Boundary Fragment Matching and Articulated Pose Under Occlusion

Nicholas R. Howe

Smith College, Northampton, Massachusetts, USA

Abstract. Silhouette recognition can reconstruct the three-dimensional pose of a human subject in monocular video so long as the camera's view remains unoccluded by other objects. This paper develops a shape representation that can describe and compare partial shapes, extending the silhouette recognition technique to apply to video with occlusions. The new method operates without human intervention, and experiments demonstrate that it can reconstruct accurate three-dimensional articulated pose tracks from single-camera walking video despite occlusion of one-third to one-half of the subject.

1 Introduction

Intense research interest has focused lately on the recovery of articulated pose information from monocular video [11, 19, 20]. Despite great progress, current methods commonly assume that subjects remain fully visible apart from self-occlusion of one body part by another. Yet outside of controlled studio conditions, extraneous objects can often block a camera's view either momentarily or for an extended period of time. This paper develops techniques to handle situations where some portion of a subject's body passes either behind a stationary object or out of the video frame.

Silhouette shape matching has already proven itself well suited for recovering articulated pose in a variety of applications [7, 14].[1] Previous work with silhouette-based pose recovery has assumed that full silhouettes are available, and employed similarity measures that require complete shape information. This paper develops a novel approach to shape comparison that can operate with either full or partial shape information. Under occlusion, the visible portion of the shape boundary can thus be extracted and used for pose recovery. Reconstructions based upon this technique can recover the articulated pose trace (3D joint positions over time) of a walking human subject undergoing either episodic or extended partial occlusion by stationary objects.

1.1 Related Work

Much research has looked at recovery of articulated three-dimensional pose from monocular video without external occlusion [11]. Recent work in this area in-

[1] Although the silhouette-to-pose relationship is many-to-one, enforcing temporal continuity can disambiguate the true pose.

cludes several alternatives to silhouette-based reconstruction [19, 20], as well as other related results that stop short of recovering full 3D pose [3, 15]. Another body of research has examined the problem of non-articulated object tracking under intermittent occlusion [4, 9]. The combination problem of single-camera articulated pose reconstruction with occlusion has received very little attention. One alternative to the method presented here would be to track the silhouettes using an occlusion-resistant deformable shape tracker [16] followed by ordinary silhouette recognition, but this would require a good prior model on the possible shapes of human silhouettes. Prior models on human motion can also guide tracking under occlusion [21]

A number of works have examined techniques specialized for partial shape matching; one recent paper gives an excellent survey [17]. The approach used herein resembles the B-spline technique of Salari and Balaji [18], but the use of the EMD embedding here improves on that work by allowing arbitrarily dense sampling of the shape boundary without increasing the complexity of the final representation. The method also somewhat resembles the curvature scale space [12, 13] and the fast correspondence of Adamek and O'Connor [1], but these do not address partial matching. The use of EMD embedding herein is adapted from work using the shape context and various other features.[5, 6].

2 Shape Matching: Sets of Boundary Fragments

This paper develops a measurement of shape similarity based upon matching many small, overlapping fragments of the shape boundary. Because each boundary fragment can be parameterized in only one dimension, the set of fragments is potentially simpler than other sets of localized descriptors such as the shape context [2], and also less affected by occlusions.

Simple metrics like Euclidean distance cannot properly compute similarity between sets. Fortunately, recent work provides a means of embedding sets of fragment descriptors in a high-dimensional metric space, such that the L_1 distance in the embedding space approximates the earth-movers distance (*EMD*) for the local shape or boundary descriptors [5]. Such an embedding will be referred to as an *EMD embedding*. The EMD corresponds to the sum error in the best global matching between the boundary fragments of the two images. The EMD embedding of a set of boundary fragments thus constitutes a practical representation for computing a meaningful similarity measure on binary shape images.

2.1 Extracting Boundary Fragments

To describe a binary image using boundary fragments, begin by representing the boundary as a sequence of points with roughly equal spacing. Extract multiple overlapping subsequences spaced uniformly along the boundary and of similar length, then express these in a uniform representation. The EMD embedding

transforms the set of fragment representations into a form offering better computational efficiency, as described in Section 2.2.

The specific application will determine the best choice of boundary segment lengt, based upon one or more heuristics. A fixed scaled length is best when the scale of the shape can be known; most pose tracking applications fall into this category once tracking has begun. Alternately, properties of the shape itself may be used to estimate a scale. For example, the fragment length may be set at some fraction of the total perimeter length.

