Dual-Camera 3D Head Tracking for Clinical Infant Monitoring

Ronald W.J.J. Saeijs
Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands
r.w.j.j.saeijs@tue.nl

Walther E. Tjon a Ten
Department of Pediatrics
Máxima Medical Center Veldhoven
Veldhoven, The Netherlands
w.tjonaten@mmc.nl

Peter H.N. de With
Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands
p.h.n.de.with@tue.nl

***Abstract*—This paper presents a new algorithm for dual-camera 3D head tracking, intended for clinical infant monitoring. The paper includes a brief motivation with reference to the state-of-the-art in face-related image analysis. The proposed algorithm uses a clipped-ellipsoid head model and 3D head pose recovery by joint alignment of paired templates based on dense-HOG features. In the algorithm, template pairs are dynamically extracted and a limited number of template pairs are stored and re-used for drift reduction. We report experimental results on real-life videos of infants in bed in a hospital, captured in visual light as well as near-infrared light. Results show consistently good tracking behavior. For challenging video sequences, the mean tracking error in terms of endocanthion location error relative to the innercanthal distance remains below 30%. This error has proven to be sufficiently low for 3D head tracking to support infant face analysis. For this reason, the proposed algorithm is used successfully in an infant monitoring system under development.**

***Index Terms*—3D head tracking, dual camera, dense HOG, infant monitoring**

I. Introduction

Monitoring children for pain and discomfort is important in many clinical contexts. For example, in the context of gastro-esophageal reflux disease (GERD), infants suspected of GERD routinely undergo 24-hour reflux monitoring, and pain monitoring will allow to analyze detailed time relations between pain and reflux to improve diagnosis. Recent work on detecting discomfort [11] and acute pain [19] of infants shows that automatic continuous monitoring can be based on video analysis of facial expression (cf. [17]).

Our objective is to develop a video analysis system for in-bed monitoring of infants in a clinical setting. This system has two significant characteristics with respect to video capture. Firstly, we use dual-camera capture, in order to ensure that the face is visible for a large range of poses of an infant's head. Secondly, we capture video from visual light and, alternatively, from near-infrared light (when infants are sleeping in the dark). For an example of infrared images and a typical dual-camera configuration, see Figure 1.

This paper concentrates on the problem of 3D head tracking in the context of dual-camera monitoring. In this context, the task of 3D head tracking is to determine the 3D pose of the head for all images of a synchronized dual-camera video sequence, given an estimated shape of the head and an estimated initial pose. Here, 3D pose has 6 degrees of freedom with reference to the world coordinate system, which is used to specify the calibrated fixed positions of the two cameras.

In order to explain the relevance of the work reported here, we briefly present a perspective with reference to the state-of-the-art in face-related image analysis. In recent years, large progress has been made in the areas of face detection [22], facial landmark localization (also known as face alignment or facial feature detection) [9], and facial landmark tracking [18] [21]. As discussed in [5], the current practice for landmark tracking is to combine a generic face detection algorithm and a generic landmark localization algorithm. With recent advances in using deep learning for these tasks, remarkable performance can now be achieved. As argued in [3], accuracy for 2D and 3D landmark localization now seems close to saturating existing datasets for in-the-wild faces. However, as yet we cannot straightforwardly solve facial landmark tracking for our monitoring application in this manner. This is because the domain of face appearances in our setting is not sufficiently covered by existing in-the-wild datasets (see Section II for the main differences). At the same time, we only have a relatively small (though slowly growing) dataset from our setting, which only very sparsely (and hence also insufficiently) covers our domain of potential face appearances.

For the system that we are developing, we consider two ways to tackle the above-mentioned problem of the appearance domain for facial landmarking. Firstly, we investigate possibilities for enlarging the set of appearances for training (so that transfer learning becomes possible). For this, we try to artificially augment our data, a.o. by means of 'face profiling' as introduced in [23]. Secondly, in our current system, we condense the domain of appearances for both training and testing, by factoring out some of the pose-related variations. For this, we employ components for landmark localization in a slightly altered manner, making use of additional 3D-pose information obtained from 3D head tracking. In this paper we focus exclusively on how to reliably obtain this additional information. Therefore, we discuss 3D head tracking here as a stand-alone problem.

Ronald Saeijs was supported by STW in project 13335 GARDIAN.

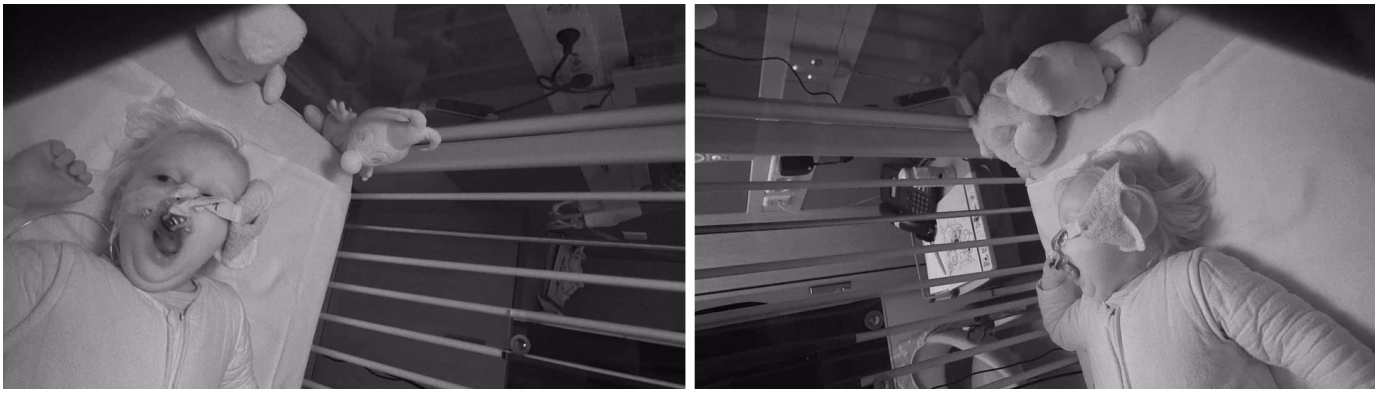

Fig. 1: Example of a pair of infrared input images (and tube, plasters, and pacifier) from sequence 2 in Table I. Here the left image is the face-view image. Cameras are positioned as follows: the camera of the left image is positioned between the bars in the lower left corner of the right image, and vice versa.

