Tracking Using a Hierarchical Structural Representation*

Anonymous OAGM submission

Abstract
Acquiring a highly specific target representation is a major challenge in the task of visual object tracking. High specificity substantially lowers the inherent ambiguity of the data association task and leads to improved tracking accuracy in presence of clutter. In this paper we propose a method generating a specific representation of the image structure for a tracked target in a hierarchically organized statistical process. Starting with simple and generic low-level local features (oriented edgelets) increasingly specific feature combinations are generated based on a temporal and spatial statistical analysis. The analysis delineates feature combinations with a frequent joint occurrence in the spatio-temporal domain. The detected relatively few specific combinations can efficiently guide a spatio-temporal association step of coherently moving image regions, and delineate the tracked target reliably. The proposed approach is demonstrated and evaluated in several experiments.

1. Introduction

Object representations which are (i) specific and (ii) at the same time invariant with respect to photometric, view and pose changes are essential components of reliable visual object recognition and tracking systems. Representations capturing the specific structure of an image object can be conceived, nevertheless, structural variations are hard to represent using models based on image statistics. In case of tracking, articulation, targets undergoing substantial pose variations and partial occlusions might lead to failures. Part-based approaches [2, 16] 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).

Hierarchical systems ensure an efficient way to represent exponential variability present in the visual data. Recent works [12, 8] 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.

In this paper we propose a tracking approach, which (i) builds a target-specific part-based model on concepts of compositionality [12] and (ii) employs the generated model for tracking. Given the ill-posed nature of data association in presence of ambiguous features, our motivation is to hierarchically combine massive amounts of ambiguous data into less parts of increasing complexity and specificity. Emerging specificity enables the tracking system to efficiently partition the data based on structural similarity and motion coherence, since the combinatorial problem of matching involves fewer parts with less ambiguity within and across frames of a sequence.

*Partially supported by the Austrian Science Fund under grants P18716-N13 and S9103-N13.

Robust statistics can be employed to estimate the underlying trends in the spatio-temporal distribution of temporally aggregated parts. Motion models estimated at higher levels of the hierarchy are propagated to lower levels of the hierarchy – guiding the motion estimation process for parts with less specificity – resulting in a spatially dense representation of the moving target.

The paper is organized as follows. Section 2. introduces related work. Section 3. gives an overview of our approach and Section 4. explains the algorithmic steps generating the hierarchical representation. Section 5. describes the association step enabling tracking. In Section 6. we discuss experimental results for several image sequences and in Section 7. we draw conclusions.

2. Related work

Edge segments are attractive low-level features offering a high degree of geometric and photometric invariance and encompassing a rich pool for building part-based models. Early works, as [11] by Gao, employ pre-designed rules to partition and group edge segments into more complex and distinctive entities. Partially driven by significant advances in visual object recognition, recent frameworks propose statistical, learning-based methodologies.

Opelt et al. [15] employ object boundary fragments to detect multiple object classes in presence of clutter and partial occlusions. Boosting is used to extract and select class-discriminative boundary fragments. A novel multilevel visual representation called "hyperfeatures" is introduced by Agarwal and Triggs in [3]. They iteratively organize local sets of image descriptors into more compley and spatially sparse parts. This results in a system able to localize several object categories. Crandall et al. [6] 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. present in [5] a hierarchical part-based description encoding geometry and appearance of object parts. The learned part-based models vote in a bottom-up manner for possible object locations.

In the field of object tracking, hierarchical representations of the structure of spatially extended targets have not been investigated in great detail yet. Ommer et al. [14] 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 relationships between parts to learn compositional structured object models. In comparison to our work, Ommer et al. do not use a hierarchical representation and propagate hypothesized part compositions over consecutive frames. In this paper similar concepts as presented by Fidler et al. [8, 7] are applied to hierarchically group and organize spatio-temporally aggregated low-level features. The aim is to generate compositions of low-level features, which can be associated and tracked in an unambiguous manner.

