2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Space-time *A Contrario* Clustering for Detecting Coherent Motions

Thomas Veit, Frédéric Cao and Patrick Bouthemy

Abstract— **This paper presents a method for detecting independent temporally-persistent motion patterns in image sequences. The result is a description of the dynamic content of a video sequence in terms of moving objects, their number, image position and approximate motion. For each detected motion pattern a local trajectory as well as a confidence level is provided. The method is based on local motion measurements extracted from short video segments. These measurements are mapped in an adequate grouping space where independent trajectories correspond to distinct clusters. The automatic cluster detection is handled in an *a contrario* framework, which is general and involves no parameter tuning. The method was validated on real video sequences featuring rigid and non-rigid moving objects, static and mobile cameras, and distracting motions. The output of this method could initialize tracking algorithms. Applications of interest are robot navigation, car-driver assistance, surveillance and activity recognition.**

I. Introduction

A. Problem setting

A general problem in motion analysis is the early reliable detection of pieces of trajectories of moving objects in image sequences. Accurately and efficiently solving this problem is of crucial interest for applications such as robot navigation, car-driver assistance, video-surveillance and human activity recognition. It seems to us that there is a gap to be filled between two types of issues. On the one hand, there are motion detection methods. Most methods are actually closer to change detection, since they make decision on very local time intervals, with no real search of any spatio-temporal coherence. As a consequence, significant moving objects cannot be distinguished from "parasitical" motion. The temporal content alone is usually very noisy; hence, local spatial (and possibly temporal) regularity is usually introduced, which is the simplest mean to enforce temporal coherence. On the other hand, if the position of a given moving object is known, efficient methods allow one to track them. Many algorithms are variations or extensions of the Kalman filter. Recent progress based on the non-linear particle filtering approach led to very impressive results able to handle occlusions and shape deformation. The weak point of these methods is their usually supervised initialization.

The method proposed in this paper adresses simultaneously coherent motion detection and track initialization. The purpose is to decide on the existence of small pieces of trajectories on short durations (typically 10 or 20 frames). Detection thresholds for extracting these pieces of trajectories are automatically computed. It is clear that such thresholds exist also from a perceptual point of view. As an example, a slowly moving object has to be observed for a long time to be detected. Hence, there should be a relation between the size of an object, its velocity, the duration of observation and its detectability. When dealing with digital image sequences, detectability is also influenced by image quality. The method described in this paper uses a detection principle, intuited by Helmholtz and formulated by Desolneux, Moisan and Morel [1]. It states that a particular configuration is perceptually relevant if it cannot occur by chance, i.e., it contradicts a general random structure of the observations.

This work was partly supported by Région Bretagne and the European Network of Excellence MUSCLE.

T. Veit, F. Cao and P. Bouthemy are with IRISA / INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France thomas.veit@inrets.fr

B. Overall strategy

The purpose of this work is to extract geometrical evidence for moving objects from a set of successive digital images (about 10-20). More precisely, is it possible to decide that image parts along a sequence display locally a coherent motion, and define a piece trajectory? With what degree of confidence?

The strategy is the following. First, local motion measurements are extracted from successive pairs of images. These measurements are based on characteristic image features such as affine invariant pieces of level lines [2], SIFT descriptors [3] or KLT features [4]. These features have to be local enough, because of partial occlusions, shadows, etc. If the duration of observation is short enough, the motion of objects is approximately rectilinear with a constant velocity. This velocity, as well as the position of the shape element at time $t = 0$ is, in this simple case, completely determined by the displacement between two images. This results in a point in $\mathbb{R}^4$: two real coordinates for the velocity and two for the initial position. Now, if these pairs correspond to the same moving object in different frames, then the corresponding points form clusters in $\mathbb{R}^4$. As a consequence, the detection of pieces of trajectories results in a grouping problem.

Let us consider M data points, X_1, ..., X_M in $\mathbb{R}^4$, each corresponding to a couple (initial position,velocity), possibly detected at different instants. Following the same argument as in [5], an *a contrario* method is adopted: assume all the pairs are casual, and do not correspond to a coherent trajectory. Then, it is sound to assume that the X_i are independent and identically distributed according to a law to be specified. It is very unlikely that an important proportion of the X_i can be observed in a single small region of $\mathbb{R}^4$. Whenever this is actually observed, then the hypothesis that the X_i are random is certainly false, and some of them should be grouped. Natural questions arise, that are answered in this paper: how many groups are there (if any)? Which groups

are relevant ? Is it possible to quantify the meaningfulness of a group of points? How to select among nested groups?

