Rigid Part Decomposition in a Graph Pyramid*

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

[1] AIT Austrian Institute of Technology, Vienna, Austria
nicole.artner@ait.ac.at
[2] PRIP, Vienna University of Technology, Austria
{ion,krw}@prip.tuwien.ac.at

Abstract. This paper presents an approach to extract the rigid parts of an observed articulated object. First, a spatio-temporal filtering in a video selects interest points that correspond to rigid parts. This selection is driven by the spatial relationships and the movement of the interest points. Then, a graph pyramid is built, guided by the orientation changes of the object parts in the scene. This leads to a decomposition of the scene into its rigid parts. Each vertex in the top level of the pyramid represents one rigid part in the scene.

1 Introduction

Tracking articulated objects and their rigid parts is an important and challenging task in Computer Vision. There is a vast amount of work in this field as can be seen in the surveys [1,2,3]. Possible applications are the analysis of human motion for action recognition, motion based diagnosis and identification, motion capture for 3D animation and human computer interfaces.

The first step in tracking articulated objects is the initialization. This step is important, because it can strongly influence the success of the tracking method. There are three possibilities for the initialization: (1) manually by the user, (2) solving the task as a recognition problem with the help of a training set [4], and (3) employing a segmentation method.

This paper presents an approach to segment the rigid parts of articulated objects from a video (third category). It is related to the concept of *video object segmentation* (VOS), where the task is to separate foreground from background in an image sequence. VOS methods can be divided into two categories [5]:

(1) Two-frame motion/object segmentation: Altunbasak et al. present in [6] a combination of pixel-based and region-based segmentation methods. Their goal is to obtain the best possible segmentation results on a variety of image sequences. Castagno et al. [7] describe a system for interactive video segmentation. An important key feature of the system is the distinction between two levels of segmentation: (1) regions and (2) object segmentation. Chen et al. [8] propose an

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

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 758–765, 2009.

approach to segment highly articulated objects by employing weak-prior random forests. The random forests are used to derive the prior probabilities of the object configuration for an input frame.

(2) Multi-frame spatio-temporal segmentation/tracking: Celasun et al. [5] present VOS based on 2D meshes. Tekalp et al. [9] describe 2D mesh-based modeling of video objects as a compact representation of motion and shape for interactive video manipulation, compression, and indexing. Li et al. [10] propose to use affine motion models to estimate the motion of homogeneous regions.

There is related work explicitly dealing with the segmentation of articulated objects (e.g. Chen et al. [8]), but the result of these approaches is still only a separation of foreground and background. To initialize tracking methods for articulated object parts, it would be convenient to have a decomposition of the articulated foreground object into its rigid parts (e.g. for the method in [11]).

In this paper, we achieve the decomposition of the rigid parts of an articulated object. The scene is observed and analyzed over time. Depending on the spatial relationships and movements in the scene the input for the grouping process is selected. The grouping itself is done in a graph pyramid and is controlled by the orientation variation resulting out of the articulated movement of the object parts in the scene. This approach is a continuation of the work in [12].

The paper is organized as follows: Sec. 2 recalls graph pyramids. In Sec. 3 the spatio-temporal filtering is explained and Sec. 4 describes how the rigid parts are identify. Sec. 5 presents experiments and in Sec. 6 we give conclusions.

2 Irregular Graph Pyramids

A *region adjacency graph* (RAG), encodes the adjacency of regions in a partition. A vertex is associated to each region, vertices of neighboring regions are connected by an edge. Classical RAGs do not contain any self-loops or parallel edges. An *extended region adjacency graph* (eRAG) is a RAG that contains the so-called *pseudo edges*, which are self-loops and parallel edges used to encode neighborhood relations to a cell completely enclosed by one or more other cells [13]. The *dual* graph of an eRAG G is called *boundary graph* (BG) and is denoted by $\bar{G}$ (G is said to be the *primal* graph of $\bar{G}$). The edges of $\bar{G}$ represent the boundaries of the regions encoded by G, and the vertices of $\bar{G}$ represent points where boundary segments meet. G and $\bar{G}$ are planar graphs. There is a one-to-one correspondence between the edges of G and the edges of $\bar{G}$, which also induces a one-to-one correspondence between the vertices of G and the 2D cells (denoted by *faces*[1]) of $\bar{G}$. The dual of $\bar{G}$ is again G. The following operations are equivalent: edge contraction in G with edge removal in $\bar{G}$, and edge removal in G with edge contraction in $\bar{G}$.

