different direction from [22, 14, 10], as it is based on motion cues, believing that pedestrians have extremely variable appearances. Similarly to [4, 17, 18, 7, 8, 5] our work is based on the clustering of spatio-temporal trajectories.

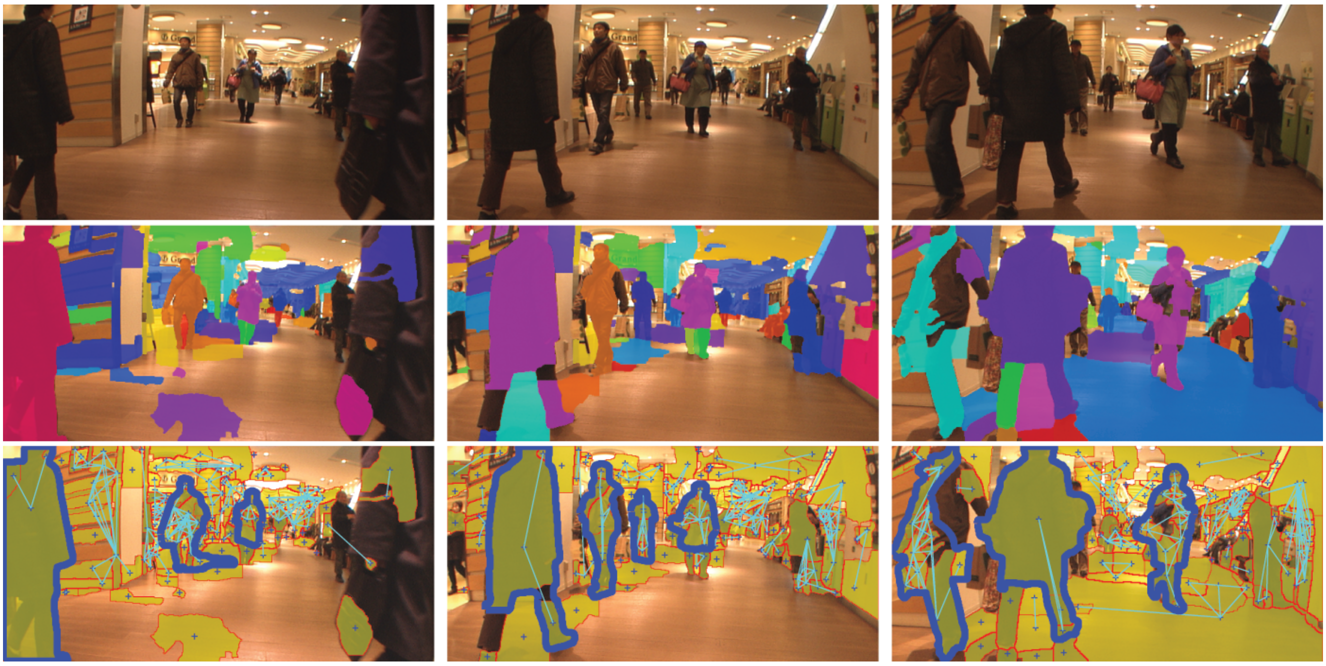

Figure 1. (*Top*) Video sequence Lazona07 from our new dataset. Walking pedestrians are captured by a camera mounted at street-view level on a mobile equipment. (*Middle*) The clusters of region trajectories computed with our algorithm are represented with different shaded colours (some colours may be repeated). (*Bottom*) A graph representation of the result provides insights into the algorithm: pixels belonging to region trajectories are shaded in yellow; boundaries of regions are marked with red lines; centroids of regions are marked with little blue crosses; pale blue lines connect region trajectories in the same cluster (a connection between region trajectories regards the regions as a whole, centroids are just used for visualization purposes). To enhance the readability, thicker blue lines are fitted to the contours of the clustered region trajectories from pedestrians with active contours [6]. It is desirable that all pixels on pedestrians be shaded in the same colour in the middle illustrations, and that all region trajectories belonging to pedestrians be connected by pale blue lines in the bottom, *i.e*. clustered together. Unshaded pixels do not belong to any region trajectory and should not penalize the result, as a video segmentation is beyond the purpose of the paper. The figure shows an appropriate clustering of most region trajectories, but also some flaws, *i.e*. the cluster on the person walking forward is re-initialized, some region trajectories on limbs are misclassified.

However we define a probabilistic framework, allowing for a task-oriented clustering, and adopt more informative region trajectories. Furthermore, our testing sequences are of unprecedented difficulty for tasks of spatio-temporal clustering: complex crowd scenes acquired by a mobile camera at the height of a meter from the ground, therefore with notable perspective effects.

The proposed method for extracting region trajectories is closely related to the works of [13] and [21]. The former extracts probabilistic point trajectories of feet from a graph of temporal point correspondences. Similarly to [21], we use DP for extracting region trajectories from a graph of interconnected image segments. However [21] aims to unsupervised segmentation of video sequences and emphasizes the Conditional Random Field model, while the extracted trajectories are based on simple photometric similarities. By contrast, we allow for encoding a task and aim to cluster the stable and semantically meaningful region trajectories. Our results are close to a video segmentation at some frames, but this should be considered a side effect of a temporally consistent clustering. Our work may also be related to [2], which provides a generative formulation of region trajectories, limited however by an excessive computational load.