The paper is organized as follows. In Sect. II, a few related works are reviewed. Section III briefly describes how the data to be clustered are extracted. Section IV is the technical core of the paper, describing the *a contrario* detection of clusters. The theory is validated on experiments in Sect. V.

II. Related work

Yuille and Grzywacz [6] also proposed a clustering approach after suitably representing visual patterns, and attempted to classify the typical configurations of visual motion. A complex observation would be a combination of these elementary motion templates, which should be detected by a grouping procedure. However, their work remains formal with no computational theory. More applied works are [7], [8] in which some motion structures are sought in spatio-temporal slices. Still more recently, Gryn, Wildes and Tsotsos [9] have specified even more precise motion templates, driven by the application, in other words trading generality for better computational efficiency.

The similarity of the ingredients involved in our method with those involved in Structure From Motion (SFM) methods might be misleading. The focus of SFM methods is more on characterizing the 3D geometry of the scene than on detecting coherent motion patterns [10]. The presence of one or several moving objects is assumed and therefore the detection issue is not addressed. Furthermore, the features detected in the image sequences need to be tracked through all the sequence. This requirement is obviously difficult to meet in the presence of occlusions or noisy image sequences. Factorization methods usually rely on spectral clustering for the clustering step. This clustering method based on algebraic matrix manipulations is known to be very sensitive to noise. Other methods rely on iterative optimization methods to build clusters, for example Expectation-Maximisation or K-means [11]. These methods require the number of clusters to be specified. Moreover, the results are sensitive to initialization. An alternative is to resort to model selection to determine the number of moving objects. Torr and Murray [12] propose a stochastic clustering method to group local motion measurements from several moving objects based on 3D geometry. They address the different issues of clustering, namely cluster validity assessment and merging of clusters. Their method relies on the combination of several heterogeneous criteria involving several parameters. Their method is based on two frames and the clustering is therefore rather based on shape than on motion coherence in time.

III. Extracting local motion measurements from images

The features to be extracted from images must be local (because of possible partial occlusion), stable, and invariant enough to the deformations an object may encounter through a sequence (approximate rigid motion, contrast change...). Different type of features meet these requirements: Affine Invariant Pieces of Level Lines (AIPLL) [2], SIFT descriptors [3] or KLT features [4]. The reader is referred to these articles for details. Given a pair of successive images of the sequence at time instant t and $t+1$, any of these features enables to compute local motion measurements. In the case of AIPLL or SIFT descriptors, a displacement measurement is obtained by matching a feature in the first frame with its best corresponding feature in the next frame. The difference between the position x_t at time instant t and x_{t+1} at time instant $t+1$ provides the displacement v. For KLT features, the displacement v is directly computed by an optimization process involving both frames [4]. Let us define the vector $(x^{\text{ref}}, v) \in \mathbb{R}^4$ by $x^{\text{ref}} = x_t - t\ v$. By first order approximation, the velocity v is constant and x^{ref} would be the initial position of feature at time instant $t = 0$. This hypothesis is sound if the duration of observation is short.

Now, a part of the same object at different time instants, or different parts of the same object should lead to approximately the same values of initial position and velocity. Figure 1 displays the two-dimensional projections of the couples $(x^{\text{ref}}, v) \in \mathbb{R}^4$ extracted from 10 successive frames. The plot in the middle corresponds to x^{ref}, i.e., the vertical position vs. the horizontal position. The right plot corresponds to the polar coordinates of v, orientation vs. magnitude. Three clusters in $\mathbb{R}^4$ can be distinguished corresponding to the three moving objects that appear in the scene displayed in the left image. Local motion measurements corresponding to the background of the scene are scattered in position and velocity direction but highly concentrated at velocity magnitude 0. Automatically detecting clusters in this four-dimensional grouping space allows us to extract the independent motion patterns that are temporally coherent, in other words the three moving objects.

In order to deal with mobile cameras, dominant motion estimation and motion compensation is applied. A general and robust dominant motion estimation algorithm is used [13]. The dominant motion is identified with camera motion. This identification is possible under some hypothesis such as the size of the moving objects in the image and the absence of significant depth discontinuities in the background. These hypotheses are usually verified in typical surveillance videos. Once the camera motion is compensated, local motion measurements corresponding to the background display almost null velocity exactly as in the static camera case.

