Tracking Objects Beyond Rigid Motion*

Nicole Artner[1], Adrian Ion[2], and Walter G. Kropatsch[2]

[1] Austrian Research Centers GmbH - ARC, Smart Systems Division, Vienna, Austria
nicole.artner@arcs.ac.at
[2] PRIP, Vienna University of Technology, Austria
{ion,krw}@prip.tuwien.ac.at

Abstract. Tracking multiple features of a rigid or an articulated object, without considering the underlying structure, becomes ambiguous if the target model (for example color histograms) is similar to other nearby regions or to the background. Instead of tracking multiple features independently, we propose an approach that integrates the underlying structure into the tracking process using an attributed graph. The independent tracking processes are driven to a solution that satisfies the visual as well as the structural constraints. An approach for rigid objects is presented and extended to handle articulated objects consisting of rigid parts. Experimental results on real and synthetic videos show promising results in scenes with considerable amount of occlusion.

1 Introduction

Tracking multiple features belonging to rigid as well as articulated objects is a challenging task in computer vision. Features of rigid parts can change their relative positions due to variable detection precision, or can become occluded. To solve this problem, one can consider using part-based models that are tolerant to small irregular shifts in relative position - non-rigid motion, while still imposing the global structure, and that can be extended to handle articulation.

One possibility to solve this task is to describe the relationships of the parts of an object in a deformable configuration - a spring system. This has already been proposed in 1973 by Fischler et al. [1]. Felzenszwalb et al. employed this idea in [2] to do part-based object recognition for faces and articulated objects (humans). Their approach is a statistical framework minimizing the energy of the spring system learned from training examples using maximum likelihood estimation. The energy of the spring system depends on how well the parts match the image data and how well the relative locations fit into the deformable model. Ramanan et al. apply in [3] the ideas from [2] in tracking people. They model the human body with colored and textured rectangles, and look in each frame for likely configurations of the body parts. Mauthner et al. present in [4] an approach using a two-level hierarchy of particle filters for tracking objects described by spatially related parts in a mass spring system.

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

In this paper we employ spring systems, but in comparison to the related work we try to stress solutions that emerge from the underlying structure, instead of using structure to verify statistical hypothesis. The approach presented here refines the concepts in [5] and extends them to handle articulation. Initial thoughts related to this work have been presented in the informal workshop [6]. The aim is to successfully track objects, consisting of one or more rigid parts, undergoing non-rigid motion. Every part is represented by a spring system encoding the spatial relationships of the features describing it. For articulated objects, the articulation points are found through observation of the behavior/motion of the object parts over time. The articulation points are integrated into the spring systems as additional distance constraints of the parts connected to them.

Looking at related work in a broader field, the work done in tracking and motion analysis is also related to our approach. There is a vast amount of work in this field, as can be seen in the surveys [7–10]. It would go beyond the scope of this paper mentioning all of this work. Interesting to know is that early works even date back to the seventies, where Badler and Smoliar [11] discuss different approaches to represent the information concerning and related to the movement of the human body (as an articulated object).

The paper is organized as follows: Sec. 2 introduces tracking rigid parts with a spring system. In Sec. 3 this concept is extended to tracking articulated objects. Experiments on real and synthetic videos and a discussion are in Sec. 4. Conclusion and future plans can be found in Sec. 5.

2 Tracking rigid parts

To identify suitable features of a rigid object, the Maximally Stable Extremal Regions (MSER) detector [12] is used to extract regions in a manually delineated region of interest. An attributed graph (AG) represents the structural dependencies. It is created by associating a vertex to each region. The corresponding 3D color histograms of the underlying regions are the attributes of the vertices. In this approach, a Delaunay triangulation is employed to insert edges between the vertices (color regions) and to define the spatial relationships between the regions. A triangulation can model a rigid structure just by imposing distance constraints between connected vertices. On each vertex of the AG, a feature tracker, in our case the Mean Shift tracker [13], is initialized and the color histograms of the vertices in the initial state become the target models $\hat{q}$.

During object tracking the color histograms of the AG and "spring-like" edge energies of the structure are used to carry out gradient descent energy minimization on the joint distribution surface (color similarity and structure).

