Physiological Modelling for Improved Reliability in Silhouette-Driven Gradient-Based Hand Tracking

Paris Kaimakis and Joan Lasenby
Signal Processing Group, Department of Engineering, University of Cambridge
{pk228, jl}@eng.cam.ac.uk http://www-sigproc.eng.cam.ac.uk/~pk228

Abstract

We present a gradient-based motion capture system that robustly tracks a human hand, based on abstracted visual information — silhouettes. Despite the ambiguity in the visual data and despite the vulnerability of gradient-based methods in the face of such ambiguity, we minimise problems related to misfit by using a model of the hand's physiology, which is entirely non-visual, subject-invariant, and assumed to be known a priori. By modelling seven distinct aspects of the hand's physiology we derive prior densities which are incorporated into the tracking system within a Bayesian framework. We demonstrate how the posterior is formed, and how our formulation leads to the extraction of the maximum a posteriori estimate using a gradient-based search. Our results demonstrate an enormous improvement in tracking precision and reliability, while also achieving near real-time performance.

1. Introduction

Markerless motion capture has traditionally been approached as a data-fitting optimisation problem: given a choice of visual information, motion capture systems are usually asked to obtain the one state of the articulated body that best conforms to the data, leading to a search for the maximum likelihood (ML) estimate for the subject's state. However, the visual data often constitute an ambiguous source of information, and several different states often provide a good fit. Hence, the associated likelihood is multimodal, and the state suggested by the data often represents a pose impossible for the subject to ever attain.

Several authors have concentrated on the formulation of better-conditioned likelihood functions, via consideration of higher-level 3D data whose degree of ambiguity is substantially reduced. Indeed, some of the most impressive tracking results that the markerless community has seen were achieved by systems running on correlation data [4, 11, 18], voxel data [3, 8, 9], and depth maps obtained with a structured light sensor [1]. Unfortunately dedicated hardware is needed for the acquisition of such data, meaning that tracking may be expensive, difficult to set up, and that it remains, for the case of those systems relying on use of a structured light sensor, to some extent intrusive.

Instead, by introducing insight on the process that generates the data, Bayesian inference provides an alternative approach to tackling the likelihood's problematic multimodality while avoiding the use of expensive sensors. The idea of using non-visual information for circumventing the ambiguity in the visual data is also supported by recent findings in neuroscience: it has been noted [2, 17] that the middle temporal cortex believed to be responsible for higher-level visual processing in the brain receives a large proportion of its inputs from other non-visual cortices, implying that vision should be regarded as an inference problem, not as one of data fitting. Process models have been used substantially lately in the literature in order to better refine the estimate for the state. In the context of motion capture, Bayesian inference takes the role of consideration of the subject's physiology, but systems that significantly benefit from such formulation are invariably sampling-based [14, 15, 16]. Instead, gradient-based systems tend to use process models that are too general as they only consider the subject's inertia [1, 9], while the more recently introduced regression-based systems use models that are too specific as they directly consider the subject's motion through training [6, 10].

In this paper we take the position that the degree of generality of the process model is not a subjective one, and is instead dictated by the subject's physiology. We acknowledge the lack of gradient-based systems that also model this physiology, and argue that design of such systems will improve robustness against divergence, while avoiding the use of dedicated hardware and the high computational costs of sampling-based methods.

2. Background

Let $\mathbf{x}_k \in \mathbb{R}^S$ and $\mathbf{Z}_k \in \mathbb{R}^P$ be, respectively, the subject's hidden state and silhouette observation at time k, where

