PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

# Dense-HOG-based drift-reduced 3D face tracking for infant pain monitoring

Ronald W.J.J. Saeijs
Walther E. Tjon A Ten
Peter H. N. de With

SPIE.

# Dense-HOG-Based Drift-Reduced 3D Face Tracking for Infant Pain Monitoring

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. The algorithm includes extensions for drift reduction, using re-registration in combination with multi-pose state estimation by means of a square-root unscented Kalman filter. The paper reports experimental results on videos of moving infants in hospital who are relaxed or in pain. Results show good tracking behavior for poses up to 50 degrees from upright-frontal. In terms of eye location error relative to inter-ocular distance, the mean tracking error is below 9%.

**Keywords:** Face tracking, pain monitoring, cylinder head model, dense HOG

## 1. INTRODUCTION

Continuous pain monitoring of infants will benefit many clinical contexts. For example, in the context of gastro-esophageal reflux disease (GERD), infants suspected of GERD 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 [1] and acute pain [2] of infants shows that automatic monitoring can be based on video analysis of facial expressions (cf. [3]).

In the previous context, we study infant face tracking in this paper, as a prerequisite for facial expression analysis. Face tracking, in general, has many applications and many proposed solutions. Most solutions aim at tracking faces of adults with limited pose variations in front of a single camera (e.g. in human-computer interaction and vehicle driver monitoring). However, in our context the tracking problem has three characteristics that existing solutions cannot handle. Firstly, for infants, and especially very young ones, facial texture is less pronounced than for adults. For example, infants have no prominent eyebrows, wrinkles or creases, and their eyes are 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, there are plasters on the face and a tube into the nose for a pH probe in the esophagus. Also, infants may have a pacifier in their mouth to reduce stress, with a large variety of appearances. In addition, cuddles, toys or blankets 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 freely shift and turn their head. As a result, tracking has to handle much larger pose variations.

Here, we propose a face tracking algorithm that accommodates the above-mentioned characteristics. In order to show results focusing on the first two characteristics (less pronounced texture and partial occlusion), we present its single-camera version here. However, it is purposely developed for extension to multiple cameras, so as to fully accommodate the third characteristic. The algorithm is based on tracking 3D head pose using dense-HOG features, and it extends our work of [4] by including drift reduction. Below, Section 2 discusses our approach in relation to other work. Section 3 describes the basic algorithm, while Section 4 presents its extensions for drift reduction. Section 5 provides experiments and results.

## 2. APPROACH AND RELATED WORK

Our face tracking approach models the face as part of the head and solves the more general problem of tracking 3D head pose. Our main motivation is that this allows maximum use of image information for robustness, because visible features of both face and non-face parts of the head can be used. This also allows to extend to multi-camera setups. As a further motivation, 3D head tracking informs us about head movements, which may serve as extra parameter for pain detection.

For general face tracking, state-of-the-art approaches are based on aligning a deformable 2D or 3D face model to input images. Facial alignment methods, e.g. the Supervised Descent Method (SDM) [5] or the Constrained Local Model (CLM) method of [6], all require a form of training, and it is essential that the training result can capture all variations of shape and

Ronald Saeijs was supported by STW in project 13335 GARDIAN (r.w.j.j.saeijs@tue.nl).

Ninth International Conference on Machine Vision (ICMV 2016), edited by Antanas Verikas, Petia Radeva, Dmitry P. Nikolaev, Wei Zhang, Jianhong Zhou, Proc. of SPIE Vol. 10341, 103411U
 · doi: 10.1117/12.2268522

appearance. In our case this is hardly feasible, because of wide pose ranges, unknown appearances of plasters, pacifiers, etc. For this reason, we need an approach without a pre-determined model of appearance. Many of such approaches [7] recover head motion, using an assumed 3D head shape. Some, e.g. [8], recover motion from keypoints. For infant heads this is not feasible, as they have few and unstable points. Others recover motion by Lucas-Kanade template alignment [9], e.g. with a cylinder [10] or ellipsoid head model and for multi-camera setups [11]. We adopt the same principle for our algorithm.

In [10, 11] , the principle of head pose recovery by 3D alignment is used with pixel intensity templates. This may give problems with less-pronounced texture and less-uniform illumination, as e.g. in case of a sleeping infant in bed. For this reason, we use densely sampled Histogram-of-Oriented-Gradient ('dense-HOG') features [12], which can improve Lucas-Kanade alignment [13]. In addition, we use insights from [14] to introduce a probabilistic algorithm for drift reduction. Our main contribution is the use of dense-HOG features for 3D tracking, in a refined algorithm with drift reduction, and its evaluation for real-life infant monitoring conditions. The paper offers a comprehensive description of the algorithm for the single-camera case, with main innovations at the end of Section 3.1 and in 4.2, and minor innovations in 3.2, 3.3 and 4.1.

