Tracking by Hierarchical Representation of Target Structure*

Nicole M. Artner[1], Salvador B. López Mármol[2], Csaba Beleznai[1], and Walter G. Kropatsch[2]

[1] Austrian Research Centers GmbH - ARC, Smart Systems Division, Vienna, Austria
[2] PRIP, Vienna University of Technology, Vienna, Austria

Abstract. Tracking of spatially extended targets with variable shape, pose and appearance is a highly challenging task. In this work we propose a novel tracking approach using an incrementally generated part-based description to obtain a specific representation of target structure. The hierarchical part-based representation is learned in a generative manner from a large set of simple local features. The spatial and temporal density of observed part combinations is estimated by performing statistics over temporally aggregated data. Detected stable combinations consisting of multiple simpler parts encompass local, specific structures, which can efficiently guide a spatio-temporal association step of coherently moving image regions, which are parts of the same target. The concept of our approach is proved and evaluated in several experiments.

Key words: hierarchical representation, edge segments, spatial statistics, temporal statistics, tracking

1 Introduction

Variable structure is hard to represent using models based on image statistics. Therefore, many tracking algorithms encounter difficulties in scenes with articulated objects, targets undergoing pose variations and partial occlusions. Part-based approaches [1, 2] avoid some of these problems by decomposing variable structure into simpler parts, but often lack the flexibility to represent complex deformable structures, given that they either rely on an *a priori* model or the part-based model is too general (for example defined by a uniform grid of blocks).

This paper proposes an approach which (i) builds a target-specific part-based model using a compositional framework [3] and (ii) employs the generated model for tracking. Our main motivation is to achieve reliable data association and thus robust tracking using a hierarchical system combining simple features into increasingly more complex parts and capturing the statistics of parts. Obtained parts with growing complexity represent specific entities occurring in images more rarely. Partitioning of the parts into trajectories based on structural similarity and motion coherence becomes feasible, since the combinatorial problem of matching involves fewer parts with less ambiguity across frames of an sequence.

* Supported by the Austrian Science Fund under grants P18716-N13.

Hierarchical systems ensure an efficient way to represent exponential variability present in the visual data. Recent works [3, 4] in visual object recognition demonstrate that hierarchical grouping of simple features generates a finite set of increasingly specific groups, able to represent structure of many object categories in a compact form. By applying the same concept to images of a video sequence, tracking can be formulated in a statistically-driven manner recovering specific parts occurring across multiple frames.

The paper is organized as follows. Sec. 2 introduces related work. Sec. 3 describes the basic concepts behind the proposed approach and Sec. 4 explains the algorithmic steps generating the hierarchical representation. Sec. 5 describes the association step enabling tracking. In Sec. 6 we discuss experimental results for several image sequences and in Sec. 7 we draw conclusions.

2 Related work

Finding compact and informative representations able to handle the large complexity encountered when representing image content (structure, texture, appearance, etc.) has been an active field of research for many decades. Hierarchical representations organizing low-level visual information into more complex parts offer means to accomplish this task. Edge segments have been attractive low-level features offering a high degree of geometric and photometric invariance and encompassing a rich pool for building potential parts. Early works (e.g. [5]) employ pre-designed rules to partition and group edge segments into more complex entities. Partially driven by significant advances in visual object recognition recent frameworks propose statistical, learning-based methodologies. Opelt et al. [6] use object boundary fragments to detect multiple object classes in presence of clutter and partial occlusions, where class-discriminative boundary fragments are extracted and selected by Boosting. Agarwal and Triggs [7] introduce a novel multilevel visual representation called "hyperfeatures", where local sets of image descriptors are gradually organized into more complex and spatially sparse parts resulting in a system localizing several object categories. Crandall et al. [8] propose a recognition approach where appearance models of parts and spatial relations between parts are simultaneously estimated and used to localize objects. Bouchard et al. [9] present a hierarchical part-based description encoding geometry and appearance of object parts, where learned models vote in a bottom-up manner for object locations.

