Dynamic Data Driven Coupling of Continuous and Discrete Methods for 3D Tracking*

Dimitris Metaxas and Gabriel Tsechpenakis

Center for Computational Biomedicine, Imaging and Modeling (CBIM),
Computer Science Dept., Rutgers University,
110 Frelinghuysen Rd, Piscataway, NJ 08854
{dnm, gabrielt}@cs.rutgers.edu

Abstract. We present a new framework for robust 3D tracking, using a dynamic data driven coupling of continuous and discrete methods to overcome their limitations. Our method uses primarily the continuous-based tracking which is replaced by the discrete one, to obtain model re-initializations when necessary. We use the error in the continuous tracking to learn off-line, based on SVMs, when the continuous-based tracking fails and switch between the two methods. We develop a novel discrete method for 3D shape configuration estimation, which utilizes both frame and multi-frame features, taking into account the most recent input frames, using a time-window. We therefore overcome the error accumulation over time, that most continuous methods suffer from and simultaneously reduce the discrete methods complexity and prevent possible multiple solutions in shape estimation. We demonstrate the power of our framework in complex hand tracking sequences with large rotations, articulations, lighting changes and occlusions.

1 Introduction

There are generally two major types of approaches to deformable and articulated shape and motion estimation: (i) the continuous ones that exploit the static and the temporal information in images, and (ii) the discrete ones that use only static information, i.e., they estimate the objects configuration based on a single frame. Continuous approaches are usually faster and more accurate than discrete approaches, but when they loose track they cannot easily recover due to error accumulation. On the other hand, discrete approaches can give a good approximation of an objects configuration without error accumulation over time. However, they have high computational cost and are based on searching in databases with limited number of object configurations.

In this paper, we introduce a new framework for robust 3D object tracking, to achieve high accuracy and robustness. Focusing on a specific case of tracking,

* This research has been funded by an NSF-ITR/NGS-0313134 and an NSF-ITR-[ASE+ECS]-0428231 Collaborative Project to the first author.

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3515, pp. 712–720, 2005.

i.e., the $3D$ hand tracking, our approach is based on a *dynamic data driven* coupling of continuous and discrete methods; when our existing continuous tracking fails based on an error measure derived from the data, we can obtain efficient object configuration re-initialization using the discrete method presented in this work.

This paper is organized as follows. In the next subsection, we give a brief description of the previous work, including the existing continuous hand tracking method we used. In section 2 we describe the proposed discrete tracking scheme. In section 3 it is explained how the coupling between the two methods is achieved. In section 4 we present our results on the $3D$ hand tracking, including the case of sign language. Finally, section 5 describes our conclusions and our future work.

1.1 Previous Work

This paper focuses on hand articulations, where several techniques exist that treat the hand configuration estimation as a continuous $3D$ tracking problem [6, 18, 5, 8]. A possible drawback of some approaches is that they introduce additive errors over time, leading to the loss of the track, and when this occurs, they cannot usually recover. This is the reason why some discrete techniques have been developed in the last few years [3, 9], treating each frame independently from the previous ones, although they usually require higher computational time.

Both continuous and discrete methods for $3D$ hand tracking can be divided in two main classes: (a) the model-based ones [6, 12, 4, 12], where $3D$ hand models are constructed and a matching takes place between the input image features and the respective features of the model projection onto the image plane, and (b) the appearance-based approaches [10, 15], which involve mapping of the image feature space to the hand configuration space. Another problem that is tackled by some methods [15, 16, 3], is the background complexity, i.e. the discrimination between the hand and the background edges, when using edges as the visual cues for hand configuration estimation.

In the last few years, some approaches that use hand configuration databases have been proposed [13, 2] and the $3D$ hand pose estimation problem is converted into a database indexing one. The main problem that arises in these methods, apart from the computational complexity, is that multiple matches between the input hand image and the database samples may occur.

In the model-based continuous tracking of [6] that we use, $2D$ edge-driven forces, optical flow and shading are computed. They are converted into $3D$ ones using a perspective camera model, and the results are used to calculate velocity, acceleration and the new position of the hand. A Lagrangian second order dynamic hand model is used to predict finger motion between the previous and the current frame. A model shape refinement process is also used, based on the error from the cue constraints to improve the fitting of the $3D$ hand model onto the input data.

