Deformable Models for Segmentation, 3D Shape and Motion Estimation and Recognition

Dimitris Metaxas
VAST Laboratory
Department of Computer and Information Science
University of Pennsylvania, Philadelphia, PA 19104-6389
dnm@central.cis.upenn.edu

Abstract

We present our framework for segmentation, 3D shape and motion estimation and recognition. We first present physics-based modeling techniques for segmentation and 3D shape and motion estimation based on single and multiple views as well as the integration of visual cues such as edges and optical flow. We then present extensions to address the reliable recognition of American Sign Language (ASL), using 3D tracking data, ASL phonology and modifications to the traditional use of Hidden Markov Models. We demonstrate the usefulness of this framework in computer vision and medical image analysis applications.

1 Introduction

Physics-based modeling [11] offers the methodology and techniques to address in a unified way difficult problems in computer vision. One of the advantages of using a deformable model in vision problems is that such models can be used to estimate the shapes and motions of objects (rigid and/or nonrigid) with unknown initial shapes. The deformations allow the model to take many different shapes, therefore avoiding the need to use a different model depending on the application. In addition, the formulation allows the simultaneous object segmentation, shape and motion estimation as well as the integration of multiple visual cues that improves 2D and 3D shape and motion estimation.

In this paper we present shape modeling modeling methods for the representation of complex models with a small number of parameters [4, 15]. We first describe our theory of shape blending [4] aimed at increasing the vocabulary of objects that we can represent using a small number of global shape parameters such as tapering and bending. For example, using such shape techniques we can combine spherical and cylindrical primitives into a single shape and we can achieve the transition of shape from a sphere to a torus. Based on this approach we were able to compactly represent complex shapes such as a cup starting from a sphere. Furthermore, this new theory allows topological changes in the shape of an initial primitive (e.g., the formation of holes), thus significantly increasing the shape coverage of our deformable primitives. Second, we present [15] a class of deformable models whose global parameters are functions. These parameters are able to represent the local shape variations based on intuitive global parameter functions, instead

BMVC 1999 doi:10.5244/C.13.1

of local deformations. We used such models in medical image analysis applications to estimate the shape and motion of the heart's ventricles.

Shape representation is the basis for model-based segmentation and tracking. Using deformable models we present segmentation as well as shape estimation and tracking methods based on the integration of visual cues [16, 17, 2, 5]. Depending on the application our methods use single or multi-view image sequences as input.

The accurate 3D tracking of the human body and face with physics-based modeling (PBM) techniques, provides a foundation for numerous dynamic recognition applications which include human-computer interaction (HCI), monitoring systems, sports medicine and athlete performance assessment. In HCI, gesture recognition and sign language recognition are of particular interest. In this paper we present general dynamic recognition methods with particular emphasis on American Sign Language (ASL) recognition. We demonstrate that the 3D gesture and facial shape and motion estimation based on PBM methods can be used as input for the reliable recognition of ASL by taking into account the phonological aspects of ASL and by modifying the traditional use of Hidden Markov Models.

In the following sections we elaborate more on the above techniques and we present our results.

2 Deformable Models: Geometry and Dynamics

We have developed [18, 11] a physics-based framework which provides deformable models with broad geometric coverage along with robust techniques for inferring shape and motion from noise-corrupted data.

In this framework, the positions of points on the model relative to an inertial frame of reference Φ in space are given by $\mathbf{x}(\mathrm{u},t) = (x_1(\mathrm{u},t), x_2(\mathrm{u},t), x_3(\mathrm{u},t))^T$, where ^{T} denotes transposition. We set up a noninertial, model-centered reference frame ϕ and express the position function as $\mathbf{x} = \mathbf{c} + \mathbf{R}\mathbf{p}$, where $\mathbf{c}(t)$ is the origin of ϕ at the center of the model and the rotation matrix $\mathbf{R}(t)$ gives the orientation of ϕ relative to Φ. Thus, $\mathbf{p}(\mathrm{u},t)$ gives the positions of points on the model relative to the model frame. We further express $\mathbf{p} = \mathbf{s} + \mathbf{d}$ as the sum of a global reference shape $\mathbf{s}(\mathrm{u},t)$ and a local displacement $\mathbf{d}(\mathrm{u},t)$.