A variety of techniques are available for clustering in unsupervised learning. A recent trend has been using an eigendecomposition for obtaining a lower-dimensional embedding of the data onto a non-linear manifold. This includes Isomap [19] and many variants of spectral clustering [16, 11]. These techniques may capture non-convex and variously shaped clusters and they all require additional parameters, often the number of clusters looked for. Our algorithm is not constrained to use any particular method, but Isomap is chosen for its attractive mathematical properties.

3. Probabilistic Region Trajectories

The extraction of region trajectories has received comparatively very little attention in vision because of the well-known non repeatability of image segmentation over the frames of a video sequence. This motivates the introduc-

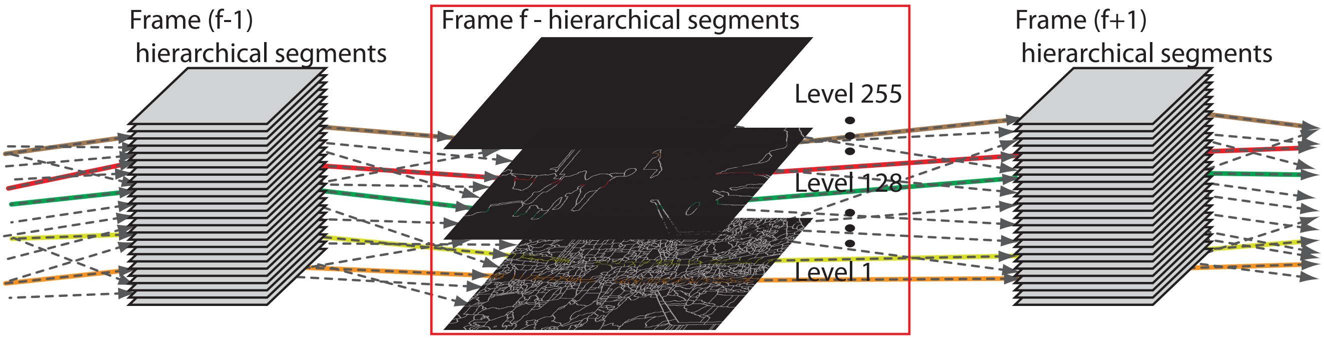

Figure 2. Extraction of probabilistic region trajectories. Hierarchical segments (levels 1-128-255 represented in the central *red box*) are connected from adjacent frames at multiple levels in the hierarchy into a directed acyclic graph (*dotted lines*), with edges weighted using an optical flow-based propagation. The probabilistic region trajectories are extracted from the graph with DP optimization (*colored lines*).

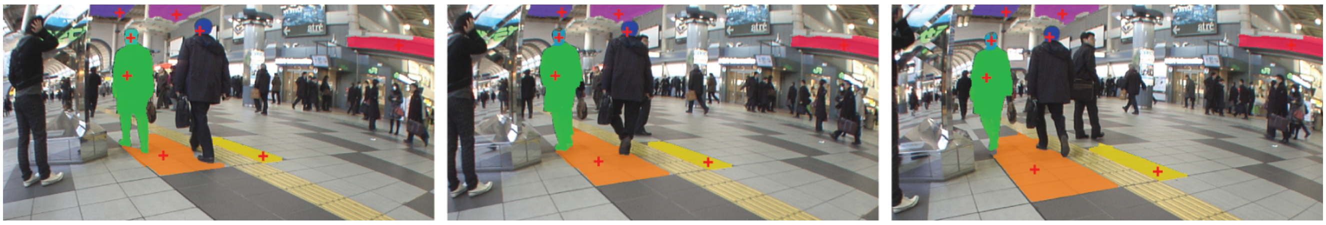

Figure 3. Colored areas represent 8 sample probabilistic region trajectories. *Red crosses* represent the propagated centroids of the regions.

tion of novel probabilistic region trajectories.

We use the hierarchical segmentation algorithm of [1] to segment each frame of the video sequence into segments, or regions, at 255 coarse-to-fine levels, according to several criteria, *i.e*. illumination, color, texture *etc*. Subsequently, a directed acyclic graph $G = (V, E)$ is built by connecting the segments from adjacent frames across all hierarchical levels, as illustrated in Fig. 2. The frame-based segmentations are generally non-repeatable at a particular level. However parts of the image are consistently segmented at certain coarse-to-fine levels. We term the region trajectories "probabilistic" because we extract the most likely paths (the probabilistic region trajectories) at the most convenient level of the hierarchy, alongside a measure of confidence.