2 Discrete Tracking

For an input frame of the examined hand sequence, we extract 2D features, which will be used for describing the current frame, but will also be integrated with the respective features of a number of past frames to serve as multi-frame descriptors. Instead of matching between single images, we perform multi-frame matching between the most recent input frames and the samples from our synthetic hand database. As will be explained in section 3, the database search is efficient, when the discrete tracking is used in our integration scheme. We search our database in two steps: (i) according to general features, we find the most appropriate cluster, and (ii) using more detailed features we search for the best matching sample sequence inside the chosen cluster. The last hand configuration is chosen as the solution for the input frame. In this way, we avoid multiple matches, taking into account the most recently estimated hand configurations of the input video, without any additional computation load.

Hand Gestures Database. Our database contains configuration sequence samples, instead of single configurations as in [2, 13]. Our synthetic hand model has 20 $dofs$, as shown in Fig. 1 and its advantage is the good skin texture, that can be used for hand edge extraction.

We created 200 configuration sequences, under 29 views, and each sequence has $N_{max} = 15$ frames, which are enough to include tracking failures in the overall coupling scheme (for a $30fps$ input video). For each sample we have stored the N_{max} joint angle sets corresponding to its successive configurations. We have also extracted and stored (i) the single frame and (ii) the multi-frame descriptors of each configuration sequence, as described below.

The database is organized according to which side of the hand is visible (projection information) in the last frame, and how many fingers are visible in the first and last frame of each sample sequence. Thus, we have divided our database into 108 clusters, each one containing 54 samples on average.

2D Hand Features. For every input frame, we use as descriptors both boundary and region-based information of the captured hand.

Single Frame Features. In order to estimate the $3D$ hand configuration by searching in our database for the best matching configuration sequence, we use the following cues. *(i) Boundary-based features:* For each input frame we ex-

Fig. 1. Our virtual hand and the 20 $dofs$ describing all possible articulations

tract the contour of the hand and the corresponding curvature function and Curvature Scale Space (CSS) map [7, 1]. The CSS peaks indicate the most important zero-crossings Z_c of the contour. In this way curvature is an efficient shape descriptor for shape changes under complex movements and scaling. *(ii) Region-based features:* We extract the edge map of each input frame, using the canny edge detector, and we calculate the edge orientation histogram of the hand, as in [15, 3], with $B = 45$ bins. The orientation histogram can provide us with information about the edges in the interior of the hand. *(iii) Projection information:* From the currently estimated configuration, we obtain the pose information for the next frame, i.e. which side of the hand is visible (palm, side or knobs view), assuming that the hands general pose does not change significantly in two successive frames. *(iv) Finger counting:* For each input frame we count the clearly visible fingers F, by calculating the most important zero-crossings extracted in (i): $F = \frac{Z_c}{2}$.

Multi-frame Features. Instead of searching and matching hand configurations in the database, we search for configuration sequences, by taking into account the N_{max} most recent frames of the input sequence. To reduce the computational complexity, we integrate the extracted $2D$ single frame features into two vectors. For P points of an object contour and its curvature (B bins of the edge orientation histogram), over N_{max} successive frames of the input video segment, we can assume that we have P points (B points) in an N_{max}-dimensional space, while we need to have P points (B points) in the $1D$ space. Thus, the problem is transformed into a dimensionality reduction task. For the hand tracking application, we used the nonlinear local Isomap embeddings proposed by Tenenbaum et. al. [14], keeping $(P, B) << N$. We chose to use Isomap, instead of using a linear embedding e.g. PCA (Principal Component Analysis), because we have nonlinear degrees of freedom, and we are interested in a global *hand movement signature*, i.e. a globally optimal solution [11].

Thus, if $\overline{K} = [K_n | n = 1, \ldots, N_{max}]$ and $\overline{H}_\vartheta = [H_{\vartheta,n} | n = 1, \ldots, N_{max}]$ are the sets of N_{max} curvature functions K_n and edge orientation histograms $H_{\vartheta,n}$, extracted over N_{max} frames, the embedded $1D$ multi-frame descriptors are respectively,

$$\tilde{k} = M^{(N,1)}(\overline{K}), \;\; and \;\; \tilde{h_\vartheta} = M^{(N,1)}(\overline{H}_\vartheta), \tag{1}$$

where $M^{(N,1)}$ represents the Isomap embedding from the N to the 1 dimensional space. Fig. 2 illustrates four examples of gesture signature using the embedded CSS descriptor. Each row represents one of the four illustrated hand movement cases, whereas the first column shows the respective embedded CSS descriptors.