Since the computational load of the grouping procedure directly depends on the number of considered local motion measurements, discarding local motion measurements that obviously belong to the background dramatically saves computation time. Two simple strategies to discard background measurements can be adopted. If, for each image of the sequence, a detection map is available that indicates which regions of the image belong to the background and which regions are moving, only features corresponding to moving regions are processed. For example, such a detection map is obtained by applying an automatic moving region detection as proposed in [14]. This strategy is preferred when working with AIPLL or SIFT descriptors. The other strategy

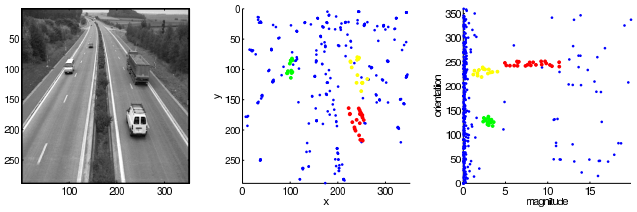

Fig. 1. Left image : three moving objects are perceptible, in the left lane a white van, in the right lane a white van and a gray truck. Middle and right plots : two-dimensional projections of four-dimensional couples (x^{ref}, v), column vs. line of the initial position and velocity direction vs. velocity magnitude. The three moving objects form three distinctive clusters in $\mathbb{R}^4$ (green, red and yellow). Elements belonging to the static background appear as a large elongated cluster (blue) with almost zero velocity magnitude and no distinct direction.

consists in discarding all features with an estimated inter-frame velocity magnitude v smaller than a given threshold, typically 1 pixel. This threshold is not very demanding. This second strategy is preferred when working with KLT features. Features remaining after discarding those belonging to the background are termed *moving features*. When applied to *moving features*, the task of the clustering procedure is to detect groups of features corresponding to each object moving independently and consistently over time.

IV. A CONTRARIO DETECTION OF SPACE-TIME COHERENCE

Let us now consider a set of points $\{X_1, ..., X_M\}$ in $\mathbb{R}^4$. Does this set contain any group? How many, and how meaningful are they? This problem is one of the numerous forms of cluster analysis. The above questions do not have a definitive answer. In particular, it is difficult to make a robust decision about the existence of a group (known as the problem of *validity*), or whether it should be cut into subgroups or not. This is precisely the problems this section deals with. Some ideas presented here have been inspired by Bock [15]. A parallel work [5] develops a theory of grouping for planar shape recognition. The main results of this theory are developed here and extended to motion analysis.

A. Number of false alarms of a group and validity

The fact that some of the X_i's may constitute a group reveals a lack of independence of these points. Since the cause of the dependence is unknown, modeling the probability of such an event is difficult. Hence, the following *a contrario* point of view is adopted. Let us assume that the X_i's are identically distributed variables in $\mathbb{R}^4$, following a probability distribution π specified later. Assume also that there is no group in the data. In this case, the X_i's are assumed to be independent. Let $R \subset \mathbb{R}^4$, independent of the X_i's. The probability that at least k out of the M data points $\{X_1, ..., X_M\}$ belong to R is given by the tail of a binomial law with parameters k, M, and $\pi(R)$

$$B(M, k, \pi(R)) = \sum_{j=k}^{M} \binom{M}{j} \pi(R)^j (1 - \pi(R))^{M-j}. \quad (1)$$

Assume that a region R containing k data points is observed. If the probability above happens to be very low, the observed data points certainly contradict the i.i.d. hypotheses. Of course, R must be given before observing the data points. From now on, an *a priori* finite set of regions $\mathcal{R}$ with cardinality $|\mathcal{R}|$ is considered, typically hyper-rectangles, assumed centered on the origin.

Let us introduce the following measure of meaningfulness.

Definition 1: Let G be a subset of $\{X_1, ..., X_M\}$ of cardinality k, $2 \leqslant k \leqslant M$. The Number of False Alarms (NFA) of a group G is defined as

$$NFA(G) = M^2 \cdot |\mathcal{R}| \min_{\substack{x \in G, R \in \mathcal{R} \\ G \subset x+R}} B(M-1, k-1, \pi(x+R)). \quad (2)$$

A group G is said to be ε-meaningful if $NFA(G) \leqslant \varepsilon$.

Before giving a mathematical result explaining why this number is introduced, let us explain how it is computed. Let us examine the term in the minimum: $x + R$ is one of the possible regions of $\mathcal{R}$, after centering at x, which is a point of G. Hence, $B(M-1, k-1, \pi(x+R))$ is the probability that at least k points (including x) are inside $x + R$ under the hypotheses of the *a contrario* model. Then, x and R are chosen to minimize this probability. Let us remark that there are at most $M|\mathcal{R}|$ possible choices of the couple (x, R). The second factor M is explained in the following.