In more details, let us consider two regions, r_f and r_{f+1}, respectively at frames f and $f+1$. The weight of the corresponding graph edge captures the similarity of r_f and r_{f+1}. It is given by the cross-correlation product $s_{r_f,r_{f+1}}$ between the "propagated" mask of the region r_f to frame $f+1$, m'_{r_f}, and the mask of r_{f+1} at that frame, $m_{r_{f+1}}$:

$$s_{r_f,r_{f+1}} = \frac{2|m'_{r_f} \cap m_{r_{f+1}}|}{|m'_{r_f}| + |m_{r_{f+1}}|} \tag{1}$$

The mask propagation consists on projecting each pixel of the region to the following frame by using an optical flow. In particular, we use the optical flow algorithm of [23] and smooth its output by bilateral filtering to preserve the boundary sharpness [15]. The computed flow at pixel $\mathbf{x}$, $\mathbf{u}(\mathbf{x})$, is averaged with the flow at its neighbouring pixels *within the region*, $\text{Ne}_{r_f}(\mathbf{x})$, weighted by spatial proximity and motion coherency ($\sigma_x = 6.9$, $\sigma_m = 3.6$):

$$\bar{\mathbf{u}}(\mathbf{x}) = \frac{\Sigma_{\mathbf{x}' \in \text{Ne}_{r_f}(\mathbf{x})} \mathbf{u}(\mathbf{x}') w(\mathbf{x}, \mathbf{x}')}{\Sigma_{\mathbf{x}' \in \text{Ne}_{r_f}(\mathbf{x})} w(\mathbf{x}, \mathbf{x}')} ,$$
$$w(\mathbf{x}, \mathbf{x}') = N(\mathbf{x}; \|\mathbf{x} - \mathbf{x}'\|, \sigma_x)\ N(\mathbf{x}; \|\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{x}')\|, \sigma_m) \ . \tag{2}$$

The region trajectories are extracted as the shortest paths in the graph by using a DP optimization, in a forward and backward pass to remove spurious correspondences. As similarly noted in [21], the extraction of shortest paths may be thought as drawing region trajectories from an undirected graphical model, where each frame is associated to a variable, z_f, to represent the selection of a segment from that frame, and is connected to its adjacent-frame variables, z_{f-1} and z_{f+1}. By defining binary potentials $\Psi(z_f, z_{f+1})$ as the similarities $s_{r_f,r_{f+1}}$, the probability of a region trajectory is expressed as the joint distribution of the corresponding sequence of segments $z_0, z_1, \ldots, z_F$:

$$P(z_0, z_1, \ldots, z_F) \propto \prod_{f=0}^{F-1} \Psi(z_f, z_{f+1}) \ , \tag{3}$$

Figure 3 shows some region trajectories spanning different resolutions. Compared to point trajectories, our region

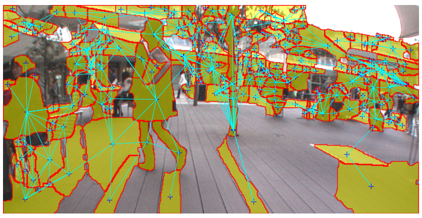

Figure 4. Spanning tree for region trajectories of length 5 frames ($L = 2$). Note that: (i) the connectivity is given by the whole regions, touching regions yield distance 1 and "continuous" connectivity, centroids are just used for visualization; (ii) not connected regions, *e.g.* due to partial occlusions, may still be clustered together if they are connected at any other frame (graph transitivity).

trajectories present numerous advantages, most notably: (i) they are dense, so we may associate several statistics to them, as we discuss in Sec. 4; (ii) they are generally robust to deformation, partial occlusion, scale change and perspective effects, *e.g.* see the disoccluding floor tiles in Fig. 3; (iii) they track "semantic" image parts, as in most cases stable image segments are semantically consistent.

4. Bayesian Framework

Let us consider a set of region trajectories $\{X_i\}_1^N$ and a symmetric distance measure $Z(X_i, X_j)$. Our task is to assign each X to a cluster $\{c\}_1^C$ and to define the number of clusters C. Given C, the best clustering hypothesis Θ is the maximum a-posteriori (MAP) estimate:

$$\hat{\Theta}_{\text{MAP}} = \arg\max_{\Theta} P(\Theta|Z)\ . \tag{4}$$

With Bayes's rule, we express the posterior in terms of the likelihood and prior probabilities:

$$P(\Theta|Z) = \frac{P(\mathbf{Z}|\Theta)P(\Theta)}{P(\mathbf{Z})} = \frac{P(\mathbf{Z}|\Theta)P(\Theta)}{\int_{\Theta} P(\mathbf{Z}|\Theta')\,P(\Theta')\,d\Theta'}\ . \tag{5}$$

Estimating the MAP, would require the evaluation of all possible clustering hypotheses Θ, a combinatorial of the number of region trajectories. Furthermore the estimation should be repeated for every possible number of clusters C and model selection should be taken into account. The process is intractable even for few trajectories. Alternatively, we use a Bayesian framework to determine the optimal, in a maximum likelihood sense, probability of clustering between region trajectories based on motion-based video features, and construct a unique graph from tracks of different frames, lengths and hierarchical levels. Then we employ Isomap [19] for effective spatio-temporal clustering.