3. Overview of our approach

This approach starts with extracting simple oriented edge segments from a sequence of images. Then local spatial configurations of multiple oriented edge segments – denoted further on as *combinations* – are built. This combinations encode the local structure of objects (i.e. rigid parts of an articulated object). The aim of this approach is to select a set of temporally invariant combinations by statistical analysis, and to reliably associate them to form trajectories. During the association process the *stability* (temporally invariant structure) and the high *specificity* (low occurrence frequency) properties of the selected combinations are exploited. As can be seen in Figure 1 the hierarchical description is

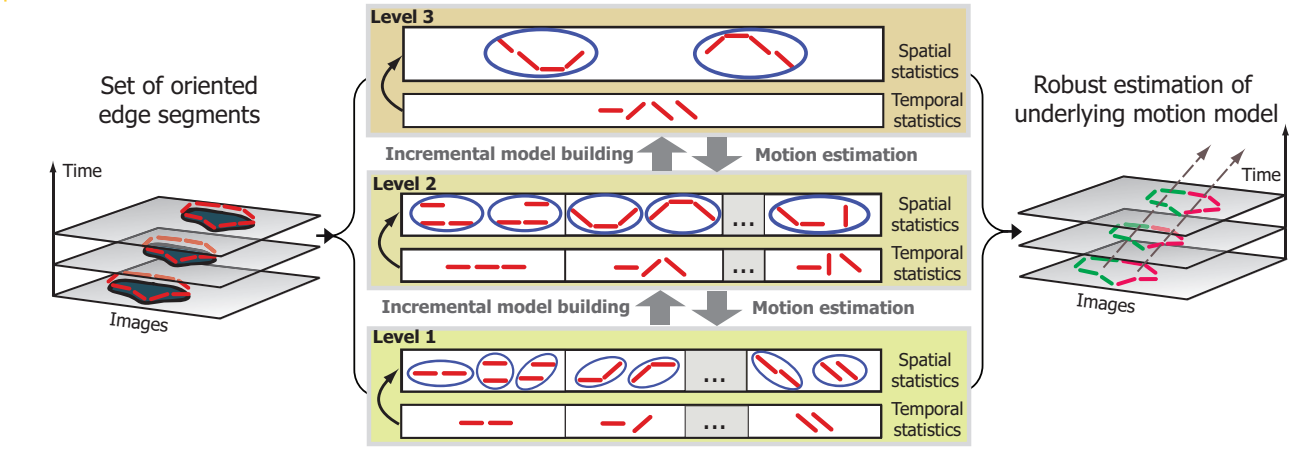


Figure 1. Concept of our approach.

built in a bottom-up process, while tracking (motion estimation) is done in a top-down manner.

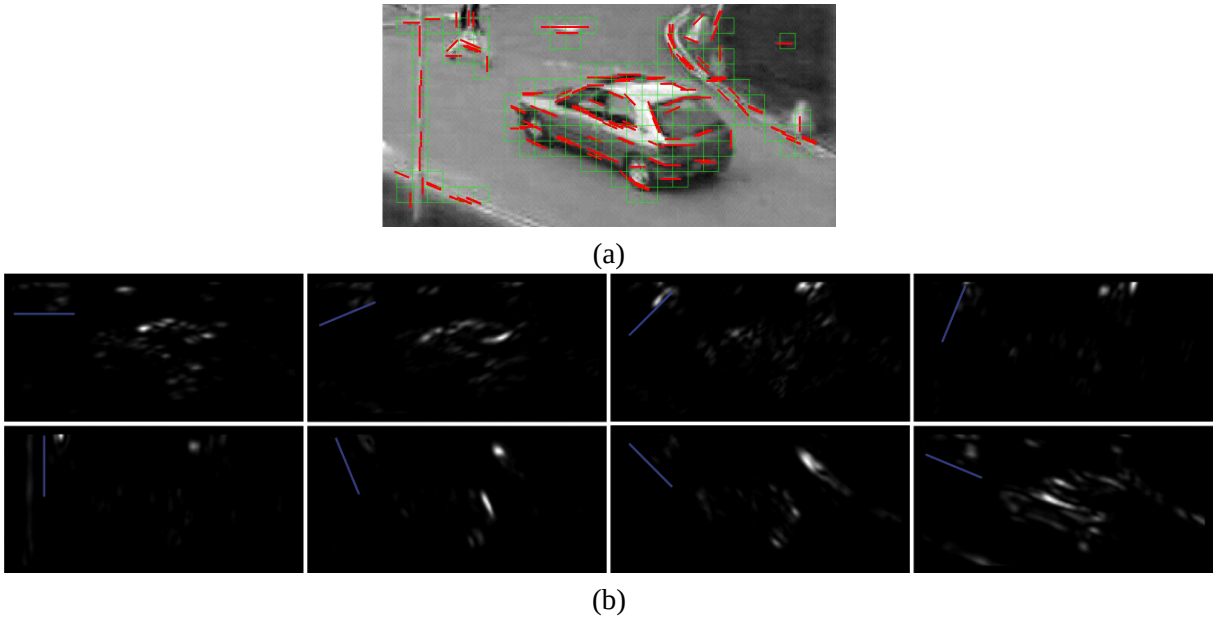
4. Building the hierarchical structural representation