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

$P = \mathcal{O}(10^4)$ is the number of pixels in the data and $S = 26$. Furthermore, let $\mathbf{H}(\mathbf{x}_k)$ be a model of the silhouette observation as described *e.g.* in [7]. Then, the likelihood is usually assumed to be a Gaussian in observation-space, *i.e.*,

$$p(\mathbf{Z}_k|\mathbf{x}_k) = \mathcal{N}\big(\mathbf{H}(\mathbf{x}_k)|\mathbf{Z}_k, \sigma^2\mathbf{I}\big) \tag{1}$$

where we assume $\sigma = 1$. The multimodality of the likelihood in state-space is induced by the strong non-linearity of the mapping $\mathbf{H}$. Given an appropriate initialisation, an iterative gradient-based search for the ML state is given by the Gauss-Newton method for non-linear least squares (LS):

$$\hat{\mathbf{x}}_k^{i+1} = \hat{\mathbf{x}}_k^i + \left[\mathbf{J}_{\mathbf{H}}^{\mathrm{T}}\big(\hat{\mathbf{x}}_k^i\big)\ \mathbf{J}_{\mathbf{H}}\big(\hat{\mathbf{x}}_k^i\big)\right]^{-1} \mathbf{J}_{\mathbf{H}}^{\mathrm{T}}\big(\hat{\mathbf{x}}_k^i\big) \left[\mathbf{Z}_k - \mathbf{H}\big(\hat{\mathbf{x}}_k^i\big)\right] \tag{2}$$

where $\hat{\mathbf{x}}_k^i$ is the estimate for the current state at the i^{th} iteration, and $\mathbf{J}_{\mathbf{H}} \equiv [\nabla_{\mathbf{x}_k} \mathbf{H}]^{\mathrm{T}}$ is the Jacobian of $\mathbf{H}$. Instead, a Bayesian framework replaces the likelihood with the posterior as the basis for inference, whose convexity is much stronger. This is given by Bayes' theorem as

$$p(\mathbf{x}_k|\mathbf{Z}_{1:k}) \propto p(\mathbf{Z}_k|\mathbf{x}_k)\ p(\mathbf{x}_k|\mathbf{Z}_{1:k-1}) \tag{3}$$

where $p(\mathbf{x}_k|\mathbf{Z}_{1:k-1})$ is the prior derived by a consideration of the state-generating process. We assume an improper prior imposed by the physiology of the subject's articulated body:

$$p(\mathbf{x}_k|\mathbf{Z}_{1:k-1}) = \begin{cases} 1 & \text{if} \quad \mathbf{x}_k \in \mathbb{V} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbb{V} \subset \mathbb{R}^S$ is the region of validity for the state, such that the subject's physiology is respected.

Despite its obvious non-Gaussianity in state-space, in Section 3 we show how this improper prior can be formulated as a Gaussian in *physiology*-space, a space described through the non-linear mapping $\boldsymbol{\eta}(\mathbf{x}_k)$. Using this definition of the physiological prior, we also formulate the posterior as a Gaussian in a space defined through the mapping $\boldsymbol{\mathcal{H}}(\mathbf{x}_k)$. Hence, the Gauss-Newton method for non-linear LS may be used to converge to the maximum *a posteriori* (MAP) estimate, meaning that tracking becomes more reliable, while maintaining the advantages of a gradient-based search. Section 4 investigates the hand's physiology in greater detail, and in Section 5 we demonstrate the superiority of this formulation via a comparative analysis of our physiological modelling versus a purely data-fitting system. A brief discussion and conclusions follow in Section 6.

3. Bayesian Framework

Beyond the state's dynamics due to the subject's inertia, the state is also subject to limitations arising from the non-convex shape of the articulated body, and from different kinds of functional limitations due to the evolutionary design of the hand as a prehensile tool. Clearly, the nature of the physiological aspects that need to be modelled varies greatly. The prior addresses the diversity of these physiological limitations by computing

$$p(\mathbf{x}_k|\mathbf{Z}_{1:k-1}) = \prod_{\varphi=1}^{\Phi} \pi_\varphi(\mathbf{x}_k) \tag{5}$$

where $\Phi = 7$ represents the total number of physiological aspects φ addressed here, and π_φ is the prior specific to one such type of physiology. In general there is a total of K_φ constraints κ associated with each physiological aspect φ, and each constraint regarding any part of the subject's physiology can be unambiguously represented as $\{\varphi, \kappa\}$. The physiology-specific priors can be further decomposed as