We define the global reference shape as $\mathbf{s} = \mathbf{T}(\mathbf{e}(\mathrm{u};\ a_0, a_1, \ldots);\ b_0, b_1, \ldots)$. Here, a geometric primitive $\mathbf{e}$, defined parametrically in u and parameterized by the variables a_i, is subjected to the *global deformation* $\mathbf{T}$ which depends on the parameters b_i. Examples of global deformations are the parameterized bending, tapering and twisting deformations.

Through the application of Lagrangian mechanics, we have developed a method [11] to convert systematically the geometric parameters of the solid primitive, the global (parameterized) and local (free-form) deformation parameters, and the six degrees of freedom of rigid-body motion into generalized coordinates or dynamic degrees of freedom.

3 Shape Abstraction Based on Shape Blending

In order to represent complex shapes (e.g., cup, light bulb, donut) with a small number of parameters we have developed a theory of shape representation based on shape blending.

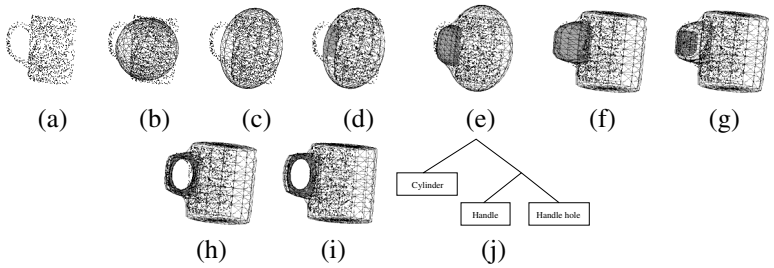


Figure 1: Fitting of a mug (a) range data (b) initial model (c-j) fitting (k) final fit of mug (l) symbolic representation

We use shape blending to increase the vocabulary of objects that we can represent using a small number of global shape parameters. Based on this approach we can estimate the shapes of complex objects starting from a sphere, undergoing topological changes in its shape if necessary (e.g., formation of a hole). In this way we can achieve the shape transition of a sphere to a torus. We first experimented [11] with the use of axial blending between parameterized primitives.

Blended shapes of that type offer great promise in terms of the variety of objects they can represent with only a few parameters. Apart from representing shape compactly it also offers shape abstraction since the shape components (e.g., sphere, cylinder) are integrated into the representation. Furthermore, they offer the capability of genus change (e.g., a sphere turns into a torus through appropriate blending), which was not possible with any of the previous parameterized models used in computer vision.

Extending axial blending, we have developed a theory of shape blending which allows the adaptive blending of shape primitives in complex ways (not just axially) to be able to recover non-trivial shapes such as a mug. As opposed to other shape representation schemes, the recovered shapes are represented with a small number of parameters (compactness) [4]. Our method also achieves shape abstraction by representing the extracted shape as a graph of components (nodes in the graph) where geometric information is also stored (as links between graph nodes).

Fig. 1 shows the fitting of a mug using this approach. The blending region which corresponds to the mug handle forms in (d), and after rough fitting is shown in (e). After further fitting of the handle in (f), a hole blend is added in (g). After rough fitting and hole opening (h), the final fit is obtained (i). A *symbolic* description of a mug is shown in (j).

4 Deformable Models with Parameter Functions

We have introduced a new class of models [15] to capture the local time varying shape of complex objects with a few parameters. We have so far used these models in medical image analysis, even though they can be used in many other applications.

