A Brief Review of 3D Face Reconstruction Methods for Face-Related Product Design

Jie Zhang[1], Kangneng Zhou[2] and Yan Luximon[1]*

[1] School of Design, The Hong Kong Polytechnic University, Hong Kong SAR
[2] School of Computer and Communication Engineering,
University of Science and Technology Beijing, China
peterzhang1130@163.com;
elliszkn@163.com;
*yan.luximon@polyu.edu.hk

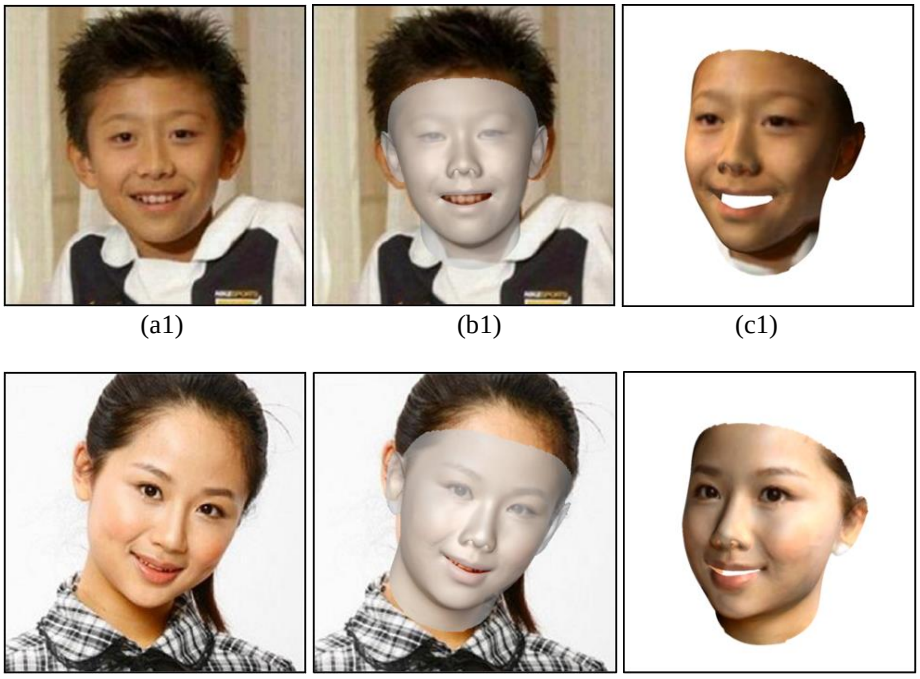
Abstract. 3D face reconstruction is highly important in the ergonomics study of 3D face, especially in terms of designing face-related products. With the development of machine vision and deep learning, it becomes feasible to reconstruct the 3D face from a single image, which can make it practical to obtain a large scale data of 3D face shape instead of using the 3D scanning technology. The 3D face reconstruction methods, to recover the 3D facial geometry under unconstrained situations from 2D images, are roughly classified into two categories, namely (1) 3D Morphable Model (3DMM) fitting based method and (2) End-to-end deep convolutional neural network (CNN) based method. The 3DMM as a general face representation is introduced emphatically and two kinds of 3DMM fitting based methods are introduced when improving the 3DMM modeling mechanism. Four representative CNN based methods are compared when regressing from pixels of face image to the 3D face coordinates in different gird-like data structures. Finally, six common face datasets largely used in the training and testing are listed.

Keywords: Face-related product design, 3D face reconstruction, 3D morphable model

1 Introduction

More and more people are using face-related products such as helmets and respirators. Fit is a very crucial issue for achieving best performance when designing this type of products. Researchers have started using 3D data for product design when dealing with the complexity of human face shape. Chu et al. [1] developed a mass-customization method for respiratory masks based on 3D human face. Lacko et al. [2] established the digital mannequins by using a human 3D head shape and designed brain-computer interfacing (BCI) headset frame based on them. Skals et al. [3] improved the fitting level of bicycle helmet liners by using 3D anthropometric data. These studies show that 3D face reconstruction can be widely used in many applica-

tions and it is highly important in the ergonomics study of 3D face [4]. The recent development of 3D laser scanning technology provides an opportunity to create a more accurate human model for product design and evaluation. High resolution 3D scans can record a massive amount of detailed geometrical information of the human face shape. However, it is unpractical to collect large-scale data of 3D face shape because it takes too much time and manpower. Fortunately, with the development of machine vision and deep learning, it becomes feasible to reconstruct the 3D face from a single image with the help of the supervised learning. The examples of 3D face reconstruction from a single image directly are shown in Figure 1.