$$\pi_\varphi(\mathbf{x}_k) = \prod_{\kappa=1}^{K_\varphi} \pi_{\{\varphi,\kappa\}}(\mathbf{x}_k) \tag{6}$$

where the constraint-specific priors $\pi_{\{\varphi,\kappa\}}$ are now the building blocks of our physiological prior. Each constraint in general refers either to an individual bone b, or to an individual pair of bones $\{b_1, b_2\}$ from the subject's skeleton; the bone(s) relevant to all the constraints of physiological aspect φ are stored in $\mathbf{B}_\varphi \in \mathbb{N}^{K_\varphi}$ (or $\mathbf{B}_\varphi \in \mathbb{N}^{K_\varphi \times 2}$ accordingly).

For every constraint $\{\varphi, \kappa\}$ addressed, we use the state in order to establish a measure of validity $\lambda_{\{\varphi,\kappa\}}(\mathbf{x}_k)$. For the purposes of assessing whether a given state $\mathbf{x}_k$ adheres to an individual constraint, it is also necessary to define upper and lower bounds $\lambda^{\max}_{\{\varphi,\kappa\}}$ and $\lambda^{\min}_{\{\varphi,\kappa\}}$ respectively, within which the measure of validity has to fall. Thus, a state $\mathbf{x}_k$ is only deemed valid if compatible with *every* constraint $\{\varphi, \kappa\}$ concerning the subject's physiology. In other words,

$$\left\{ \left\{ \lambda^{\min}_{\{\varphi,\kappa\}} \leq \lambda_{\{\varphi,\kappa\}}(\mathbf{x}_k) \leq \lambda^{\max}_{\{\varphi,\kappa\}} \right\}_{\kappa=1}^{K_\varphi} \right\}_{\varphi=1}^{\Phi} \iff \mathbf{x}_k \in \mathbb{V} \tag{7}$$

The constraint-specific priors are defined in a way that rewards valid states $\mathbf{x}_k \in \mathbb{V}$ as follows:

$$\pi_{\{\varphi,\kappa\}}(\mathbf{x}_k) = \begin{cases} a_1 & \text{if} \quad \lambda_{\{\varphi,\kappa\}} < \lambda^{\min}_{\{\varphi,\kappa\}} \\ 1 & \text{if} \quad \lambda^{\min}_{\{\varphi,\kappa\}} \leq \lambda_{\{\varphi,\kappa\}} \leq \lambda^{\max}_{\{\varphi,\kappa\}} \\ a_2 & \text{if} \quad \lambda_{\{\varphi,\kappa\}} > \lambda^{\max}_{\{\varphi,\kappa\}} \end{cases} \tag{8}$$

where

$$a_1 = \exp\left\{ -\tfrac{1}{2\alpha^2_{\{\varphi,\kappa\}}} \left(\lambda_{\{\varphi,\kappa\}}(\mathbf{x}_k) - \lambda^{\min}_{\{\varphi,\kappa\}} \right)^2 \right\} \tag{9}$$

$$a_2 = \exp\left\{ -\tfrac{1}{2\alpha^2_{\{\varphi,\kappa\}}} \left(\lambda_{\{\varphi,\kappa\}}(\mathbf{x}_k) - \lambda^{\max}_{\{\varphi,\kappa\}} \right)^2 \right\} \tag{10}$$

and a plot for this equation is shown in Figure 1. It is easy to see that the binary definition of the prior in (4) would imply

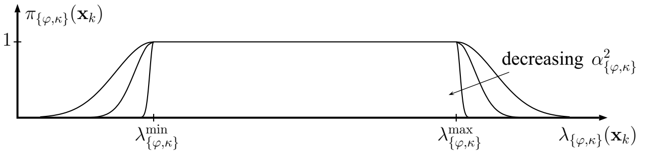

Figure 1. Prior specific to physiological aspect φ and constraint κ, as defined in equation (8).