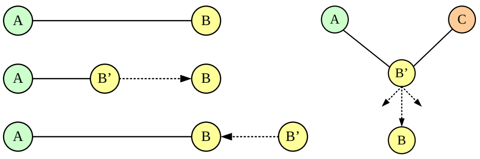
2.1 Realizing the spring system by graph relaxation

The objective is to link the processes of structural energy minimization of the graph, and color histogram similarity maximization by Mean Shift tracking. Graph relaxation is a possibility to realize the spring system. It introduces a mechanism, which imposes structural constraints on the mode seeking process of Mean Shift. As the tracked objects are rigid, the objective of the relaxation is to maintain the tracked structure as similar as possible to the initial structure. Thus the aim of graph relaxation is to minimize the total energy of the structure.

The variations of the edge lengths in the AG and their directions are used to determine a structural offset for each vertex. This offset vector is the direction where a given vertex should move such that its edges restore their initial length and the energy of the structure is minimized. This structural offset vector $\boldsymbol{O}$ is calculated for each vertex v as follows:

$$\boldsymbol{O}(v) = \sum_{e \in E(v)} k \cdot (|e'| - |e|)^2 \cdot (-\boldsymbol{d}(e, v)), \tag{1}$$

where $E(v)$ are all edges e incident to vertex v, k is the elasticity constant of the edges in the structure, e is the edge length in the initial state and e' at a different point in time. $\boldsymbol{d}(e, v)$ is the unitary vector in the direction of edge e that points toward v. Fig. 1 shows two simple examples for graph relaxation.



(a) (b)

Fig. 1. Graph relaxation examples. B is the initial state of the vertex and B' the deformed one. The arrows visualize the structural offset vectors $O(B')$.

2.2 Combining iterative tracking and graph relaxation

For every frame, Mean Shift and structural iterations are performed until a maximum number of iterations is reached ϵ_i, or the graph structure attains equilibrium, i.e. its total energy is beneath a certain threshold ϵ_e (see Algorithm 1).

The ordering of the regions during the iterations of the optimization process depends on the correspondence between the candidate models $\hat{p}$ of the regions in the current frame and the target models $\hat{q}$ from the initialization. Both models are normalized 3D color histograms in the RGB color space. The similarity between the models can be determined by the *Bhattacharyya* coefficient

$$B = \sum_{u=1}^{m} \sqrt{\hat{p}_u \cdot \hat{q}_u}. \tag{2}$$

For more details on the Bhattacharyya coefficient see [13]. The regions are ordered descending by the Bhattacharyya coefficient and with this the iterations start with the most confident regions.

To compute the position of each region (vertex in AG), Mean Shift offset and structure-induced offset are combined using a mixing coefficient

$$g = 0.5 - (B - 0.5). \tag{3}$$

g weights the structural offset vector and $1 - g$ the offset of Mean Shift. This gain ensures that the offset vector of Mean Shift has a greater influence on the resulting offset vector if the Bhattacharyya coefficient B is high, meaning that candidate and target model are similar. If the Bhattacharyya coefficient is low the gain leads to an increased influence of the structural offset.

3 Imposing articulation

Articulated motion is a piecewise rigid motion, where the rigid parts conform to the rigid motion constraints, but the overall motion is not rigid [10]. An *articulation point* connects several rigid parts. The parts can move independent to each other, but their distance to the articulation point remains the same. This paper considers articulation in the image plane (1 degree of freedom).

As described in Sec. 2, the rigid parts of an articulated object are tracked combining the forces of the deterministic tracker and the graph structure. To integrate articulation, two vertices of each rigid part are connected with the common articulation point[3]. These two *reference vertices* constrain the distance of all other vertices of the same part to the articulation point. The reference vertices are directly influenced by the articulation point and propagate the "information" from the other connected parts during tracking.

Each rigid part is iteratively optimized as explained in Sec. 2 and for articulated objects the articulation points are integrated into this process through their connection to the reference vertices.