Once identified, boundary fragments must be described in a concise numeric format. Suppose that a fragment is parameterized by s, where $s = 0$ at one end of the fragment and $s = 1$ at the other end, and $\tau(s)$ gives the tangent to the boundary at s. Sample $\tau(s)$ at uniform intervals, compute the discrete cosine transform (DCT), then zero the constant term and truncate high-order terms beyond k. The inverse DCT then yields a k-dimensional representation of the fragment shape that is rotation invariant and effectively low-pass filtered.

Rotational information can be restored to the descriptor if desired by including $\sin(\tau(0.5))$ and $\cos(\tau(0.5))$ as additional features. Using two features for rotation information avoids circular anomalies when comparing values such as 0 and 2π. Including rotation information makes sense when shapes can be oriented *a priori*, as for example in video where the the y axis aligns with gravity.

2.2 Embedding the Boundary Fragments

In order to efficiently compute the best matching between shapes, the EMD embedding takes the set of fragment descriptors and creates a high-dimensional vector describing the shape as a whole. Each component of the embedded vector covers a region of the fragment descriptor space, with the magnitude of the component depending upon the size of the region and on how many fragments lie within it. The full embedding algorithm is too involved to give here, but follows Grauman and Darrell [5], with several modifications. To ensure that new shapes can be embedded within the same framework as existing ones, the embedding covers a fixed region of the fragment descriptor space: $[-2\pi, 2\pi]$ for the sampled features, and $[-1, 1]$ for the rotation-dependent features. Furthermore, the hierarchical subdivision of the feature space into component regions stops after five levels.

2.3 Partial Shape Matching with the Boundary Fragments

Under occlusion, some portion of a shape is unobservable. Define a *partial shape* as the result of hiding some part of a binary shape image with a mask. Note that a partial shape is not equivalent to the smaller shape made by deleting the masked portion; instead, the masked portion is simply undefined. Instead of a closed contour, the boundary of a partial shape becomes one or more open curves.

Many traditional shape descriptors simply cannot be computed for shape fragments due to the undefined region. Those comprising sets of localized descriptors, such as boundary fragments and the shape context, can still compute descriptors for regions that do not overlap with the undefined area. However, the shape context runs into problems: because it is defined on a circular area, its descriptors easily overlap unknown areas even when centered on a visible point. Consider frames 55-60 in Figure 2, where the entire shape lies close to the occlusion, but there are nonetheless long boundaries visible. Without sufficient local descriptors, the shape context cannot match accurately and becomes inviable for occluded pose recovery.

Given a partial shape as described above, straightforward computation yields an EMD-embedded descriptor that incorporates all fully-defined boundary fragments. For matching purposes, the components of this descriptor should be normalized to sum up to the fraction α of contour segments fully visible, while descriptors of complete shapes are all normalized to sum to one. For pose recovery, Section 3 describes how to estimate α when it is unknown. Experiments on retrieval tasks show low sensitivity to errors in α: retrieval sets overlap by up to 80% for α values varying by as much as 30%.

2.4 Benchmark Experiments with Boundary Fragment Matching

Boundary fragments shows reasonable performance as a general shape matching tool on standard test sets. For example, on the MPEG7 CE-Shape-1 test set, boundary fragments score 68.01% on the standard "bullseye" criterion [10]. Experiments in the same framework show a similar result for the shape context (68.11%). Other work has reported better results for shape context when combined iteratively with geometric warping [2]. Presumably boundary fragment matching would also benefit from such a treatment, although the procedure is too slow for application to pose recovery.

3 Pose Recovery Under Occlusion

Monocular video provides only limited cues for reconstructing the full 3D pose of a subject. Silhouette recognition offers a simple yet effective way to apply background knowledge to the problem. Silhouettes observed in the video serve as keys to look up known 3D poses with similar silhouettes for further consideration [7]. The approach does have drawbacks; most commonly noted is the difficulty of identifying accurate silhouettes. Although this remains an area of research, current performance is adequate in some applications [8].

This paper addresses a different problem, occurring when part of a subject's body is occluded by stationary objects situated along the camera's line of sight, or by the frame boundary. In either case, it becomes impossible to determine the shape of the entire silhouette, and thus to retrieve the 3D pose directly.

3.1 Boundary Fragments for Silhouette Lookup

EMD-embedded boundary fragment matching provides the framework for silhouette lookup. Because other sources describe silhouette-based pose recovery in detail [7], only an outline appears here. After background stabilization (if necessary) and modeling, change detection yields a silhouette in each video frame [8]. Each silhouette becomes a query into a database of silhouettes with known 3D poses, acquired via motion capture. The most likely pose-silhouette pairs are retained and registered to the video frame. A temporal synchronization step then searches for the sequence of poses that simultaneously maximizes the similarity to the observed silhouettes while minimizing the energy of pose changes from frame to frame. Further postprocessing smooths the results and optimizes their fidelity to the observations. Boundary fragment matching provides a convenient mechanism for the silhouette lookup stage of the algorithm; other portions remain unchanged.