a binary definition for the constraint-specific priors $\pi_{\{\varphi,\kappa\}}$, which is not exactly the case in equation (8) as Figure 1 illustrates. However, by adjusting the variance $\alpha^2_{\{\varphi,\kappa\}}$ to a small value, we can force the above definition to become binary, this way yielding an overall prior that closely approximates (4). The benefit of using (8) rather than a binary definition for $\pi_{\{\varphi,\kappa\}}$ is twofold. Firstly, it ensures that a gradient-based approach for the extraction of the MAP estimate may be established. Secondly, the variance $\alpha^2_{\{\varphi,\kappa\}}$ now controls the level to which the constraint-specific prior is allowed to influence any potentially erroneous suggestion coming from the likelihood, *i.e.* it indicates the importance of meeting the requirements of constraint $\{\varphi,\kappa\}$ over the measurement model's compatibility with the visual data. Hence, a big variance $\alpha^2_{\{\varphi,\kappa\}}$ indicates that the limitation imposed by constraint $\{\varphi,\kappa\}$ is a soft one, and can therefore be compromised to some extent for the sake of good compliance with the observations $\mathbf{Z}_k$. Conversely, a small $\alpha^2_{\{\varphi,\kappa\}}$ indicates that the constraint's limitations *must* be met regardless of whether the resulting state $\mathbf{x}_k$ is compatible with the observations or not. In general, the severity of each constraint depends on the aspect of physiology examined. It is useful, however, to first rewrite (8) as

$$\pi_{\{\varphi,\kappa\}}(\mathbf{x}_k) = \exp\left\{-\frac{1}{2}\Big(\eta_{\{\varphi,\kappa\}}(\mathbf{x}_k) - \zeta_{\{\varphi,\kappa\}}\Big)^2\right\} \quad (11)$$

where the new measure of validity $\eta_{\{\varphi,\kappa\}}$ is now given by

$$\eta_{\{\varphi,\kappa\}}(\mathbf{x}_k) = \begin{cases} \frac{\lambda_{\{\varphi,\kappa\}}(\mathbf{x}_k)}{\alpha_{\{\varphi,\kappa\}}} & \text{if } \lambda_{\{\varphi,\kappa\}} < \lambda^{\min}_{\{\varphi,\kappa\}} \\ 0 & \text{if } \lambda^{\min}_{\{\varphi,\kappa\}} \leq \lambda_{\{\varphi,\kappa\}} \leq \lambda^{\max}_{\{\varphi,\kappa\}} \\ \frac{\lambda_{\{\varphi,\kappa\}}(\mathbf{x}_k)}{\alpha_{\{\varphi,\kappa\}}} & \text{if } \lambda_{\{\varphi,\kappa\}} > \lambda^{\max}_{\{\varphi,\kappa\}} \end{cases} \quad (12)$$

and the new bounds $\zeta_{\{\varphi,\kappa\}}$ are now

$$\zeta_{\{\varphi,\kappa\}} = \begin{cases} \frac{\lambda^{\min}_{\{\varphi,\kappa\}}}{\alpha_{\{\varphi,\kappa\}}} & \text{if } \lambda_{\{\varphi,\kappa\}} < \lambda^{\min}_{\{\varphi,\kappa\}} \\ 0 & \text{if } \lambda^{\min}_{\{\varphi,\kappa\}} \leq \lambda_{\{\varphi,\kappa\}} \leq \lambda^{\max}_{\{\varphi,\kappa\}} \\ \frac{\lambda^{\max}_{\{\varphi,\kappa\}}}{\alpha_{\{\varphi,\kappa\}}} & \text{if } \lambda_{\{\varphi,\kappa\}} > \lambda^{\max}_{\{\varphi,\kappa\}} \end{cases} \quad (13)$$