B. A set of candidate groups

There are 2^M subsets of $\{X_1, ..., X_M\}$. It is not possible to compute the NFA of every possible group. Most of them are anyway certainly irrelevant. In order to drastically reduce the number of candidate groups, a classical single linkage hierarchical clustering procedure is applied. The result is a binary tree, each node being a candidate group. The root of the tree contains all the M data points.

Remark This does not solve the two problems at hand: number of clusters, meaningfulness or validity of each cluster. It only proposes a hierarchy of partitions of the data set. From this procedure, $M-1$ candidates with more than 2 points are proposed. It is then possible to prove the following result.

Proposition 1: If X_1, ..., X_M are i.i.d. points following π, then the expectation of the number of ε-meaningful groups among any set of M candidate groups is less than ε.

In particular, the result holds for the set of candidate groups provided by the clustering procedure, since there are less than M candidates. We refer the reader to [5] for a complete proof.

The interpretation of this result is more important than its proof. Set ε to a small value, less than 1. If an ε-meaningful group is observed, then chance alone is certainly not a good explanation for it, since there are less than $\varepsilon < 1$ such meaningful groups on average. The lower the NFA, the less likely it is that such a group has been generated by the *a contrario* model. Hence, the NFA provides a validity measure. In general, the NFA of a meaningful group is much lower than 1 (see Sect. V).

C. Merging criterion

The hierarchy provided by the tree of candidate groups allows us to simplify the problem of merging small groups into a larger one. Indeed, since the tree of clusters is binary, this question can be answered for two sibling nodes. The merging method is then applied recursively. Following an *a contrario* argumentation, the two groups G_1 and G_2 are separated if it is more unlikely to observe two groups G_1 and G_2 than a single group containing them. It is natural to define a number of false alarms for a pair of sibling groups: $NFA_g(G_1, G_2)$. For technical details and exact definition of NFA_g, the reader is referred to [16]. Using the same kind of arguments as for Proposition 1, one can prove that, on average, there are less than ε pairs with NFA_g less than ε. More interestingly, the normalization of probabilities into NFAs makes it possible to compare events of different nature, such as groups and pairs of groups, because the numbers of false alarms have comparable magnitudes.

Definition 2: Let G be a subset of the M data points. A group G is said *indivisible* if, and only if, for all pairs G_1 and G_2 such that $G_1 \cap G_2 = \emptyset$ and $G_1 \cup G_2 \subset G$,

$$NFA(G) < NFA_g(G_1, G_2).$$

D. Practical algorithm

So far, a group validity criterion and a merging criterion have been defined. The last point is that a group can be slightly enlarged by adding a few points. Again, what is best? This question is easily answered by comparing the NFAs of the groups through the inclusion tree.

Definition 3: A group G is said to be *maximal* ε-meaningful if

1) G is ε-meaningful.
2) G is indivisible.
3) G is more meaningful than all its indivisible descents.
4) for all indivisible ascent G', either $NFA(G) < NFA(G')$ or there exists another indivisible descent G'' of G' such that $NFA(G'') < NFA(G')$.

The last condition only reflects that the tree is an asymmetric graph and ensures that a group can eliminate smaller groups in the tree only if it is more meaningful than *all* of them.

All these definitions may seem a bit formal. Actually, the implementation basically reduces in counting points in hyper-rectangles. Let us sum up the meaningful group detection algorithm.

1) **Clustering step**. Given M data points, compute the binary tree by a single linkage algorithm. Each node corresponds to a candidate group.
2) **Validity step**. For each candidate group G,
   a) compute the region $x + R$, $x \in G$, $R \in \mathcal{R}$ containing all the points of G and such that $\pi(x + R)$ is minimal.
   b) compute $NFA(G)$ and tag G as valid if $NFA(G) \leqslant \varepsilon$.
3) **Merging step**. For each sibling pair G_1 and G_2:
   a) Compute the intersection of $x_1 + R_1$ and $x_2 + R_2$, obtained in the computation of $NFA(G_1)$ and $NFA(G_2)$.
   b) Remove the points of G_1 and G_2 in this intersection.
   c) Compute $NFA_g(G_1, G_2)$.
4) **Final step**. Explore the tree and detect maximal meaningful groups according to Def. 3.