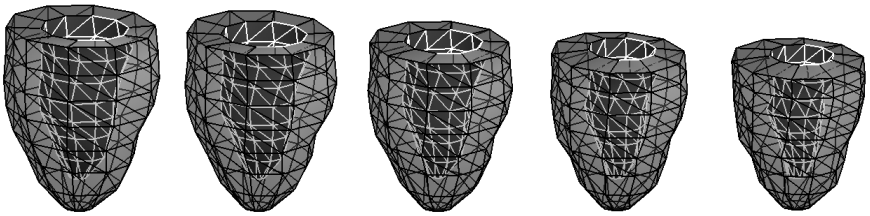
In these models, the global parameters $\mathbf{q}_s = (a_0(\mathrm{u}), a_1(\mathrm{u}), .., b_0(\mathrm{u}), b_1(\mathrm{u}), b_1(\mathrm{u}), ..)^T$ are functions of u, instead of being constants [11]. This definition allows us to generalize definitions of primitives (e.g., superquadrics, cubes) and parameterized deformations (e.g., twisting) as will be shown in the following example.

Our technique for creating primitives with parameter functions can be applied to any

parametric primitive, by replacing its constant parameters with differentiable parameter functions. For example, using this method we can transform an ellipsoid primitive [11] to a primitive with parameter functions. The definition of such a generalized primitive $\mathbf{e} = (e_1, e_2, e_3)^T$ is given as follows:

$$\begin{aligned} \mathbf{e} = (e_1, e_2, e_3)^T, e_1 = a_0 a_1(u) \ \cos u \ \cos v, \\ e_2 = a_0 a_2(u) \ \cos u \ \sin v, \ e_3 = a_0 a_3(u) \ \sin u, \end{aligned} \tag{1}$$

where $-\pi/2 \leq u \leq \pi/2$, $-\pi \leq v < \pi$. Here, $a_0 \geq 0$ is a scale parameter, and $0 \leq a_1(u), a_2(u), a_3(u) \leq 1$, are aspect ratio parameter functions. We can also define an open parameterized primitive given by the above definition by restricting the ranges of the u and v parameters to a subset of the above definition. This formulation of deformations with continuous parameter functions is general and can be applied to any underlying shape $\mathbf{e}$.



(a) time 1 (b) time 2 (c) time 3 (d) time 4 (d) time 5

Figure 2: LV fitted models during systole

By extending this method we have developed a volumetric model of the LV [15] which allows the estimation of its shape and motion on and within the walls. Fig. 2 shows fitting results of a volumetric model to MR tagging data over 5 time frames. The top row shows a view from the base of the LV of the fitted model. The twisting of the inner wall (shown in white) is obvious. The middle row shows a side view of the model, while the last row is similar to the first row and shows a view of the model from the apex. We can easily observe the longitudinal contraction as well as the radial contraction.

5 Coupling Deformable Models and Illumination Constraints

In an effort to estimate shape from shading information we have developed a theory for the numerically robust integration of nonlinear holonomic constraints within a deformable model framework, regardless of their origin [16]. The method is based on the use of Lagrange multipliers and Baumgarte stabilization within a deformable model framework.

Any type of illumination constraint can be incorporated in our physics-based modeling framework [11], provided that the illumination law is a differentiable function of the normal or tangents to the surface. This encompasses practically all the illumination models used in computer vision, from the simple Lambertian model to more complex highly nonlinear models [13, 14]. Instead of extracting the shape parameters directly

Methods	Penny			Sombrero			Mozart		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
Best from [24]	7.2	4.7	4.4	6.9	5.4	5.6	15.1	8.8	7.7
Average Error	6.3	3.9	3.4	4.9	4.5	5.1	8.9	7.1	6.0
Best-fit Error	4.1	3.7	3.4	4.3	3.8	3.9	8.5	7.1	5.8

Table 1: Average Z error for synthetic images. For comparison, "Best from [24]" provides for each image, the best one of the results obtained by the algorithms surveyed in [24]. Thus it includes results from several algorithms — no single algorithm performed best on all the images.

from these illumination constraints (which is not always possible), we use them to provide the necessary generalized forces that will deform our model and estimate the object's 3D shape. Our methodology obviates the need for commonly used approximations (e.g., linearization) to these equations, or the need to solve partial differential equations requiring boundary conditions.