Because fragment matching handles both complete and partial shapes, the method applies easily to the partial shapes that arise during external occlusion. However, the algorithm must know what portions of the shape are occluded so that it can distinguish the real silhouette boundaries from the occlusion edges. The discussion below begins by assuming that an operator provides a manually created "occlusion map" identifying areas containing potential occluding objects (Figure 1). Section 3.3 addresses how to generate such maps automatically.

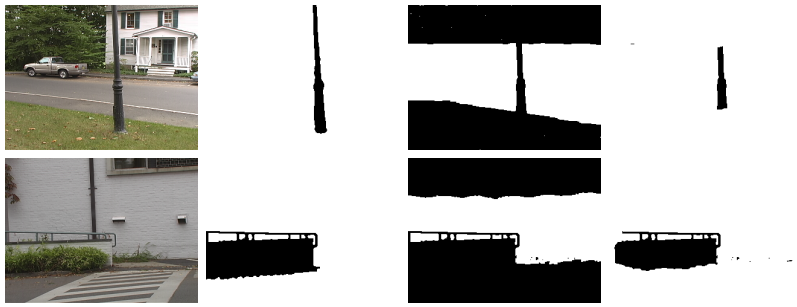

Fig. 1. Static background with manually provided and automatically generated occlusion maps for two sample videos. The automatic masks are more conservative in the visible areas they identify. Rightmost column shows confirmed occluded areas after algorithm has run.

Given an occlusion map, the lookup process first determines whether the observed silhouette touches any occluded areas. If it does not, then normal lookup proceeds. If occluded areas overlap the silhouette, then the visible portions of the silhouette generate a partial shape consisting of one or more boundary contours. The scale of the figure (for boundary section length) and the visibility α may be

estimated in most cases from neighboring frames with silhouettes already registered. In this case, the partial shape query returns candidate poses from the database just as a full image query would. Section 3.3 discusses how to bootstrap scale and visibility estimates for clips consisting entirely of occluded frames.

Registering a retrieved silhouette with its video frame becomes slightly more complicated when working with occlusion. One or more boundary contours will be visible in the frame. These must be matched to equivalent portions on the border of the retrieved silhouette. Phase matching between sequences of equally-space points extracted along both boundaries yields the desired correspondence. The registration scale and translation then minimize the Euclidean distance between the corresponding sequence points. Once the retrieved silhouettes are registered, the remainder of the computation proceeds as before.

3.2 Experiments with Occluded Video

The evaluation test set comprises two short videos involving significant occlusion. The first, *Pole*, shows a walking subject passing completely behind a lightpost. The second, *Ramp*, shows the subject walking up a handicapped access ramp. A low wall at the edge of the ramp obscures the view of the subject's legs in the latter half of this video, making pose recovery much more challenging.

The reconstruction of the *Pole* video entirely avoids any major errors. Occlusions by the post and the frame edges are handled gracefully. There is some small error in the arm positions and in the hip orientation, similar to those occurring in silhouette-based reconstruction without occlusion. Space constraints preclude reproducing the results here, as they are similar to those presented later. This clip shows boundary fragment matching easily handling transient external occlusions of this sort. Note that the figure can be tracked outside the frame boundary only so long as a sufficient amount of the boundary is visible; the outer limit for getting reasonable results seems to be about $\alpha = 0.3$.

The extended occlusion of the legs in the *Ramp* video makes pose reconstruction much more difficult in the second half of the clip. The system must infer what the legs are doing from the motions of the upper body. When the arms are visible this is somewhat easier, but there are points in the stride where the upper body shape appears more or less as an undifferentiated pillar. Despite this, the shape-fragment matching reconstructs the *Ramp* motion with only one significant error, a stutter-step near the very end of the clip. Errors near the beginning and end of a clip sometimes occur due to the lack of corroborating observations from neighboring frames on one side.

3.3 Fully Automatic Reconstruction

Manual occlusion maps are a crutch that preclude fully automatic operation. This section describes an algorithm to automatically determine the non-occluded areas, using no more information than the silhouette observations already employed for pose retrieval (as derived from background modeling and change detection). The composition of all the areas where the subject silhouette is observed

over time forms a map of known unoccluded areas. The complement of this set is the union of two pieces: true occlusion zones and areas of background that are indistinguishable from occlusion zones because the subject was never observed there.

