3D Face Tracking for Infant Monitoring Using Dense HOG and Drift Reduction

Ronald W.J.J. Saeijs[a], Walther E. Tjon a Ten[b], and Peter H.N. de With[a]
[a]Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands
[b]Department of Pediatrics, Máxima Medical Center Veldhoven, The Netherlands

Abstract

This paper presents a new algorithm for 3D face tracking intended for clinical infant pain monitoring. The algorithm uses a cylinder head model and 3D head pose recovery by alignment of dynamically extracted templates based on dense-HOG features. Drift reduction is obtained from re-registration in combination with multi-pose state estimation by a square-root unscented Kalman filter. Results on videos of moving infants in hospital show good tracking for poses up to 50 degrees from upright-frontal, with mean eye-location error relative to inter-ocular distance below 9% .

1. Introduction

This paper considers infant face tracking for video analysis of facial pain expression in a clinical context, e.g. for monitoring infants suspected of gastro-esophageal reflux disease (GERD). In this context, the tracking problem has three characteristics that existing solutions cannot handle. Firstly, with infants, and especially very young ones, facial texture is less pronounced than with adults. For example, infants do not have prominent eyebrows and their eyes are often closed. Secondly, parts of the face may be covered. For example, in case of monitoring for GERD, there are plasters on the face and a tube into the nose. Also, infants may have a pacifier in their mouth, with a large variety of appearances, and cuddles, blankets, etc. may partly occlude the face. Thirdly, infants being monitored are not oriented towards a camera, and more cameras may be needed to keep the face in view. For example, in case of monitoring for GERD, infants are in bed where they can shift and turn their head. As a result, tracking has to handle wide pose ranges.

We propose a face tracking algorithm to accommodate the above characteristics. Focusing on the first two (texture and occlusion), we give its single-camera version. However, it is developed to extend to multiple cameras so as to fully accommodate the third characteristic. For details see [3, 4].

Ronald Saeijs is supported by STW in project 13335 GARDIAN.

2. Approach and related work

We model the face as part of the head and solve the more general problem of tracking 3D head pose. This enables maximum use of image information for robustness, because visible features of both face and non-face parts of the head can be used. It also allows to extend to multi-camera setups.

For general face tracking, state-of-the-art approaches are based on aligning a deformable face model to input images, with models covering all variations of shape and appearance. This is not feasible here, because of wide pose ranges, unknown appearances of plasters, pacifiers, etc. Instead, we follow other approaches that recover head motion using an assumed 3D head shape. In particular, we follow [5] where Lucas-Kanade template alignment is used with a cylinder head model. As innovation, we use densely sampled HOG features (instead of pixel intensities), which can improve Lucas-Kanade alignment [1]. In addition, we use insights from [2] to define a probabilistic version of drift reduction.

3. Basic tracking algorithm

We assume full-perspective image projection, while modeling the infant head as a 3D cylinder, with its pose defined by a 3D rigid-body transformation (using a 6-dim. vector $[\omega_x, \omega_y, \omega_z, t_x, t_y, t_z]$ of exponential coordinates).

Our aim for tracking is to estimate the pose of the head in each new image by aligning templates derived from older images. Here, a template is a set of 3D points on the head surface with associated appearance values and the intuition that it represents a textured part of the head. We can derive a template from a given image with a given pose, by reverse-projecting pixel locations (with their appearance values) onto the head surface. For initialization, we need the pose in the initial image and a specification of the head surface.

As building block for our algorithm, we use an alignment step that takes an image and a template and a start pose. It yields a new pose that aligns the template with the image by minimizing the sum-of-weighted-squared-errors of its 2D projection. To implement this, we use the Lucas-Kanade (LK) method for gradient descent optimization with

template point weights $w = w_D \cdot w_R$ adapted per LK-iteration. Here, w_D relates to image projection of the head (so that points seen from the side contribute less than points seen from the front), and w_R relates to the IRLS method of [5] to handle noise, non-rigid motion and occlusion (so that points with large errors for the current pose contribute less). For robustness, every alignment step is preceded by outlier removal. For removal, we pick a reference image and use the pose computed earlier for this image to project the template and remove template points with a large error. In the basic algorithm we use one alignment step for each image, with a template derived from the preceding image and using its predecessor as reference for outlier removal.

As a main innovation from [5], we have re-defined all of the above to use dense-HOG features instead of pixel intensities. As shown in [3], this significantly improves accuracy.

