ROBUST FACE TRACKING WITH A CONSUMER DEPTH CAMERA

Fei Yang[1]*, Junzhou Huang*[2]*, Xiang Yu*[1]*, Xinyi Cui*[1]*, Dimitris Metaxas*[1]

[1]Rutgers University [2]University of Texas at Arlington

ABSTRACT

We address the problem of tracking human faces under various poses and lighting conditions. Reliable face tracking is a challenging task. The shapes of the faces may change dramatically with various identities, poses and expressions. Moreover, poor lighting conditions may cause a low contrast image or cast shadows on faces, which will significantly degrade the performance of the tracking system. In this paper, we develop a framework to track face shapes by using both color and depth information. Since the faces in various poses lie on a nonlinear manifold, we build piecewise linear face models, each model covering a range of poses. The low-resolution depth image is captured by using Microsoft Kinect, and is used to predict head pose and generate extra constraints at the face boundary. Our experiments show that, by exploiting the depth information, the performance of the tracking system is significantly improved.

Index Terms— Face tracking, Depth camera

1. INTRODUCTION

Reliable tracking of 3D deformable faces is a challenging task in computer vision. For one, the shapes of faces change dramatically with various identities, poses and expressions. For the other, poor lighting conditions may cause a low contrast image or cast shadows on faces, which will significantly degrade the performance of the tracking system.

Most previous face tracking systems uses a single optical camera. Existing methods can be divided into two categories: appearance based methods and feature based methods. The appearance based methods use generative appearance models to capture the shape and texture variations of faces, such as active appearance models (AAMs) [1] [2] and 3D morphable models [3]. The deformation parameters can be estimated using gradient descent optimization. These methods may suffer from weak generalization capability due to lighting and texture variations.

The feature based methods track local facial features and apply global shape models to constrain the feature locations. Vogler et al. [4] developed a system that detects facial landmarks by using active shape models (ASMs) [5]. Zhang et al. [6] proposed a method using a combination of semantic features, silhouette features, and online tracking features, and estimated model parameters in the form of an energy minimization problem. In addition to linear shape subspaces, sparse learning based methods [7] [8] are also applied to model shape priors to handle complex shape variations and non-Gaussian errors. The feature based methods have better generalization for new faces, but they may still lose tracking due to the lack of semantic features and occlusions.

One of the main difficulties of visual descriptors on RGB data is the insufficient discrimination on shapes, textures and foreground objects. The problem gets even worse with poor lighting condition and changing of viewpoints. On the other hand, the depth information is invariant to color, texture and lighting, making it easier to differentiate foreground objects from background. Based on recent development of inexpensive depth cameras, there is rapidly growing interest in exploring depth information in vision systems.

A large body of previous work with depth cameras focus on estimating the gestures of human bodies [9] [10]. As 3D face models attract more attention for face recognition and animations [3] [11], some recently studies also use depth cameras to recover 3D face shapes and other characteristics. Fanelli et al. [12] developed a random forest algorithm to estimate head orientations from the range data. Cai et al.[13] developed a maximum likelihood solution to track face shapes from the noisy input depth data. Weise et al. [14] developed a realtime system to reconstruct 3D head shapes from the range data, which are used to generate face animations. Lai et al. [15] proposed a method for object recognition by combining both RGB and depth data. Other applications include 3D scene reconstruction [16] and indoor robotics [17], etc.

In this paper, we develop a framework to track face shapes by using both color and depth information. Since the faces in various poses lie on a nonlinear manifold, we build piecewise linear face models, each model covering a range of poses. The low-resolution depth image is captured by using Microsoft Kinect, and is used to predict head pose and generate extra constraints at the face boundary. KLT trackers [18] are employed to track individual facial landmarks, under global shape constraints.