(a1) (b1) (c1)

(a2) (b2) (c2)

Fig. 1. Examples of 3D face reconstruction from a single image (a): 2D facial image (b) and (c): Reconstructed 3D face.

2 3D Face Reconstruction Method

3D face reconstruction is to recover the 3D facial geometry under unconstrained situations from 2D images, which is also an important task in the field of computer vision and graphics. With the development of computer vision and machine learning, there are many methods that try to handle this inherently ill-posed problem. These approaches toward this problem can be roughly classified into two categories, namely

(1) 3D Morphable Model (3DMM) fitting based method and (2) End-to-end deep convolutional neural network (CNN) based method.

2.1 3DMM fitting based methods

3D Morphable Model: The 3D Morphable Model (3DMM) was proposed by Blanz and Vetter [5, 6] as a general face representation and a principled approach to image analysis. They found that it is possible to reconstruct face shape by solving a non-linear optimization problem which is constrained by linear statistical models of facial shape and textures [7].

In the 3DMM, the geometry of a face is represented as a shape-vector, $S=(x_1,y_1,z_1,\ldots,x_n,y_n,z_n)^T \in R^{3n}$, that contains the x, y, z coordinates of its n vertices. For simplicity, it is assumed that the number of valid texture values in the texture map is equal to the number of vertices. Therefore, we can use a texture-vector $T=(R_1,G_1,B_1,\ldots, R_n,G_n,B_n)^T \in R^{3n}$ to represent the texture of a face where R, G and B are values of colors. A morphable face model can be constructed using a data set of exemplar faces, each M represented by its shape-vector S_i and texture-vector T_i: $M=(S,T)$. A 3D face can be described with the a basis transformation of Principal component analysis (PCA) [8]:

$$T = \overline{T} + A_t \alpha_t \tag{1}$$

$$S = \overline{S} + A_s \alpha_s \tag{2}$$

Where, S is a 3D face, $\overline{S}$ is the mean shape, A_s is the shape principle basis trained on the 3D face scans with neural expression and α_s is the shape representation coefficient; $\overline{T}$ is the mean texture, A_t is the texture principle basis trained on the 3D face scans with neural expression and α_t is the texture representation coefficient.

The facial expression component is considered to improve the original 3DMM, because of the expression variance of the shape data [9]:

$$S = \overline{S} + A_s \alpha_s + A_{exp} \alpha_{exp} \tag{3}$$

Where A_{exp} is the expression principle basis trained on the offset between expression scans and neural expression and α_{exp} is the expression representation coefficient.

Usually, the Basel Face Model (BFM)[10] and the Face Warehouse dataset[9] can be used to generate the shape (A_t) and expression (A_{exp}) principle basis. The target of the single-image based 3D face modeling is to predict the coefficients α_s and α_{exp}. After obtaining the 3D face shape S, it can be projected onto the 2D image panel with the scale orthographic projection from a specified viewpoint:

$$V = f * P * R * \left(\overline{S} + A_s \alpha_s + A_{exp} \alpha_{exp} \right) + t_{2d} \tag{4}$$

Where V is the 2D coordinates of the 3D vertices projected onto the 2D plane, f is the scale factor, P is the orthographic projection matrix [[1,0,0],[0,1,0]], R is the rotation matrix that consists of 9 parameters, and t_{2d} is the translation vector. The collec-

tion of all the model parameters is $\alpha=[f, R, t_{2d}, \alpha_s, \alpha_{exp}]^T$. However, the inverse process of recovering the 3DMM parameters from a single image is quite challenging.