The problem of 3D head tracking in our context is special because of the use of two cameras, and it is complicated by challenges relating to facial texture, illumination, occlusion, etc., as discussed in Section II. In this paper, we present a new algorithm that handles all of these complications, based on simultaneous alignment of dynamically extracted dense-HOG templates. Our main contributions are the definition of the refined algorithm and its evaluation for real-life infant monitoring conditions.

Below, Section II points out the main challenges posed by our setting and discusses our 3D head tracking approach in relation to other work. Section III contains definitions, models, and notations that are used for a compact description of the 3D tracking algorithm in Section IV. Finally, Section V presents experiments and results, with conclusions in Section VI.

II. Challenges, approach, and related work

In comparison to typical in-the-wild conditions for face-related image analysis, our setting of clinical infant monitoring poses several additional challenges. Below we discuss them in two groups, viz. challenges related to facial appearance in individual images and those related to image sequences.

Considering individual images, there are three main aspects in which infant faces in our clinical context may be different. Firstly, in comparison with adults and older children, faces of young infants have less pronounced texture (in addition to somewhat different shape proportions). For example, they have no prominent eyebrows, wrinkles or creases. Also, infants in bed will have their eyes closed for long periods of time. Secondly, in clinical settings parts of the face may be covered. For example, in case of monitoring for GERD, the face has plasters and a tube in the nose for a probe in the esophagus. In addition, infants may have a pacifier in their mouth, resulting in a large variety of appearances. More generally, cuddles, toys, or blankets may also partly occlude the face. Thirdly, as seen from bedside cameras, the range of head poses is very wide, with sometimes non-uniform illumination from directions that do not occur in typical indoor or outdoor in-the-wild settings. For example, camera viewpoint or illumination may also be from below (with reference to the face).

Considering image sequences, there are two additional challenges in our setting. Firstly, sometimes infant head movements (and especially head turns) may be very fast. For monitoring pain and discomfort, episodes of fast movements are often very relevant, as for example episodes of coughing in case of monitoring for GERD. Secondly, occasionally fast-changing occlusions may occur, for example when an infant moves a hand or arm in front of its face.

In view of the above, the essence of the tracking approach that we need is that it does not rely on a pre-determined model of appearance. Among existing approaches that meet this requirement (cf. [13]), many use a pre-determined model of a rigid 3D head shape for recovering head motion. We adopt the same principle, because it allows maximum use of image information for robustness, since visible features of both face and non-face parts of the head can be used. Existing approaches based on this principle have broad variations, both in terms of head shape and in terms of motion recovery.

With respect to head shape, the variations of rigid-head 3D tracking approaches range from simple geometric shapes such as cylinders (e.g. [20] [8] [7]) and ellipsoids (e.g. [10]) to more complicated generic and morphable 3D models (e.g. [4]). For our case, we use a clipped ellipsoid, for three reasons. Firstly, it is a simple shape that can approximately match the actual head shape, even when posed at a slightly inaccurate angle. This is important because small pose errors may result from pose initialization, especially when initialization starts from a difficult, non-frontal, camera view. Secondly, an ellipsoid approximates the most relevant part of an infant head (in between top and bottom) well, reflecting that it is more curved (less elongated) than an adult head. Thirdly, by clipping we exclude the neck and the top of the head, where simple shape approximation is least accurate.

With respect to motion recovery (for a given head shape), two main variants exist. Some concepts for tracking (e.g. [8] for single-camera tracking and [4] for multi-camera tracking) recover motion by key point matching. For our application, this is not feasible, because infant heads often yield few and unstable key points. Other tracking systems (e.g. [20] [10] [7]) recover motion by Lucas-Kanade template alignment [12]. In [15] and [16] we have introduced a new single-camera variant

that uses densely sampled Histogram-of-Oriented-Gradient ('dense-HOG') features [6], instead of pixel intensities for a template. This improves Lucas-Kanade alignment [1] [15]. For our case here, we also use a dense-HOG-based variant, because it yields better tracking for less-pronounced texture and less-uniform illumination, e.g. for a sleeping infant in bed.

For practical systems, the number of cameras is an important factor for 3D head tracking. Most approaches were developed for a single camera or, in a few cases, two (stereo) or three cameras positioned close to each other. This is because many applications (e.g. in human-computer interaction or vehicle driver monitoring) are concerned with tracking an adult with a natural orientation (e.g. towards a screen or a steering wheel). Very few 3D head tracking approaches were developed for multiple cameras positioned far from each other. This was done in [4] for uncalibrated positions and in [7] for known camera positions. In our case, we use calibrated positions and we apply the principle of joint template alignment from [7], because it allows to use image information from both cameras simultaneously.

Overall, our dual-camera 3D head tracking algorithm uses joint camera alignment from [7] and is based on [16] in its use of dense HOG, weighting, trimming, and drift reduction.

III. Definitions, models, and notations

A. Cameras, images, and camera projection

For in-bed monitoring, we need cameras positioned such that the face is visible for a maximal pose range. For this reason, we use two cameras with wide-angle lenses viewing between the bars of the bed sides. In order to eliminate lens distortions, images are pre-processed first, and we only consider their corrected versions from here on. For an image I, we represent 2D image locations as $\mathbf{u} = [u, v]^\mathsf{T} \in \mathbf{U}$, where $\mathbf{U} = \mathbb{R}^2$, and we define $\mathbf{U}(I) \subset \mathbf{U}$ to denote its set of pixel locations (and, in general, we may add indices $c \in \{1, 2\}$ to refer to a specific camera, e.g. I_c). We use gray-level images only, in order to treat visual-light capture and infrared capture alike, and images are captured by the two cameras as pairs at the same time instant.