The input of the bottom-up process are oriented edge segments detected in the space-time volume (several consecutive frames) of a video sequence (see Figure 1). Beginning at level 1 up to a desired top level, combinations of local features are built by grouping edge segments together in combinations of increasing size. Temporal and spatial statistics are employed to reduce the number of all possible edge segment combinations in a level to a set of *stable* combinations. The stable combinations of level L_i represent the basis for the combinations of the next level L_{i+1}. Each stable combination of level L_i is extended by an additional segment selected out of a local neighborhood and the statistical analysis (temporal and spatial) is repeated to produce the stable combinations for level L_{i+1}. This incremental building process is carried out until a desired top level is reached, where only few, specific combinations remain, e.g. level 3 in Figure 1, or no more stable combinations are found.

4.1. Extraction of edge segments

In this approach the local structure of foreground objects and background is described by oriented edge segments as in [8, 7]. Oriented edges characterize the local geometry in a spatially localized and more selective manner than local histograms. By using a filter bank consisting of oriented Gabor filters (8 orientations in $0°$-$180°$, $\sigma = 0.7$) the local edge segments are detected.

Each frame of the current space-time volume is filtered with the oriented filter bank and for each orientation the magnitude of filter responses (without sign) is calculated. To extract edge segments, each image is divided into a set of non-overlapping rectangular neighborhoods of a size D. The size of the rectangular neighborhoods is small enough to capture important shape details such as the shoulder silhouette of a human. In each rectangular neighborhood the locally dominant orientation is determined by analyzing all filter responses within the neighborhood and finding the orientation with the maximum response. If a local maximum response is smaller than T_m it is considered as noise and ignored. For the experiments in this paper we used the values T_m=15 and D=10 pixels, the latter being approximately $\frac{1}{10}$ of the foreground object's height. Figure 2 shows an example for the extraction of edge segments and the responses of the oriented filter bank.



(a)

(b)

Figure 2. Application of oriented filter bank on one frame of sequence 3 (PETS 2001 dataset). (a) Extracted edge segments (red) and grid of local neighborhoods (green). (b) Responses of oriented filter bank (blue lines visualize orientations).

4.2. Building Levels of hierarchy

Each level L_i of the hierarchical structural representation contains a set of M_{L_i} stable combinations $C = \{c_1, c_2, \ldots, c_M\}$ consisting of $i+1$ edge segments. Each edge segment s_l represents one of O possible edge orientations $S = \{s_1, s_2, \ldots, s_O\}$. As we aim to avoid exponential complexity (O^N for N levels), the combinations of edge segments are built as follows.

Each combination of edge segments c_k in level L_i consists of i segments forming the primary part p_{prim} and one segment representing the marginal part p_{mar}. For expample 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. The primary parts of the combinations in level L_i are the stable combinations from the previous level L_{i-1} (level L_1 is an exception of this property). So the number of possible combinations of edge segments for level L_i is $M_{L_{i-1}} \cdot O$.

A combination is uniquely defined by the orientation of its edge segments e.g. $\{s_1, s_2, s_5, s_7\}$. Each combination can occur several times in a frame – denoted further on as *combination occurrence*.

