MVA'98 IAPR Workshop on Machine Vision Applications, Nov. 17-19, 1998, Makuhari, Chiba, Japan

13—14

Matching of 3D Graphs for Human Motion Analysis

Carlos Yániz Jairo Rocha
Department of Mathematics and Computer Science *
University of the Balearic Islands, Spain

Abstract

Recognition of a human and its parts in a scene is described in this paper. Two views from two cameras in ortogonal position are used. A 3D regular region graph is defined to gather possible poses based on the two 2D regular (approximate uniform thickness) regions, one for each view, at a given frame. The 3D graph edges represent regular volumes such that its projections appear in the silhouettes, and the nodes represent connections of these volumes. Some of these are "phantom" volumes, that is protrusions not corresponding to real limps. The algorithm we present here finds the pose of the human subject by using criteria of size, geometry and position to get a good distribution of the human parts considering temporal coherence and avoiding the phantom volumes. The method performs an optimal search of the pose that minimizes a cost function that covers the criteria. Experimental results and error analysis, and an application to automatic computer animation are shown.

1 Introduction

It is a common practice in human motion analysis to fit *a priori* shape models into image data [AC97]. In our approach, the data is first interpreted into 3D volumes, as in [BLZ98], and then matched to a 3D model. Unlike [BLZ98], our system uses a general model whose size does not have to coincide with the subject and only two cameras.

Our system finds this description of human motion in three steps, starting with silhouettes. First, the boundary of the smooth body surface is used to recover candidate human parts, in each view. Then, 3D part positions hypotheses are carried out, and finally, constrained matching with a model gives the best interpretation. The process is able to interpret human silhouette sequences and to give an approximate description of part movements.

The recognition module finds the best interpretation under a combination of criteria for geometrical relationships within and between model parts, avoiding "phantom" volumes while interpreting the data as much as possible.

Section 2 defines the 3D regular part graph concept. Section 3 describes the matching with a human model. Section 4 shows some results and we present in Section 5 some conclusions and current work.

2 The 3D Skeleton Graph

We assume that human parts (head, trunk, arms, legs) are elongated and approximately conical. In other words, each part is symmetrical with respect to an axis. The human parts considered are explained in Section 3. When parts are put together in space, they create connection regions which are not as regular as the original parts. There are also regions which are still regular because they are not affected by the connections, so they preserve their symmetry. We attempt to recover these regular regions, and also the connection regions.

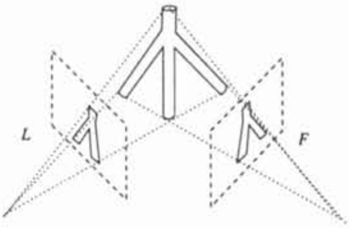

Figure 1: 3D Volume intersection.

In each frame, we can find the projections of the symmetrical volumes as (almost) symmetrical regions, that we call regular regions. These regular regions are found using an algorithm which decomposes the shape into regular and singular regions [SM93]. Regular regions correspond to elongated regions limited by opposite contour pieces. These are quasi-parallel and near to each other. Therefore, arms, legs, head and trunk are easily identified as regular regions. Regions which are not regular are considered singular. They are parts of the shape which connect or end up the regular regions, such as, shoulders, heap, regions where arms cross each other, feet tips, etc. Thus, the human shape can be represented as a graph, in which nodes correspond to singular regions and edges correspond to regular regions.

As shown in Figure 1, it is possible to construct

*Address: E-07071 Palma de Mallorca, e-mail: carlos@anim.uib.es, jairo@ipc4.uib.es a 3D graph of volumes and connections from the two 2D graphs of regular regions, using the camera calibration parameters. The formal and practical details are explained elsewhere [YRP98].

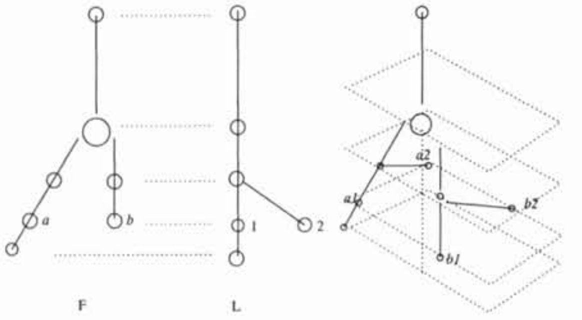

Figure 2: Scheme of the 3D skeleton of two graphs.