Hierarchical representation of structure for spatially extended targets in the context of tracking has not been investigated in great detail yet. Ommer et al. [10] present an approach where simple interest points are tracked in a frame-by-frame manner. Interest points as simplest parts are represented by local descriptors, which are used to analyze spatio-temporal relations between parts to learn compositional structured object models. In contrast to our work, Ommer et al. do not use a hierarchical representation and propagate hypothesized part compositions over consecutive frames. In our approach we apply similar concepts as presented by Fidler et al. [4, 11] to hierarchically organize spatio-temporally aggregated low-level features in order to create compositions which can be associated and tracked in an unambiguous manner.

3 Concept of our approach

Our approach starts with extracting simple oriented edge segments from a sequence of images. A local spatial configuration of multiple oriented edge segments – denoted further on as *combination* – encodes local structure. The main objective of the proposed approach is to select a set of temporally invariant combinations by statistical analysis, and to reliably associate them to form trajectories. Association exploits the *stability* (temporally invariant structure) and the high *specificity* (low occurrence frequency) properties of selected combinations. As can be seen in Fig. 1 the hierarchical representation is built in a bottom-up process, while the motion estimation (tracking) is carried out in a top-down manner.

The input of the bottom-up process are oriented edge segments detected in several consecutive frames of a video sequence. Starting at level 1, the edge segments are grouped together in combinations of two. By examining the temporal statistics of combinations, their number is reduced by retaining only the stable combinations. Then spatial statistics of edge segments for each remaining combination is examined, and combinations whose segments represent stable structures (i.e. occurring multiple times at the same relative positions) are kept. After the steps of temporal and spatial statistics, the stable combinations for level 1 are found and the process continues with level 2. Each stable combination is extended by an additional segment (selected from the pool of neighboring segments one-by-one) and the statistical analysis (temporal and spatial) is repeated. The incremental generation of edge combinations is carried out until a desired top level is reached, where only few, specific combinations remain, e.g.: level 3 in Fig. 1, or no more stable combinations are found (see Sec. 4).

The generated set of specific combinations at the top level is used in a top-down process to estimate the motion model of foreground objects. Combinations of segments at the top level are distinctive and in the best case they appear only once in an image. Hence, an association of combinations occurring at different time instances becomes feasible, providing an estimate of the underlying motion model. The motion models estimated by associating combinations at the highest level of the hierarchy, can be used to guide the association step of combinations at lower levels (with less specificity) yielding a dense structural description of foreground objects. Foreground objects are delineated by grouping stable combinations, which obey the same motion model (for details see Sec. 5).

4 Building the hierarchical representation

Each level L_i of the hierarchical representation contains M_{L_i} stable combinations $C = \{c_1, c_2, \ldots, c_M\}$ of $i+1$ segments. A segment s_l represents one of O possible edge orientations $S = \{s_1, s_2, \ldots, s_O\}$. To avoid an exponential complexity of O^N for N levels, the combinations of segments are built in the following way:

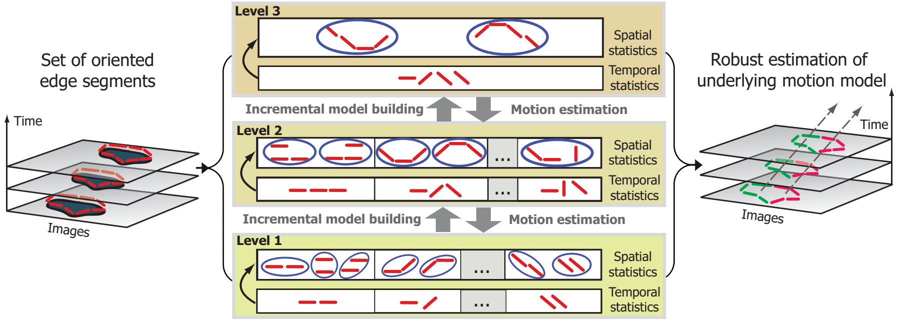


Fig. 1. Concept of our approach.

Each combination c_k of level L_i consists of i edge segments forming the primary part p_{prim} and one segment representing the marginal part p_{mar}. For example in level L_2 the primary part p_{prim} consists of a set of segments $\{s_1, s_2\}$ and the marginal part p_{mar} of the segment s_5. Except level L_1, the primary parts of the combinations are the stable combinations of level L_{i-1}. So the number of possible combinations for level L_i is $M_{L_{i-1}} \cdot O$.