4. Algorithm extensions for drift reduction

The basic algorithm from Section 3 may drift because alignment errors accumulate during tracking. We introduce two options for extending the algorithm for drift reduction. Both options add a second alignment step for each image. This second step uses another template that is derived from a key image with a pose that is 'close' to the pose resulting from the first alignment step. For the second step, we use the pose resulting from the first alignment step as start pose, and the preceding image as reference for outlier removal. In both options, we select a limited set of key images during tracking, in such a way that their poses are just 'far' enough apart. (To define 'close' and 'far', we set thresholds on the differences of both the angles and the positions of poses.) The two options differ in treating key and reference poses as either fixed values or probabilistic estimates, as follows:

For the first algorithm option, we output poses per image and consider them fixed forever. As a result, templates from key images are derived once and then stored and re-used.

For the second option, we consider poses as probabilistic estimates. Based on Method 3 in [2], the idea is to keep improving estimates for key images and reference images that may still be used in the future. For this, we define a 'multi-pose' vector composed from the exponential coordinates of the poses associated with the (dynamically updated) set of key images and reference images. We model this multi-pose as a multivariate Gaussian distribution and we maintain the combination of its mean and covariance as state variable in our algorithm. We now consider our probabilistic algorithm as a recursive state estimator that uses alignment steps as its measurement steps. With a simple state transition model (Brownian head motion in exponential-coordinate space from the most recent image to its successor), we can use Kalman filtering to predict (after each alignment step) a new multi-pose state. Because of the non-linearities involved, we have to use a square-root unscented Kalman form.

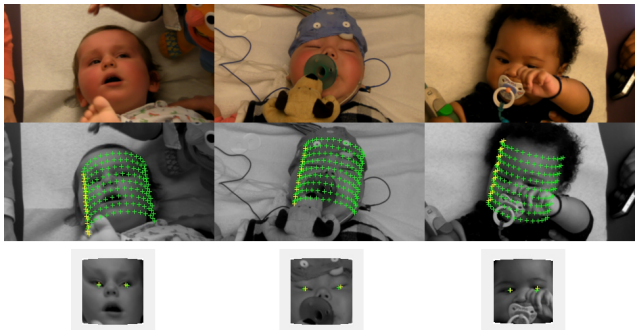

Figure 1: Example frames. Left to right: Sequence 2, 8, 9. Top: original. Mid: input with front-of-posed-cylinder output (green visible, yellow not). Bottom: normalized front view (green eye location, yellow ground truth).

5. Experimental results and conclusions

For experiments, we used video from handheld cameras of infants in bed, relaxed or in pain (see Figure 1). Table 1 shows accuracy for 10 sequences with angle θ of rotation from upright-frontal (with respect to the camera) limited to 50 degrees so that both eyes are visible. (Note: one θ-value corresponds to many combinations of yaw, pitch, and roll.) For this, we use ELE (eye location error relative to inter-ocular distance) in a normalized front view (cf. Figure 1). It shows good tracking behavior, with mean ELE below 0.09.

nr.	length (#frames)	estimated θ (degrees)		with re-use (ELE)		probabilistic (ELE)	
		min	max	mean	max	mean	max
1	982	7.9	48.0	**0.026**	**0.073**	**0.026**	**0.073**
2	750	13.4	47.1	**0.046**	**0.137**	0.049	0.142
3	430	5.8	34.0	**0.071**	**0.141**	0.085	0.193
4	739	7.0	22.5	**0.037**	0.097	**0.037**	**0.095**
5	458	8.4	43.6	0.044	0.221	**0.033**	**0.113**
6	1274	0.7	41.9	0.051	0.180	**0.047**	**0.158**
7	1007	3.7	48.6	0.021	0.087	**0.020**	**0.083**
8	466	21.5	32.4	**0.039**	**0.080**	0.042	0.087
9	760	13.7	32.9	**0.054**	**0.126**	0.055	0.127
10	480	20.8	31.9	0.048	0.110	**0.044**	**0.103**

Table 1: Tracking accuracy.

References

[1] E. Antonakos, J. Alabort-i Medina, G. Tzimiropoulos, and S.P. Zafeiriou. Feature-Based Lucas-Kanade and Active Appearance Models. *IEEE Trans. Image Process.*, 24(9):2617–2632, 2015. 1

[2] A. Rahimi, L.-P. Morency, and T. Darrell. Reducing drift in differential tracking. *Computer Vision and Image Understanding*, 109(2):97–111, feb 2008. 1, 2

[3] R.W.J.J. Saeijs, W.E. Tjon a Ten, and P.H.N. de With. Dense-HOG-based 3D face tracking for infant pain monitoring. In *Int. Conf. Image Processing ICIP*, pages 1719–23, 2016. 1, 2

[4] R.W.J.J. Saeijs, W.E. Tjon a Ten, and P.H.N. de With. Dense-HOG-Based Drift-Reduced 3D Face Tracking for Infant Pain Monitoring. In *Int. Conf. Machine Vision ICMV*, 2016. 1

[5] J. Xiao, T. Moriyama, T. Kanade, and J.F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *Int. Journ. Imaging Systems and Technology*, 13(1):85–94, 2003. 1, 2