Prior We model in the prior the connectedness of the graph. We assume that only neighboring segments may be clustered together and extend this to the region trajectories. Let us consider all the region trajectories which exist at frame f and in the range $[f - L_{\text{min}}, f + L_{\text{min}}]$. A spanning tree is constructed according to the criterion of minimum and equal distances (we use Kruskal's algorithm, modified to include equal distances at each iteration). The prior probability $P(\Theta)$ is uniform over the connected nodes and zero otherwise. At the same frame, this connectivity is preserved for graphs of longer region trajectories, existing in the ranges $[f - L, f + L]$, $L \in (L_{\text{min}}, L_{\text{max}}]$ (in the experiments $L_{\text{min}} = 2$, $L_{\text{max}} = 16$, corresponding to region trajectories of lengths 5 to 33 frames).

Likelihood Let us assume $P_Z(X_i, X_j)$, the probability that two region trajectories X_i and X_j be clustered together. We define the likelihood of a clustering hypothesis Θ, $P(\mathbf{Z}|\Theta)$, as

$$P(\mathbf{Z}|\Theta) = \prod_{i,j \in c_k, k \in [1,C]} P_Z(X_i, X_j) \quad \times \prod_{i \in c_k, j \in c_m, k,m \in [1,C], k \neq m} (1 - P_Z(X_i, X_j))\ . \tag{6}$$

$P(\mathbf{Z}|\Theta)$ is computed along the spanning tree and is the product of the probabilities between region trajectories clustered together, according to Θ, times the product of the complementary probabilities between trajectories assigned to different clusters. Eqn. 6 naturally penalises separating trajectories with high linking probability. From a model selection point of view, this corresponds to penalizing models with higher complexity (larger number of clusters C).

Video features The estimation of the probability of clustering $P_Z(X_i, X_j)$ between region trajectories motivates the introduction of novel video features.

A symmetric distance measure Z between two region trajectories may be based on a variety of statistics, which may be computed on the segments, or parts of them, forming the trajectories. In the same way a dense set of region trajectories can compactly represent the whole video sequence, computing *video* features on these trajectories may theoretically provide all the statistics measurable from the video itself, not just limited to corners and blobs, nor to single frames. The video features, made available from representing the video in terms of region trajectories, can describe low- and higher-level motion cues, as well as colour-, texture- and, in general, appearance-based statistics.

We are interested in *motion-based clustering with low-level cues* and therefore define video features between two region trajectories as combinations of: SUPPORTS (a) the entire area of the regions in the trajectories, (b) the neighbouring portions of areas of the regions, (c) the propagated centroids; LOW-LEVEL MOTION STATISTICS (1) mean, (2) median, (3) min, (4) max, (5) standard deviation, (6) variance, (7) Fourier transform; QUANTITIES (i) position, (ii) veloc-

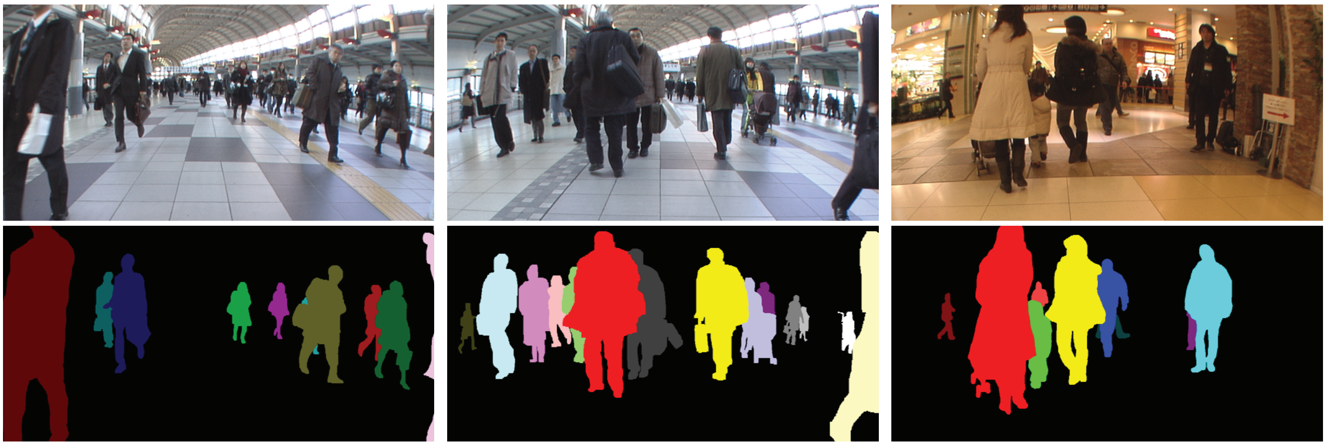

Figure 5. (*top row*) Sample training video sequences and (*bottom row*) the provided ground truth labelling. This work focuses on pedestrians, neglecting all other objects and the background (represented with a black/zero value).

ity, (iii) velocity normalized by the local flow, (iv) acceleration, (v) acceleration normalized by the local flow.