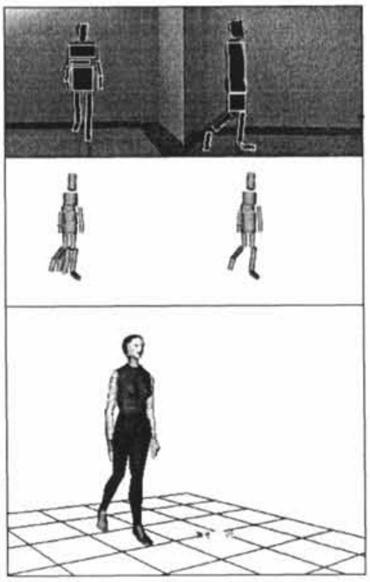

Figure 3: Motion recovery of frame 3. From top to bottom: (a) original frames with regular regions; (b) 3D graph and matched subgraph; (c) result driving *Nancy.*

Figure 3(b) displays a 3D skeleton graph generated during the analysis process. A 3D graph edge is represented as a cylinder, where its radius depends on the observed widths of the 2D edges. On the left part of the figure, there are four legs because those are all the possible positions generated by the views. See also the scheme in Figure 2, where point a could be at point 1 or 2, in the other view. For simplicity, nodes drawn at the same height are assumed to be in the same epipolar plane.

In Figure 1, notice how the direct reconstruction of the 3-D shape of the body by the volume intersection of the projection rays can produce "phantom" volumes that do not correspond to real limps.

Any subgraph of the 3D graph such that its projections are the original view graphs is a valid 3D interpretation of the views. Valid interpretations should also respect the human model structure and mechanics. Therefore, a subgraph which matches a human model and projects onto the views gives a 3D pose for the views. When this process is applied to each frame, a sequence of 3D graphs is generated.

3 Optimal Fitting of a Model

Each 3D graph contains segments for all possible positions (except for segments facing the camera) of 2D skeleton graphs. To choose one interpretation, knowledge of the human structure is used throughout a graph model which defines the parts of a person, and some quantitative relations among their lengths and widths. The human model considered is a graph which consists of 17 parts connected accordingly.

The model is structurally matched with the 3D graph and the numerical relationships are used to constrain the possible matchings. There are three types of constraints: aspect constraints (relate maximal length to the width of a part), width constraints among parts, and maximal length with respect to a previously defined part length (in our case, the trunk). All constraints are very relaxed to cover the sizes of most humans but avoid considering matchings which are not possible. Hence, the system is designed to recognize any person.

The process matches groups (paths) of segments in the 3D graph with each model part, matching as much as possible and rejecting the matching not satisfying the constraints.

The grouping into segment paths is required since several segments in the graph can form a single part, and vice versa, several model parts can appear as a single segment. A path generation process finds approximate collinear paths (sequences of adjacent edges) which are as long as possible and are called *maximal paths.* Next, for each maximal path, it generates *all* possible sub-paths, which will be called simply *paths* of the graph, since no other graph paths will be considered in this paper. This is a procedure which allows the future grouping of segments which actually belong to the same human part. Two paths which do not share segments are called *compatible.*

Let P and P_H be path sets. Given $G(V, E, P)$ and $H(V_H, E_H, P_H)$, a 3D graph and a model graph, respectively, an *interpretation*

$$\phi : P_H \to P$$

is a partial one-to-one mapping such that

- the domain and the range consist of compatible paths, and
- it respects the path relations, so that if two paths in H are related in certain locations, their images in G are also related in the same locations.

If ϕ is an interpretation of H in G, we also say that ϕ is a *shape instance* of H in G.

The conditions ensure that paths are mapped into paths at the same relative positions. Therefore, an interpretation has the same flavour of *homeomorphic* subgraphs [Har72], but it is a more relaxed concept.

The search is initialized with the matching of all possible paths in G with the trunk. Only the paths that respect the aspect constraints are considered and from those the largest 50 are chosen, so the interpretation is biased to find a human as big as possible in the image. Each of those single matchings ϕ defines a relative *scale*, that is fix for future extensions of ϕ, and represented by S_ϕ.

Assume we have a partial matching between H and G. We select the next part in H to be matched and calculate the possible paths in G that can be matched to it, including none. There are some alternative cases considered when a part connection is not visible but appears elsewhere, for example, when the shoulder connection is not visible but a piece of arm is visible coming out from the middle of the trunk.