4.1 Detection of segments

The structure of foreground objects and background is described by oriented edge segments. A filter bank consisting of oriented Gabor filters (8 orientations in 0°-180°, $\sigma = 0.7$) is used for detecting local edge segments. Each frame is filtered using the oriented filter bank and the magnitude of filter responses (without sign) is computed for each orientation. The image is divided into a set of non-overlapping rectangular neighborhoods along a dense grid. In a given rectangular neighborhood of a size D – small enough to capture important shape details such as the shoulder silhouette of a human – the locally dominant orientation is determined by analyzing all filter responses within the neighborhood for all orientations and finding the orientation with the maximum response. Local maxima with a value smaller than T_m are considered as noise and are ignored. In our approach we used the values T_m=15 and D=10 pixels, the latter being approximately $\frac{1}{10}$ of the foreground object's height.

4.2 Building Levels

As mentioned in Sec. 3 the hierarchical representation is built in a bottom-up process. The stable combinations for each level are selected by temporal and spatial statistics as described in the following.

Level i: For each of the F frames all possible combinations of p_{prim} and p_{mar} are enumerated within local windows B consisting of multiple local neighborhoods

D. In order to avoid prohibitive complexity due to the combinatorial nature of the enumeration task, the size of the local analysis window is defined to be small at lower levels of the hierarchy and increased at higher levels (see Tab. 2).

The local window is centered over each stable combination of level L_{i-1}, representing the primary part p_{prim}. Then all possible combinations of p_{prim} with an additional segment p_{mar} are formed. After all new combinations are built, temporal and spatial statistics are applied to select the stable combinations of level L_i.

Level 1: As there are no stable configurations from a previous level for level L_1 at the bottom of the hierarchy, local analysis is performed differently than in the other $N-1$ levels. The local analysis window is slid over all local neighborhoods in all frames starting in the top left corner with a step size equal to D. Within the sliding window at the actual position, all possible combinations are built, where the edge segment with the lower index defines p_{prim} (e.g.: s_1) and the segment with the higher index becomes p_{mar} (e.g.: s_5).

Temporal statistics: The task of temporal statistics is to estimate the density of combination occurrences over F frames by binning and to retain most frequently occurring combinations.

The density of combination occurrences is captured in form of a 3D histogram H, where the histogram is spanned by the primary part indices p_{prim}, the marginal part indices p_{mar} and the frame numbers f. Combinations frequently appearing across multiple frames are selected by a threshold T_f – representing a certain percentage of frames – and retained. The result of temporal statistics is a set of temporally stable combinations consisting of edge segments in an arbitrary spatial arrangement (see Fig. 1).

Spatial statistics: The combinations remaining after temporal analysis form the input for spatial statistics. Fig. 2 visualizes the creation of a co-occurrence histogram of segments which encodes spatial relations. For each temporally stable combination the primary part p_{prim} of the combination is centered at $(0, 0)$ and the spatial distribution of corresponding marginal part p_{mar} – relative to p_{prim} – is built in form of a two-dimensional histogram.

The obtained set of spatial distributions is used to select combinations with frequently occurring spatial edge configurations within F frames. Mean shift mode seeking is used to locate the most (up to four) significant modes of the distribution. Modes with densities below a threshold T_p and corresponding combinations are discarded. The outcome of spatial statistics is a set C of combinations corresponding to spatially stable edge configurations.

Fig. 3 shows bar diagrams displaying the effect of temporal and spatial statistics on the number of stable combinations and all occurrences of those combinations. Tab. 1 complements the information from Fig. 3(a) with the number of all possible combinations. For all experiments in Sec. 6 the parameters described in this section are set to the values of Tab. 2.