Important features of the structure of an object do not necessarily correspond to easily trackable visual features, e.g. articulation points can be occluded, or can be hard to track and localize. Articulation points are thus not associated to a tracked region (as opposed to tracked features of the rigid parts). The position of the articulation points is determined in an initial frame (see Sec. 3.1) and used in the rest of the video (see Sec. 3.2).

3.1 Determining the articulation point

For discrete time steps the motion of rigid parts connected by an articulation point can be modeled by considering rotation and translation separately:

$$\mathbf{p}' = \text{translate}(\text{rotate}(\mathbf{p}, \mathbf{c}, \theta), \mathbf{o}),$$

[3] One could consider connecting all points of a part, but this would unnecessarily increase the complexity of the optimization process.

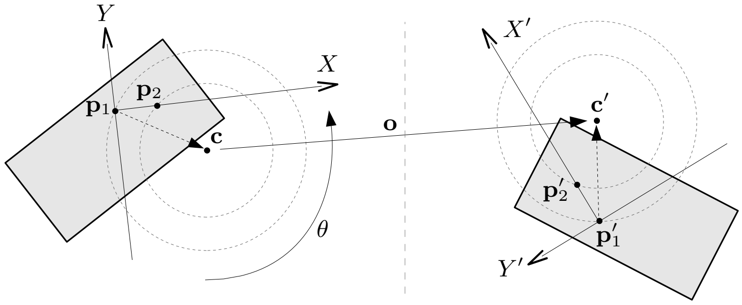


Fig. 2. Encoding and deriving of an articulation point in the local coordinate system, during two time steps: t and $t+\delta$.

where $\mathbf{p} = (x, y)$ is the vertex at time t and $\mathbf{p}' = (x', y')$ is the same vertex at time $t+\delta$. $\mathbf{p}'$ is obtained by first rotating $\mathbf{p}$ around $\mathbf{c} = (x_c, y_c)$ with angle θ and then translating it with offset $\mathbf{o} = (x_o, y_o)$. More formally,

$$\mathbf{p}' = (R * (\mathbf{p} - \mathbf{c}) + \mathbf{c}) + \mathbf{o}, \tag{4}$$

where R is the 2D rotation matrix with angle θ.

To compute the position of $\mathbf{c}$ at time t it is enough to know the position of two rigid parts A and B. Each of them is represented by two reference vertices, at times t and $t+\delta$: $\mathbf{p}_i, \mathbf{p}_i'$, $0 < i \leqslant 4$, where $\mathbf{p}_i$ is the position of a vertex at time t and $\mathbf{p}_i'$ is the position at time $t+\delta$. The vertices of part A are identified by $i \in \{1,2\}$ and of B by $i \in \{3,4\}$. The previous relations produce a system of eight equations in eight unknowns: x_c, y_c, x_o, y_o, $\sin(\theta_A)$, $\cos(\theta_A)$, $\sin(\theta_B)$, $\cos(\theta_B)$, where θ_A, and θ_B are the rotation angles of the two parts.

The position of the articulation point $\mathbf{c}$ is computed in the first frames and used further on as mentioned below.

3.2 Integration into spring system

To derive the position of the articulation point in each frame of the video, the following procedure is applied. In the frame in which the position of the articulation point was computed (see Sec. 3.1), a local coordinate system is created for each adjacent rigid part and aligned to the corresponding reference vertices. In Fig. 2 this concept is shown, where $\mathbf{p}_1, \mathbf{p}_2, \mathbf{c}, X, Y$ are the tracked vertices, articulation point (rotation center) and coordinate system at time t; $\mathbf{p}_1', \mathbf{p}_2', \mathbf{c}', X', Y'$ at time $t+\delta$; $\mathbf{o}$ is the offset (translation), and θ is the rotation angle. The positions of the articulation point in the local coordinate systems of each connected part are determined and associated to the respective rigid part.