## 3. MODELING AND BASIC TRACKING ALGORITHM

### 3.1 Modeling

We represent 2D image locations as $\mathbf{u} = [u, v]^\mathsf{T} \in \mathbf{U}$, where $\mathbf{U} = \mathbb{R}^2$. For an image $I$, we use $\mathbf{U}(I) \subset \mathbf{U}$ to denote its pixel locations. We represent 3D points in terms of homogeneous coordinates as $\mathbf{x} = [x, y, z, 1]^\mathsf{T} \in \mathbf{X}$, where $\mathbf{X} = \mathbb{R}^3 \times \{1\}$. We model the infant head as an oriented 3D surface with centroid at $[0, 0, 0, 1]^\mathsf{T}$. The head pose is defined by applying a 3D rigid-body transformation $g \in SE(3)$, while using the vector $\mathbf{p} = [\omega_x, \omega_y, \omega_z, t_x, t_y, t_z]^\mathsf{T} \in \mathbb{R}^6$ of exponential coordinates [15] of $g$ as our representation of pose. From $\mathbf{p}$ we obtain the homogeneous matrix $\mathbf{G}$ for $g$ as follows [15]:

$$\mathbf{G} = \mathrm{e}^{\widehat{\mathbf{p}}} = \mathbf{I} + \widehat{\mathbf{p}} + \frac{\widehat{\mathbf{p}}^2}{2!} + \frac{\widehat{\mathbf{p}}^3}{3!} + \cdots, \qquad \text{where} \quad \widehat{\mathbf{p}} = \begin{bmatrix} 0 & -\omega_z & \omega_y & t_x \\ \omega_z & 0 & -\omega_x & t_y \\ -\omega_y & \omega_x & 0 & t_z \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{1}$$

Here, $\mathbf{I}$ is the $4\times4$ identity matrix. For practical calculation of the exponential mapping $\mathrm{e}^{\widehat{\mathbf{p}}}$ and its inverse $\widetilde{\log}\,\mathbf{G}$, see [15]. We assume full-perspective image projection as a function $W$ from points $\mathbf{x}$ to locations $\mathbf{u}$:

$$\mathbf{u} = W(\mathbf{x}; \mathbf{C}) = [u', v']^\mathsf{T}/s', \qquad \text{where} \quad [u', v', s']^\mathsf{T} = \mathbf{C}\mathbf{x}. \tag{2}$$

Here, $\mathbf{C}$ is the $3\times4$ camera projection matrix that combines the intrinsic and extrinsic characteristics of the camera. Because parts of a projected surface may not be visible, we also assume a function $w(\mathbf{x}; \mathbf{C})$ that yields positive values for visible surface points and zero for all other points. For posed-head points $\mathrm{e}^{\widehat{\mathbf{p}}}\mathbf{x}$, we write $W(\mathbf{x};\mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}})$ instead of $W(\mathrm{e}^{\widehat{\mathbf{p}}}\mathbf{x};\mathbf{C})$ etc. to associate the rigid-body transformation $\mathrm{e}^{\widehat{\mathbf{p}}}$ with the camera. As a result, head point references $\mathbf{x}$ remain in the un-posed coordinate system centered at $[0, 0, 0, 1]^\mathsf{T}$. The equivalence follows from (2).

For an image $I$, we use $I(\mathbf{u})$ to denote the appearance value $\mathbf{a} \in \mathbf{A}$ associated with a pixel location $\mathbf{u} \in \mathbf{U}(I)$. As a main innovation over [10, 11] , we use dense-HOG feature vectors as appearance values instead of pixel intensities. We use 36-dimensional vectors, from a HOG variant with blocks of $2\times2$ cells, cells of $8\times8$ pixels, and 9-bin histograms for signed orientations. The variant uses trilinear interpolation of location and orientation, and $L^2$-*Hys* normalization [12]. In [4], we have experimentally shown how dense-HOG yields more accurate tracking than intensity, for a similar basic algorithm. We also use $I(\mathbf{u})$ for non-pixel locations $\mathbf{u} \in \mathbf{U}$. It is then defined by linear inter-/extrapolation of pixel appearance values.