We use deformable models or grids with both global and local deformations. During shape estimation, we first fit the model's global parameters given the illumination constraints, and then we refine its shape based on the model's local deformations, using a coarse to fine grid. Use of a deformable model-based approach offers shape flexibility and the additional advantage of the numerically robust computation of the necessary derivatives, producing improved shape estimation results. This is demonstrated in a series of experiments with real and synthetic data. The experiments consist of the standard test images used in the thorough comparative survey by Zhang et al.[24].

Following [24], we also report the average absolute error for each image. For each image, we compare our result to the best result on the same test image as reported in [24]. Our results are in almost all cases clearly better than the results computed using the SFS-based only algorithms reviewed in [24], and in no case worse.

Visual inspection of our results shows that our method manages to recover surface information and detail in most cases. Discontinuities are handled well, thanks to the elasticity of the model. The amount of detail recovered depends on the number of elements in the model. When fitting a fine mesh, local detail will appear in the first few iterations. As the fitting progresses, the deformation will deepen and broaden but without significant loss of detail. This is an advantage of the hard constraints approach.

In [17], we extend our method to the case when the lightsource direction is not known. In that we estimate both lightsource and shape in a coupled iterative process. For each improved estimate of the shape we can derive an improved estimate of the lightsource and then in turn improve our estimate of the shape. The lightsource is estimated using the Levenberg - Marquardt method, the initial estimate can be obtained using the method by [25]. Convergence of the lightsource is typically within 5 degrees of the true solution.

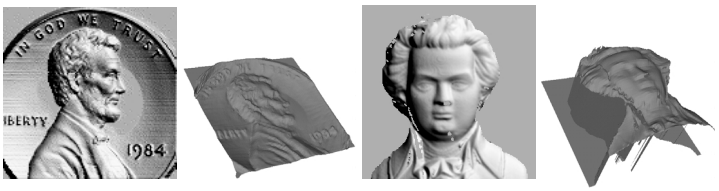
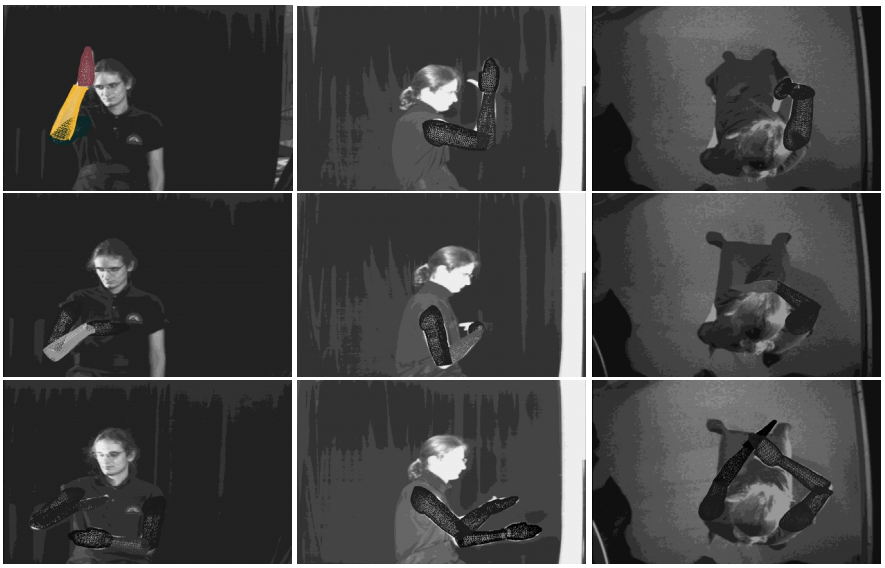


Figure 3: Results on Synthetic Images



(a) (b) (c)

Figure 4: Experiments of multiview-based arm tracking.

6 3D Tracking of Humans

In the following sections we present our physics-based modeling approach and results for the 3D human shape and motion estimation. We present edge based methods that use multi-view image sequences as input as well as the integration of optical flow and edges within a deformable model framework that uses a single image sequence as input.