4.2.1. Level i

For each frame in the current space-time volume all possible combinations of p_{prim} and p_{mar} are enumerated. This is done 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 B is defined to be small at lower levels of the hierarchy and increased at higher levels (see Tabular 2).

The local window B is centered over each stable combination of level L_{i-1}, representing the primary part p_{prim} of the new combination. Then all possible combinations of p_{prim} with an additional segment

out of B representing p_{mar} are formed. Every combination occurrence has to be unique, meaning that no couple of combinations is allowed to contain the same edge segments, they have to differ at least by one segment.

4.2.2. Level 1

To process of building the combinations for the bottom level of the hierarchy differs from the other levels as there are no stable configurations from a previous level. The local analysis window B 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. As there is no stable combination from a previous level to represent the primary part it is necessary to come up with a different assignment rule. The combinations of the first level consist of two edge segments, where the segment with the lower index simply defines p_{prim} (e.g. s_2) and the segment with the higher index becomes p_{mar} (e.g. s_5).

4.3. Statistical analysis

After all possible combination occurrences of a level are found, temporal and spatial statistics are applied to select the stable combinations.

4.3.1. Temporal statistics

The task of temporal statistics is to estimate the density of the occurrence of each combination within in the current space-time volume by binning and to retain most frequently occurring combinations.

The density of the occurrence of each combination is captured with the help of a 3D histogram H, where the histogram is spanned by the primary part indices p_{prim} (unique index assigned by algorithm), the marginal part indices p_{mar} (orientation index) and the frame numbers f. Combinations which appear in a certain percentage of frames, defined by threshold T_f, are retained. The outcome of temporal statistics is a set of temporally stable combinations. Temporal statistics do not consider the spatial arrangement of the edge segments in the image space, this is done by spatial statistics (see following paragraph).

4.3.2. Spatial statistics

The combinations remaining after temporal statistics form the input for spatial statistics. Spatial statistics focus on analyzing the spatial arrangement of the edge segments. Figure 3 visualizes the creation of a co-occurrence histogram of edge segments, which encodes the spatial relationships. For each temporally stable combination all occurrences are analyzed. The primary part p_{prim} of the combination is centered at $(0, 0)$ in a local coordinate system and the spatial distribution of the corresponding marginal parts p_{mar} – relative to th primary part 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 the F frames of the current space-time volume. Mean shift mode seeking is employed on each distribution to locate the most (up to four) significant modes. Modes with densities below a threshold T_p and its corresponding combinations are discarded. The result

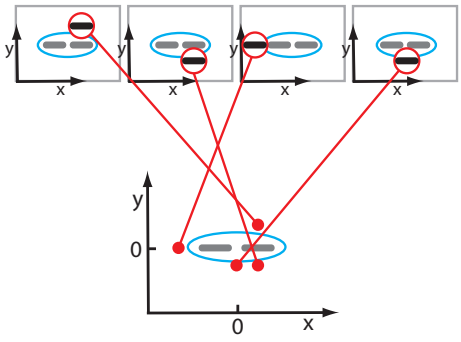
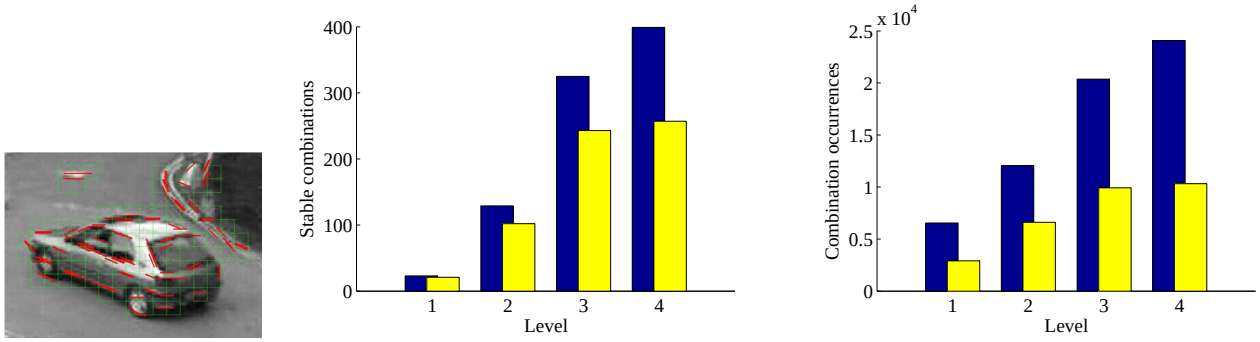