For the undistorted images, we model the output image of the camera as a full-perspective projection of the 3D scene onto the focal plane. Representing 3D points in homogeneous coordinates as $\mathbf{x} = [x, y, z, 1]^\mathsf{T} \in \mathbf{X}$, with $\mathbf{X} = \mathbb{R}^3 \times \{1\}$, we define this camera projection as a function W from 3D points to 2D image locations, as follows:

$$W(\mathbf{x};\mathbf{C}) = [u', v']^\mathsf{T}/s', \quad \text{where} \quad [u', v', s']^\mathsf{T} = \mathbf{C}\mathbf{x}. \tag{1}$$

Here, $\mathbf{C}$ is the 3×4 camera projection matrix that combines the intrinsic and extrinsic characteristics of a given camera.

B. Head, pose, visibility, and moves

As our model for the head, we choose a clipped ellipsoid. We define the un-posed head as an oriented 3D surface with centroid at $[0, 0, 0, 1]^\mathsf{T}$. From this, we obtain the posed head by applying a 3D rigid-body transformation $g \in SE(3)$. To represent a head pose, we use the 6-dimensional vector $\mathbf{p} = [\omega_x, \omega_y, \omega_z, t_x, t_y, t_z]^\mathsf{T} \in \mathbb{R}^6$ of exponential coordinates of g. From $\mathbf{p}$ we obtain the homogeneous transformation matrix $\mathbf{G}$ of g as follows [14]:

$$\mathbf{G} = \mathrm{e}^{\widehat{\mathbf{p}}} = \mathbf{I} + \widehat{\mathbf{p}} + \frac{\widehat{\mathbf{p}}^2}{2!} + \frac{\widehat{\mathbf{p}}^3}{3!} + \cdots, \tag{2}$$

where $\mathbf{I}$ is the 4×4 identity matrix and operator $\widehat{\ }$ is defined by

$$\widehat{\mathbf{p}} = \begin{bmatrix} 0 & -\omega_z & \omega_y & t_x \\ \omega_z & 0 & -\omega_x & t_y \\ -\omega_y & \omega_x & 0 & t_z \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3}$$

For practical calculation of the exponential mapping $\mathrm{e}^{\widehat{\mathbf{p}}}$ and its inverse $\widetilde{\log}\,\mathbf{G}$, see [14].

A pose $\mathbf{p}$ transforms a surface point $\mathbf{x}$ to a posed-head point $\mathrm{e}^{\widehat{\mathbf{p}}}\mathbf{x}$, so that projection by a camera $\mathbf{C}$ yields an image location $W(\mathrm{e}^{\widehat{\mathbf{p}}}\mathbf{x};\mathbf{C})$. For our algorithm we re-write this as $W(\mathbf{x};\mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}})$, which associates the rigid-body transformation with the camera (instead of the head points) and allows to keep all head point references $\mathbf{x}$ in the un-posed coordinate system centered at $[0, 0, 0, 1]^\mathsf{T}$. The equivalence follows from (1).

For every combination of camera and pose, some part of the head will not be visible. Therefore, camera projections of 3D points on the invisible side of the head surface should be ruled out. For compact description of our algorithm, we rule out such points by introducing a non-negative weight function $w_D(\mathbf{x};\mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}},..)$ that yields zero for invisible points (with values for other points defined later in Section IV-B). As for W, we keep references $\mathbf{x}$ for w_D in the un-posed coordinate system. Note that the ',..'-notation indicates that more parameters are involved, as w_D is shape-dependent.

For convenient description of tracking, we also use the un-posed coordinate system as reference for head motion. For this, we introduce the concept of a *move* with value $\Delta\mathbf{p}$. We define this as a pose update that results from applying a rigid-body transformation $\mathrm{e}^{\widehat{\Delta\mathbf{p}}}$ to the un-posed head. For a pose variable $\mathbf{p}$, it corresponds to changing $\mathbf{p}$ as implied by

$$\mathrm{e}^{\widehat{\mathbf{p}}} \leftarrow \mathrm{e}^{\widehat{\mathbf{p}}}\mathrm{e}^{\widehat{\Delta\mathbf{p}}}. \tag{4}$$

C. Dense-HOG appearance and templates

For our tracking, we consider head appearance in terms of dense-HOG features instead of pixel intensities. In this, we follow [15] where dense HOG was shown to yield more accurate single-camera tracking.

We use the same HOG variant as [15], based on blocks of 2×2 cells, cells of 8×8 pixels, and 9-bin histograms for signed orientations. Using trilinear interpolation of location and orientation, plus L^2-*Hys* normalization [6], it yields a normalized 36-dimensional vector for every pixel location. We use this vector as our definition for the appearance value $\mathbf{a} \in \mathbf{A}_1$ at pixel locations, so that $\mathbf{A}_1 = \{\mathbf{a} \in \mathbb{R}^{36} \mid \|\mathbf{a}\|_2 = 1\}$. For convenience, we will use the notation $I(\mathbf{u})$ to denote the appearance value for an image I at pixel location $\mathbf{u} \in \mathbf{U}(I)$. Generalizing this to non-pixel locations, we define $I(\mathbf{u})$ for

$\mathbf{u} \in \mathbf{U} \setminus \mathbf{U}(I)$ by means of linear inter-/extrapolation of pixel appearance values. As a consequence, we define the general set of appearance values as $\mathbf{A} = \mathbb{R}^{36}$.