Model fitting based methods: After the 3DMM was developed, large amount of research has been done to improve 3DMM modeling mechanism. The model fitting based methods can be mainly classified into two groups: the template fitting based methods [6, 11, 12] and the regression based methods [13].

The template fitting based methods are to establish a 3D face appearance model to fit the 2D face images, such as Analysis-by-Synthesis 3DMM [6] and Active Appearance Model (AAM) [12]. However, the performance of the template fitting based methods is affected mainly by the 2D image patterns reside within the variations described by the 3D face appearance model, which shows limited robustness in unconstrained environment.

The regression methods are to estimate the 3DMM parameters by solving a nonlinear optimization problem to establish a correspondence of the points between the single image and the established 3D face model, such as facial landmarks, local features. However, these methods mainly depend on the detection accuracy of landmarks or other feature points. The Cascaded Regression [14] and CNN-based [15-17] methods have been largely used in the landmark detection and face alignment and nearly achieved the state-of-the-art performance in 2D facial landmarks detection.

These two achievements of Cascaded Regression and CNN are combined into a Cascaded NN as a regressor to estimate the new parameters in the 3D Dense Face Alignment (3DDFA) method [13]. In the first stream, a Projected Normalized Coordinate Code (PNCC) is staked with the inputted image and sent to the CNN with an intermediate parameter, while in the second stream, some anchors with consistent semantics are generated and conducted by Pose Adaptive Convolution (PAC). This method can solve the high nonlinearity of face alignment across large poses and the self-occlusion challenge in the face modeling. However, these model-based methods are limited in the face shape of the original model, thereby generating limited geometry with facial expression and occlusions [18].

2.2 End-to-end CNN based methods

The CNN can be directly trained to regress from pixels of 2D face image to the 3D face coordinates in gird-like data structure (such as UV position map [18], volumetric representation [19], depth map [20] and graph mesh [21]) by inputting a single image. These methods can reconstruct 3D face from the completely unconstrained facial images with facial expression and occlusions.

In the end-to-end Position map Regression Network (PRN) [18], UV position map is used to record the 3D facial structure of a complete face in UV shape and a CNN is used to regress the UV position map to simultaneously predict dense alignment and reconstruct 3D face shape. This network is model-free and light-weighted, thereby directly producing the 3D face reconstruct results in real time. However, there is geometric distortion between the UV position map and the 3D coordinates, which make it produce less precise results.

In the Volumetric Regression Network (VRN) [19], a CNN is used to directly learn a mapping from pixels of the RGB image to a 192 x 192 x 200 volume of 3D coordinates. Their experiment results demonstrate that VRN-Guided architecture can achieve better performance that firsts detects the 2D projection of the 3D landmarks and stacks these with the original image to regress the volume. The biggest advantage of this network is generating the full 3D facial geometry from completely unconstrained facial images. However, the discretization limits the resolution of point cloud and the quantization errors are introduced during mesh voxelization.

The Graph Convolution Networks (GCNs) is used to regress 3D face coordinate by processing 3D face mesh. The 3D facial mesh is recorded in $M=(V,W)$ where V has 3D vertices on the 3D face surface and W is a sparse adjacency matrix of V to denote their relationships edge [21]. This network can capture coarse-to-fine features of face mesh in a hierarchical manner and generate 3D face coordinate directly.

The deep Dense-Fine-Finer Network (DF^2Net) [20] integrates three sub-networks to map the input single image to a dense depth image, thereby realizing subtle 3D facial details. A face image is first input to a dense depth network for dense but rough 3D reconstruction, and then the output depth map along with the original image are fed into a hypercolumn network for refinement, finally the refined depth map and the multi-scale face images are jointly processed with a multi-resolution hypercolumn network to estimate a residual depth map for finer detail reconstruction.