A (dual) irregular graph pyramid [13,14] is a stack of successively reduced planar graphs $P = \{(G_0, \bar{G}_0), \ldots, (G_n, \bar{G}_n)\}$. Each level $(G_k, \bar{G}_k), 0 < k \leq n$

[1] Not to be confused with the vertices of the dual of a RAG (sometimes also denoted by the term *faces*).

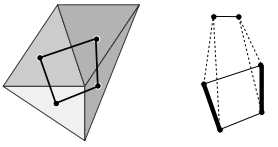

Fig. 1. Left: triangulation and associated adjacency graph. Right: contraction of two edges (thick) in a two level pyramid.

is obtained by first contracting edges in G_{k-1} (removal in $\bar{G}_{k-1}$), if their end vertices have the same label (regions should be merged), and then removing edges in G_{k-1} (contraction in $\bar{G}_{k-1}$) to simplify the structure. The contracted and removed edges are said to be *contracted* or *removed* in $(G_{k-1}, \bar{G}_{k-1})$. In each G_{k-1} and $\bar{G}_{k-1}$, contracted edges form trees called *contraction kernels*. One vertex of each contraction kernel is called a *surviving vertex* and is considered to have been 'survived' to $(G_k, \bar{G}_k)$. The vertices of a contraction kernel in level $k-1$ form the *reduction window* $W(v)$ of the respective surviving vertex v in level k. The *receptive field* $F(v)$ of v is the (connected) set of vertices from level 0 that have been 'merged' to v over levels $0 \dots k$.

For the sake of simplicity, the rest of the paper will only use the adjacency graph G, but for correctly encoding the topology, both G and $\bar{G}$ have to be maintained. Figure 1 shows an example triangulation and pyramid.

3 Spatio-temporal Filtering

The aim of the spatio-temporal filtering is to define the input for the grouping process. As mentioned in Sec. 1, the filtering is carried out by observing the scene in sequence of frames. This observation is realized by tracking interest points.

The filtering focuses on the spatial relationships of the interest points over time. A planar, triangulated graph G is used as representation for the filtering. The vertices V of the graph are the interest points and the edges E, which encode the spatial relationships, are inserted with a Delaunay triangulation [15].

As the aim is to find rigid parts of articulated objects, the task of the filtering is to select interest points corresponding to rigid parts. To identify these points, the changes of the edge lengths in the triangulated graph over time is considered. A triangle is *potentially rigid* for the decomposition process if its edges lengths do not vary remarkably in the observation period. In every frame of the video sequence the edge lengths $||e||$ can be calculated. To decide which triangles are *potentially rigid* the *edge length variation* is determined.

Definition 1. *The* ***edge length variation*** *of an edge is the difference between the minimum and the maximum length of the edge in the observation period.*

A triangle is labeled as *potentially rigid* if the edge length variations of all three edges are beneath a certain threshold ϵ. This threshold is necessary, because noise, discretization, and small imprecisions in the localization ability of the tracker[2] can affect the outputted positions of the interest points.

[2] e.g. the detection vs. localization problem in edge detection.

The result of the spatio-temporal filtering is a triangulation, where each triangle is labeled *potentially rigid* or *not rigid* (see Sec. 5 Fig. 2(b)). The *potentially rigid* triangles are then passed on as input to the grouping process (see Sec. 4).

4 Rigid Part Decomposition

The task of the rigid part decomposition is to split the *potentially rigid* triangles from the spatio-temporal filtering into groups of triangles, each describing one rigid part.

As this paper focuses on articulated objects, the triangles describe a locally deformable object, which follows a globally articulated motion. Due to the local deformation freedom the *edge length variation* of triangles belonging to a rigid part and those connecting such parts might not differ to much. E.g. even though the bones of a human perform an articulated motion, the flesh and skin are elastic and thus a smooth and continuous deformation can be observed in the triangles going from the lower arm to the upper arm and then to the torso. Another aspect is that if the region around an articulation point is densely sampled (tracked by many points), the *edge length variation* resulting out of a rotation of e.g. 90 degrees is going to be insignificant. For all these reasons, the *edge length variation* is not a sufficient property for the decomposition task.