In every frame, having the tracked reference vertices enables determining the local coordinate system and the position of the articulation point. For each frame Algorithm 1 is executed. When determining the current position of the

Algorithm 1 Algorithm for tracking articulated objects.

```
 1: PROCESSFRAME
        $\epsilon_e$ threshold total energy of structure
        $\epsilon_i$ threshold maximum number of iterations
 2:     $i \leftarrow 1$                                        ▷ iteration counter
 3:     while ($i < \epsilon_i$ and $E_t > \epsilon_e$) do
 4:         for every rigid part do
 5:             define region order depending on $B$
 6:             for every region do
 7:                 do Mean Shift iteration
 8:                 do structural iteration
 9:                 calculate mixing gain $g$ (Eq. 3)
10:                 mix offsets depending on $g$ and set new position
11:             end for
12:         end for
13:         calculate current position of articulation point (Eq. 5)
14:         for every rigid part do
15:             define region order depending on $B$
16:             for every region do
17:                 do Mean Shift iteration
18:                 do structural iteration including articulation point
19:                 calculate mixing gain $g$
20:                 mix offsets depending on $g$ and set new position
21:             end for
22:         end for
23:         $i \leftarrow i + 1$
24:         $E_t \leftarrow$ determine total energy of spring system
25:     end while
26: end
```

articulation point (line 13 in Algorithm 1), the hypothesis of the adjacent parts for the position of the articulation point are combined using the gain a:

$$a_i = \frac{Z_i}{\sum\limits_{k=1}^{m} Z_k}, \quad Z_i = \sum_{j=1}^{v_i} B_{ij} \tag{5}$$

where Z_i is the sum of all Bhattacharyya coefficients (see Eq. 2) of part i with v_i regions/vertices, m is the number of adjacent regions, and a_i is the gain for part i weighting its influence on the position of the articulation point. a_i depends on the correspondence of the color regions of a rigid part with the target models $\hat{q}$ of this regions from the initial frame. This results in high weights for the hypothesis of parts which are confident (e.g. not occluded).

4 Experiments

The following experiments show sequences with one articulation point. More articulation points can be handled by pairwise processing of all adjacent rigid parts (a more efficient strategy is planned). In all experiments we employ a priori knowledge about the structure of the target object (number of rigid parts and articulation points). A method like in [14] could be used to automatically delineate rigid parts and articulation points of an object. The elasticity constant k (see Equ. 1) is set to 0.2 for all experiments (this value was selected empirically).

Experiment 1: Fig. 3 shows an experiment with a real video sequence, where the challenge is to track the partly non-rigid motion of the pattern on the t-shirt. The pattern is not only translated and rotated, but also squeezed and expanded (crinkles of the t-shirt). The idea behind this experiment is to show how the proposed approach handles independent movement of the features of a single rigid part. As can be seen in the second row of Fig. 3 and Tab. 1, Mean Shift combined with structure is superior to Mean Shift alone. The graphs in Fig. 3 (and all other experiments) represent a visual support to easily see the spatial arrangement of the tracked regions. In the results without the spring system there is no inter-relation between the trackers, and the graphs show the deformation of the structure of the object.