Matching Between Input Frames and Database Samples. The matching criterion is the undirected chamfer distance. In general, given two point sets A and B, their undirected chamfer distance $d(A, B)$ is defined by the forward $d^f(A, B)$ and backward $d^b(A, B)$ distances:

$$d^f(A,B) = \frac{1}{\|A\|} \cdot \sum_{a^i \in A} \min_{b^j \in B} \|a^i - b^j\|, \;\; d^b(A,B) = \frac{1}{\|B\|} \cdot \sum_{b^i \in B} \min_{a^j \in A} \|a^j - b^i\|, \quad (2)$$

$$d(A,B) = d^f(A,B) + d^b(A,B) \quad (3)$$

Replacing A and B with the multi-frame descriptors $\tilde{k} = [\tilde{k}^i | i = 1, \ldots, P]$, $\tilde{k}_s = [\tilde{k}_s^i | i = 1, \ldots, P]$, we obtain the chamfer distance d_k between the embedded curvature functions of the input video segment and the database configuration sequence, respectively. Similarly, replacing A and B with $\tilde{h}_\theta = [\tilde{h}_\theta^i | i = 1, \ldots, B]$, $\tilde{h}_{\theta,s} = [\tilde{h}_{\theta,s}^i | i = 1, \ldots, B]$, we obtain the respective chamfer distance d_h between the embedded edge orientation histograms. In Fig. 2 we present four gestures (four key-frames are shown), where the first one ((b1)-(e1)) is completely different from the other three, the second one ((b2)-(e2)) is similar to the last two, which are the same gesture performed twice. The undirected chamfer distances between the corresponding embedded curvature functions of Fig. 2(a1)-(a4) are $d_k^{(a_1),(a_2)} = 10.33$, $d_k^{(a_1),(a_3)} = 7.87$, $d_k^{(a_1),(a_4)} = 8.54$, $d_k^{(a_2),(a_3)} = 6.93$, $d_k^{(a_2),(a_4)} = 6.31$, and $d_k^{(a_3),(a_4)} = 1.21$, where superscripts indicate the corresponding cases compared. It can be seen also numerically that cases (a3) and (a4) of Fig. 2 are the most similar.

For a given database cluster $S \in \mathbf{S}$, where $\mathbf{S}$ is the set of all clusters, i.e. the entire database, and a set of input frames u, the best matching sample $\hat{s} \in S$ $(\in \mathbf{S})$is given by,

$$\hat{s} = \arg\max_{s \in S} p(s|d_h(u,s)) \cdot p(s|d_k(u,s)), \quad (4)$$

where $d_h(u,s)$ and $d_k(u,s)$ are the extracted chamfer distances. The two matching probability functions $p(s|d_h(u,s))$ and $p(s|d_k(u,s))$ are gaussian distributions.

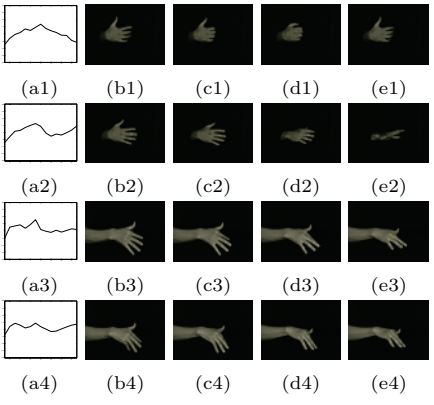

Fig. 2. Embedded CSS descriptors: (a1)-(a4) are the $1D$ embedded CSS, and (b1)-(e1), (b2)-(e2), (b2)-(e2) and (b2)-(e2) are key-frames of the respective sequences

3 Dynamic Data Driven Coupling of Continuous with Discrete Tracking

Our approach is based primarily on the continuous tracking method of [6], briefly described in 1.1. During each tracked frame we dynamically search in the synthetic hand database and choose the cluster that best matches the corresponding descriptors of our input sequence segment. In particular, our cluster search is based on the pose of the hand in the last (current) frame and the number of fingers visible in the first, and the last frame of the video segment. In this way, for each time step we know which database cluster best matches with the N_{max} most current frames of the input sequence. When there is an indication that the continuous method is about to loose track, we use the discrete approach, where we search in the best matching cluster, for the best matching sample sequence, using the multi-frame descriptors.