Figure 3. Concept of spatial statistics. At the top different spatial edge segment arrangements of a combination are shown. In the bottom the creation of the spatial distribution (relative to the centered primary part) is visualized.



(a) (b) (b)

Figure 4. An example of the numerical results of temporal and spatial statistics. (a) Extracted edge segments. (b) Number of stable combinations after temporal (blue, dark) and spatial statistics (yellow, bright). (c) Occurrences of stable combinations after temporal (blue, dark) and spatial statistics (yellow, bright).

after spatial statistics is a set of stable combinations in space and time. A stable combination is now uniquely defined by the orientation of its edge segments and their spatial arrangement.

Figure 4 shows two bar diagrams, where (b) displays the effect of temporal and spatial statistics on the number of stable combinations and (c) shows the effect of the statistics on the occurrences of those combinations. The numbers of the bar diagrams are from the statistics of a space-time volume of sequence 3 (see experiments Section 6.). Tabular 1 complements the information from Figure 4(b) with the number of all possible combinations. For all experiments in Section 6. the parameters described in this section are set to the values of Tabular 2.

Table 1. The numerical data of the combinations of bar diagram (b) in Figure 4 (third and forth column) and the number of all possible combinations (second column).

Level	All combinations	After temporal statistics	After spatial statistics
1	36	23	21
2	288	129	102
3	2304	325	243
4	18432	399	257

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

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

5. Tracking using the built hierarchy

The hierarchical representation is used in a top-down manner to estimate the motion models of foreground objects. While going up in the hierarchy the combinations of edge segments become more distinctive and in the best case the combinations appear only once in an image. The idea behind the top-down process is that the association of combinations at the top level is less ambiguous and can be used to guide the association step of combinations at lower levels (with less specificity). Using the combinations of all levels of the hierarchy results in a dense structural description of the foreground objects. Foreground objects are delineated by grouping stable combinations, which obey the same motion model.

The proposed top-down tracking consists of three steps: (1) temporal association between the obtained combinations using robust statistical estimation, (2) grouping of combinations following the same motion model (see Section 5.1.) and (3) association of trajectory segments in overlapping space-time volumes (see Section 5.2.).

5.1. Temporal association

Reliable association of combinations requires that the combinations are stable and highly specific. The previously described temporal and spatial statistics capture combinations which occur in multiple frames of the analyzed space-time volume. Combinations in the higher levels of the hierarchy are more specific and occur less frequently. The estimation task of the underlying motion models of the combinations can be solved as a regression problem. To keep the complexity of the motion model estimation low, linear motion models are assumed within the analyzed space-time volume (typically 20 frames).

The RANSAC algorithm [9] is used to carry out the regression task. Estimation is started at the top level of the 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 top level the best fitting motion model is estimated. 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 [4], in a two-dimensional vector space spanned by velocity components along x and y. Mode seeking is performed in the velocity space to find the underlying trends – peaks defined by velocity components of frequently occurring motion models. Usually there is a mode around the origin of the velocity space encompassing combinations belonging to the stationary background (no movement). Other modes represent moving foreground objects.

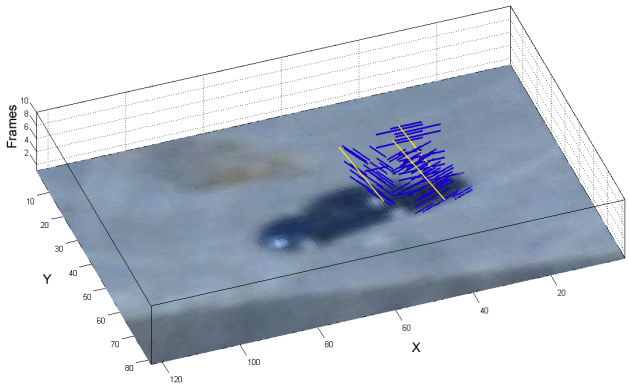


Figure 5. Space-time plot showing all stable combinations describing a moving object (car from sequence 1 of the experiment section) and obeying the same motion model estimated by robust regression.