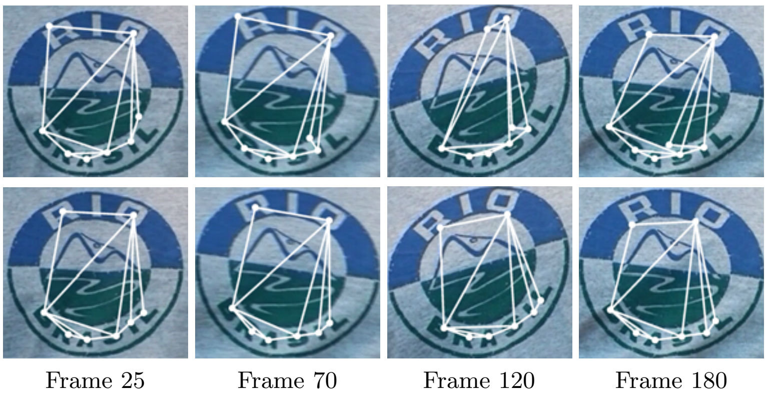


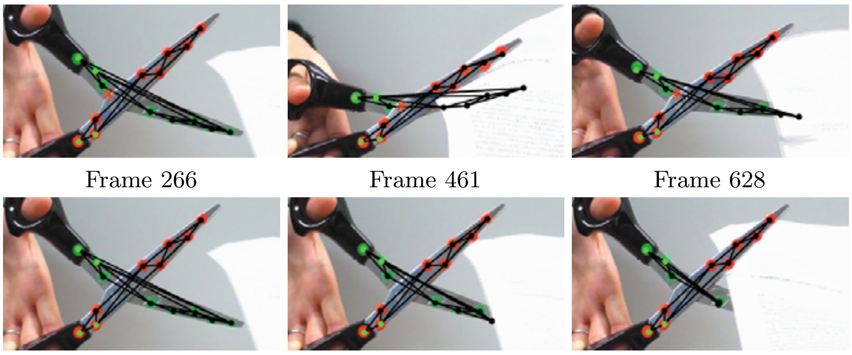
Fig. 3. Experiment 1. Tracking non-rigid motion without (top row) and with structure (bottom row). Frame 25 in bottom row shows how the graph should look like.

Table 1. Sum of spatial deviations in pixels from ground truth for experiment 1.

Frame	25	70	120	180
spatial deviation without structure	122.18	152.66	269.86	196.96
spatial deviation with structure	58.99	66.49	140.64	124.96

Experiment 2: In experiment 2 the task is to track scissors through partial occlusions. The employed Mean Shift tracking tracks color regions. It was necessary

to put color stickers on the scissors, to create features to track. Fig. 4 shows that the additional information provided by structure helps to successfully overcome the occlusion. Without the support of the spring system the Mean Shift trackers mix up the regions.


Frame 266 Frame 461 Frame 628

Frame 266 Frame 274 Frame 286

Fig. 4. Experiment 2. Top row: with structure and articulation point. Bottom row: without structure. The red star-like symbol represents the estimated articulation point.

Experiment 3: In the following experiment (see Fig. 5) a synthetic sequence is used to accurately analyze the behavior of the approachThe synthetic pattern contains 7 color regions (region size: height 10 to 20 pixels, width 10 to 20 pixels) and is 50×100 pixels, the occlusion is 100×100 pixels. The patterns are translated by a x-offset of 6 pixels per frame and rotated by 4 degrees. Due to the big movement between the frames and the full occlusion of the left pattern in frame 8, separately tracking the patterns fails. Using the estimated articulation point, it is possible to successfully track the regions through this sequence. The distance constraint imposed by the articulation point is the reason why, even though there are big to full occlusions, the positions of the occluded regions can be reconstructed without visible features. Fig. 6 shows the deviation from ground truth of experiment 3. We did several of these synthetic experiments and found out that tracking including the articulation point is in all cases superior to tracking the parts separately.

4.1 Discussion

The Mean Shift tracker fits very well into our approach as the spring system optimization is also iterative, and we are able to re-initiate Mean Shift at any given state of a vertex in the spring system. Another tracker with the same properties could also be used. As tracking with Mean Shift is used to solve the association task (avoiding complex graph matching), the success of this approach is highly dependent on the results of the trackers. It is necessary that at least part of the vertices of the spring system can be matched.

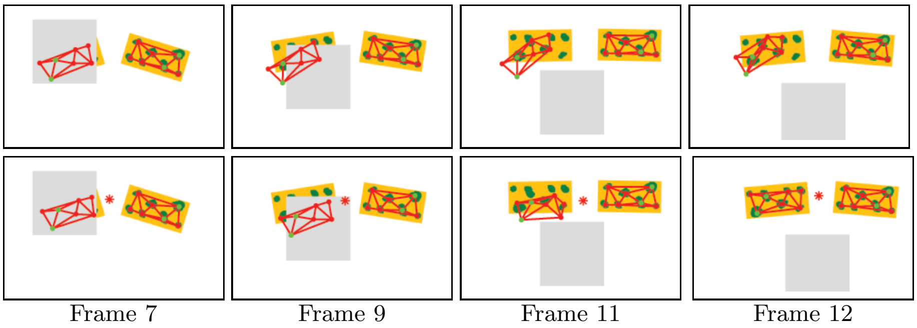
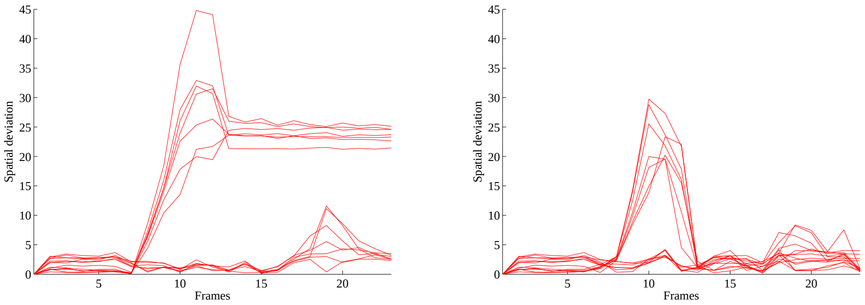