3 Face Datasets

The CNN-based methods can achieve great accuracy in both 2D face alignment and 3D face reconstruction. However, they need a huge amount of face datasets, including 2D annotated face images and 3D face information. When doing research on 3D face reconstruction, there are many publicly available human face datasets and them can be classified into three categories based on their format: (1) face images with landmarks, (2) 2D face images and 3D shape data and (3) 2D face images and 3DMM-based reconstructed shape data, shown in Tables 1, 2 and 3 respectively. The datasets in Tables 1 and 2 can usually be used to evaluate the accuracy of 2D face alignment and 3D face reconstruction respectively and the datasets in Table 3 can usually be used to evaluate the accuracy of 3D face alignment and train the established network of the CNN-based methods.

Table 1. Common publicly available appearance images with landmarks of human faces.

Database	No. images	No. landmarks	Comments
CelebA[22]	202,599	5	
AFLW[23]	25,993	21	
COFW[24]	1345 (training)+507(testing)	29	
LFPW[25]	1432	29	
AFLW-LFPA[26]	1299	34	face images with extended 11 landmarks from AFLW
300W[27]	4000+	68	standardized databases with 68 land-

			marks. including AFW[28], LFPW[25], HELEN [29]and XM2VTS[30].
300W-LP[13]	122,450	68	synthesized large-pose face images from 300W
300VW[31]	114 videos	68	
3DFAW[26]	23,000	68	multi-view images with 3D annotation
WFLW[32]	7500 (training)+2500(testing)	98	
Helen[29]	2000 (training)+330 (testing)	194	

Table 2. Common publicly available 3D shape data and appearance images of human faces.

Database	Format and resolution	Coverage	No. samples	Scanner
BU-4DFE[33]	triangle mesh (35k vertices), texture image (1,040×1,329)	face, neck, sometimes ears	101 individuals×six 100 frame expression sequences	Dimensional Imaging
Florence[34]	triangle mesh (40k vertices), 3341×2027 texture image	face, neck, sometimes ears	100+ individuals	3dMD
BP4D-Spontaneous[35]	triangle mesh (30k-50k vertices), texture image (1,040 × 1,329)	face, neck, sometimes ears	41 individuals×eight one minute dynamic sequences	Dimensional Imaging
FaceWarehouse[9]	raw: 640 × 480 RGBD; processed: triangle mesh (11k vertices, consistent topology)	face, some-times ears	150 individuals×20 expressions	Microsoft Kinect

Table 3. Common publicly available appearance images and 3DMM-based reconstructed shape of human faces.

Database	Format and resolution	No. samples	Image source
300W-3D/300W-3D-Face[27]	3DMM parameters/ fitted 3D mesh and 2d images	20,000 (training)+ 4386(testing)	300W[27]
AFLW2000-3D[15]	fitted 3D mesh and 2d images	2,000	AFLW[23]
LS3D-W[36]	fitted 3D mesh and 2d images	230,000	AFLW[23], 300VW[31], FDDB[37]

4 Discussion

The error of the 3D face reconstruction method is significant for face-related product design. Therefore, it is necessary to investigate the performance of different methods across poses and datasets, The Normalized Mean Error (NME) is proposed as the evaluation metric, which is:

$$\mathrm{NME} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|X_i - \hat{X}_i\right\|}{d} \tag{5}$$

where X_i is the i-th coordinate of the ground truth 3D face and $\hat{X}_i$ denotes the ith coordinate of the reconstructed 3D face. d is the normalization factor. Detailly, d can be the bounding box size and the outer interocular distance when comparing the performance of face alignment and face reconstruction respectively. Another evaluation metric is the Mean Error (ME), which can be calculated as the following formulation:

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}\left\|X_i - \hat{X}_i\right\| \tag{6}$$

The accuracy of the dense face alignment results on AFLW2000-3D is compared among 3DDFA[13], PRN[18] and GCNs[21]. 45K points are taken from the face shape reconstructed by these methods. The quantitative results of NME are illustrated in Table 4. The bounding box size is used as the normalization factor *d*. The RRN and GCNs have the similar accuracy, which outperform 3DDFA on the 3D face alignment tasks by a large margin.