Using the above, we introduce a dense-HOG template with the intuition that it represents a textured part of the head. We define a template $T \subset \mathbf{X} \times \mathbf{A}$ as a set $\mathbf{X}(T)$ of 3D points situated on the un-posed head surface with associated appearance values. For a template point $\mathbf{x} \in \mathbf{X}(T)$, we will use $T(\mathbf{x})$ to denote its associated appearance value.

We extract a template from an image, using reverse projection of head pixels for a presumed value of the head pose. To define this, we use the fact that camera projection maps visible 3D points on the head surface one-to-one to 2D head locations in the image. Reverse camera projection is then the inverse mapping, which takes 2D head locations to 3D points on the head surface. When applied to pixel locations only, this inverse mapping yields a finite set of 3D points on the un-posed head surface. The extracted template is defined by associating this set of 3D points with the appearance values of the originating pixels.

For a given combination of camera $\mathbf{C}$ and pose $\mathbf{p}$, some part of a template may be invisible. In order to rule out this part, we define a convenient notation for the set Ω of visible points of a template T, as follows:

$$\Omega(T, \mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}}) = \big\{ \mathbf{x} \big|\ \mathbf{x} \in \mathbf{X}(T)\ \wedge\ w_D(\mathbf{x}; \mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}}) > 0 \big\}. \quad (5)$$

IV. Tracking algorithm

A. Dual-camera alignment move

Our aim for tracking is to estimate the pose of the head in each new pair of images by aligning pairs of templates extracted from earlier image pairs. For this, we introduce an alignment move as a main building block for the algorithm.

The intuition for an alignment move is that it moves the head from a coarsely estimated head pose to a new pose that better approximates the true pose in a pair of images. For this, it takes a pair of templates to serve as an ad-hoc textured 3D approximation of the head, with each template approximating the part that is visible for its corresponding camera. The paired templates are then jointly moved, using 6 degrees of freedom, to obtain the best match with the true head appearance in the paired images.

More formally, we define a dual-camera alignment move (for a pair of images I_c of cameras $c \in \{1, 2\}$) as a process that takes a pair of templates T_c and an estimated pose $\mathbf{p}$. Its desired output is an updated value for $\mathbf{p}$ such that, for all template points $\mathbf{x}$, image appearances $I_c(\mathbf{u})$ at projected locations $\mathbf{u} = W(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}})$ are close to template appearances $T_c(\mathbf{x})$. We formalize this as an optimization for minimum sum-of-weighted-squared-errors cost that yields a move $\Delta\mathbf{p}$ (using the concept of 'move' that we defined in Section III-B):

$$\Delta\mathbf{p} = \underset{\Delta\mathbf{p}}{\operatorname{argmin}} \sum_{c=1}^{2} \sum_{\mathbf{x} \in \Omega(T_c, \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}\mathrm{e}^{\widehat{\Delta\mathbf{p}}})} w_c(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}\mathrm{e}^{\widehat{\Delta\mathbf{p}}}, ..) \cdot \Big\| I_c(W(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}\mathrm{e}^{\widehat{\Delta\mathbf{p}}})) - T_c(\mathbf{x}) \Big\|_2^2. \quad (6)$$

Here, w_c is a non-negative weight function that adjusts the contribution of individual template points from camera c in the alignment process. It is further defined in Section IV-B.

To implement an alignment move, we employ the Lucas-Kanade (LK) method of gradient descent optimization [12]. For this, we adopt the standard LK formulation in [2] with two modifications. Firstly, we use a 3D→2D motion model by defining new warp functions $\mathcal{W}_c$ with more parameters, as follows: $\mathcal{W}_c(\mathbf{x}; \Delta\mathbf{p}, \mathbf{p}, \mathbf{C}_c) = W(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}\mathrm{e}^{\widehat{\Delta\mathbf{p}}})$. Secondly, we use a 3D-compatible parameter update, as defined in (4).

The intuition for LK is that it approximates alignment by iteration, resulting in a sequence of smaller moves. Each iteration updates $\mathbf{p}$ by a move $\Delta\mathbf{p}_{LK}$, which is defined as the approximation for $\Delta\mathbf{p}$ that results from (6) after first-order Taylor expansion of $W(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}\mathrm{e}^{\widehat{\Delta\mathbf{p}}})$ for small $\Delta\mathbf{p}$. This is:

$$\Delta\mathbf{p}_{LK} = -\mathbf{H}^{-1} \sum_{c=1}^{2} \sum_{\mathbf{x} \in \Omega(T_c, \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}})} w_c(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}, ..) \cdot \Big[\nabla I_c \frac{\partial \mathcal{W}_c}{\partial \Delta\mathbf{p}} \Big]^{\mathsf{T}} \Big[I_c(W(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}})) - T_c(\mathbf{x}) \Big]. \quad (7)$$

Here, ∇I_c is the gradient $[\frac{\partial I_c}{\partial u}, \frac{\partial I_c}{\partial v}]$ of image I_c, and $\frac{\partial \mathcal{W}_c}{\partial \Delta\mathbf{p}}$ is the Jacobian of the warp (for $\Delta\mathbf{p} = 0$ and varying $\mathbf{x}$ and given $\mathbf{p}$), and $\mathbf{H}$ is the 6×6 Gauss-Newton approximation of the joint-camera Hessian, defined as follows:

$$\mathbf{H} = \sum_{c=1}^{2} \sum_{\mathbf{x} \in \Omega(T_c, \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}})} w_c(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}, ..) \cdot \Big[\nabla I_c \frac{\partial \mathcal{W}_c}{\partial \Delta\mathbf{p}} \Big]^{\mathsf{T}} \Big[\nabla I_c \frac{\partial \mathcal{W}_c}{\partial \Delta\mathbf{p}} \Big]. \quad (8)$$

B. Weighting and iterative re-weighting

The use of weights for alignment serves three purposes. This is reflected by defining w_c as a product of three terms:

$$w_c(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}, ..) = w_D(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}, ..)\, w_{R,c}(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}})\, w_{C,c}. \quad (9)$$

The three terms change per LK-iteration. Below, we define them from left to right, because the definitions depend on each other.