6.1 Multi-View based Human Tracking

In cases where the frame-rate is not high enough or when the optical flow constraint is not valid we have developed an algorithm for human shape and motion estimation that uses edges extracted from multiple image sequences [7, 8, 9, 10, 11]. A novel aspect of the algorithm was the active selection of a subset of the available input images at any given time instant based on criteria related to the visibility and motion of a tracked human part with respect to each of the cameras.

In Fig. 4 we demonstrate three arm tracking experiments from images sequences captured using three cameras. Each row corresponds to images sequences from the same arm tracking experiment and each column to the views captured from one of the cameras for each of the three experiments.

6.2 Integration of Visual Cues for Facial Tracking

We have recently proposed a theory for the integration of optical flow information within a deformable model framework [1, 2, 5]. Based on this approach, the optical flow is treated as a nonholonomic constraint, i.e., it constrains the velocity of the model parame-

(a) (b) (c) (d) (e) (f)

Figure 5: A face motion and expression tracking example.

ters, within a deformable model framework. As a result, we can simultaneously use edges and optical flow to improve our shape and motion estimation results. We have applied this framework to the estimation of facial shape and motion including expressions. Another important and novel aspect of this work is that we have constructed a deformable face mask whose parameterization is based on facial measurement statistics, taken by anthropometrists [3]. In [5] we have shown experimentally that the integration of optical flow and edges results in longer and more accurate facial tracking compared to using only one of these cues.

The following figure shows results from fitting a deformable mask (the mask's deformable parts are clearly visible) to the face of a subject performing a complex head motion and facial expressions. Notice that the shape of the deformable mask continuously changes based on the subject's facial expressions. Fig. 5 shows a subject making a series of face motions: opening the mouth in (b) and (c), smiling in (d) and (e), and raising the eyebrows in (e) and (f). In each case, the motion parameter values change appropriately, and at the correct times.

7 ASL Recognition

The tracking of the human body with physics-based modeling, as described in the previous section, provides a foundation for applications in human-computer interaction. In this domain, gesture recognition and sign language recognition are of particular interest.

The former application is important, because together with speech, gestures are an integral part of human communication [12]. Hence, a functional user interface based on gesture and speech would make human-computer interaction much more natural. The latter application is important, because sign languages are much more structured than gestures. Thus, they offer an appealing test bed for new ideas and algorithms before they are applied to gesture recognition. In addition, a working sign language recognition system can make deaf-hearing interaction easier. In the following, we describe our framework for American Sign Language (ASL) recognition.

We use the output from our 3D tracking methods to extract suitable features, that we

in turn use as the input to a Hidden Markov Model (HMM) recognition framework, so as to recognize continuous ASL utterances. The output of the HMM framework is a representation of the individual signs and phonemes in the signed sentence, according to the structure of ASL, that we translate to English on a word-per-word basis for convenience. In the future, a machine translation system could use this representation to translate ASL to English.

7.1 The Importance of 3D Data

Using 3D data — as opposed to 2D data — is essential for robust ASL recognition systems. Not only does it significantly improve recognition rates even with simple Cartesian coordinates as the feature set [21], but it also allows further preprocessing and analysis of the data that greatly increase the robustness of the recognition system.

In particular, using 3D data allows us to extract the global properties of the signal, such as whether the hands are moving in a straight line, in a circle, or in a plane. These global properties are valuable features for HMMs when the signs are broken down into their constituent phonemes, because often a single phoneme corresponds to a single movement of the hands in a line, plane, or circle [22].

7.2 The Importance of Phonology

The breakdown of the signs into their constituent phonemes is necessary to make ASL recognition scalable. ASL is a highly inflected language; that is, signs can be modified in many ways to indicate subject, object, and numeric agreement, among others [19]. The mapped vocabulary of ASL consists of approximately 6000 signs, of which most can be inflected in some form, yielding a large number of different possible appearances of signs. Capturing all these different appearances with HMMs is a hopeless task, as the amount of training data required would be prohibitive.