The idea is to group triangles into rigid parts where all the triangles inside a group have a similar *orientation variation* over the whole video, and the average orientation variation of the triangles in two neighboring groups differs. This problem is similar to the single image segmentation problem, where the results should be regions with homogeneous color/texture (small internal contrast) neighbored to regions that look very different (high external contrast).

Definition 2. *The* ***orientation variation*** *over time is a 1D signal that encodes at each time step (frame of the input video) the accumulated orientation change relative to the orientation at the beginning of the video.*

E.g. turning around the axis once will give a value of 360° degrees, and turning twice will give 720°, not 0°. The direction of rotation is encoded by the sign: counter clockwise (CCW) is positive, and clockwise (CW) is negative, e.g if turning 30° CCW, then 15° CCW, and then 28° CW, the computed variations will be 30°, 45° = 30°+15°, 17° = 45°-28°.

Definition 3. *The* ***orientation variation of a triangle*** *is the 1D signal obtained by taking the average of the 1D signals of the three edges of the triangle.*

Using an irregular pyramid for the grouping task has the advantage that its structure adapts to the data. Also, using a hierarchy reduces the complexity of the grouping (global decisions become local ones), and the produced description contains information that can be used to cope with complexity in a coarse-to-fine tracking approach.

Alg. 1 creates a graph pyramid in which each top level vertex identifies a detected rigid part. The receptive field of these vertices identifies the triangles

Algorithm 1. *BuildPyr*(T): Group triangles into rigid parts.

Input: potentially rigid triangles T (see Section 3)

1: $G_0 = (V_0, E_0)$
   /*$V_0 = T$, *and* $(v, w) \in E_0 \iff$ *the corresponding triangles share an edge**/
2: $k = 0$
3: **repeat**
4: /**select edges to contract**/
   $K = \emptyset$
   $\forall v \in G_k$ **do** $K \leftarrow K \cup \arg\min_{(v,w)\in G_k}\{X(v, w)\}$
5: /**filter edges based on internal/external difference**/
   $\forall (v, w) \in K$, **if** $X(v, w) > I'(v, w)$ **then** $K \leftarrow K \setminus \{(v, w)\}$
6: **if** $K \neq \emptyset$ **then** break K into trees of radius 1
7: **if** $K \neq \emptyset$ **then** $G_{k+1} \leftarrow$ contract edges K in G_k and simplify
8: $k \leftarrow k + 1$
9: **until** $K = \emptyset$

Output: Graph pyramid $P = \{G_0, \ldots, G_{k-1}\}$.

that the respective part consists of. In the base level, one vertex is associated to each *potentially rigid* triangle. Two vertices are connected by an edge if the respective triangles share a common edge. Edges to be contracted are selected from the edges proposed by the Minimum Spanning Tree Alg. by Boruvka [16] (Line 4). The **external difference** $X(v, u)$ between two vertices is:

$$X(v, w) = \max(|V(v) - V(w)|) \tag{1}$$

where $V(v), V(w)$ are the 1D signals associated to v respectively w. They encode the average of the orientation variation of the triangles in the receptive fields. For the vertices in the base level G_0 they are the orientation variations of the corresponding triangles. For a vertex in a higher level they can be computed as:

$$V(v) = \frac{\sum_{u\in W(v)} |F(u)| \cdot V(u)}{\sum_{p\in W(v)} |F(p)|} \tag{2}$$

where $|F(v)|$ is the size of $F(v)$ and can be propagated up in the pyramid. The **internal difference** $I(v)$ of a vertex at level $k > 0$ is:

$$I(v) = \max(\max\{I(u)\}, \max\{X(p_i, p_j)\}) \tag{3}$$

where $u \in W(v)$ and $p_i, p_j \in W(v)$ s.t. p_i, p_j are connected by an edge. For the vertices in the base level $I(v) = 0$. The value $I'(v, w)$ is defined as:

$$I'(v, w) = \min(I(v) + \frac{\beta}{|F(v)|}, I(w) + \frac{\beta}{|F(w)|}) \tag{4}$$

where β is a parameter of the method that allows regions to start forming in the base level where $I(v) = 0$ for all vertices. For a discussion about β in the context of image segmentation see [17,18].