The performance of the face reconstruction on Florence dataset is evaluated among 3DDFA[13], VRN[19], PRN[18] and GCNs[21].The most common 19K points with 3D coordinates are selected from the face region of all compared methods. Because the generated point clouds of different methods are not aligned, the Iterative Closest Points algorithm is adopted to find the nearest points between the network output and the ground truth point cloud [18]. The quantitative results of NME are illustrated in Table 4. The outer interocular distance of 3D coordinates is used as the normalization factor *d*. The performance of GCNs is slightly better than PRN and both of them outperform VRN and 3DDFA [18].

Table 4. Performance comparison of different methods, including 45K points face alignment on AFLW2000-3D and 19K points face reconstruction on Florence [18]

Database	3DDFA	VRN	PRN	GCNs
AFLW2000-3D	6.56	-	4.41	4.31
Florence	6.29	5.23	3.72	3.63

Besides, these training and testing face database includes different yaw angles from 0 to 90. The comparison results demonstrate that the performance of these method for small yaw angles (0~30) is better than medium (30~60) and large (60~90) yaw angles [18]. However, it is still necessary to confirm whether the performance of the 3D face reconstruction method can meet the requirement when designing face-related products. Fortunately, when applying by using 3D face reconstruction method into designing face related

products, there are two main important advantages and developments.

Firstly, the facial feature analysis of large-scale 3D faces could become faster and practical. We just need to use the single face image to reconstruct the 3D facial shape, rather than to capture their 3D face information by using 3D scanning system. However, these existing face reconstruction methods can reconstruct the facial shape effectively, but cannot estimate the facial dimension. Hence, we need to improve the existing face reconstruction methods to make it have the ability to estimate the facial shape and dimension together. If it can be realized, the dimension features and difference in terms of genders, ages and regions can be analyzed easily and fast by reconstructing 3D face for a large-scale of face images, which can provide an excellent reference for designers.

Although there are many publicly available faces databases which are used to train the model and evaluate the performance of 3D face reconstruction methods, they cannot be used to analyze the variance of face dimensions among all age and ethnic groups since they are limited and, in most cases, the facial data of a specific population is none. If we want to analyze the dimension features for designing face-related products of a selected population, we still need to establish a database of the selected population.

Secondly, the personalized customization of face-related products can also become easier and faster. The client just needs to upload their face image to the designer or seller and then they can reconstruct its 3D face to help them to design or select its suitable face-related products. With the development of camera techniques, it is a general trend that the phone will have RGB with depth cameras (RGB-D). The RGB single face image combined with the depth information may improve the accuracy of 3D face reconstruction, which also need to be studied and compared with these 3D face reconstruction methods based on single image further.

5 Conclusion

To realize the feature analysis of a large-scale 3D faces and the personalized customization of face-related products, the 3D face reconstruction from a single image is a practical approach. In this paper, the 3D face reconstruction methods are roughly classified into two categories, namely (1) 3DMM fitting based method and (2) End-to-end CNN based method. Applying 3D face reconstruction can be used into the design of face-related products as it is practical and cost effective. Furthermore, 3D face reconstruction methods can also be used to reconstruct the surface of the human body and other specific parts of the body, such as ears, hands and foots.

Acknowledgement

This work was supported by The Hong Kong Research Grants Council (RGC PolyU. 15603419).