### 3.2 Template alignment using Lucas-Kanade gradient descent and IRLS

Our aim for tracking is to estimate the pose of the head in each new image based on pose knowledge derived from previous images. For this, we repeatedly align templates derived from previous images with new images. For initialization, we need an initial pose $\mathbf{p}_0$ and a specification of the un-posed head surface in the initial image $I_0$. We present the algorithm for any general shape, but in the sequel we employ a cylinder shape, as it is least sensitive to initial pose error [10].

We define a template $T \subset \mathbf{X} \times \mathbf{A}$ as a set $X(T)$ of 3D points situated on the un-posed head surface with associated appearance values and the intuition that it represents a textured part of the head. We use $T(\mathbf{x})$ to denote the appearance value $\mathbf{a}$, associated with 3D point $\mathbf{x}$ in template $T$. We also define a template extraction function $\tau(I, \mathbf{p}; \mathbf{C})$ that yields a template by reverse-projecting pixel locations of image $I$ (with their appearance values) onto the surface of the head, for a presumed pose $\mathbf{p}$. Here, reverse projection of a location $\mathbf{u}$ is defined to yield (if it exists) the unique 3D point $\mathbf{x} = R(\mathbf{u})$ on the un-posed head surface that is visible and mapped to $\mathbf{u}$ by image projection for pose $\mathbf{p}$, i.e. so that $W(R(\mathbf{u}); \mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}}) = \mathbf{u}$.

We define an alignment step as a process that takes an image $I$, a template $T$, and a start pose $\mathbf{p}_{start}$ and yields a stop pose $\mathbf{p}$ such that, for all template points $\mathbf{x}$, image appearance $I(\mathbf{u})$ at projected locations $\mathbf{u}=W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}})$ is close to template appearance $T(\mathbf{x})$. We formulate this as an optimization for minimum sum-of-weighted-squared-error cost, as follows:

$$\mathbf{p}=\underset{\mathbf{p}}{\operatorname{argmin}} \sum_{\mathbf{x}\in\Omega(T,\mathbf{C}e^{\widehat{\mathbf{p}}})} w(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}})\cdot\Big\|I(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}))-T(\mathbf{x})\Big\|_2^2, \quad \text{where} \quad \Omega(T,\mathbf{C}e^{\widehat{\mathbf{p}}})=\big\{\mathbf{x}\big|\mathbf{x}\in X(T)\wedge w(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}})>0\big\}. \quad (3)$$

Here, summation is performed only over those template points $\mathbf{x}$ that are visible in the new pose $\mathbf{p}$, as defined by $\Omega$, and the positive values of function $w$ specify weights to adjust the influence of those individual points in the alignment process.

To implement an alignment step, we use the Lucas-Kanade (LK) method of gradient descent optimization [9]. Starting from $\mathbf{p}=\mathbf{p}_{start}$, LK iteratively updates $\mathbf{p}$ by approximately solving a reformulation of (3) in terms of an update $\Delta\mathbf{p}$ for $\mathbf{p}$:

$$\Delta\mathbf{p}_{opt}=\underset{\Delta\mathbf{p}}{\operatorname{argmin}} \sum_{\mathbf{x}\in\Omega(T,\mathbf{C}e^{\widehat{\mathbf{p}}}e^{\widehat{\Delta\mathbf{p}}})} w(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}e^{\widehat{\Delta\mathbf{p}}})\cdot\Big\|I(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}e^{\widehat{\Delta\mathbf{p}}}))-T(\mathbf{x})\Big\|_2^2. \quad (4)$$

Comparing (4) with the standard LK formulation in [16], there is a difference in the function $W$ and its arguments, and in the $\Omega$-restriction on contributing points. Our motion model is a 3D→2D warp function $\mathcal{W}$ with three parameters $\Delta\mathbf{p}$,$\mathbf{p}$,$\mathbf{C}$ defined by $\mathcal{W}(\mathbf{x};\Delta\mathbf{p},\mathbf{p},\mathbf{C})=W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}e^{\widehat{\Delta\mathbf{p}}})$. As a result, updating also differs from [16]: we update $\mathbf{p}$ using $e^{\widehat{\mathbf{p}}}\leftarrow e^{\widehat{\mathbf{p}}}e^{\widehat{\Delta\mathbf{p}}}$. An LK-iteration approximates $\Delta\mathbf{p}_{opt}$ in (4) using first-order Taylor expansion of $W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}e^{\widehat{\Delta\mathbf{p}}})$ for small $\Delta\mathbf{p}$ to yield