Just like spoken languages, ASL has a limited number of phonemes that are the building blocks for the signs. We use the Movement-Hold phonological model [6] as the basis for modeling ASL, according to which the number of distinct phonemes is approximately 150. Finding enough training data for modeling ASL with HMMs that represent a distinct phoneme each is far easier than finding enough training data for HMMs that represent a distinct appearance of each sign. We showed in previous work that for small vocabularies the recognition performance of phoneme modeling is comparable to the recognition performance of whole-sign modeling [22]. Thus, phoneme modeling is a promising approach for future large-scale ASL recognition.

Closely related to breaking down the signs into phonemes is the problem of modeling the transitions between sequences of signs. For example, in the sequence "FATHER READ," the sign for "FATHER" is performed at the forehead, whereas the sign for "READ" starts in the space in front of the chest. Thus, the signer's hand must move downward from the forehead to the chest between these two signs. These extra movement are called **movement epenthesis**, and they are part of the phonology of ASL. We first modeled epenthesis in conjunction with whole-sign modeling and demonstrated that it improved recognition performance significantly [20]. Later we integrated epenthesis movements into our phonological framework for ASL, so as to maintain a unified view on the language [22].

In summary, taking advantage of ASL linguistics, particularly ASL phonology, is highly beneficial to ASL recognition and should be mandatory for future work in this field. We expect that taking advantage of the other aspects of ASL linguistics, such as morphology and syntax will yield similar benefits in future work.

7.3 Sequential and Parallel Aspects of ASL

The breakdown of the signs into phonemes also raises an interesting question that highlights the differences between spoken languages and signed languages. Whereas in spoken languages all phonemes occur sequentially, in ASL phonemes occur both sequentially and in parallel. For example, handshapes and hand orientations can change at the same time as the hand moves, and the left and right hands move at the same time.

HMMs, however, are a sequential framework by nature. Therefore, the naive solution to the problem of modeling the parallel aspects of ASL would consist of modeling all possible combinations of phonemes that can occur in parallel. However, the sheer number of possible combinations, which is on the order of 10^{10} [23], makes this solution impractical.

Instead, we classify the phonemes that occur in parallel into independent channels; for example, we assume that the left and right hands move in two separate channels independently from each other. This independence assumption may not be entirely valid, but there is linguistic evidence that the hands move at least partially independently from each other [6]. Under this assumption, it becomes possible to model the independent channels in separate HMMs in parallel. Combining the channels then consists simply of multiplying the probabilities of the parallel HMMs.

We showed that in the case of the left and right hands, parallel HMMs can indeed make recognition more robust [23]. This result is promising for integrating other parallel aspects of ASL, such as handshape and facial expressions in the future. Thus, together with phoneme modeling, parallel HMMs lay the groundwork for scalability in sign language recognition.

8 Conclusions

We have presented a framework for shape estimation, tracking, analysis and recognition of dynamic events. Our approach has been based on the realization that these problems are interdependent and we use domain knowledge when necessary. In our future research we will continue to exploit the coupling between these problems to improve our methods.

In our applications, 3D model shape and motion information for event recognition is required, which often does not allow the real time tracking and recognition of these events. However, the advent of significantly faster computers in the near future will allow our algorithms to run in real time compared to their current interactive time performance. It is important to note that the simpler problem of 2D shape and motion based recognition can already be addressed in real time.

9 Acknowledgments

This research was supported in part by a NSF Career Award NSF-9624604, NSF IRI-97-01803, NSF-EIA98-09209, ARO-DAAH04-96-1-007, NASA-96-OLMSA-01-147, ONR-YIP, and ONR-DURIP'97 N00014-97-1-0385.

References

[1] D. DeCarlo and D. Metaxas. "The integration of optical flow and deformable models with applications to human face shape and motion estimation". Procs. of the IEEE Computer Society on Computer Vision and Pattern Recognition, pp. 231-238, June 1996.

[2] D. DeCarlo and D. Metaxas. "Deformable model-based shape and motion analysis from images using motion residual error". Procs 6th International Conference on Computer Vision, India, January 1998.