The search tree is expanded with the possible matching of the next part and a cost function allows to order these partial matchings.

3.1 Evaluation Function

The cost of a matching is defined so it allows us to compare different transformations that map the input graph into the model. In the same order of Definition 4, the cost includes costs for geometrical difference between a model part and its image, cost of movement relative to the previous frame, and for segments in each view which are projections of several matched segments (to penalize the use of phantom volumes), length of parts non matched in the model, and length of segments non matched in the input.

The *initial* interpretation is the the empty one. An interpretation is *final* if it cannot be completed more while respecting the interpretation conditions, i.e., it is maximal.

The evaluation function $f(\phi)$ for a partial interpretation ϕ is the sum of two functions $g(\phi)$ and $h(\phi)$. $g(\phi)$ is the cost accumulated for each pair matched by ϕ, in other words, the part of the cost that can be calculated precisely. In contrast, $h(\phi)$ is an estimation of the minimum cost needed to complete ϕ, e.g., to reach a final interpretation from ϕ.

Let ϕ be an interpretation. We say that a segment $s \in \mathsf{dom}\phi$ if there is a path in the domain of ϕ that contains s, and, similarly, for the range we use the the expression $s \in \mathsf{ran}\phi$. To measure the multiple projection, let e be a (2D) edge of a graph view, Lateral(L) or Frontal(F). Define $3D_\phi(e)$ as the set of 3D segments whose projections on the view correspond to e and that are in the range of ϕ, e.g., all the matched 3D segments whose projection is the same e.

Definition 1 (costs) *The cost for multiple projections of a 2D segment e is*

$$mult_\phi(e) = max(0, length(e)\,(|3D_\phi(e)| - 1)).$$

The value of the geometric cost of matching α to β is:

$$geo(\alpha,\beta) = |length(\beta) - S_\phi\, length(\alpha)|.$$

The value of the movement cost $mov(\alpha,\beta)$ is the distance from β to the part matched in the previous frame to α.

Definition 2 (accumulated cost) *The accumulated cost of a interpretation ϕ is*

$$g(\phi) = \sum_{\phi(\alpha)=\beta} geo(\alpha,\beta) + mov(\alpha,\beta) \;+\; \sum_{e\in L\cup F} mult_\phi(e).$$

Definition 3 (heuristic function) *The value of the heuristic function of an interpretation ϕ is*

$$h(\phi) = \sum_{s\notin \mathsf{ran}\phi} length(s) \;-\; S_\phi \sum_{p\notin \mathsf{dom}\phi} length(p).$$

Definition 4 (total cost) *The cost of a final interpretation Φ is*

$$f(\Phi) = g(\Phi) + S_\Phi \sum_{p\notin \mathsf{dom}\Phi} length(p) + \sum_{s\notin \mathsf{ran}\Phi} length(s)$$

Algorithm A^* is effective in finding the optimal solution provided that the heuristic function is *admissible* [Pea84]. The following proposition states the admissibility but the proof is omitted for paper limitations.

Proposition 1 (admissibility) *If $geo(\alpha,\beta) \geq length(\beta) - S_\phi\, length(\alpha)$ then h is admissible, e.g., h never over-estimates the real cost of the best completion of ϕ.*

The best matching in each frame is chosen. The matchings recognize the human parts in the sequence.

4 Experimental Results

Several sequences were filmed by two cameras. We show here the results on a walking man sequence. Each 288 × 216 frame view was filtered to remove background and noise and converted into a bilevel image.

2D skeletons from each view were converted into 3D graphs, which were matched against a human model. Angles between the recognized parts are given to an avatar, so that the results are displayed with the aid of a virtual humanoid, named *Nancy*, a VRML humanoid. The experiment was performed using the C programming language on a SPARCstation IPC and it takes 1.2 cpu seconds per frame.

We only show the first 12 frames separated approximately $\frac{1}{8}$ s. from each other. See Figure 4.

From the figures, a qualitative evaluation can be carried out. Most of the part orientations are correct and the humanoid seems to walk normally. The pose of legs is recovered better than arms' because they are less affected by occlusion in this sequence. We plan to use optical flow to guide the matching more precisly and obtain more robustness.

5 Conclusions and Future Work

An analysis vision system for human motion detection is presented in this paper. The global system analyses the two-view input image sequence to discover the 3D spatial information in each image frame. The subject could be *any* human wearing tight clothes, and the type of motion is *any* in which the users' appendages are most of the time visible on the silhouette. No other assumptions are made with respect to the subject or the type of motion. Currently, our method is limited by the fact that we assume the person is moving in front of a simple background.