$$\Delta\mathbf{p}_{approx}=-\mathbf{H}^{-1}\sum_{\mathbf{x}\in\Omega(T,\mathbf{C}e^{\widehat{\mathbf{p}}})} w(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}})\cdot\Big[\nabla I\frac{\partial\mathcal{W}}{\partial\Delta\mathbf{p}}\Big]^{\mathsf{T}}\Big[I(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}))-T(\mathbf{x})\Big], \quad (5)$$

where $\nabla I$ is the gradient $[\frac{\partial I}{\partial u},\frac{\partial I}{\partial v}]$ of image $I$, $\frac{\partial\mathcal{W}}{\partial\Delta\mathbf{p}}$ is the Jacobian of the warp (for $\Delta\mathbf{p}=0$ and given $\mathbf{x}$, $\mathbf{p}$ and $\mathbf{C}$), and $\mathbf{H}$ is the 6×6 Gauss-Newton approximation of the Hessian:

$$\mathbf{H}=\sum_{\mathbf{x}\in\Omega(T,\mathbf{C}e^{\widehat{\mathbf{p}}})} w(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}})\cdot\Big[\nabla I\frac{\partial\mathcal{W}}{\partial\Delta\mathbf{p}}\Big]^{\mathsf{T}}\Big[\nabla I\frac{\partial\mathcal{W}}{\partial\Delta\mathbf{p}}\Big]. \quad (6)$$

Partly similar to [10], we adapt $w$ per LK-iteration as product of a density term $w_D$ and a robustness term $w_R$: $w=w_D\cdot w_R$. The density term relates to image projection of the head surface. For this, consider an infinitesimal area around a template point $\mathbf{x}$ and its projection in the image plane. We define $w_D(\mathbf{x};\mathbf{C})$ as the area ratio of the latter divided by the former. It varies with the direction of the surface normal at $\mathbf{x}$ and the distance of $\mathbf{x}$ to the image plane. As a result, points seen from the side contribute less than points seen from the front (and, unlike in [10], our definition here extends to multiple cameras). The robustness term relates to the IRLS method of [10], which is used to handle noise, non-rigid motion and occlusions. At each LK-iteration, this method adapts $w_R$ based on the error resulting from the current pose estimate $\mathbf{p}$:

$$w_R(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}))=e^{\frac{-\left\|I(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}))-T(\mathbf{x})\right\|_2^2}{2\sigma^2}}, \qquad \text{where} \quad \sigma=1.4826\cdot m\ , \quad m=\underset{\mathbf{x}\in\Omega(T,\mathbf{C}e^{\widehat{\mathbf{p}}})}{\operatorname{median}}\big\|I(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}}}))-T(\mathbf{x})\big\|_2. \quad (7)$$

### 3.3 Template use and outlier removal

For tracking, we supply sequentially images $I_k$ for $k=0,1,2,\ldots$. With $\mathbf{p}_0$ given as input, poses $\mathbf{p}_k$ are produced as outputs by repeating a single routine for each $k$, $k\geq 1$. In the basic algorithm, this routine performs one alignment step per image $I_k$. The template and start pose for this step are derived from the previous image: $T=T_{k-1}=\tau(I_{k-1},\mathbf{p}_{k-1};\mathbf{C})$, $\mathbf{p}_{start}=\mathbf{p}_{k-1}$.

For robustness, and especially for handling temporary occlusions, each alignment step is preceded by outlier removal. Outlier removal takes the candidate template $T$ for alignment plus an outlier reference image index $r$ (for which pose $\mathbf{p}_r$ was already computed) and replaces $T$ by a smaller template $T'$ by removing template points $\mathbf{x}$ with a large difference between their appearance $I_r(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}_r}}))$ in the outlier reference image $I_r$ and their template appearance $T(\mathbf{x})$, as follows:

$$T'=\big\{\ \langle\mathbf{x},\mathbf{a}\rangle\ \big|\ \langle\mathbf{x},\mathbf{a}\rangle\in T\ \wedge\ \|I_r(W(\mathbf{x};\mathbf{C}e^{\widehat{\mathbf{p}_r}}))-T(\mathbf{x})\|_\infty\leq\max(\min(c\cdot\sigma_r,d),m_r)\ \big\}, \quad (8)$$

where $\sigma_r$ and $m_r$ are defined as in (7) with $I$ and $p$ indexed by $r$ and using $L_\infty$-norm. This is similar to outlier removal for intensity appearance in [10], but we use max-min thresholds because dense-HOG appearances are normalized vectors.