Figure 1 shows for each test video the regions containing no observation of the subject. Although these masks cover much more area than the manually created occlusion masks, these masks may nevertheless function in the same role. Treating zero-silhouette regions as occlusion zones will disregard some valid silhouette boundaries that cannot be verified as real. A typical silhouette will now have undefined regions above the head and below the feet, because the subject was never observed in these areas. This increases the challenge of database retrieval: since occlusions cannot be distinguished from unidentified background, in practice all frames must use partial shapes for retrieval.

Atomatic occlusion maps may not be error-free, and the penalties for error are not symmetric. Marking a visible area as occluded merely makes retrieval slightly more difficult by reducing the length of boundary available as a query. This is generally much less serious than counting an occluded area as visible, which will usually introduce spurious boundaries that are more likely to interfere with both retrieval and registration. For example, the railing at the top of the wall in the *Ramp* clip is not always segmented properly in the silhouette of every frame. To prevent such problems, a special high-threshold foreground segmentation generates the occlusion map, biasing the result against mistakenly labeling occlusion zones as visible areas.

Estimating the parameter relating library scales to observed silhouette dimensions becomes more difficult without frames known to be unoccluded. This work adopts a heuristic approach, assuming that the silhouettes with maximal vertical extent are unoccluded or nearly so, and estimating scale based upon their height. Allowing slow (0.5%) changes in scale between neighboring frames causes each silhouette to impose a minimum scale on all other frames. Silhouettes whose vertical extent indicates a scale greater than the minimum imposed by all other frames are considered reliable indicators of the true scale. Interpolation gives the estimated scale of the remaining frames. While effective for the clips tested here, this heuristic is not universally reliable and might be less successful in some cases than a technique based upon boundary curvature or limb thickness.

Estimating the visibility α is also more difficult without known occlusion-free frames. The silhouette dimensions for the indicator frames described above can give a very rough estimate of α. Unoccluded silhouette perimiters average around four times the figure height, although this ratio can vary by up to 30% in either direction. In most cases this suffices for adequate retrieval. Nevertheless, once an initial frame has been solved, using the visibility of registered silhouettes in a neighboring frame is generally more accurate and therefore preferable.

Figures 2 and 3 show the reconstruction results under fully automatic operation. Despite the increased difficulty of retrieving and registering correct candidate silhouettes from partial shapes at every frame, the boundary fragment

matching reconstructs both clips without major errors. The automatic reconstructions capture the qualitative features of the walk nearly as well as the result using the manual occlusion map, but exhibit somewhat larger transient deviations in scale and body orientation. Interestingly, this reconstruction avoids stutter-steps in the *Ramp* clip.

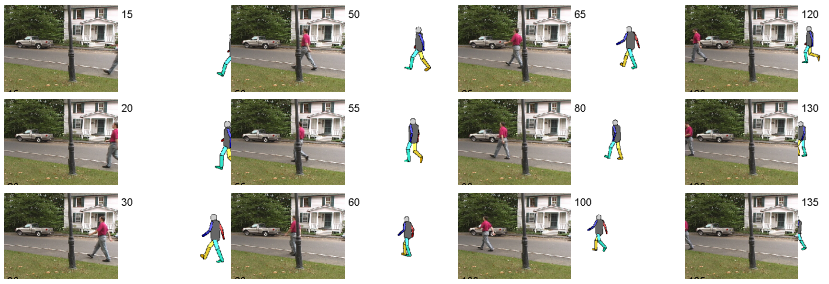

Fig. 2. Automatic reconstruction of *Pole* clip shown at selected frames.

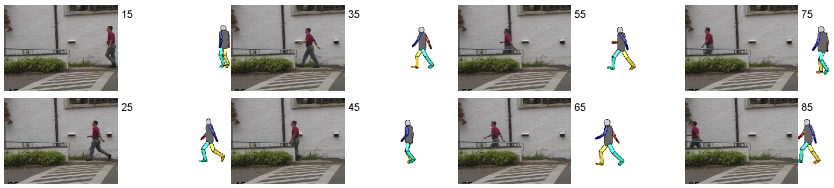

Fig. 3. Automatic reconstruction of *Ramp* clip shown at periodic frames.