The last details to be specified are the choice of the *a priori* distribution π and the set of regions $\mathcal{R}$. Although the grouping method described so far is generic, the choice of π is more problem-specific. In the case at hand, the position and velocity of objects are considered independent. Of course this is not true for real objects. However, the *a contrario* hypotheses describe the absence of correlation of all the observations. Moreover, unless it has been specified by the application, the position of a moving object is arbitrary, hence the position distribution is assumed to be uniform. No direction plays a particular role either. Hence, the velocity direction distribution is taken uniform in $(0°, 360°)$. The only problem is the norm of the velocity. A simple solution is to learn it on the data itself: the distribution of the velocity magnitudes is given by the empirical histogram of the observed velocities. This provides the right order of magnitude and a fair enough distribution profile. The distribution of the data points is simply the product of these four marginal distributions. Since these dimensions are assumed uncorrelated, it does make sense to consider regions whose main directions are parallel to the axes of coordinates. This results in a set $\mathcal{R}$ of hyper-rectangles with quantized sizes in each dimension.

V. Experimental results

We report experimental results for the proposed coherent motion detection method applied to various image sequences. The first experiment illustrates the grouping of local motion measurements obtained with SIFT descriptors. Results with AIPLL are very similar and not displayed here. The second experiment relies on local displacement measurements computed with the KLT technique. The last experiment shows how the method enables to group local displacement measurements corresponding to the same moving object undergoing occlusion. The framerate for all experiments is 25 frames/second.

A. Experiments with SIFT descriptors

This typical surveillance sequence (Fig. 2) figures rigid and non-rigid moving objects, respectively a car and a pedestrian. Local motion measurements are computed by matching SIFT descriptors. The automatic cluster detection procedure described in the previous section is applied once to all local motion measurements (second row) and once to *moving features* only (cf. Sec. III) (lower row). In both cases, a structured description of the dynamic content of the scene is correctly recovered: two coherent motion patterns are detected. Both are located in the left part of the image. The lower group (pedestrian) displays a leftward slightly upward

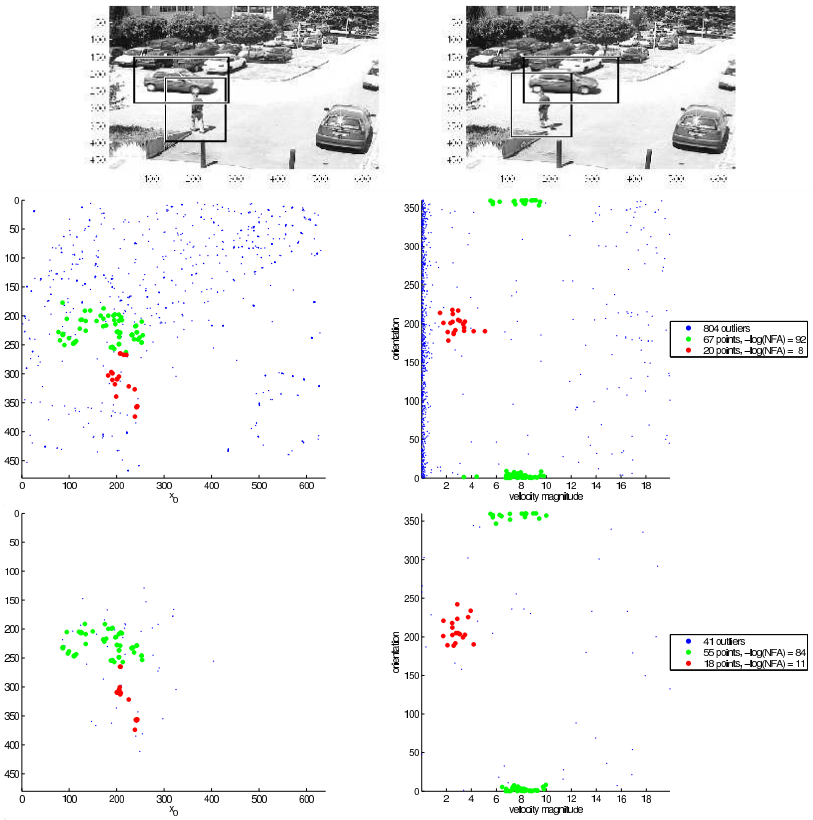

Fig. 2. First row: first (t=1) and last (t=10) input frames. In the left image, the black rectangles delineate the regions associated to the clusters when grouping *moving SIFT descriptors*. In the right image, the regions (black rectangles) extracted in the first frame are simply moved according to the mean motion of the cluster points. They correctly fit the moving content of the image sequence. The second row presents the two-dimensional projections of the four-dimensional motion space when considering all the SIFT descriptors of each image. The third row contains the clustering results when considering only *moving SIFT descriptors*. The confidence levels $-\log(NFA)$ of the detected groups appear in the legend on the right.