Threshold $c \cdot \sigma_r$ (with $c = 1.5$) is based on robust statistics as in [10] (for low-median cases), $d$ (set at 0.35) avoids ineffectively high values (for mid-median cases), and $m_r$ guarantees that never more than half of the template is removed (for high-median cases). In the basic algorithm, outlier removal of template $T_{k-1}$ in step $k$ ($k \geq 2$) uses outlier reference $r = k-2$, which refers to the image preceding image $I_{k-1}$ from which $T_{k-1}$ was derived (for $k = 1$ outlier removal does not apply).

## 4. ALGORITHM EXTENSIONS FOR DRIFT REDUCTION

### 4.1 Selective storage and re-use of templates

The basic algorithm from Section 3.3 may drift because alignment errors accumulate during tracking. As a first measure for drift reduction, we store a limited number of templates as key references, and re-use them to correct this. This is partly similar to re-registration as mentioned (but not detailed) in [10]. We use $S \subset \mathbb{N}_0$ as a shorthand for the set of stored key references, where an element $s \in S$ is an index of an input image $I_s$ from which a key reference template was extracted (using pose $\mathbf{p}_s$). Also, for selection of key references for storage and re-use, we define a boolean function *close* for a pair of poses $\mathbf{p}, \mathbf{p}'$ and a factor $f$, such that

$$\mathit{close}(\mathbf{p}, \mathbf{p}'; f) \equiv \big( |\mathit{angle}(\mathrm{e}^{-\widehat{\mathbf{p}}}\mathrm{e}^{\widehat{\mathbf{p}'}})| \leq \alpha \ \wedge \ |\mathit{dist}(\mathrm{e}^{-\widehat{\mathbf{p}}}\mathrm{e}^{\widehat{\mathbf{p}'}})| \leq f \cdot \rho \big), \tag{9}$$

where *angle* and *dist* denote rotation angle and translation distance, with $\alpha$ set at 10 degrees and $\rho$ set at the head radius.

For storage, we initialize the key reference set from the initial image: $S = \{0\}$. During tracking, we add a new reference $s$ whenever we output a pose $\mathbf{p}_s$ that is far from any of the references stored sofar, viz. when $\forall s' \in S : \neg\, \mathit{close}(\mathbf{p}_s, \mathbf{p}_{s'}; 2)$. For re-use, we extend the routine for image $I_k$ from Section 3.3 with a second alignment step (with preceding outlier removal). The second alignment step re-uses a key reference template $T_s$ and uses the stop pose $\mathbf{p}$ of the first step as its start pose, and its preceding outlier removal uses outlier reference $r = k-1$. We select key reference $s \in S$ such that its pose $\mathbf{p}_s$ is close to the current pose, viz. so that $\mathit{close}(\mathbf{p}, \mathbf{p}_s; 1)$ holds. (If more candidates exist, we select the most recent one. In the rare cases where none exist, we skip the second alignment step. This is based on the intuition that texture from $T_s$ may not sufficiently resemble head texture in the current image if there is a large difference in pose.)

### 4.2 Probabilistic poses and square-root unscented Kalman filter for multi-pose estimation

As a second measure for drift reduction, we introduce a probabilistic version of the algorithm-with-re-use from Section 4.1. Based on method 3 in [14], the idea is to consider pose outputs as estimates, and to keep improving estimates from the past that may be used as references for the future. For this, consider the multi-pose state vector $\mathbf{p}_{*_k}$ that represents the collection of pose variables that had to be stored in the routine for image $I_k$ in the algorithm-with-re-use. This collection consists of poses of key references in $S = \{s_1, s_2, \ldots, s_{|S|}\}$ plus poses $\mathbf{p}_{k-2}$ and $\mathbf{p}_{k-1}$ of outlier references plus output pose $\mathbf{p}_k$. From these, we define $\mathbf{p}_{*_k}$ (with dimension varying with $k$) by vertical concatenation, in order of increasing image index:

$$\mathbf{p}_{*_k} = \big[\, \mathbf{p}_{s_1}^{\mathsf{T}}, \mathbf{p}_{s_2}^{\mathsf{T}}, \ldots, \mathbf{p}_{|S|}^{\mathsf{T}}, \mathbf{p}_{k-2}^{\mathsf{T}}, \mathbf{p}_{k-1}^{\mathsf{T}}, \mathbf{p}_k^{\mathsf{T}} \,\big]^{\mathsf{T}} \tag{10}$$

In the probabilistic algorithm, we maintain a probabilistic counterpart of this multi-pose state. For this, we model the same collection of poses as a multivariate Gaussian distribution $\mathit{Pr}(\mathbf{p}_{*_k}) = \mathrm{Norm}_{\mathbf{p}_{*_k}}[\boldsymbol{\mu}_{*_k}, \boldsymbol{\Sigma}_{*_k}]$, and we use the combination of mean $\boldsymbol{\mu}_{*_k}$ and covariance $\boldsymbol{\Sigma}_{*_k}$ as new state. By analogy with (10), it has parts $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_{i,j}$ etc. relating to image indices $i, j$.