Fig. 5. Experiment 3. Top row without articulation point and bottom row with.



(a)

(b)

Fig. 6. Spatial deviation for each region. (a) without and (b) with articulation point. The big deviations are a result of the full occlusion in frame 8 in Fig. 5.

The current approach extends the rigid structure to handle articulation. This only imposes a distance constraint and does not consider any information related to the motion of the parts. During an occlusion the articulation point improves the reconstruction of the positions of the occluded regions. Nevertheless, the distance constraint brought in by the articulation point is not always enough to successfully estimate the positions (it is sufficient for translations, but not for rotations of parts). For example if one of two rigid parts of an object is completely occluded and there is a big rotation of the occluded part between adjacent frames this approach may fail.

At the moment the two reference vertices are selected with no special criteria. This criteria could be the connectivity of the vertices or their visual support.

5 Conclusion

This paper presents a structural approach for tracking objects undergoing non-rigid motion. The focus lies on the integration of articulation into the spring systems describing the spatial relationships between features of the rigid parts of an object. The position of the articulation points is derived by observing the movements of the parts of an articulated object. Integrating the articulation point into the optimization process of the spring system leads to improved tracking results in videos with big transformations and occlusions. A weakness of this approach is that it cannot deal with big rotation during occlusions. Therefore, we plan to consider higher level knowledge like spatio-temporal continuity to observe the occluded part reappearing around the borders of the visible occluding object. Another open issue is dealing with scaling and perspective changes. Future work is also to cope with pose variations and the resulting changes in the features representing the object.

References

1. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. Transactions on Computers **22** (January 1973) 67–92
2. Felzenszwalb, P.F.: Pictorial structures for object recognition. IJCV **61** (2005) 55–79
3. Ramanan, D., Forsyth, D.: Finding and tracking people from the bottom up. Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on **2** (June 2003) II–467–II–474 vol.2
4. Mauthner, T., Donoser, M., Bischof, H.: Robust tracking of spatial related components. In: ICPR, IEEE (December 2008) 1–4
5. Artner, N., Mármol, S.B.L., Beleznai, C., Kropatsch, W.G.: Kernel-based tracking using spatial structure. In: 32nd Workshop of the AAPR, OCG (May 2008) 103–114
6. Artner, N.M., Ion, A., Kropatsch, W.G.: Tracking articulated objects using structure (accepted). In: Computer Vision Winter Workshop 2009, PRIP, Vienna University of Technology, Austria (February 2009)
7. Gavrila, D.M.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding **73**(1) (January 1999) 82–980
8. Moeslund, T.B., Hilton, A., Krger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding **104**(2–3) (2006) 90–126
9. Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. Computer Vision and Image Understanding **73**(3) (March 1999) 428–440
10. Aggarwal, J.K., Cai, Q., Liao, W., Sabata, B.: Articulated and elastic non-rigid motion: A review. In: IEEE Workshop on Motion of Non-Rigid and Articulated Objects. (1994) 2–14
11. Badler, N.I., Smoliar, S.W.: Digital representations of human movement. ACM Comput. Surv. **11**(1) (1979) 19–38
12. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. Image and Vision Computing **22**(10) (September 2004) 761–767
13. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. PAMI **25**(5) (2003) 564–575
14. Mármol, S.B.L., Artner, N.M., Ion, A., Kropatsch, W.G., Beleznai, C.: Video object segmentation using graphs. In: 13th Iberoamerican Congress on Pattern Recognition, Springer (September 2008) 733 –740