We use a density term w_D to ensure that template points contribute in proportion to the amount of visible head area that they represent. For invisible points, this yields $w_D = 0$. For visible points, we consider an infinitesimal surface area around a template point $\mathbf{x}$ and its camera projection, and we define $w_D(\mathbf{x}; \mathbf{C}_c\mathrm{e}^{\widehat{\mathbf{p}}}, ..)$ as the area ratio of the latter divided by the former. This follows from the direction of the surface normal at $\mathbf{x}$ and the distance of $\mathbf{x}$ to the image plane. For points seen by the same camera, it implies that points seen from the side contribute less than points seen from the front.

We use robustness terms $w_{R,c}$ to reduce contributions of template points that cause large appearance errors, typically due to local noise, non-rigid motion or occlusion. For this, we employ the IRLS (iteratively re-weighted least squares) method of [20] for robust optimization. The method adapts weights prior to every LK-iteration, based on error statistics

for the current pose estimate. Applying this independently for each camera, we define $w_{R,c}$ for camera c as follows:

$$w_{R,c}(\mathbf{x};\mathbf{C_c}e^{\widehat{\mathbf{p}}}) = \mathrm{e}^{\frac{-\left\|I_c(W(\mathbf{x};\mathbf{C_c}e^{\widehat{\mathbf{p}}}))-T_c(\mathbf{x})\right\|_2^2}{2{\sigma_c}^2}}, \qquad (10)$$

where $\sigma_c = 1.4826 \cdot m_c$ and

$$m_c = \underset{\mathbf{x}\in\Omega(T_c,\mathbf{C_c}e^{\widehat{\mathbf{p}}})}{\mathrm{median}} \left\|I_c(W(\mathbf{x};\mathbf{C}_c e^{\widehat{\mathbf{p}}}))-T_c(\mathbf{x})\right\|_2. \qquad (11)$$

We use camera terms $w_{C,c}$ to balance contributions of cameras. For this, we set $w_{C,c}$ for $c \in \{1,2\}$ such that the sum of weights over all points per camera yields a constant:

$$w_{C,c} \cdot \sum_{\mathbf{x}\in\Omega(T_c,\mathbf{C}_c e^{\widehat{\mathbf{p}}})} w_D(\mathbf{x};\mathbf{C}_c e^{\widehat{\mathbf{p}}},..) w_{R,c}(\mathbf{x};\mathbf{C}_c e^{\widehat{\mathbf{p}}}) = \text{constant}. \qquad (12)$$

C. Template trimming

For robustness, we normally do not use a template in its entirety for an alignment move. This is because a template reflects the appearance of the head at some moment in time, and over time appearance may change locally, for example, as a result of a change of facial expression or a temporary occlusion. For this reason, we normally omit areas of large local changes from a template. Here we call this trimming.

In order to judge where local changes may have occurred, we compare a template T with a reference image I' of the same camera at another time moment. For this, we use the pose estimate $\mathbf{p}'$ that was already computed for image I'. For each template point $\mathbf{x}$, we compare its appearance $T(\mathbf{x})$ and the appearance $I'(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}'}}))$ at its projected location in I'. We omit $\mathbf{x}$ as outlier from the trimmed version if

$$\|I'(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}'}}))-T(\mathbf{x})\|_\infty > \max(\min(b\cdot\sigma', d), m'), \qquad (13)$$

where σ' and m' are defined as in (11) with I and p primed and using L_∞-norm. This is similar to outlier removal for intensity appearance in [20], but we use max-min functions because dense-HOG appearances are normalized vectors. Threshold $b\cdot\sigma'$ (with $b = 1.5$) is based on robust statistics as in [20] (for low-median cases), threshold d (with $d = 0.35$) avoids inappropriately high values (for mid-median cases), and m' guarantees that never more than half of the template is omitted (for high-median cases).

D. Complete dual-camera tracking algorithm

The tracking algorithm sequentially inputs pairs of images and it maintains an internal variable for the estimated head pose in the most recent input. After a new input, the pose variable is updated by one or two alignment moves, and the resulting value is the output pose for this input. Internally, this output pose is also used to extract a pair of templates from the corresponding input images. For convenience, we call this pair the current templates.

For the first input, we assume that its output pose results from initialization by face detection, etc. This initialization also yields the specification of the un-posed head surface.

For the second input, the output pose is updated by one alignment move. This move uses the current templates.

For all other inputs, we perform two alignment moves in succession to produce the output pose. We refer to them as the approach move and the refinement move, respectively. The approach move uses trimmed versions of the current templates. For input i, the trimming reference is input $i-2$ (which is the input preceding the input from which the current templates are extracted). The refinement move uses trimmed versions of a pair of key templates. For the trimming of key templates, we use input $i-1$ as reference.

The purpose of the refinement move is to reduce drift. Its use is partly similar to re-registration as mentioned (without detail) in [20]. Our method for drift reduction here is almost the same as in the first extension for drift reduction of [16]. While tracking, we store a limited number of templates as key templates, and we re-use them later to refine poses for output and thereby correct accumulated alignment errors. For a pair of key templates, we also store the pose that was used to extract them. Below, we refer to this pose as the key pose that corresponds with the key templates.

For storage of key templates, we use the following rule. After initialization, we store the current templates as the first pair of key templates. During tracking, we store a new pair of current templates if their corresponding key pose is far from all the key poses that have already been stored.