References

1. Chu, C.-H., et al., Design customization of respiratory mask based on 3D face anthropometric data. International Journal of Precision Engineering and Manufacturing, 16(3): 487-494 (2015).
2. Lacko, D., et al., Ergonomic design of an EEG headset using 3D anthropometry. Applied ergonomics, 58: 128-136 (2017).
3. Skals, S., et al., Improving fit of bicycle helmet liners using 3D anthropometric data. International Journal of Industrial Ergonomics, 55: 86-95 (2016).
4. Long, J., M. Helland, and J. Anshel. A vision for strengthening partnerships between optometry and ergonomics. in HFESA 47th Annual Conference. (2011).
5. Blanz, V. and T. Vetter. A morphable model for the synthesis of 3D faces. in Siggraph. (1999).
6. Blanz, V. and T. Vetter, Face recognition based on fitting a 3d morphable model. IEEE Transactions on pattern analysis and machine intelligence, 25(9): 1063-1074 (2003).
7. Egger, B., et al., 3D Morphable Face Models--Past, Present and Future. arXiv preprint arXiv:1909.01815, (2019).
8. Jolliffe, I., Principal component analysis. Springer (2011).
9. Cao, C., et al., Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics, 20(3): 413-425 (2013).
10. Paysan, P., et al. A 3D face model for pose and illumination invariant face recognition. in 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. Ieee (2009).
11. Cristinacce, D. and T. Cootes, Automatic feature localisation with constrained local models. Pattern Recognition, 41(10): 3054-3067 (2008).
12. Cootes, T.F., G.J. Edwards, and C.J. Taylor, Active appearance models. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6): 681-685 (2001).
13. Zhu, X., et al., Face alignment in full pose range: A 3d total solution. IEEE transactions on pattern analysis and machine intelligence, 41(1): 78-92 (2017).
14. Dollár, P., P. Welinder, and P. Perona. Cascaded pose regression. in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE (2010).
15. Sun, Y., X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. in Proceedings of the IEEE conference on computer vision and pattern recognition. (2013).
16. Zhang, Z., et al. Facial landmark detection by deep multi-task learning. in European conference on computer vision. Springer (2014).
17. Jourabloo, A. and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. in Proceedings of the IEEE conference on computer vision and pattern recognition. (2016).
18. Feng, Y., et al. Joint 3d face reconstruction and dense alignment with position map regression network. in Proceedings of the European Conference on Computer Vision (ECCV). (2018).

19. Jackson, A.S., et al. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. in Proceedings of the IEEE International Conference on Computer Vision. (2017).
20. Zeng, X., X. Peng, and Y. Qiao. DF2Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction. in Proceedings of the IEEE International Conference on Computer Vision. (2019).
21. Wei, H., S. Liang, and Y. Wei, 3D Dense Face Alignment via Graph Convolution Networks. arXiv preprint arXiv:1904.05562, (2019).
22. Liu, Z., et al. Deep learning face attributes in the wild. in Proceedings of the IEEE international conference on computer vision. (2015).
23. Koestinger, M., et al. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. in 2011 IEEE international conference on computer vision workshops (ICCV workshops). IEEE (2011).
24. Burgos-Artizzu, X.P., P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. in Proceedings of the IEEE international conference on computer vision. (2013).
25. Belhumeur, P.N., et al., Localizing parts of faces using a consensus of exemplars. IEEE transactions on pattern analysis and machine intelligence, 35(12): 2930-2940 (2013).
26. Jeni, L.A., et al. The first 3d face alignment in the wild (3dfaw) challenge. in European Conference on Computer Vision. Springer (2016).
27. Sagonas, C., et al. 300 faces in-the-wild challenge: The first facial landmark localization challenge. in Proceedings of the IEEE International Conference on Computer Vision Workshops. (2013).
28. Zhu, X. and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. in 2012 IEEE conference on computer vision and pattern recognition. IEEE (2012).
29. Le, V., et al. Interactive facial feature localization. in European conference on computer vision. Springer (2012).
30. Messer, K., et al. XM2VTSDB: The extended M2VTS database. in Second international conference on audio and video-based biometric person authentication. (1999).
31. Shen, J., et al. The first facial landmark tracking in-the-wild challenge: Benchmark and results. in Proceedings of the IEEE international conference on computer vision workshops. (2015).
32. Wu, W., et al. Look at boundary: A boundary-aware face alignment algorithm. in Proceedings of the IEEE conference on computer vision and pattern recognition. (2018).
33. Yin, L., et al. A high-resolution 3d dynamic facial expression database. in IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands. (2008).
34. Bagdanov, A.D., A. Del Bimbo, and I. Masi. The florence 2D/3D hybrid face dataset. in Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. ACM (2011).
35. Zhang, X., et al., Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing, 32(10): 692-706 (2014).

36. Bulat, A. and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). in Proceedings of the IEEE International Conference on Computer Vision. (2017).
37. Jain, V. and E. Learned-Miller, Fddb: A benchmark for face detection in unconstrained settings. UMass Amherst technical report (2010).