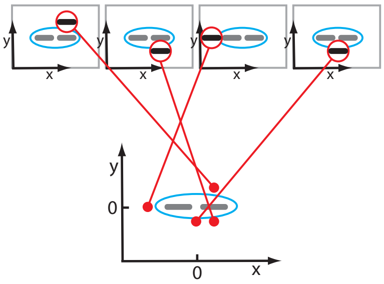
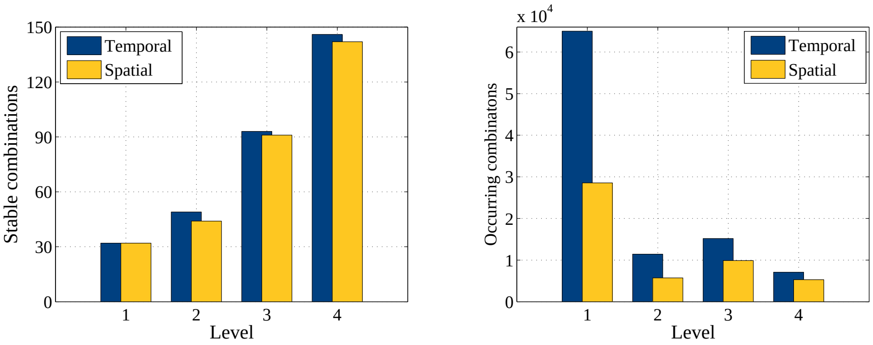


Fig. 2. The concept of spatial statistics. At the top different spatial edge configurations of a given combination are shown. At the bottom the obtained spatial distribution (relative to the centered primary part) of marginal parts is shown.



(a) Stable combinations.

(b) All occurrences of combinations.

Fig. 3. Example numerical results of statistics. (a) Number of stable combinations after temporal (blue, dark) and spatial statistics (yellow, bright). (b) All occurrences of stable combinations after temporal (blue, dark) and spatial statistics (yellow, bright).

Level	All combinations	After temporal statistics	After spatial statistics
1	36	32	32
2	288	49	44
3	2304	93	91
4	18432	146	142

Table 1. The numerical data of the combinations of bar diagram (a) in Fig. 3 (third and forth column) and the number of all possible combinations (second column).

Level	D	B	T_f	T_p
1	10	$3 \cdot D \times 3 \cdot D$	70%	1.0
2	10	$3 \cdot D \times 3 \cdot D$	70%	1.0
3	10	$5 \cdot D \times 5 \cdot D$	50%	1.0
4	10	$5 \cdot D \times 5 \cdot D$	50%	1.0

Table 2. Values of parameters for experimental results for each level.

5 Tracking using the built hierarchy

Our proposed object tracking approach involves three steps. First, a temporal association between the obtained combinations is carried out using robust statistical estimation (see Sec. 5.1). Secondly, combinations following the same motion model are grouped together spatially, thus delineating the tracked object in each image frame (see Sec. 5.2). Thirdly, hierarchical construction of combinations in overlapping space-time volumes is repeated (see Sec. 5.3).

5.1 Temporal association of combinations

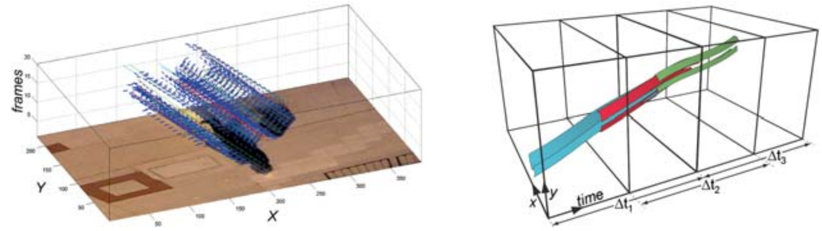
Reliable association of combinations requires that combinations are (i) stable and (ii) highly specific. The previously described statistical analysis captures combinations which occur in multiple image frames of the analyzed spatio-temporal volume. Combinations with an increasing number of segments are more specific and occur less frequently. Given the set of formed combinations, we consider the estimation task of an underlying motion model as a regression problem. In order to keep the complexity of motion model estimation step low, we assume a linear motion model within the analyzed time span (typically 20 frames).