For a formal definition of 'far from' for a pose $\mathbf{p}$ and a key pose $\mathbf{p}'$, we introduce a boolean condition *close*$(\mathbf{p},\mathbf{p}';f)$ with a threshold parameter f. This condition is defined as:

$$|\mathit{angle}(\mathrm{e}^{-\widehat{\mathbf{p}}}e^{\widehat{\mathbf{p}'}})| \le \alpha \ \wedge \ |\mathit{dist}(\mathrm{e}^{-\widehat{\mathbf{p}}}e^{\widehat{\mathbf{p}'}})| \le f\cdot\rho, \qquad (14)$$

where *angle* and *dist* denote rotation angle and translation distance, $\alpha = 10$ degrees, and ρ equals half the head width. We then define 'far' for key pose storage as $\neg$ *close*$(\mathbf{p},\mathbf{p}';2)$.

For usage in a refinement move, we search for a pair of stored key templates with a key pose that is close to the estimated pose resulting from the approach move. This is based on the intuition that the appearance of the head in key templates can only resemble the appearance in the input if the corresponding poses are close. Here, we formally define 'close' as *close*$(\mathbf{p},\mathbf{p}';1)$. If more than one key pose is found, we select the most recent one. In the rare cases where none is found, we select the key templates based on minimizing an ad-hoc alternative measure for the proximity of a pose and a key pose.

V. Experiments

A. Dataset and evaluation criteria

For our experiments, we used two types of dual-camera sequences recorded at Máxima Medical Center Veldhoven. The first type show infants in bed, with a tube and plasters and occasionally a pacifier. They were captured during 24-hour reflux monitoring for GERD, partly in visual light and partly in infrared. For quantitative evaluation, we selected 5 short episodes with significant motion, changes of facial expression, and partial occlusions.

The second type of sequences show realistic anatomical dummies of very young infants in motion, captured in visual

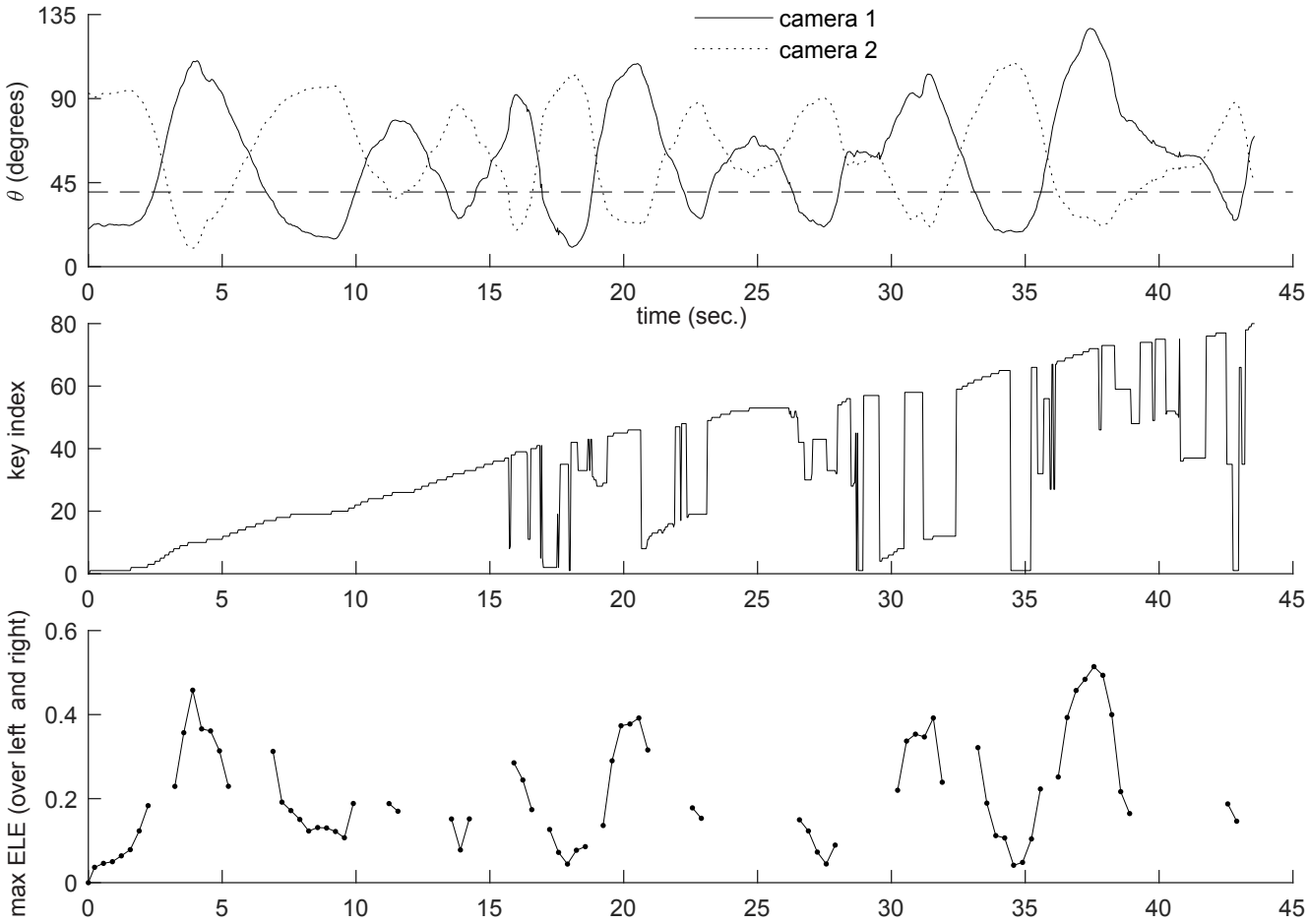

Fig. 2: Tracking results for sequence 13: *top* = head angle θ; *mid* = key index (index of key templates used for refinement); *bottom* = maximum ELE.

light in the same hospital beds. In these videos the dummies are manipulated to move the head in a trajectory that rapidly covers a large range of realistic poses, while trying to avoid frequent repetition of poses. For quantitative evaluation, we selected 9 short videos, featuring 2 dummies with different versions (with/without tube and plasters, and with/without pacifier).