Hence, the constraint-specific prior, although non-Gaussian in state-space, may be expressed as a Gaussian in the space defined by $\eta_{\{\varphi,\kappa\}}(\mathbf{x})$. Using (11) we can build up a similar representation for the physiology-specific priors π_φ and, in turn, for the overall prior. From equation (6) we have

$$\begin{aligned} \pi_\varphi(\mathbf{x}_k) &= \prod_{\kappa=1}^{K_\varphi} \pi_{\{\varphi,\kappa\}}(\mathbf{x}_k) \\ &= \exp\left\{-\frac{1}{2}\Big(\boldsymbol{\eta}_\varphi(\mathbf{x}_k) - \boldsymbol{\zeta}_\varphi\Big)^{\mathrm{T}}\Big(\boldsymbol{\eta}_\varphi(\mathbf{x}_k) - \boldsymbol{\zeta}_\varphi\Big)\right\} \quad (14) \end{aligned}$$

where the type-specific physiology model and bounds, $\boldsymbol{\eta}_\varphi$ and $\boldsymbol{\zeta}_\varphi$ respectively, are given by

$$\boldsymbol{\eta}_\varphi(\mathbf{x}_k) = \Big[\eta_{\{\varphi,1\}}(\mathbf{x}_k) \ldots \eta_{\{\varphi,\kappa\}}(\mathbf{x}_k) \ldots \eta_{\{\varphi,K_\varphi\}}(\mathbf{x}_k)\Big]^{\mathrm{T}} \quad (15)$$

$$\boldsymbol{\zeta}_\varphi = \Big[\ \zeta_{\{\varphi,1\}} \ \ \ldots \ \ \zeta_{\{\varphi,\kappa\}} \ \ \ldots \ \ \zeta_{\{\varphi,K_\varphi\}}\ \Big]^{\mathrm{T}} \quad (16)$$

In a similar fashion, the type-specific priors are combined in order to give the overall prior as per (5):

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{Z}_{1:k-1}) &= \prod_{\varphi=1}^{\Phi} \pi_\varphi(\mathbf{x}_k) \\ &= \exp\left\{-\frac{1}{2}\Big(\boldsymbol{\eta}(\mathbf{x}_k) - \boldsymbol{\zeta}\Big)^{\mathrm{T}}\Big(\boldsymbol{\eta}(\mathbf{x}_k) - \boldsymbol{\zeta}\Big)\right\} \quad (17) \end{aligned}$$

where the physiology model $\boldsymbol{\eta}$ and bounds $\boldsymbol{\zeta}$ now cover the complete physiology of the body and are given by

$$\boldsymbol{\eta}(\mathbf{x}_k) = \Big[\boldsymbol{\eta}_1^{\mathrm{T}}(\mathbf{x}_k) \ \ldots \ \boldsymbol{\eta}_\Phi^{\mathrm{T}}(\mathbf{x}_k)\Big]^{\mathrm{T}} \quad \boldsymbol{\zeta} = \Big[\boldsymbol{\zeta}_1^{\mathrm{T}} \ \ldots \ \boldsymbol{\zeta}_\Phi^{\mathrm{T}}\Big]^{\mathrm{T}} \quad (18)$$

Finally using (1), (17) and (3), the posterior becomes

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{Z}_{1:k}) &\propto p(\mathbf{Z}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{Z}_{1:k-1}) \\ &\propto \exp\left\{-\frac{1}{2}\Big(\boldsymbol{\mathcal{H}}(\mathbf{x}_k) - \boldsymbol{\mathcal{Z}}_k\Big)^{\mathrm{T}}\Big(\boldsymbol{\mathcal{H}}(\mathbf{x}_k) - \boldsymbol{\mathcal{Z}}_k\Big)\right\} \quad (19) \end{aligned}$$

where $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{Z}}$ are obtained by augmenting the measurement model and observations with the physiology model and bounds respectively, as follows:

$$\boldsymbol{\mathcal{H}}(\mathbf{x}_k) = \Big[\mathbf{H}^{\mathrm{T}}(\mathbf{x}_k) \quad \boldsymbol{\eta}^{\mathrm{T}}(\mathbf{x}_k)\Big]^{\mathrm{T}} \quad \boldsymbol{\mathcal{Z}}_k = \Big[\mathbf{Z}_k^{\mathrm{T}} \quad \boldsymbol{\zeta}^{\mathrm{T}}\Big]^{\mathrm{T}} \quad (20)$$