Any combination of supports (a-c), statistics (1-7) and quantities (i-v) defines a simple distance Z between two region trajectories based on their overlapping frames. We do not arbitrarily select a particular Z (*e.g.* compare [4, 5]). Instead we associate a vector of all available Z's to each pair of region trajectories and learn a Random Forests classifier [3] on the two class problem: two region trajectories belong to the same Vs different clusters. The optimal probability P_Z is given by the voting ratio of the decision trees (300 trees with minimum node size 36).

In Sec. 5 we consider the complex motion patterns of pedestrians and the *task* of clustering their relevant trajectories.

Graph-based clustering A graph allows clustering non-convex and variously shaped structures. We assemble the graphs for each frame and length into a unique graph.

The unique graph has the region trajectories from the whole sequence as nodes. Edges are drawn in the graph according to the links in the single graphs and assigned the probability of clustering between the trajectories. As two region trajectories X_i and X_j may be linked over several frames and lengths, the evidence needs to be accumulated. Two or more instances of clustering probabilities between X_i and X_j, $P'_{Z_{ij}}$ and $P''_{Z_{ij}}$, are therefore multiplied and renormalised, so that the resulting $P_{Z_{ij}}$ is consistent with Eqn. 6:

$$P_{Z_{ij}} = \frac{P'_{Z_{ij}} P''_{Z_{ij}}}{P'_{Z_{ij}} P''_{Z_{ij}} + (1 - P'_{Z_{ij}})(1 - P''_{Z_{ij}})} \qquad (7)$$

Clustering technique The spatio-temporal clustering of the region trajectories is achieved using the Isomap algorithm [19]: this embeds the trajectories into a lower dimensional non-linear manifold where the actual clustering is obtained by the use of the K-means clustering technique. The proposed algorithm processes all the frames "in batch", as a sufficient number of trajectories needs be considered for capturing the curvature of the manifold.

5. Experimental validation

Here we discuss experiments on challenging outdoor video sequences. Results are task-oriented aggregation steps, *e.g.* relevant to subsequent detection and recognition.

Dataset description Recently, [5] has provided a well designed dataset of 26 video sequences, 12 of which were chosen from the Hopkins 155 database [20]. For evaluation purposes, [5] provides meaningful error metrics and an evaluation software, alongside a ground-truth labelling of sparse frames, with labels including pedestrians, cars, and various other objects. As discussed in Sec. 4, our model may be employed to learn different motion patterns but this study emphasizes the most challenging case of pedestrians. We select therefore the 16 sequences containing people, as detailed in Tab. 1. To limit the computational load we consider the first 100 frames only.

Our main purpose is to test our algorithm on complex crowded scenes. We introduce therefore a new dataset of 15 ground-truthed video sequences, captured in the Tokyo undergrounds of Shinagawa and Kawasaki, from a camera mounted on a mobile equipment (fish-eye lens and HDi resolution). Samples of the extracted images (de-interlaced, cropped and resized to 600x300) and the respective labels are illustrated in Fig. 5. Notably the camera height from the ground plane is about a meter, generating strong perspective effects on the approaching pedestrians. The labelling poses the attention on the pedestrians and their belongings (briefcases, bags *etc.*) and neglects all other objects and the background, to which a zero value is assigned.

We train on 10 sequences from our dataset, and test on the

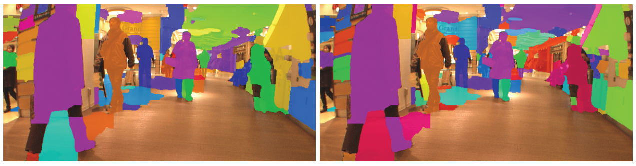

Figure 6. The clusters of region trajectories on the pedestrians are well determined and, to a large extent, the input number of clusters does not affect the segmentation output. no.clsuters=180 (*left*) and the no.clusters=380 (*right*) yield similar results as in Fig. 1(*middle column*), obtained for no.clsuters=90: as is desired, more clusters only fragment the objects and background.

remaining 5 and on all selected videos from [5]. Notably, the camera setup, and the camera and pedestrian motions in the videos of [5] differ significantly from ours, so testing on those sequences also provides testing for generalization.

Error metrics We take care to provide a comprehensive quantitative evaluation of our algorithm. To this purpose we gather the available error metrics from relevant works on spatio-temporal clustering [5] and video segmentation [21]. Tab. 1 reports five metrics *computed over the clusters of pixels from pedestrians*: (i) **density**: the percentage of correctly classified pixels over the total number of pixels, from [5], it may be paired with *recall* [1]; (ii) **overall clustering error**: the percentage of bad pixel labels over the total number of labels, from [5], it complements *precision* [1]; (iii) **average clustering error**: similar to the overall error, but given by averaging over each cluster first, from [5]; (iv) **segmentation covering**: a general measure of how well the computed clusters of pixels superpose on the true segmentation, given by the Dice coefficient between the sets, from [21], it relates to *F-measure* [1]; (v) **head count**: the number of pedestrians considered (still persons or people further away with respect to the main actors are neglected). Note that in [5] only a global evaluation of metrics over the all video sequences is available (for variable lengths) and that this includes other objects alongside pedestrians. We present detailed metrics for each sequence separately to offer closer evaluation and future comparison, but *our metrics only regard pedestrians*, which must be kept in mind in a comparison to [5]. Additionally, we do not report a metric on over-segmentation error because we only allow one cluster to explain a single pedestrian at each frame.