motion at about 3 pixels per frame. The upper group (car) moves rightward at about 8 pixels per frame. The method adapts to the presence of moving objects with different speeds. The lower speed limit for detection is approximately 1 pixel/frame. Here, the confidence levels reflect the nature of the moving objects. The cluster corresponding to the car which is a large rigid object has a confidence level, given by $-\log_{10}(NFA)$, close to 100. This high confidence value is due to the large number of points in the cluster and the steady velocity direction. The cluster corresponding to the smaller non-rigid moving pedestrian contains less points. Moreover, their corresponding directions are less steady. Therefore, the confidence level is only about 10. Let us point out that the trees in the background of the scene are moving because of a strong wind. This motion is correctly not detected as coherent when applying the clustering to all features.

Applying the algorithm to all features or only to *moving features* impacts only the computation time. Computation time greatly depends on the number of features involved which usually increases with the size of the image. As an example, for 10 frames of size 352×288, it takes about 3 seconds to extract the *moving SIFT descriptors* and to cluster the *moving features*. Extracting all the SIFT descriptors and clustering takes about 20 seconds (Pentium, 3Mhz).

Discarding features corresponding to the static background of the scene decreases the computational cost of the clustering procedure in reducing the number of considered observations. It has almost no impact on the performance of the method but greatly simplifies the grouping task.

B. Experiments with KLT features with background subtraction

In this section, KLT features provide the local motion measurements. These features are less descriptive than AIPLL or SIFT descriptors. However, their simplicity and the low computational complexity of KLT features are very attractive. The sequence (Fig. 3 and Fig. 4) contains a pedestrian walking on a sidewalk and illustrates the behavior of the detection method on non-rigid objects. The camera is tracking the pedestrian. The tree and the bushes in the foreground are moving due to the wind. Local motion measurements are accumulated over 10 frames.

In the first part of the video, the unoccluded torso of the pedestrian is detected as a moving region. The NFA is rather low and thus the confidence in the detection is high ($-\log_{10}(NFA) = 168$). In the second part of the sequence,

the pedestrian is partially occluded by the branches of the tree. Only a few motion measurements are still available. However, the pedestrian is still detected. Of course, the confidence in the detection is then smaller ($-\log_{10}(NFA) = 24$) reflecting the fact that there is less evidence in favour of coherent motion. Computation time for processing a video segment of 10 frames is about 3s.

C. Occlusions

The last part of this experimental section is concerned with moving objects undergoing severe occlusion. The proposed coherent motion detection method succeeds in grouping together local motion measurements before and after occlusion. The number of frames involved in this experiment is larger (15 frames) in order to observe the objects before and after occlusion. The image sequence, Fig. 5, shows a pedestrian crossing another one and getting occluded. The camera is hand-held and is tracking the first pedestrian. Local motion measurements are accumulated through 15 successive frames. Both pedestrians are detected as undergoing coherent motions. The local motion measurements belonging to each of them are clustered into two separate groups. Outliers correspond to measurements due to noise or measurements on the arms and legs having a periodic motion that does not display sufficient coherence.

D. Number of frames involved in the detection process

The number of frames during which motion information is accumulated can vary. Part of this work was to study how long an image sequence has to be examined in order to detect groups of coherent motion. The conclusion is that several factors influence the minimal observation time for detection: size of objects, image quality, quality of the first order approximation (constant velocity). It turns out that under favorable conditions, the required number of frames can be as small as 3 or 5. The number of frames involved in the coherent motion detection process can be tuned according to the specific application and the experimental conditions:

- 3-5 frames: "instantaneous" motion detection while ensuring motion coherence;
- 5-10 frames: short-term coherent motion detection;
- 10-30 frames: long-term coherent motion detection, especially in the case of occlusions.

In all cases the observation time remains short: about one second for 25 frames/second videos.

VI. Conclusion

This paper presents a method to detect independent coherent motion patterns in image sequences. The automatic clustering of local motion measurements leads to a general *coherent motion* detection algorithm. The result is a structured description of the dynamic scene content: number of moving objects, position, magnitude and direction of their displacements, i.e., local trajectories. The proposed framework enables to control the number of false alarms and associates a confidence level to each detected independent motion pattern. The local motion measurements are extracted by means of characteristic image features. Possible types of image features are: affine invariant pieces of level lines, SIFT descriptors and KLT features. Results on various real image sequences illustrate the ability of the method to detect temporally consistently moving objects (cars, pedestrians) without being distracted by moving textures (water, leaves). Future work will aim at extending the proposed method to 3D motion models. If the scene geometry is known it could be incorporated into the model to take into account variation of the projected velocity due to depth changes of moving objects. As for the local trajectories provided as an output of the method, they could become useful for long term trajectory analysis. Further work on the clustering algorithm itself consists in processing partial trees in order to reduce computation time. The description of the scene provided by this method could become useful for surveillance, activity recognition, as well as robot navigation.