In our tracking experiments, all inputs consisted of pairs of distortion-corrected downscaled gray-level images with 960×540 pixels at 30 fps.

For judging tracking quality, we concentrate on the face because it is the part of the head that is most relevant in the context of our monitoring system. For this, we select, for every pose, from every pair of images one image as the face-view image for that pose. This is defined as the image with the smallest head angle. Here the concept of the head angle $\theta_c(\mathbf{p})$ for a pose $\mathbf{p}$ and a camera c is defined as the angle of rotation from pose $\mathbf{p}$ to a pose that is upright-frontal with respect to camera c (note that a single θ_c-value can correspond with many rotation axes and therefore with many combinations of yaw, pitch and roll).

For qualitative judgment, we usually consider the face in a normalized version of the face-view image (NFV). This view is obtained by projecting the visible head texture into an image that corresponds with a virtual camera positioned upright-frontally before the head.

For quantitative judgment, we consider two landmarks in the NFVs that are related to the eyes. For these, we choose the left and right endocanthion points, because they can be identified consistently for both open and closed eyes. For ground truth, endocanthion annotations in input images are reverse-projected as 3D points on the head and then NFV-projected. We compare them with NFV-projections of fixed 3D references, viz. the reverse-projected annotations of the first image of the sequence. As comparison metric, we use the endocanthion location error (ELE), defined for an NFV as the distance between locations of same-eye endocanthion points divided by the innercanthal distance for the ground truth in the NFV of the first face-view image. This metric allows to compare results for different sequences, as well as alternative ground-truth annotations. We use annotations made at full camera resolution in each 10-th frame, or $\frac{1}{3}$sec.

Note that the ELE metric defined here differs from that in [15] [16], which is based on eye centers with location errors divided by the eye-center distance. Elsewhere, e.g. in [18], location errors of landmarks are divided by the outercanthal

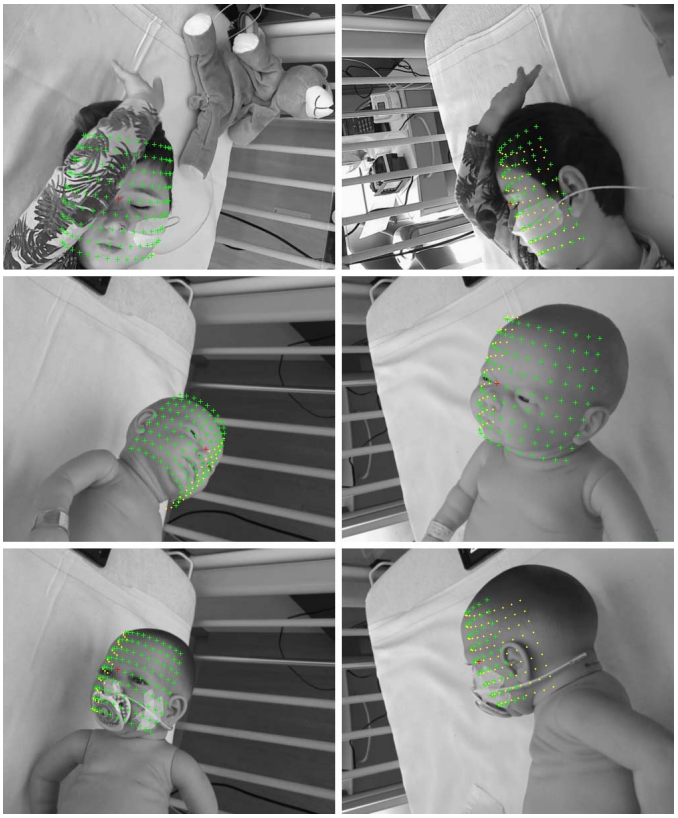

Fig. 3: Examples of tracking results in cropped versions of paired inputs. The overlay dots indicate the front half of the clipped-ellipsoid surface. Green dots denote the visible part, yellow dots denote the invisible part. (*top* = sequence 3; *mid* = sequence 8; *bottom* = sequence 12.)

distance. Such divisors would yield roughly 2 and 3 times smaller numbers here, respectively.

B. Tracking results

Our qualitative impression from many experiments with both types of sequences is that tracking is consistently good, for both visual light and infrared light.

Figure 2 gives an example of tracking, showing results for an input sequence that features an anatomical dummy. In this figure, the θ-graph illustrates the fast head motion. The index graph illustrates that the head pose does not often return near values from much earlier on, as new values for the key index are added steadily. The ELE graph shows that the errors are largest when the face is completely turned away (with head angle up to 127 degrees) from Camera 1, which is the camera of the initial face-view image.

We have excluded ELE values for poses with head angles $\theta \geq 40$ degrees. This is because of an artifact of ELE computation, which is due to the fact that the 3D reference points derived from an image of a sequence, do not usually coincide with the true 3D endocanthion points. This causes ELE offsets that increase disproportionally for increasing θ.

Table I shows quantitative results on selected videos. All of these sequences are challenging, as illustrated by Figs. 1 and 3, and for all of them the mean ELE remains below 0.3.

TABLE I: Tracking accuracy.

nr.	time	ELE		type	light	extras
		mean	max			
1	24s	0.279	0.825	infant 1	infrared	tube/plaster/pacifier
2	31s	0.294	1.073	infant 1	infrared	tube/plaster/pacifier
3	1m06s	0.164	0.369	infant 2	visual	tube/plaster
4	19s	0.288	0.538	infant 2	visual	tube/plaster
5	1m12s	0.070	0.255	infant 2	infrared	tube/plaster
6	24s	0.078	0.204	dummy 1	visual	
7	44s	0.107	0.230	dummy 1	visual	
8	43s	0.045	0.091	dummy 2	visual	
9	46s	0.119	0.278	dummy 1	visual	pacifier
10	1m07s	0.272	0.623	dummy 1	visual	pacifier
11	1m03s	0.090	0.355	dummy 2	visual	tube/plaster
12	51s	0.175	0.710	dummy 1	visual	tube/plaster/pacifier
13	44s	0.189	0.514	dummy 1	visual	tube/plaster/pacifier
14	1m13s	0.210	0.787	dummy 1	visual	tube/plaster/pacifier

Note: mean/max ELE are over all left and right endocanthion points with ground truth, but excluding images for which head angle $\theta \geq 40$ degrees.