The probabilistic algorithm uses templates as before, but template extraction now takes a pose mean $\boldsymbol{\mu}_i$ as an estimate where it used a value $\mathbf{p}_i$ as parameter before. Also, key templates are not stored but re-extracted from stored image parts, so that they can improve as estimates $\boldsymbol{\mu}_i$ improve. Similarly, alignment steps are used as before, but an alignment step now starts from a pose mean $\boldsymbol{\mu}_{\mathit{start}}$. Based on Appendix A in [14], we can approximate the uncertainty in its stop pose $\mathbf{p}$, viz. in terms of a zero-mean Gaussian distribution for the optimization variable $\Delta\mathbf{p}$ from (4), with covariance $\boldsymbol{\Sigma}_{\Delta\mathbf{p}}$ given by

$$\boldsymbol{\Sigma}_{\Delta\mathbf{p}} = \varsigma^2 \cdot \mathbf{H}^{-1}\,, \qquad \text{where} \quad \varsigma^2 = \big|\Omega(T, \mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}})\big|^{-1} \cdot \sum_{\mathbf{x} \in \Omega(T, \mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}})} w(\mathbf{x}; \mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}}) \cdot \Big\| I(W(\mathbf{x}; \mathbf{C}\mathrm{e}^{\widehat{\mathbf{p}}})) - T(\mathbf{x}) \Big\|_2^2 . \tag{11}$$

Using the above, we define the algorithm as a recursive state estimator that uses an alignment step as its measurement step. The underlying world states for this estimator are composite multi-pose vectors $\mathbf{p}_*$ as in (10). As the temporal model for the estimator, we model the relationship between two successive world states as follows: pose components referring to the same image in both world states remain unchanged, and the value of a new pose component $\mathbf{p}_k$ depends only on the value

of $\mathbf{p}_{k-1}$ in the preceding world state. For the latter, we use a simple transition model $Pr(\mathbf{p}_k|\mathbf{p}_{k-1}) = \mathrm{Norm}_{\mathbf{p}_k}[\mathbf{p}_{k-1}, \mathbf{\Sigma}_{trans}]$, which represents Brownian head motion in exponential-coordinate space. We can now use Kalman techniques, as follows:

Before input of a new image $I_k$, we predict a new state $\langle \boldsymbol{\mu}_{*_k}, \mathbf{\Sigma}_{*_k} \rangle$ from the previous state by adding new state parts for image index $k$, and by removing (through function *reduce*) parts that are no longer needed (viz. for index $k-3$ if $k-3 \notin S$):

$$\boldsymbol{\mu}_{*_k} = \left[\, reduce(\boldsymbol{\mu}_{*_{k-1}}^{\mathsf{T}}) \,,\, \boldsymbol{\mu}_{k-1}^{\mathsf{T}} \,\right]^{\mathsf{T}}, \qquad \mathbf{\Sigma}_{*_k} = \begin{bmatrix} reduce(\mathbf{\Sigma}_{*_{k-1}}) & 0 \\ 0 & \mathbf{\Sigma}_{trans} \end{bmatrix}. \tag{12}$$

After an alignment step for image $I_k$, we update state $\langle \boldsymbol{\mu}_{*_k}, \mathbf{\Sigma}_{*_k} \rangle$ by means of a Kalman measurement incorporation step. Because a measurement (i.e. stop pose $\mathbf{p}$) is highly non-linear as a function of both measurement noise (i.e. alignment uncertainty $\Delta\mathbf{p}$) and world state (viz. start pose $\mathbf{p}_{start}$), we have to use a version of a Square-Root Unscented Kalman Filter for a system in general form (see [17]). This filter uses an unscented transform that first creates so-called sigma points to approximate the probability distribution of the current multi-pose world state (plus the noise state $\Delta\mathbf{p}$), then applies the non-linear function to the set of sigma points, and finally computes a new $\langle \boldsymbol{\mu}_{*_k}, \mathbf{\Sigma}_{*_k} \rangle$ from the function outputs. Using a square-root implementation, we actually compute and store roots of $\mathbf{\Sigma}$'s, to avoid non-positive-definite $\mathbf{\Sigma}$-results.