References

[1] A. Desolneux, L. Moisan, and J. Morel, "A grouping principle and four applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 4, pp. 508–513, 2003.

[2] J. Lisani, L. Moisan, P. Monasse, and J. Morel, "On the theory of planar shape," *SIAM Multiscale Modeling and Simulation*, vol. 1, no. 1, pp. 1–24, 2003.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Jal. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] J. Shi and C. Tomasi, "Good features to track," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Seattle, June 1994, pp. 593–600.

[5] F. Cao, J. Delon, A. Desolneux, P. Musé, and F. Sur, "A unified framework for detecting groups and application to shape recognition," *Jal. of Mathematical Imaging and Vision*, September 2006, , published online.

[6] A. Yuille and N. Grzywacz, "A theoretical framework for visual motion," in *High-Level Motion Processing*, T. Watanabe, Ed. MIT Press, 1998.

[7] Y. Ricquebourg and P. Bouthemy, "Real-time tracking of moving persons by exploiting spatio-temporal image slices," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 797–808, 2000.

[8] S. Sarkar, D. Majchrzak, and K. Korimilli, "Perceptual organization based computational model for robust segmentation of moving objects," *Computer Vision and Image Understanding*, vol. 86, pp. 141–170, 2002.

[9] J. Gryn, R. Wildes, and J. Tsotsos, "Detecting motion patterns via direction maps with applications to surveillance," in *7th IEEE workshop on Applications of Computer Vision*, 2005, pp. 202–209.

[10] A. W. Fitzgibbon and A. Zisserman, "Multibody structure and motion: 3-D reconstruction of independently moving objects." in *6th European Conference on Computer Vision*, ser. LNCS, vol. 1843. Dublin: Springer, June 2000, pp. 891–906.

[11] R. Hartley and R. Vidal, "The multibody trifocal tensor: Motion segmentation from 3 perspective views," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. I, June 2004, pp. 769–775.

[12] P. H. S. Torr and D. W. Murray, "Stochastic motion clustering," in *Proc. of the 3rd European Conference on Computer Vision*, ser. LNCS, vol. 2. Stockholm: Springer, 1994, pp. 328–337.

[13] J. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *Jal. of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, 1995, software available at http://www.irisa.fr/vista/Motion2D.

[14] T. Veit, F. Cao, and P. Bouthemy, "An a contrario decision framework for region-based motion detection," *Int. Jal. of Computer Vision*, vol. 68, no. 2, pp. 163–178, 2006.

[15] H. Bock, "On some significance tests in cluster analysis," *Jal. of Classification*, vol. 2, pp. 77–108, 1985.

[16] T. Veit, F. Cao, and P. Bouthemy, "An a contrario space-time grouping framework for the detection of coherent motions," INRIA, Tech. Rep. 6061, december 2006.

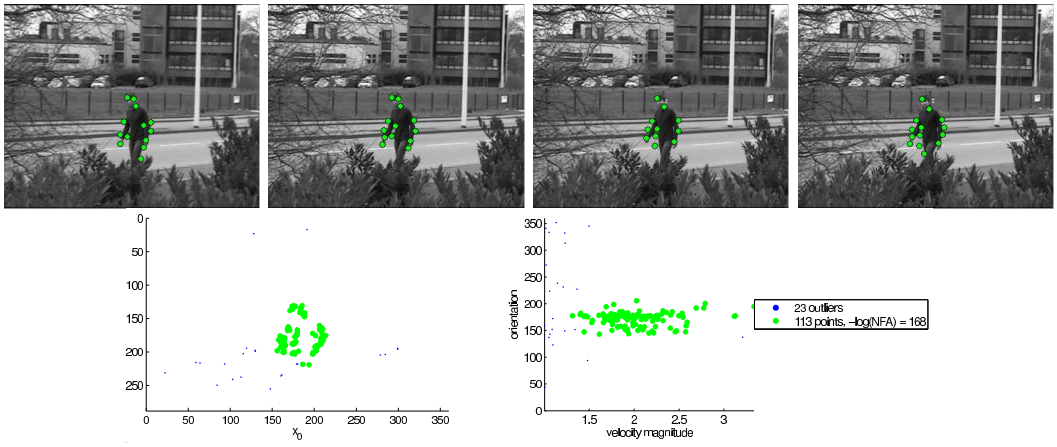

Fig. 3. Pedestrian sequence. Local motion measurements are accumulated on 10 successive frames. The pedestrian is detected as a coherent moving region. The oscillating motion of the twigs of the tree and of the bushes is not detected as coherent in time