Given the practical situation of our case study with infants, it can be concluded that the ELE of below 0.3 is sufficiently low to sustain the tracking. This is in alignment with our earlier qualitative observations on consistently good tracking from many experiments. More specifically, the 3D pose information obtained is reliable enough to enable and support landmark localization in the context of our monitoring system.

VI. Conclusions

We have presented a new algorithm for dual-camera 3D head tracking, using a clipped-ellipsoid rigid head model and 3D head pose recovery. The algorithm employs joint alignment of paired templates based on dense-HOG features. In the algorithm, template pairs are dynamically extracted and a limited number of template pairs are stored and re-used for drift reduction.

We have also evaluated the algorithm on video sequences of infants in bed in a hospital, captured in visual light as well as in near-infrared light, and also on sequences of realistic anatomical dummies of very young infants. Results show consistently good tracking behavior. For challenging video sequences, the mean tracking error in terms of endocanthion location error relative to the innercanthal distance remains below 30%. This error has proven to be sufficiently low for 3D head tracking in the context of infant monitoring.

With this performance, the proposed algorithm is able to provide reliable 3D pose information that can be used to support landmark localization for infant face analysis. This support allows to circumvent the problem of absence of sufficient facial data of infants addressing the practical conditions of clinical monitoring (e.g. nighttime, pacifiers, etc.). This absence has been preventing the use of state-of-the-art deep-learning-based techniques.

The experiments have shown that the stand-alone problem of 3D head tracking can be solved with the proposed algorithm, as it can handle all tracking challenges for clinical infant monitoring (e.g. lack of texture, occlusions, etc.). For this reason, the proposed algorithm is used successfully in an infant monitoring system under development.

REFERENCES

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

[2] S. Baker and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *Int. Journ. Computer Vision*, 56(3):221–255, 2004.

[3] A. Bulat 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 *IEEE Int. Conf. Computer Vision ICCV*, 2017.

[4] Q. Cai, A. Sankaranarayanan, Q. Zhang, Z. Zhang, and Z. Liu. Real time head pose tracking from multiple cameras with a generic model. In *IEEE Conf. Computer Vision and Pattern Recognition Workshops CVPRW*, pages 25–32, 2010.

[5] G. Chrysos, E. Antonakos, P. Snape, A. Asthana, and S. Zafeiriou. A Comprehensive Performance Evaluation of Deformable Face Tracking "In-the-Wild". *Int. Journ. Computer Vision*, 2017.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. Computer Vision and Pattern Recognition CVPR*, volume 1, pages 886–893, 2005.

[7] J. Harguess, C. Hu, and J. Aggarwal. Full-motion recovery from multiple video cameras applied to face tracking and recognition. In *IEEE Int. Conf. Computer Vision Workshop ICCVW*, pages 1889–1896, 2011.

[8] J. Jang and T. Kanade. Robust 3d head tracking by online feature registration. In *IEEE Conf. Automatic Face and Gesture Recognition*, 2008.

[9] X. Jin and X. Tan. Face alignment in-the-wild: A Survey. *Computer Vision and Image Understanding*, 162:1–22, 2017.

[10] H. Kwang and J. Myung. 3D head tracking and pose-robust 2D texture map-based face recognition using a simple ellipsoid model. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems IROS*, pages 307–312, 2008.

[11] C. Li, S. Zinger, W. Tjon a Ten, and P. de With. Video-based Discomfort Detection for Infants Using a Constrained Local Model. In *Int. Conf. Systems, Signals and Image Processing IWSSIP*, pages 81–84, 2016.

[12] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Int. Joint Conf. Artificial Intelligence IJCAI*, volume 81, pages 674–679, 1981.

[13] E. Murphy-Chutorian and M. M. Trivedi. Head Pose Estimation in Computer Vision: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):607–626, 2009.

[14] R. Murray, Z. Li, and S. Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[15] R. Saeijs, W. Tjon a Ten, and P. de With. Dense-HOG-based 3D face tracking for infant pain monitoring. In *IEEE Int. Conf. Image Processing ICIP*, pages 1719–1723, 2016.

[16] R. Saeijs, W. Tjon a Ten, and P. de With. Dense-HOG-Based Drift-Reduced 3D Face Tracking for Infant Pain Monitoring. In *Proc. SPIE 10341 Int. Conf. Machine Vision*, page 103411U, 2016.

[17] S. Sathyanarayana, K. Satzoda, S. Sathyanarayana, and S. Thambipillai. Vision-based patient monitoring: a comprehensive review of algorithms and technologies. *Journ. Ambient Intelligence and Humanized Computing*, 2015.

[18] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results. In *IEEE Int. Conf. Computer Vision Workshop ICCVW*, pages 1003–1011, 2015.

[19] B. Slaats, S. Zinger, P. de With, W. Tjon a Ten, and S. Bambang Oetomo. Video analysis for acute pain detection in infants. In *5th joint WIC/IEEE Symp. Information Theory and Signal Processing*, pages 50–57, 2015.

[20] J. Xiao, T. Moriyama, T. Kanade, and J. 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.

[21] S. Zafeiriou, G. Chrysos, A. Roussos, E. Ververas, J. Deng, and G. Trigeorgis. The 3D Menpo Facial Landmark Tracking Challenge. In *IEEE Int. Conf. Computer Vision ICCV*, pages 2503–2511, 2017.

[22] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015.

[23] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face Alignment Across Large Poses: A 3D Solution. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016.