[3] D. DeCarlo, D. Metaxas and M. Stone. "An Anthropometric Face Model using Variational Techniques". Procs. of ACM Siggraph, Orlando, FL. July, 1998.

[4] D. DeCarlo and D. Metaxas. "Shape Evolution with Structural and Topological Changes using Blending". IEEE Pattern Analysis and Machine Intelligence (PAMI), 20(11), pp. 1186-1205, 1998.

[5] D. DeCarlo and D. Metaxas. "Combining Information using Hard Constraints". Procs. of the IEEE Computer Society on Computer Vision and Pattern Recognition, Fort Collins, CO, June 1999.

[6] S. K. Liddell and R. E. Johnson. American Sign Language: The phonological base. *Sign Language Studies*, 64:195–277, 1989.

[7] I. A. Kakadiaris, D. Metaxas, and R. Bajcsy. "Active part-decomposition, shape and motion estimation of articulated objects: A physics-based approach", In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 980-984, Seattle, WA, June 21-23 1994.

[8] I. A. Kakadiaris and D. Metaxas, "3D human body model acquisition from multiple views", In Proc. of the IEEE Fifth International Conference on Computer Vision, pp. 618-623, Boston, MA, June 20-23, 1995.

[9] I. A. Kakadiaris and D. Metaxas, "Model based estimation of 3D human motion with occlusion based on active multi-viewpoint selection", In Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 81–87, San Francisco, CA, June 1996.

[10] I. Kakadiaris, D. Metaxas and R. Bajcsy. "Inferring Object Structure in 2D from the Deformation of Apparent Contours". Journal of Computer Vision and Image Understanding, 65(2), pp. 129-147, February, 1997.

[11] D. Metaxas. "Physics-Based Deformable Models: Applications to Computer Vision, Graphics and Medical Imaging", Kluwer-Academic Publishers, November 1996.

[12] D. McNeill. *Hand and mind: what gestures reveal about thought*. University of Chicago Press, Chicago, 1992.

[13] S.K. Nayar, K. Ikeuchi and T. Kanade. Surface Reflection: Physical and Geometrical Perspectives In *PAMI*, 13(7):611–634, July 1991

[14] M. Oren and S.K. Nayar. Diffuse reflectance from rough surfaces. In *ICCV 1993* pages 763-764, 1993.

[15] J. Park, D. Metaxas and L. Axel. "Analysis of Left Ventricular Wall Motion Based on Volumetric Deformable Models and MRI-SPAMM". Medical Image Analysis Journal, 1(1), pp. 53-71, March 1996.

[16] D. Samaras, D. Metaxas. Incorporating Illumination Constraints in Deformable Models. In *CVPR98* pages 322-329, 1998.

[17] D. Samaras, D. Metaxas. Coupled Lighting Direction and Shape Estimation from Single Images. In *ICCV99* 1999.

[18] D. Terzopoulos and D. Metaxas. Dynamic 3D Models with Local and Global Deformations: Deformable Superquadrics. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(7):703–714, 1991.

[19] C. Valli and C. Lucas. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, Washington DC, 1995.

[20] C. Vogler and D. Metaxas. Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. *SMC,* 1997.

[21] C. Vogler and D. Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 363–369, Mumbai, India, 1998.

[22] C. Vogler and D. Metaxas. Toward scalability in ASL recognition: Breaking down signs into phonemes. *Gesture Workshop,* Gif sur Yvette, France, 1999.

[23] C. Vogler and D. Metaxas. Parallel Hidden Markov Models for American Sign Language recognition. ICCV'99.

[24] R. Zhang, P. Tsai, J.E. Cryer, M. Shah. Analysis of shape from shading techniques. In *CVPR 1994* pages 377-384, 1994.

[25] Zheng, Chellappa Estimation of illumination direction, albedo, and shape from shading. *PAMI* 13(7):680-702, 1991.

[26] O. Zienkiewicz. *The Finite Element Method*. McGraw-Hill, 1977.