Robust regression is performed using the RANSAC algorithm [12]. Estimation is started at the highest level of hierarchy, where combinations are the most specific and their spatio-temporal distribution best exhibits the underlying linear structure. Typically, despite of the high specificity of combinations, the space-time distribution of a given combination contains multiple structures, therefore the regression task is challenging.

For each combination at the highest level we estimate the best fitting motion model. The slope of motion estimate encodes motion direction and magnitude in the image space. Motion vector estimates are accumulated – in a similar manner to layered motion representations [13] – in a two-dimensional vector space spanned by velocity components along x and y. Mode seeking is performed to find the underlying trend, i.e. peaks defined by velocity components of frequently occurring motion models. Detected peaks encompass combinations belonging to stationary background and moving foreground objects. Since the number of stable combinations at the highest level is low, the obtained set of coherently moving combinations defines a spatially sparse object description. In order to obtain a spatially more dense set of stable combinations, we perform RANSAC estimation also at lower levels of the hierarchy. Sampling is guided by the motion model estimated at the highest level: motion model estimates which do not belong to any of the previously detected modes of accumulated motion vectors are discarded. In this manner, the estimation step is able to recover motion paths of less stable and less distinctive combinations at lower levels of the hierarchy and to provide a dense structural representation of targets (see example in Fig. 4(a)).

5.2 Spatial grouping of associated combinations

The obtained set of combinations obeying the same motion model is used to spatially delineate the tracked object. Since a given stable combination is not



(a) Space-time plot.

(b) Incremental spatio-temporal tracking.

Fig. 4. (a) Space-time plot showing all stable combinations describing a moving object and obeying the same motion model. (b) Illustration depicting the incremental spatio-temporal object tracking. Different colors indicate distinct stable combinations defining object trajectory segments in consecutive overlapping space-time volumes.

necessarily present in every frame, missing instances of combinations are generated by interpolating the location of each segment using the underlying motion model. Due to the interpolation step, the tracked object is described in each frame by the same number of stable combinations and spatial grouping can be carried out for each frame. Spatial delineation is performed by computing the convex hull of centroid locations of segments belonging to stable combinations.

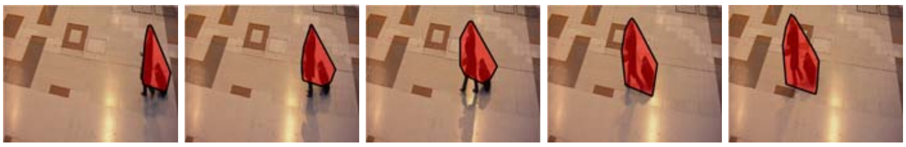
5.3 Incremental space-time processing

As the shape of a tracked target can change, usually varying sets of stable combinations represent the target structure at different time instances. Therefore, we perform the described hierarchical construction of combinations in an incremental manner. We aggregate edge segments from overlapping space-time volumes (see Fig. 4(b)) and form stable combinations in a given volume independently. Assuming kinematic smoothness of target motion, trajectory segments obtained for individual volumes are associated using the estimated motion models and spatial proximity.

6 Experiments and Discussion

We used three video sequences to prove our concept and qualitatively evaluate the performance of our approach. Video sequences 1 and 2 are parts from PETS 2006 and 2000 dataset.

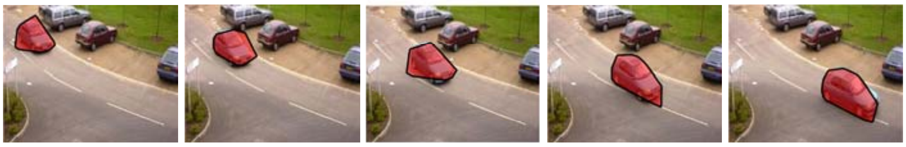
In sequence 1 a moving pedestrian is segmented and tracked as foreground object. The obtained convex hull – spanned by segments belonging to stable combinations – covers image regions, where spatio-temporal stability was found, therefore heavily articulated parts, such as the feet are not part of the obtained target representation (see Fig. 5). Stationary structures (tiling on the ground) belonging to the background are delineated as well (not shown) explaining recovered motion with a zero velocity model.