Line 6 of Alg. 1 keeps the contraction operations local (optimal for parallel processing) and avoids contracting the whole graph in a single level. It excludes edges from K to create trees of radius 1 for the current contraction. The excluded edges will be selected again in the next level. In [19] three such methods, MIES, MIS, and D3P (used in our experiments) are described. The described concept is related to the image segmentation method in [17], with the difference that:

1. we do not start form a pixel image, but from a triangulation;
2. our features are not color values but 1D signals of orientation variation;
3. and most important, we do not assign the edge weights based on the weights in the level below, but recompute them to reflect the difference between the average variation of the triangles in the two neighboring regions.

The difference at 3. has the effect that a long chain of regions that differ by a constant, small difference, will not be merged to create a single region.

5 Experiments

The moving foreground objects in the experiments are humans, but the presented approach is applicable to any arbitrary object. The Kanade-Lucas-Tomasi tracker [20] is used to track corner points to supply the necessary observations.

Both video sequences show a person undergoing articulated motion in the image plane. In Fig. 2 the result of the triangulation and the following spatio-temporal filtering are visualized. Fig. 3 presents the decomposition results. The result for sequence 1 is ideal, meaning that each rigid part and the background are one vertex in the top level of the graph pyramid. For sequence 2 the right lower arm is represented by two top vertices. Also, two additional regions, one corresponding to a part of the hair, and one connecting the left upper arm with the background have also been produced as rigid parts. The reason for this are

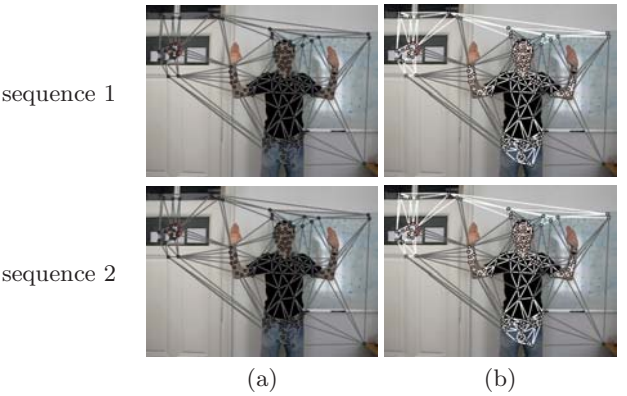

Fig. 2. Triangulation (a) without (b) with labeling. White: potentially rigid, grey: not rigid.

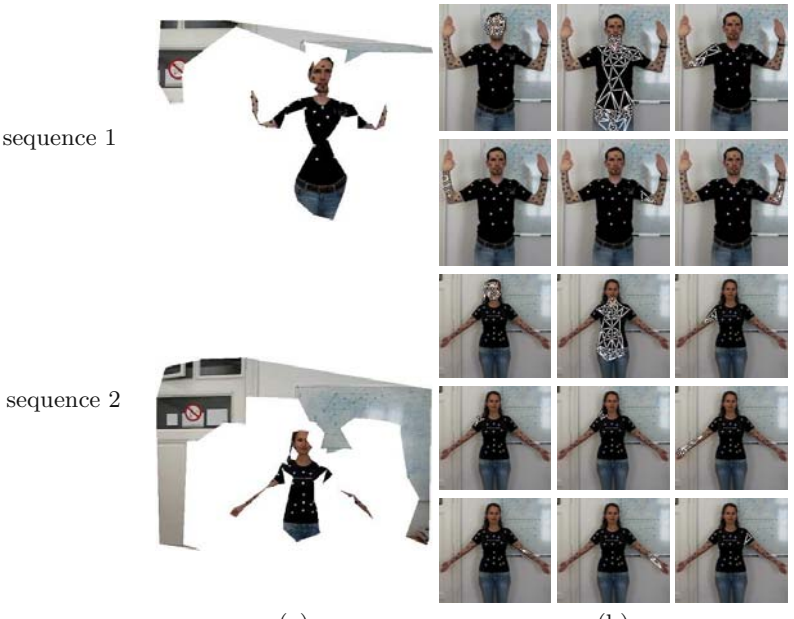

Fig. 3. (a) pixels belonging to a rigid part. (b) graphs for rigid parts of the foreground.

the difficulties mentioned at the beginning of Sec. 4 and the fact that the labeling into *potentially rigid* and *not rigid* has to allow certain variation (see Sec. 3). The torso is connected with the base of the chin in both sequences because during tracking the features at the base of the chin slide when the head is tilted and remain in the same position in the image, creating a *potentially rigid* triangle.