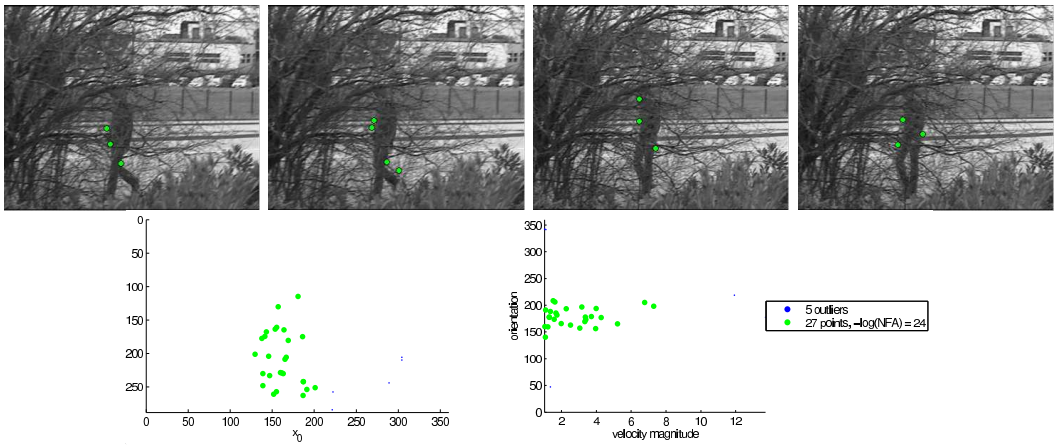

Fig. 4. Pedestrian sequence. First row, 4 frames out of the 10 processed. The pedestrian is now partially occluded. However, sufficient evidence for coherent motion is still available. The confidence in the detection decreases by a factor 10 from $-log_{10}(NFA) = 168$ without occlusion to $-log_{10}(NFA) = 24$ when the pedestrian is partially occluded by twigs.

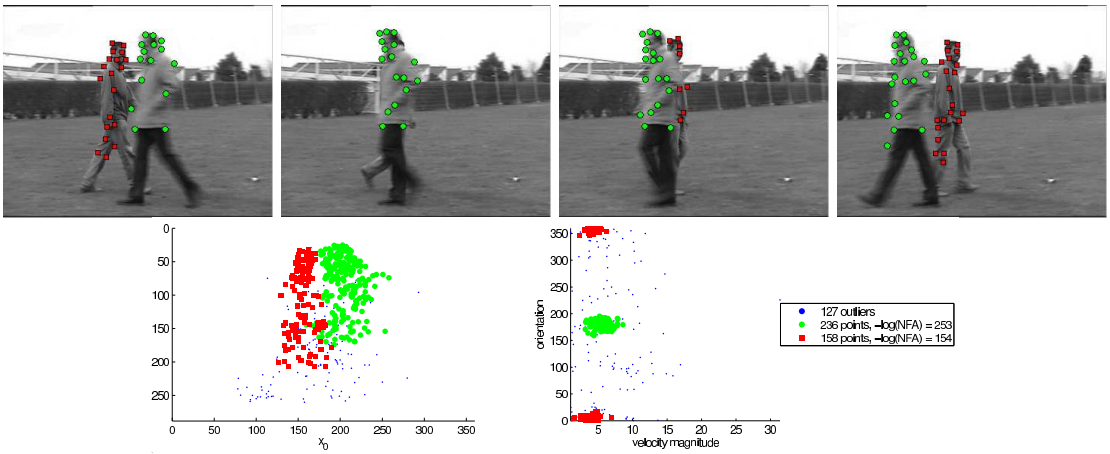

Fig. 5. Pedestrians crossing sequence. One pedestrian gets completely occluded by another. The camera is hand held and is tracking the further pedestrian. Local motion measurements are computed from 15 successive frames. Two clusters corresponding to the two pedestrians are detected with very high confidence, $-log_{10}(NFA) = 154$ and $-log_{10}(NFA) = 253$. Computation time for 15 frames: about 10s.