Frame 1 Frame 15 Frame 30 Frame 45 Frame 60

Fig. 5. Convex hull of the tracked object in sequence 1.

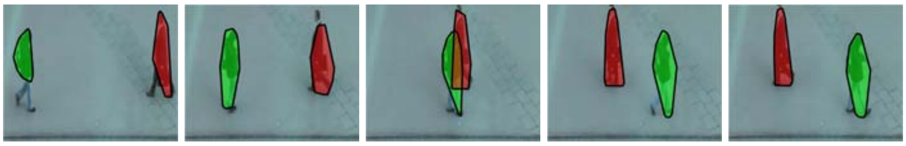
Sequence 2 shows a car tracking example. At lower levels of the hierarchy combinations formed in the image region of the moving target have similar structures as combinations detected at stationary objects in the background. Nevertheless, stable combinations constructed at higher levels are target-specific and delineate coherently moving parts in an unambiguous manner. Fig. 6 shows some frames of the sequence with the tracking results.



Frame 1 Frame 15 Frame 30 Frame 45 Frame 60

Fig. 6. Convex hull of the tracked object in sequence 2.

Sequence 3 depicts two pedestrians where one partially occludes the other for a certain number of frames. Target-specific stable combinations found before and after the occlusion event follow the same motion model and thus they become associated with the same trajectory. The interpolation step of part locations generates observations which are missing during occlusion. Both targets are segmented correctly and tracked in a stable manner, as shown in Fig. 7.



Frame 5 Frame 22 Frame 51 Frame 72 Frame 81

Fig. 7. Convex hull of tracked object of sequence 3.

As the proposed representation for tracking employs no prior model, it can be applied to track arbitrary targets, rigid and non-rigid objects. In absence of a prior model, feature selection for the tracking task is completely data-driven and unbiased, implying that all detected features and their combinations – when stable and following a common motion model – are used to estimate the structure and motion of a target.

7 Conclusion

This paper introduces first concepts of a novel tracking approach, where the structure of a target is represented by edge segment combinations, which are formed in a hierarchical analysis framework. The obtained structural models represent specific entities, which can be reliably associated between frames of a space-time volume. Combinations are formed in a fully data-driven manner and they integrate all available low-level features representing temporally invariant target structure and coherent motion. The framework is applicable to multiple interacting targets and presented grouping and tracking results show promising performance. In future we plan to achieve rotation invariance of our description to improve the performance with articulated objects and change the temporal association in a way so that nonlinear motion models are possible.

References

1. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: CVPR. Volume 1., IEEE Computer Society (2006) 798–805
2. Ramanan, D., Forsyth, D.A.: Finding and tracking people from the bottom up. In: CVPR. Volume 2., IEEE Computer Society (2003) 467–474
3. Geman, S., Potter, D., Chi, Z.: Composition systems. Quarterly of Applied Mathematics **60** (2002) 707–736
4. Fidler, S., Leonardis, A.: Towards scalable representations of object categories: Learning a hierarchy of parts. In: CVPR, IEEE Computer Society (2007) 1–8
5. Gao, Q.: Perceptual tracking of edge features. In: ICIP. Volume 1. (1994) 958–962
6. Opelt, A., Pinz, A., Zisserman, A.: A boundary-fragment-model for object detection. In: ECCV, Springer (2006) 575–588
7. Agarwal, A., Triggs, B.: Hyperfeatures multilevel local coding for visual recognition. In: ECCV, Springer (2006) 30–43
8. Crandall, D.J., Huttenlocher, D.P.: Weakly supervised learning of part-based spatial models for visual object recognition. In: ECCV, Springer (2006) 16–29
9. Bouchard, G., Triggs, B.: Hierarchical part-based visual object categorization. CVPR **1** (2005) 710–715
10. Ommer, B., Buhmann, J.M.: Compositional object recognition, segmentation, and tracking in video. In: EMMCVPR. Volume 4679., Springer (August 2007) 318–333
11. Fidler, S., Skočaj, D., Leonardis, A.: Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. PAMI **28** (March 2006) 337–350
12. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6) (June 1981) 381–395
13. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Computer Vision and Image Understanding **63** (January 1996) 75–104