Discussion The probabilistic region trajectories from pedestrians are successfully clustered at most frames. The clustering errors reported in Tab. 1 are comparable to those from [5] ($\epsilon_{\text{overall}}$: ours is 17.39% against 7.13%[1]; $\epsilon_{\text{average}}$: ours is 17.52% against 31.14%[1]) but our density indexes are an order of magnitude greater (ρ: ours is 50.62% against 3.27%[1]). Additionally, the reported average segmentation covering index is greater than in a state-of-the-art video segmentation work [21] (SC: ours is 0.62% against 0.52%), opening the way to extending this work to motion-based video segmentation, see discussion in Sec. 6.

V Sequences	ρ	$\epsilon_{\text{overall}}$	$\epsilon_{\text{average}}$	SC	HC
Dataset of [5]					
Miss Marple1	44.02%	0.25%	0.25%	0.61	1
Miss Marple2	69.62%	7.93%	7.93%	0.79	1
Miss Marple3	40.31%	16.47%	16.47%	0.54	1
Miss Marple4	32.35%	24.58%	24.58%	0.45	1
Miss Marple5	66.75%	8.69%	8.69%	0.77	1
Miss Marple6	63.19%	37.18%	37.18%	0.63	1
Miss Marple7	52.58%	5.53%	5.53%	0.68	1
Miss Marple8	13.19%	38.48%	38.48%	0.22	1
Miss Marple9	42.05%	2.21%	3.20%	0.59	2
Miss Marple10	42.84%	60.24%	60.24%	0.41	1
Miss Marple11	54.65%	4.76%	4.76%	0.69	1
Miss Marple12	54.76%	23.42%	23.42%	0.64	1
Miss Marple13	62.85%	0.29%	0.29%	0.77	1
Tennis	50.47%	46.14%	46.14%	0.52	1
People1	51.00%	1.17%	1.17%	0.67	1
People2	69.27%	0.86%	1.97%	0.82	2
Averages	50.62%	17.39%	17.52%	0.61	1.1
Our Dataset					
Lazona07	48.33%	13.65%	17.94%	0.62	5
Lazona08	63.43%	25.01%	25.77%	0.69	6
Lazona14	64.06%	32.11%	32.11%	0.66	1
Shinagawa04	48.70%	23.40%	29.02%	0.60	5
Shinagawa05	49.61%	25.01%	24.48%	0.60	5
Averages	54.83%	23.84%	25.86%	0.63	4.4
Averages on all	51.62%	18.92%	19.51%	0.62	1.9

Table 1. Evaluation results. Legend: ρ density; $\epsilon_{\text{overall}}$ overall error; $\epsilon_{\text{average}}$ average error; SC segmentation covering; HC head count.

We believe that the main advantage of our new model lies in the use of regions. These define a continuous flow variation on the objects of interest, allowing reasoning on body parts generally moving in completely different fashion, *e.g.* arms and legs. Not less important is the clustering framework. Results show that it generalizes well to video sequences where the camera and the objects move very differently from the training dataset. Furthermore it is not limited to pedestrians but may be used to learn other complex motion patterns, and it provides robust clustering results, especially on frames densely covered with region trajectories. We observe two kinds of flaws of our current model: (i) at some frames the clusters are re-initialized; (ii) some region trajectories from the arms and legs, or on pedestrians moving frontally at the same speed as the camera (no apparent motion) are misclassified. Failures on cases of no apparent motion is expected in algorithms relying on motion cues and could be addressed by integrating appearance into