For each frame of the input video, the difference between the model projection onto the image plane, and the tracked hand gives us the indication whether continuous tracking fails or not. In order to estimate this difference, we use the undirected chamfer distance d_{map} as in eqs. (2),(3), between the model and the examined and edges. Thus, what we need to learn is the joint probability of continuous tracking given the distance d_{map} between the hand and the estimated model: $p(\mathbf{c}|d_{map})$.

We applied continuous tracking off-line, including cases where it fails, and we concluded that this happens when there are complex hand movements, such as strong articulations and rotations, i.e. where optical flow estimation fails. We performed complicated movements and articulations between simple hand movements (translations or rotations parallel to the camera plane), i.e. cases where it is successively $\mathbf{c} = \{1, 0, 1\}$ (continuous tracking success-failure-success). We used five hand sequences, $3,600$ frames each, where a hand performed the same pattern ($\mathbf{c} = \{1, 0, 1\}$) under many different articulations. The joint probability $p(\mathbf{c}|d_{map})$ was then learnt using a a linear two-class SVM [17].

In summary our method is based on the use of two different types of tracking methods. At each time only one is used based on an error measured defined by the input data. This approach results in significantly more robust tracking results as described below.

4 Experimental Results

After learning off-line the probability $p(\mathbf{c}|d_{map})$, we automatically estimated the threshold for d_{map}, under which the continuous tracking can be applied safely, as $T \simeq 0.27$.

In Fig. 3 the upper images of each row illustrate the case of a strong rotation, where fingers get occluded, while the lower images of each row show the extracted final result based on coupling between continuous and discrete tracking. In this case, the hand performs a strong rotation without any significant

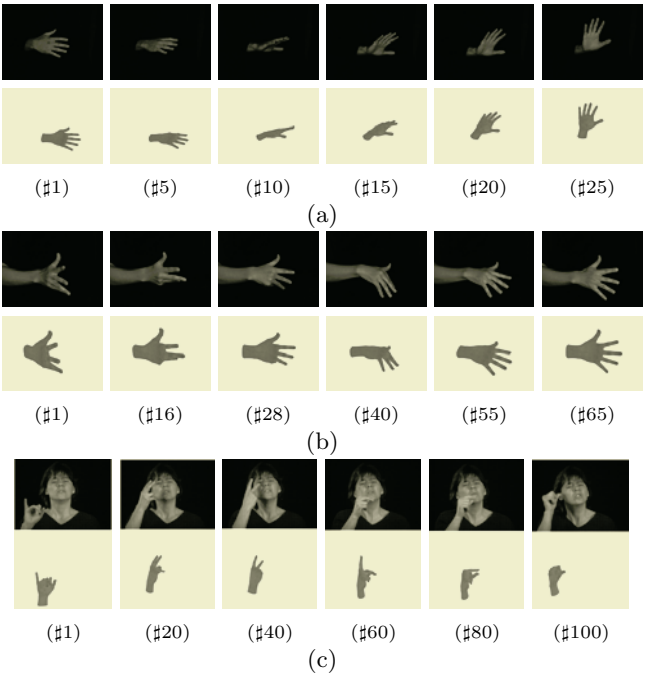

Fig. 3. Tracking results using our coupling framework

finger articulations and continuous tracking is performed in long time intervals without any significant error.

Fig. 3(b) shows a more difficult case, where the hand performs not only complicated rotations but also finger articulations. In this case, the time intervals where continuous tracking was successfully applied are smaller, i.e. in the coupling scheme discrete method initializes the hand model in more frames.

In Fig. 3(c), we illustrate the case of tracking a hand performing sign language, including both fingerspelled letters and continuous signs. Most of the fingerspelled letters of sign language cannot be captured with continuous tracking and that is why its coupling with a discrete scheme is crucial. The illustrated key-frames are taken from the phrase 'Who did John see yesterday?', which in sign language is said as 'John (fingerspelled) - see - who - yesterday (?)'. Our coupling method performs well even in this case, where there are lighting chances and complicated background, which are handled by the continuous tracking method of [6].