The overview of our system is shown in Fig. 1. A kinect sensor is used to capture both RGB and depth data, and the depth image is used to estimate the face orientation. The face subspace of the closest pose is selected to constrain the shape of the face. We estimate the landmark locations using both

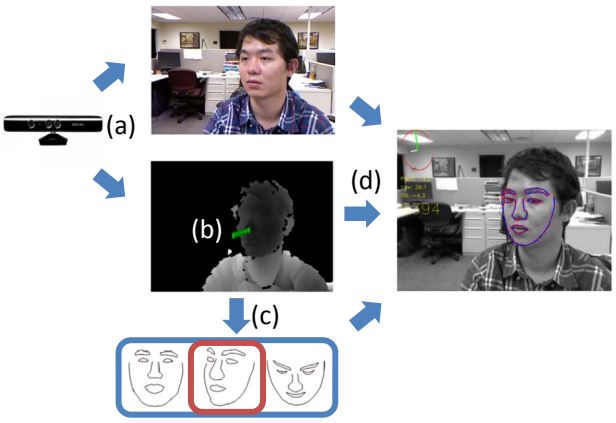

Fig. 1. Overview of our system. **(a)** A kinect sensor is used to capture both RGB and depth data. **(b)** The depth image is used to estimate the face orientation. **(c)** The face subspace of the closest pose is selected to constrain the face shape. **(d)** Both RGB and depth information are used to track the face.

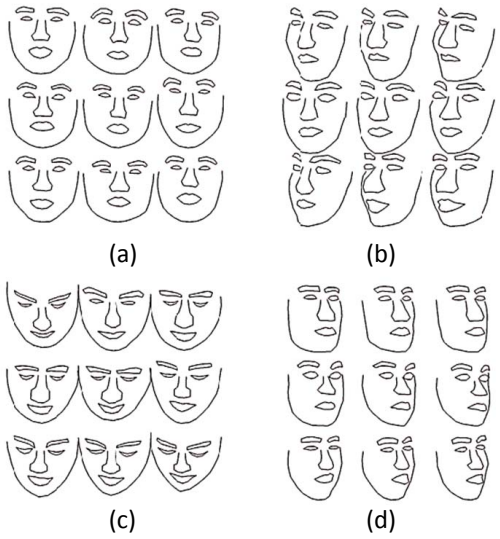

Fig. 2. Training set of the active shape models. **(a)** frontal; **(b)** left half profile; **(c)** lower; **(d)** right half profile.

the 1D gradient features described in [5] and face silhouette captured from the depth image. Finally the landmarks are tracked with the global shape constrains.

2. POSE ESTIMATION AND FACE SUBSPACES

We follow the work of Fanelli et al. [12] to estimate face orientation. In their approach, head pose estimation is formulated as a regression problem, and the pose parameters are estimated directly from the depth data. The regression is implemented within a random forest framework, learning a mapping from simple depth features to a probabilistic estimation of real-valued parameters such as 3D nose coordinates and head rotation angles. The system works in real-time on a frame-by-frame basis.

The face shapes under various poses lie on a hypersphere-like manifold. Traditional shape models including active shape models (ASMs) [5] and active appearance models (AAMs) [19] impose the variation of face shapes in a linear subspace, which cannot handle large pose changes. To solve this problem, we train piecewise linear ASM models. The model is switched during the tracking procedure, based on the head pose estimated from the depth data.

We train a total of 7 ASM models, i.e., frontal, half profile (left and right), full profile (left and right), upper, and lower. Each model covers a range of face poses. Some examples of the training faces are shown in Fig. 2.

3. FACE TRACKING

Traditional shape fitting methods use the mean shape as the initial shape to be placed on the face image. The current shape is iteratively modified by updating each landmark position and then constraining the overall shape to lie on the shape subspace. The active shape model and its recent extension [20] use 1D and 2D profile models to locate the approximate position of each landmark by template matching. Any template matcher can be used, for instance, the classical ASM forms a fixed-length normalized gradient vector (called the profile) by sampling the image along a line orthogonal to the shape boundary at the landmark. However, the gradient features often suffer from lack of discrimination, especially under poor lighting conditions and complex backgrounds. The landmarks on face boundaries are often misaligned to background points with stronger gradients.

We propose a new local fitting method to solve this problem. For face boundary landmarks, instead of only matching the Mahalanobis distance with the profile model, we also estimate the edge information in the depth map. The score of each position along normal of boundary is defined as

$$d^2 = (g - \bar{g})^T S_g^{-1} (g - \bar{g}) + \lambda || \bigtriangledown I_D ||^2 \qquad (1)$$

where g is the gradient of current position. $\bar{g}$ and S_g are the mean and covariance matrix acquired during the training phase. We use an additional term $|| \bigtriangledown I_D ||^2$ to estimate the edge intensity of the depth map I_D. λ is the parameter controlling the trade-off between two terms.