## 5. EXPERIMENTS

### 5.1 Dataset and evaluation criteria

For experiments, we used sequences captured with handheld cameras at Máxima Medical Center Veldhoven, with parental consent [2]. They show infants in bed, in various conditions: relaxed, experiencing acute pain during interventions, experiencing post-operative pain, etc. For quantitative evaluation, we also selected 10 videos of different infants with a lot of motion and, for some, also large changes of expression from acute pain. In order to judge tracking accuracy quantitatively from both eyes, we selected sub-sequences without extremely non-frontal poses. For this, we used the angle $\theta(\mathbf{p})$ of rotation from pose $\mathbf{p}$ to an upright-frontal pose (with respect to the camera) as a measure of non-frontality and limit $\theta < 50$ degrees. (Note that a $\theta$-value can correspond with many rotation axes and therefore with many combinations of yaw, pitch and roll.) Video (30 fps) was input as down-scaled gray-level signal with $480 \times 270$ pixels for Sequence 1 and $320 \times 180$ for others.

For qualitative judgment, we consider faces in normalized frontal view (NFV). For this, visible head texture is projected into an image that corresponds with a camera positioned upright-frontally before the head (cf. Figure 1). For quantitative judgment, we consider center-of-eye locations in NFVs. As ground truth, center-of-eye annotations in input images are reverse-projected as 3D points on the posed head and then NFV-projected. We compare them with NFV-projections of fixed 3D head points, viz. the reverse-projected annotations of the first image of the sequence. As comparison metric, we use the eye location error (ELE), defined for an NFV as the distance between same-eye locations divided by the distance between left/right-eye ground-truths in the NFV of the first image. This metric allows to compare results for different sequences, as well as alternative ground-truth annotations. (We annotated eye centers at full input resolution in each 10-th frame ($\frac{1}{3}$sec). Re-annotation trials have shown low annotation noise: mean ELE below 0.02, outliers up to 0.08.)

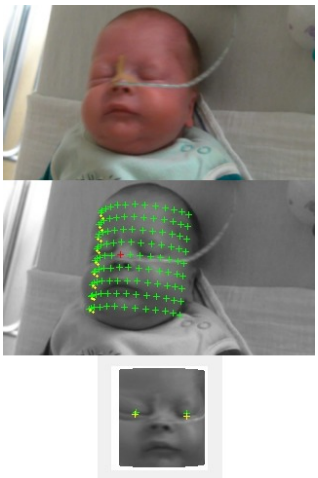

Table 1: Tracking accuracy.

| nr. | length (#frames) | estimated* $\theta$ (degrees) | | basic algorithm (ELE)† | | algor. with re-use (ELE)† | | $\lvert S\rvert$‡ | probabilistic algor. (ELE)† | | $\lvert S\rvert$‡ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | min | max | mean | max | mean | max | | mean | max | |
| 1 | 982 | 7.9 | 48.0 | 0.038 | 0.084 | **0.026** | **0.073** | 7 | **0.026** | **0.073** | 7 |
| 2 | 750 | 13.4 | 47.1 | 0.059 | 0.164 | **0.046** | **0.137** | 7 | 0.049 | 0.142 | 8 |
| 3 | 430 | 5.8 | 34.0 | 0.101 | 0.184 | **0.071** | **0.141** | 10 | 0.085 | 0.193 | 9 |
| 4 | 739 | 7.0 | 22.5 | 0.062 | 0.187 | **0.037** | 0.097 | 4 | **0.037** | **0.095** | 5 |
| 5 | 458 | 8.4 | 43.6 | 0.056 | 0.163 | 0.044 | 0.221§ | 15 | **0.033** | **0.113** | 13 |
| 6 | 1274 | 0.7 | 41.9 | 0.110 | 0.343 | 0.051 | 0.180 | 14 | **0.047** | **0.158** | 13 |
| 7 | 1007 | 3.7 | 48.6 | 0.053 | 0.161 | 0.021 | 0.087 | 14 | **0.020** | **0.083** | 13 |
| 8 | 466 | 21.5 | 32.4 | 0.042 | **0.078** | **0.039** | 0.080 | 3 | 0.042 | 0.087 | 3 |
| 9 | 760 | 13.7 | 32.9 | 0.055 | 0.128 | **0.054** | **0.126** | 13 | 0.055 | 0.127 | 12 |
| 10 | 480 | 20.8 | 31.9 | **0.040** | 0.106 | 0.048 | 0.110 | 5 | 0.044 | **0.103** | 6 |

*Values for $\theta$ are from the output of the probabilistic algorithm. †Mean/max ELE are over all left and right eyes with ground truth. ‡$|S|$ is number of key references at end of the sequence. § Sequence 5 has 1 frame where re-use output has 1 invisible eye (excluded from mean/max-ing).