[1] Best value between first 50 and 200 frames results in [5] is reported

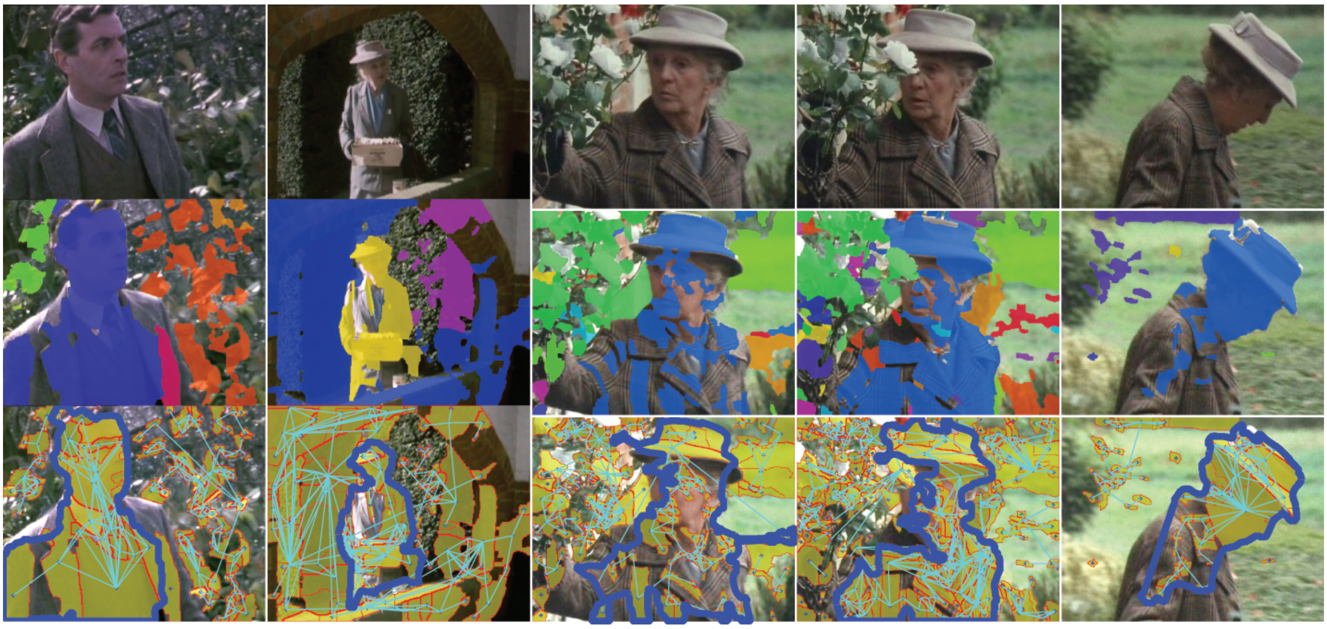

Figure 7. Video sequences from [5]: (*Columns from left to right*) Miss Marple1 frame 80; Miss Marple2 frame 80; Miss Marple7 frames 30,50,80. The figure presents the frames, the colour-coded clusters of region trajectories and the graph representation of the spatio-temporal clustering, as described in Fig. 1. A qualitative inspection reveals that the algorithm succeeds in clustering most of the region trajectories from each pedestrian, although both the camera and the pedestrians move in a completely different fashion from our training dataset, and the pedestrians are observed at very different scales. A supplementary video can be downloaded from `http://fabiogalasso.org/`

the framework, as discussed in Sec. 4. On the other hand, a common cause seems to be underlying the other flaws: a divergent image segmentation over multiple hierarchical levels at some frames. As seen in Sec. 3, a probabilistic approach addresses most issues of repeatability by choosing the appropriate hierarchical level. However a divergence over multiple levels may result in no region trajectories on pedestrians or, to a less extent, in a sparser covering. The former implies re-initialization of the cluster at the following frame, the latter may cause discontinuities in the flow variations between adjacent regions, resulting, possibly, in the misclassification of the most challenging arms and legs. Interestingly, such flaws are more frequent on our testing sequences than on those from [5], although ours are actually more similar to the training set. On the one hand, our new video sequences are of unprecedented difficulty in tasks of motion-based clustering; on the other hand, we believe that such flaws are inherent in algorithms defining video trajectories by using image segmentations, also a trend in state-of-the-art video segmentation [21]. Future research on the topic should address this issue, as we discuss in Sec. 6.

TIME COMPLEXITY Our non-optimized Matlab code runs on a single CPU in few minutes per frame. Most computational load (∼95%) is taken by constructing the graph of image segments and computing the video features. The region trajectories represent the video sequence compactly (less than 3000 are used for each of the sequences), Isomap and K-means cost therefore less than seconds per frame. Implementing the algorithm is relatively easy, as it includes simple sub-modules (*e.g.* bilateral filtering, mask propagation, DP) and others available on the web (*e.g.* [1, 23, 3, 19]) used with standard parameters, unless specified. Similarly, the video features are easy to code and extend, to include new aspects of Vision into the probabilistic framework.

6. Conclusions and Future Work

We have presented a novel approach to the spatio-temporal clustering of trajectories based on motion. We have introduced novel probabilistic region trajectories, a novel Bayesian framework, and we have applied our model to the difficult case of pedestrians in crowded outdoor video sequences, comparing to a state-of-the-art algorithm. The present work is the first of its kind in various respects, most importantly: first example of clustering of region trajectories and first example of object learning from video data.