Although 3D pose ground truth is unavailable for these clips, the results can be evaluated in comparison to 2D tracking points hand-entered by two individuals. For the *Pole* clip, the difference between the two humans in placement of control points averaged 1.9 pixels; the automatic result differed from the human mean by 6.1 pixels (roughly four inches). No increase in error was observed during the occlusion by the lamp pole. The *Ramp* clip is more difficult for humans to annotate due to the extended occlusion, and disagreement between human point placements averaged 4.4 pixels. The disagreement between the automatic and human results averaged 7.4 pixels. On this clip, occlusion causes a noticable decrease in accuracy for both the humans and the computer algorithm: the error averaged for the frames before and after frame 40 are 2.1 vs. 6.6 pixels for humans, and 5.8 vs. 8.7 pixels for the computer.

While walking motions arguably make for a simple evaluation choice, they are nevertheless of interest in many applications. Furthermore, these experi-

ments indisputably provedemonstrate the utility of shape fragment matching for handling occlusion. Perhaps this will spur the development of additional occlusion-handling techniques for other pose reconstruction modalities.

3.4 Refined Occlusion Maps

The automatic reconstruction can proceed one step further to refine the original automatic occlusion map, distinguishing between true occlusion zones and areas with no data. With pose reconstruction in hand, animation and rendering reveals the region swept out by the moving subject in the image frame. The reconstruction registration may not be entirely accurate, so the outer edge of this region should be thrown out using morphological erosion. The intersection of the remaining area and the original occlusion map yields a set of pixels held to be occupied by occluding objects. Storing these locations may help in subsequent pose reconstructions, or in other scene interpretation tasks. Figure 1 shows the refined automatic occlusion maps for the two clips.

4 Conclusion

This paper has described an extension of silhouette-based monocular 3D pose reconstruction to handle partial occlusions by stationary objects. One enabling development is EMD-embedded boundary fragments, a novel contour-based description of shape that allows comparison of partial shapes. The other key is the explicit use of an occlusion/visibility map, allowing the algorithm to discriminate between valid silhouette boundaries and spurious ones arising from occlusion. The occlusion map may be created by hand, but it can also be generated through an automatic process with surprisingly little decrease in the quality of the final result.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0328741.

References

1. T. Adamek and N. O'Connor. Efficient contour-based shape representation and matching. In *Multimedia Information Retrieval*, pages 138–143, 2003.
2. M. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(24):509–522, April 2002.
3. R. Fablet and M. Black. Automatic detection and tracking of human motion with a view-based representation. In *European Conference on Computer Vision*, pages 476–491, 2002.

4. P. Gabriel, J. Verly, J. Piater, and A. Genon. The state of the art in multiple object tracking under occlusion in video sequences. In *Advanced Concepts for Intelligent Vision Systems*, pages 166–173, 2003.
5. K. Grauman and T. Darrell. Fast contour matching using approximate earth mover's distance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume I, pages 220–227, 2004.
6. K. Grauman and T. Darrell. Efficient image matching with distributions of local invariant features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume II, pages 627–634, 2005.
7. N. Howe. Silhouette lookup for monocular 3d pose tracking. *Image and Vision Computing*, 2006. (to appear).
8. N. Howe and A. Deschamps. Better foreground segmentation through graph cuts. Technical report, Smith College, 2004. http://arxiv.org/abs/cs.CV/0401017.
9. O. Lanz and R. Manduchi. Hybrid joint-separable multibody tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 413–420, 2005.
10. L. J. Latecki, R. Lakämper, and U. Eckhardt. Shape descriptors for non-rigid shapes with a single closed contour. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 424–429, 2000.
11. T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, March 2001.
12. F. Mokhtarian. Silhouette-based object recognition with occlusion through curvature scale space. In *Proceedings of the European Conference on Computer Vision*, pages 566–578, 1996.
13. F. Mokhtarian, S. Abbasi, and J. Kittler. Robust and efficient shape indexing through curvature scale space. In *Proceedings of the British Machine Vision Conferenc*, pages 53–62, 1996.
14. G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *European Conference on Computer Vision*, 2002.
15. D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 271–278, 2005.
16. Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi. Particle filtering for geometric active contours with application to tracking moving and deforming objects. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2–9, 2005.
17. E. Saber, Y. Xu, and A. M. Tekalp. Partial shape recognition by sub-matrix matching for partial matching guided image labeling,. *Pattern Recognition*, 38:1560–1573, 2005.
18. E. Salari and S. Balaji. Recognition of partially occluded objects using b-spline representation. In *Proc. SPIE, High-Speed Inspection Architectures, Barcoding, and Character Recognition*, volume 1384, pages 115–123, 1991.
19. L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume I, pages 421–428, 2004.
20. C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Learning to reconstruct 3d human motion from bayesian mixtures of experts: A probabilistic discriminative approach. Technical Report CSRG-502, University of Toronto, October 2004.
21. R. Urtasun, D. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.