Figure 1: Example probabilistic algorithm, Sequence 6, frame 1274. Top: original image; mid: input and front-of-posed-cylinder overlay (green = visible, yellow = invisible); bottom: NFV (green = estimated eye location, yellow = ground truth eye location).

### 5.2 Tracking results

Our qualitative impression from experiments on many sequences is that faces are tracked well for poses with $\theta < 50$ degrees. Table 1 shows quantitative results on the selected 10 sequences. It shows that drift reduction is effective. For an impression of remaining tracking errors, see the last frame of the longest sequence in Figure 1, with $\max(\text{ELE}_{left}, \text{ELE}_{right}) = 0.059$. From close inspection, we found that Table 1 cannot be used to judge the relative merits of drift reduction variants (re-use vs. probabilistic). This is because differences are so small that evaluation effects start interfering, e.g. annotated-eye-center offset between open/closed eyes, and initial-pose inaccuracy (they explain minor anomalies for Sequence 8 and 3). Ongoing work for multi-camera extension may use other criteria. Here, we cautiously state that, generally, mean ELE is below 0.09.

## 6. CONCLUSIONS

We presented a new algorithm for face tracking under challenging conditions, using a cylinder head model and 3D pose tracking by alignment of dynamically extracted templates based on dense-HOG appearance. We evaluated the algorithm on single-camera videos of moving infants in bed, relaxed or in pain. The full, probabilistic, algorithm was compared with a basic version without drift reduction, as well as with a non-probabilistic version. Experiments show good tracking behavior for poses up to 50 degrees from upright-frontal, with mean ELE below 0.09. The available criteria, however, are inconclusive as to the relative merits of the probabilistic extension. The algorithm can be extended for multi-camera setups, for use as part of an infant pain monitoring system in a clinical context.

## REFERENCES


[1] C Li, S Zinger, W.E. Tjon a Ten, and P.H.N. de With, "Video-based Discomfort Detection for Infants Using a Constrained Local Model," in *Int. Conf. Systems, Signals and Image Processing IWSSIP*, 2016, pp. 81–84.

[2] B. Slaats, S. Zinger, P.H.N. de With, W.E. 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*, 2015, pp. 50–57.

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

[4] 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 (accepted for publication)*, 2016.

[5] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in *IEEE Conf. Computer Vision and Pattern Recognition CVPR*, 2013, pp. 532–539.

[6] J.M. Saragih, S. Lucey, and J.F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. Journ. Computer Vision*, vol. 91, no. 2, pp. 200–215, sep 2011.

[7] E. Murphy-Chutorian and M. M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, 2009.

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

[9] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision.," in *Int. Joint Conf. Artificial Intelligence IJCAI*, 1981, vol. 81, pp. 674–679.

[10] 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*, vol. 13, no. 1, pp. 85–94, 2003.

[11] J. Harguess, C. Hu, and J.K. Aggarwal, "Full-motion recovery from multiple video cameras applied to face tracking and recognition," in *IEEE Int. Conf. Computer Vision Workshops ICCV Workshops*, 2011, pp. 1889–1896.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Computer Vision and Pattern Recognition CVPR*, 2005, vol. 1, pp. 886–893.

[13] E. Antonakos, J. Alabort-i Medina, G. Tzimiropoulos, and S.P. Zafeiriou, "Feature-Based Lucas-Kanade and Active Appearance Models," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2617–2632, 2015.

[14] A. Rahimi, L.-P. Morency, and T. Darrell, "Reducing drift in differential tracking," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 97–111, feb 2008.

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

[16] S. Baker and I. Matthews, "Lucas-Kanade 20 Years On: A Unifying Framework," *Int. Journ. Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[17] H.M.T. Menegaz, J.Y. Ishihara, G.A. Borges, and A.N. Vargas, "A Systematization of the Unscented Kalman Filter Theory," *IEEE Trans. Automatic Control*, vol. 60, no. 10, pp. 2583–2598, 2015.