Running ASM in every frame is computationally expensive and causes jittering. To solve this problem, we track the features using the KLT tracker [18] across consecutive frames. The KLT tracker is a method for registering two local features and computes the displacement of the feature by minimizing the intensity matching cost. It efficiently gives the new location of each landmark in the next frame. The

landmark coordinates are projected into the shape subspace selected by the pose estimator. The positions are further updated by using the global shape constraints.

4. EXPERIMENTS

We test our system with a subject sitting in front the kinect sensor and performing different head poses and face expressions. The background is static, but with rich textures. As shown in Fig. 3, our system could effectively track the facial landmarks, while disabling the depth data often leads to lost tracking in this scenario.

5. CONCLUSION

In this paper, we develop a framework to track face shapes by using both color and depth information. Since the faces in various poses lie on a nonlinear manifold, we build piecewise linear face models, each model covering a range of poses. The low-resolution depth image is captured by using Microsoft Kinect, and is used to predict head pose and generate extra constraints at the face boundary. Our experiments show that, by exploiting the depth information, the performance of the tracking system is significantly improved.

6. REFERENCES

[1] Jing Xiao, Simon Baker, Iain Matthews, and Takeo Kanade, "Real-time combined 2D+3D active appearance models," in *Proc. CVPR*, 2004, pp. 535–542.

[2] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. ICCV*, 2009, pp. 1034–1041.

[3] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. SIGGRAPH*, 1999, pp. 187–194.

[4] Christian Vogler, Zhiguo Li, Atul Kanaujia, Siome Goldenstein, and Dimitris N. Metaxas, "The best of both worlds: Combining 3D deformable models with active shape models," in *Proc. ICCV*, 2007, pp. 1–7.

[5] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[6] Wei Zhang, Qiang Wang, and Xiaoou Tang, "Real time feature based 3-D deformable face tracking," in *Proc. ECCV*, 2008, pp. 720–732.

[7] Shaoting Zhang, Yiqiang Zhan, Maneesh Dewan, Junzhou Huang, Dimitris N. Metaxas, and Xiang Sean Zhou, "Towards robust and effective shape modeling: Sparse shape composition," *Medical Image Analysis*, vol. 16, no. 1, pp. 265–277, 2012.

[8] Fei Yang, Junzhou Huang, and Dimitris N. Metaxas, "Sparse shape registration for occluded facial feature localization," in *Proc. the 9th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2011, pp. 272–277.

[9] Jamie Shotton, Andrew W. Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, 2011, pp. 1297–1304.

[10] Mao Ye, Xianwang Wang, Ruigang Yang, Liu Ren, and Marc Pollefeys, "Accurate 3D pose estimation from a single depth image," in *Proc. ICCV*, 2011, pp. 731–738.

[11] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas, "Expression flow for 3D-aware face component transfer," *ACM Trans. Graphics*, vol. 30, no. 4, pp. 60, 2011.

[12] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc J. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. DAGM-Symposium*, 2011, pp. 101–110.

[13] Qin Cai, David Gallup, Cha Zhang, and Zhengyou Zhang, "3D deformable face tracking with a commodity depth camera," in *Proc. ECCV*, 2010, pp. 229–242.

[14] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly, "Realtime performance-based facial animation," *ACM Trans. Graphics*, vol. 30, no. 4, pp. 77, 2011.

[15] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox, "Sparse distance learning for object recognition combining RGB and depth information," in *Proc. ICRA*, 2011, pp. 4007–4013.

[16] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard A. Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew J. Davison, and Andrew W. Fitzgibbon, "Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. the 24th Annual ACM Symposium on User Interface Software and Technology*, 2011, pp. 559–568.

[17] João Cunha, Eurico Pedrosa, Cristóvão Cruz, António J. R. Neves, and Nuno Lau, "Using a depth camera for indoor robot localization and navigation," in *Robotics Science and Systems 2011 Workshop on Advanced Reasoning with Depth Cameras*, 2011.

[18] Jianbo Shi and Carlo Tomasi, "Good features to track," in *Proc. CVPR*, 1994, pp. 593 – 600.

Fig. 3. Face tracking results

[19] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[20] Stephen Milborrow and Fred Nicolls, "Locating facial features with an extended active shape model," in *Proc. ECCV*, 2008, pp. 504–513.