5 Summary and Conclusions

We have presented a Dynamic Data Driven Application System (DDDAS) for 3D object tracking in monocular sequences based on a novel coupling of continuous and discrete tracking methods. We have focused on 3D hand tracking, since it is

a very challenging task with a wide variety of applications. We have shown how our approach handles complex articulations, abrupt and large movements and occlusions. Our aim is to further evolve our method to be used for tracking of a much larger type of articulated and nonrigid motions. In particular we plan to further analyze signed languages such as ASL and also use our method in HCI applications.

References

1. S. Abbasi, F. Mokhtarian and J. Kittler, "Curvature Scale Space Image in Shape Similarity Retrieval," *Multimedia Systems*, 7(6), pp. 467-476, 1999.
2. V. Athitsos and S. Sclaroff, "Database Indexing Methods for 3D Hand Pose Estimation ," *Gesture Workshop*, Genova, Italy, April 2003.
3. V. Athitsos and S. Sclaroff, "Estimating 3D Hand Pose from a Cluttered Image," *IEEE Conference on Computer Vision and Pattern Recognition*, Wisconsin, June, 2003.
4. J. Lee and T. Kunii, "Model-based Analysis of Hand Posture," *IEEE Computer Graphics and Applications*, 15, pp. 77-86, 1995.
5. J. Lin, Y. Wu and T.S. Huang, "Modeling the Constraints of Human Hand Motion," *5th Annual Federated Laboratory Symposium (ARL2001)*, Maryland, 2001.
6. S. Lu, D. Metaxas, D. Samaras and J. Oliensis, "Using Multiple Cues for Hand Tracking and Model Refinement," *IEEE Conference on Computer Vision and Pattern Recognition*, Wisconsin, June 2003.
7. F. Mokhtarian and A. Mackworth, "A Theory of Multiscale, Curvature-based Shape Representation for Planar Curves," *Pattern Analysis and Machine Intelligence*, 14(8), pp. 789-805, 1992.
8. J. Rehg and T. Kanade, "Model-based Tracking of Self-occluding Articulated Objects," *IEEE International Conference on Computer Vision*, Cambridge, MA, June, 1995.
9. R. Rosales, V. Athitsos, L. Sigal and S. Sclaroff, "3D Hand Pose Reconstruction Using Specialized Mappings," *IEEE International Conference on Computer Vision*, Vancouver, Canada, July, 2001.
10. N. Shimada, K. Kimura and Y. Shirai, "Real-time 3D Hand Posture Estimation based on 2D Appearance Retrieval Using Monocular Camera," *IEEE ICCV Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, Vancouver, Canada, July 2001.
11. V. de Silva and J.B. Tenenbaum, "Global versus Local Methods in Nonlinear Dimensionality Reduction," *Advances in Neural Information Processing Systems 15*, (eds.) M.S. Baker, S. Thrun and K. Obermayer, Cambridge, MIT Press, pp. 705-712, 2002.
12. B. Stenger, P.R.S. Mendonca and R. Cipolla, "Model-based 3D Tracking of an Articulated Hand," *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, December 2001.
13. B. Stenger, A. Thayananthan, P.H.S. Torr and R. Cipolla, "Hand Pose Estimation Using Hierarchical Detection," *International Workshop on Human-Computer Interaction*, Prague, Czech Republic, May 2004.
14. J.B. Tenenbaum, V. de Silva and J.C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science Magazine*, 290, pp. 2319-2323, 2000.

15. A. Thayananthan, B. Stenger, P. H. S. Torr and R. Cipolla, "Shape Context and Chamfer Matching in Cluttered Scenes," *IEEE Conference on Computer Vision and Pattern Recognition*, Madison, June 2003.
16. J. Triesch and C. von der Malsburg, "A System for Person-Independent Hand Posture Recognition against Complex Backgrounds," *Pattern Analysis and Machine Intelligence*, 23(12), December 1999.
17. V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
18. H. Zhou and T.S. Huang, "Tracking Articulated Hand Motion with Eigen Dynamics Analysis," *IEEE International Conference on Computer Vision*, Nice, France, October 2003.