6 Conclusion

This paper presented a graph-based approach to decompose the rigid parts of articulated objects. First a spatio-temporal filtering is performed, where the spatial relationships of the interest points over time are analyzed and a triangulation is produced, with triangles labeled as *potentially rigid* and *not rigid*. The *potentially rigid* triangles are given as input to a grouping process that creates a graph pyramid s.t. in the top level each vertex represents a rigid part in the scene. The orientation variation of the input triangles controls the building process of the pyramid and is used to compute the similarity between two groups of triangles. The success of the presented approach depends on the quality and robustness of the tracking results. The presented approach fails if there are remarkable perspective changes or scaling. This is one of the open issues we are planing to deal with in the future. Additionally, we are going to find the articulation points connecting the rigid parts of the foreground objects.

References

1. Gavrila, D.M.: The visual analysis of human movement: A survey. CVIU 73(1), 82–980 (1999)
2. Moeslund, T.B., Hilton, A., Krger, V.: A survey of advances in vision-based human motion capture and analysis. CVIU 104(2–3), 90–126 (2006)
3. Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. CVIU 73(3), 428–440 (1999)
4. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV 61(1), 55–79 (2005)
5. Celasun, I., Tekalp, A.M., Gokcetekin, M.H., Harmanci, D.M.: 2-d mesh-based video object segmentation and tracking with occlusion resolution. Signal Processing: Image Communication 16(10), 949–962 (2001)
6. Altunbasak, Y., Eren, P.E., Tekalp, A.M.: Region-based parametric motion segmentation using color information. Graphical Models and Image Processing 60(1), 13–23 (1998)
7. Castagno, R., Ebrahimi, T., Kunt, M.: Video segmentation based on multiple features for interactive multimedia applications. Circuits and Systems for Video Technology 8(5), 562–571 (1998)
8. Chen, H.T., Liu, T.L., Fuh, C.S.: Segmenting highly articulated video objects with weak-prior random forests. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 373–385. Springer, Heidelberg (2006)
9. Tekalp, A., Van Beek, P., Toklu, C., Gunsel, B.: Two-dimensional mesh-based visual-object representation for interactive synthetic/natural digital video. Proceedings of the IEEE 86(6), 1029–1051 (1998)
10. Li, H., Lin, W., Tye, B., Ong, E., Ko, C.: Object segmentation with affine motion similarity measure. Multimedia and Expo, 841–844 (2001)
11. Artner, N., Ion, A., Kropatsch, W.G.: Tracking objects beyond rigid motion. In: Workshop on Graph-based Representations in PR, May 2009. Springer, Heidelberg (2009)
12. 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, September 2008, pp. 733–740. Springer, Heidelberg (2008)
13. Kropatsch, W.G.: Building irregular pyramids by dual graph contraction. Vision, Image and Signal Processing 142(6), 366–374 (1995)
14. Kropatsch, W.G., Haxhimusa, Y., Pizlo, Z., Langs, G.: Vision pyramids that do not grow too high. PRL 26(3), 319–337 (2005)
15. Tuceryan, M., Chorzempa, T.: Relative sensitivity of a family of closest-point graphs in computer vision applications. Pattern Recognition 24(5), 361–373 (1991)
16. Nesetril, J., Milková, E., Nesetrilová, H.: Otakar boruvka on minimum spanning tree problem translation of both the 1926 papers, comments, history. Discrete Mathematics 233(1-3), 3–36 (2001)
17. Haxhimusa, Y., Kropatsch, W.G.: Segmentation graph hierarchies. In: Fred, A., Caelli, T.M., Duin, R.P.W., Campilho, A.C., de Ridder, D. (eds.) SSPR&SPR 2004. LNCS, vol. 3138, pp. 343–351. Springer, Heidelberg (2004)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. IJCV 59(2), 167–181 (2004)
19. Kropatsch, W.G., Haxhimusa, Y., Ion, A.: Multiresolution Image Segmentations in Graph Pyramids. In: Applied Graph Theory in Computer Vision and Pattern Recognition. SCI, vol. 52, pp. 3–42. Springer, Heidelberg (2007)
20. Birchfeld, S.: Klt: An implementation of the kanade-lucas-tomasi feature tracker (March 2008), `http://www.ces.clemson.edu/~stb/klt/`