Since the number of stable combinations at the top level is low, the obtained set of coherently moving combinations defines only a spatially sparse object description. To obtain a spatially more dense description, RANSAC estimation is also performed at lower levels of the hierarchy. As the estimation of motion models at lower levels is considerably more ambiguous, the estimation is guided by the motion models estimated at the top level. If motion models of combinations in lower levels do not belong to any of the previously detected peaks in the velocity space they are discarded. In this way, RANSAC 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 description of foreground objects (see example in Figure 5).

As a given stable combination is not 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 (see results in experiment Section 6.).

5.2. 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, the hierarchical description is independently rebuilt in an incremental manner within overlapping space-time volumes (see Figure 6). The size of the space-time volume (number of frames) and the overlap highly depends on the video sequence. Clearly, the space-time volume will be small (e.g. 10 frames) if the object is undergoing fast motion and appearance changes.

Assuming kinematic smoothness of target motion, trajectory segments obtained for individual volumes are associated using the estimated motion models and spatial proximity. Mean shift mode seeking is applied to the distribution of the velocity space to associate the motion models from the previous space-time volume to the models in the current volume. As the motion model of a foreground object is not going to change dramatically between the overlapping space-time volumes the association task is easily solved.

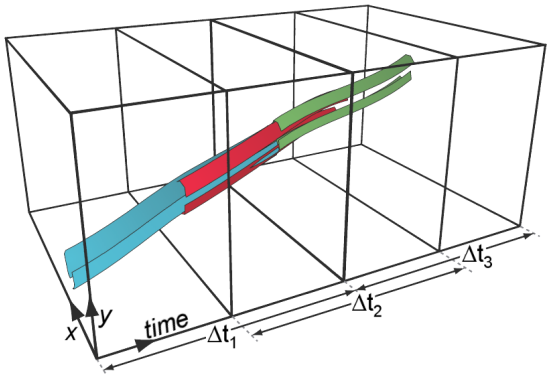


Figure 6. Illustration depicting the incremental space-time object tracking. Different colors indicate distinct stable combinations defining object trajectory segments in consecutive overlapping space-time volumes.

Figure 7. Convex hull and trajectory of a tracked object in sequence 1 (from the VIVID dataset).

6. Experiments and Discussion

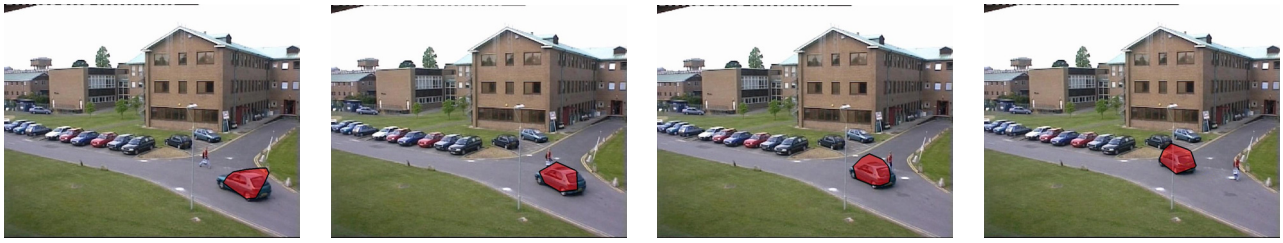
We performed tracking experiments on three publicly available video sequences to prove our concept and qualitatively evaluate the performance of our approach. Video sequence 1 is part of the Vivid Tracking Evaluation Testbed [13], sequence 2 is taken from the CAVIAR dataset [10] and sequence 3 is a video of the PETS 2001 dataset [1].