We were able to integrate different source of knowledge in order to interpret sequences. We currently keep studying different ways of combining relaxed a priori shape information with the displacements and the regions detected on each view.

Future research includes the use of colour to segment arbitrary images and manage complex backgrounds.

Acknowledgements. We would like to thank Ms. Cindy Reed-Ballreich from *3Name3D Inc.* for kindly making the humanoid Nancy available to us. This work is supported in part by the grant CICYT TIC96-0528.

References

[AC97] J K. Aggarwal and Q. Cai. Human motion analysis: A Review. In *Proceedings of IEEE Non-Rigid and Articulated Motion Workshop*, pages 90–103, Puerto Rico, USA, 1997.

[BLZ98] A. Bottino, A. Laurentini, and P. Zuccone. Toward non-intrusive motion capture. *Computer Vision-ACCV'98*, 1352:416–423, 1998.

[Har72] F. Harary. *Graph Theory.* Addison-Wesley, 1972.

[Pea84] J. Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving.* Addison-Wesley, Paris, France, 1984.

[SM93] T. Suzuki and S. Mori. Structural description of line images by the cross section sequence graph. *Int. J. PRAI*, 7(5):1055–1076, 1993.

[YRP98] C. Yaniz, J. Rocha, and F. Perales. 3D Regular region graph for reconstruction of human motion. In *Proceedings of ECCV 98 Workshop on Perception of Human Action*, Freiburg, Germany, June 1998.

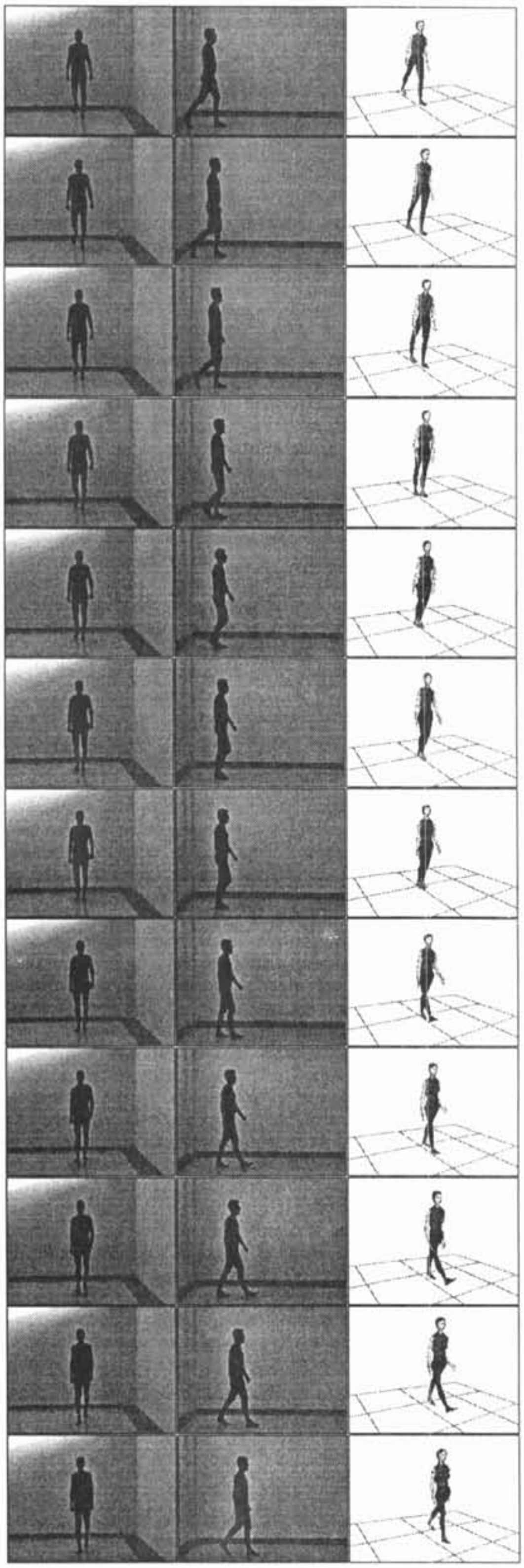

Figure 4: Recognition result of 12 frames driving *Nancy*. From top to bottom: original frontal view sequence, lateral sequence, and Nancy's one view sequence.