The use of pre-computed image segments, already seen in state-of-the-art video segmentation algorithms, seems to be computationally and biologically plausible in processing video sequences. It comes however with some limitations inherent in an appearance-based image segmentation. Besides including principled video cues into the segmentation algorithm, future research may be addressed to create

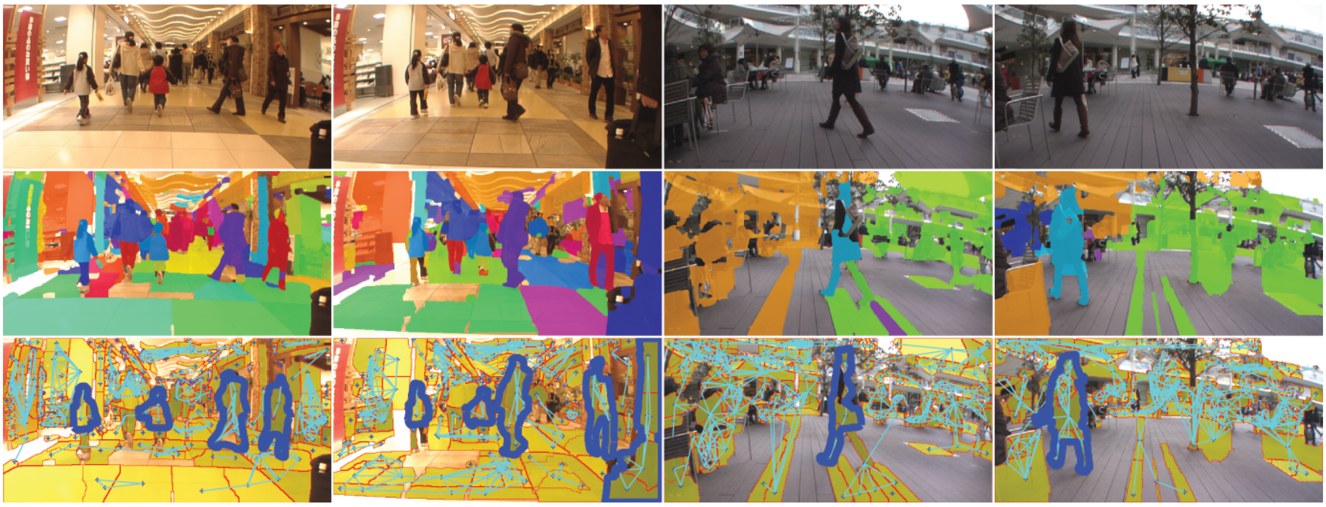

Figure 8. Video sequences from our new dataset: (*First two columns*) Lazona08; (*second two columns*) Lazona14. Frames, colour-coded clusters and graph representations are explained in Fig. 1. The examples show a successful aggregation of most region trajectories from pedestrians into single clusters, alongside some flaws of our algorithm: some region trajectories on limbs are misclassified and the lady inbetween the two children in Lazona08 is confused with the background (a case of no apparent motion as she moves forward at the same speed as the camera). These video sequences are of unprecedented complexity for tasks of spatio-temporal clustering.

temporally stable image hierarchies. Other future research will be pursued on including the trajectory extraction and clustering into a unified model and on using higher-level motion- and appearance-based video features.

Acknowledgements

The authors wish to acknowledge Vijay Badrinarayanan for valuable discussions.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *CVPR*, 2009.

[2] A. Barbu and S. Zhu. On the relationship between image and motion segmentation. In *SCVMA Workshop ECCV*, 2004.

[3] L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, 2001.

[4] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *CVPR*, 2006.

[5] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[6] T. F. Chan and L. A. Vese. Active contours without edges. *Transactions on Image Processing*, 10(2):266–277, 2001.

[7] A. Cheriyadat and R. Radke. Non-negative matrix factorization of partial track data for motion segmentation. In *ICCV*, pages 865–872, 2009.

[8] M. Fradet, P. Prez, and P. Robert. Clustering point trajectories with various life-spans. In *CVMP*, 2009.

[9] Z. Khan, T. Balch, and T. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *ECCV*, 2004.

[10] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, pages 878–885, 2005.

[11] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

[12] K. Okuma, A. Taleghani, N. D. Freitas, O. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, pages 28–39, 2004.

[13] F. Perbet, A. Maki, and B. Stenger. Correlated probabilistic trajectories for pedestrian motion detection. In *ICCV*, 2009.

[14] M. D. Rodriguez and M. Shah. Detecting and segmenting humans in crowded scenes. In *ACM Multimedia*, 2007.

[15] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80(1):72–91, 2008.

[16] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *ICCV*, pages 1154–1160, 1998.

[17] J. Sivic, F. Schaffalitzky, and A. Zisserman. Object level grouping for video shots. *IJCV*, 67:189–210, April 2006.

[18] D. Sugimura, K. M. Kitani, T. Okabe, Y. Sato, , and A. Sugimoto. Using individuality to track individuals: clustering individual trajectories in crowds using local appearance and frequency trait. In *ICCV*, 2009.

[19] J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319 – 2323, 2000.

[20] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *CVPR*, 2007.

[21] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Multiple hypothesis video segmentation from superpixel flows. In *ECCV*, 2010.

[22] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247–266, 2007.

[23] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *Pattern Recognition (Proc. DAGM)*, pages 214–223, 2007.