Sequence 1 shows multiple moving cars and the scene is viewed by a moving aerial camera. In this experiment a moving car is segmented and tracked as a foreground object. The obtained convex hull – spanned by segments belonging to stable combinations – covers image regions, where spatio-temporal stability was found (see Figure 7). Segment combinations on the moving background are detected as well and form a spatially separated structure (not shown), since they follow a different motion model than the car. Figure 7 shows the delineated and tracked car along with its trajectory.

Sequence 2 shows a pedestrian tracking example (see Figure 8). The line pattern of the ground plane tiling generates combinations, which are at lower levels of the hierarchy similar to the combinations found at the person. However, at higher levels of the hierarchy certain features combinations occur only for the person, rendering the association task feasible. The coherently moving parts of the

Frame 1057 Frame 1076 Frame 1104 Frame 1165

Figure 8. Convex hull of the tracked object in sequence 2 (from the CAVIAR dataset).



Frame 33 Frame 53 Frame 73 Frame 102

Figure 9. Convex hull of tracked object of sequence 3 (from the PETS 2001 dataset).

person are delineated in an unambiguous manner. Figure 8 shows some frames of the sequence with the tracking results.

Sequence 3 depicts another car tracking example (see Figure 9). Target-specific stable combinations are found despite of the slow velocity of the vehicle. The car is tracked in a stable manner, as shown in Figure 9.

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

In this paper we introduced a tracking approach motivated by the concept of compositionality, where specific combinations of multiple simple features are formed in a hierarchical analysis framework. The set of detected specific parts is partitioned spatially to delineate coherently moving image regions which form the tracked targets. The approach is able to cope with temporally smooth structural variations by performing the search for specific combinations in consecutive spatio-temporal volumes. Combinations are formed in a fully data-driven manner while integrating information from all available low-level features representing slowly varying target structure. The framework is applicable to multiple interacting targets and the presented initial grouping and tracking results show promising performance. Future plans involve the incorporation of rotationally invariant low level descriptions in order to enable the system to segment multiple coherently moving regions of articulated objects.

References

[1] Pets: Performance evaluation of tracking and surveillance. http://www.cvg.rdg.ac.uk/slides/pets.html, Feburary 2009.

[2] Amit Adam, Ehud Rivlin, and Ilan Shimshoni. Robust fragments-based tracking using the integral histogram. In *CVPR*, volume 1, pages 798–805. IEEE Computer Society, 2006.

[3] Ankur Agarwal and Bill Triggs. Hyperfeatures û multilevel local coding for visual recognition. In *ECCV*, pages 30–43. Springer, 2006.

[4] Michael J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63:75–104, January 1996.

[5] Guillaume Bouchard and Bill Triggs. Hierarchical part-based visual object categorization. *CVPR*, 1:710–715, 2005.

[6] David J. Crandall and Daniel P. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV*, pages 16–29. Springer, 2006.

[7] S. Fidler, D. Skočaj, and A. Leonardis. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *PAMI*, 28:337–350, March 2006.

[8] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR*, pages 1–8. IEEE Computer Society, 2007.

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[10] Robert Fisher. Caviar: Context aware vision using image-based active recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/, Feburary 2009.

[11] QiGang Gao. Perceptual tracking of edge features. In *ICIP*, volume 1, pages 958–962, 1994.

[12] S. Geman, D. Potter, and Z. Chi. Composition systems. *Quarterly of Applied Mathematics*, 60:707–736, 2002.

[13] Martial Hebert. Vivid tracking evaluation testbed. http://www-old.ri.cmu.edu/projects/project_520_text.html, Feburary 2009.

[14] Bjórn Ommer and Joachim M. Buhmann. Compositional object recognition, segmentation, and tracking in video. In *EMMCVPR*, volume 4679, pages 318–333. Springer, August 2007.

[15] Andreas Opelt, Axel Pinz, and Andrew Zisserman. A boundary-fragment-model for object detection. In *ECCV*, pages 575–588. Springer, 2006.

[16] Deva Ramanan and David A. Forsyth. Finding and tracking people from the bottom up. In *CVPR*, volume 2, pages 467–474. IEEE Computer Society, 2003.