Thus, by considering the bounds $\boldsymbol{\zeta}$ violated by the state as observations, and by augmenting the dimensionality of the measurement model to allow for the subject's physiology, it is possible to define the posterior as a Gaussian in the space defined by $\boldsymbol{\mathcal{H}}(\mathbf{x}_k)$. A similar formulation to the one presented here was previously used in [12] although not in the context of the physiological characteristics of the subject. Given an appropriate initialisation, the MAP estimate can now be located in much the same way as per the search carried out for the ML estimate, *i.e.* via a few iterations of the Gauss-Newton method for non-linear LS:

$$\hat{\mathbf{x}}_k^{i+1} = \hat{\mathbf{x}}_k^{i} + \Big[\mathbf{J}_{\boldsymbol{\mathcal{H}}}^{\mathrm{T}}(\hat{\mathbf{x}}_k^{i}) \ \mathbf{J}_{\boldsymbol{\mathcal{H}}}(\hat{\mathbf{x}}_k^{i})\Big]^{-1} \mathbf{J}_{\boldsymbol{\mathcal{H}}}^{\mathrm{T}}(\hat{\mathbf{x}}_k^{i}) \ \Big[\boldsymbol{\mathcal{Z}}_k - \boldsymbol{\mathcal{H}}(\hat{\mathbf{x}}_k^{i})\Big] \quad (21)$$

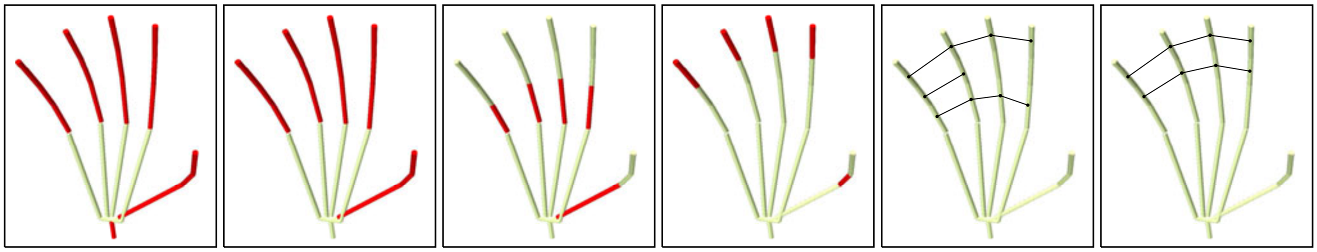

Figure 2. The highlighted bones and linked pairs of bones indicate those subject to the physiological limitations relevant to (from left to right): inertia, flexion, abduction, intradigital correlation, transdigital correlation, and friction.

where

$$\mathbf{J}_{\mathcal{H}}(\mathbf{x}_k) \;=\; \Big[\mathbf{J}_{\mathbf{H}}^{\mathrm{T}}(\mathbf{x}_k) \quad \mathbf{J}_{\boldsymbol{\eta}}^{\mathrm{T}}(\mathbf{x}_k)\Big]^{\mathrm{T}} \tag{22}$$

and $\mathbf{J}_{\boldsymbol{\eta}} \equiv [\nabla_{\mathbf{x}_k}\boldsymbol{\eta}]^{\mathrm{T}}$ is the Jacobian of the physiology model $\boldsymbol{\eta}$ consisting of the individual Jacobians $\mathbf{J}_{\boldsymbol{\eta}_\varphi}$ of all the physiology-specific measures of validity $\boldsymbol{\eta}_\varphi$:

$$\mathbf{J}_{\boldsymbol{\eta}}(\mathbf{x}_k) = \Big[\mathbf{J}_{\boldsymbol{\eta}_1}^{\mathrm{T}}(\mathbf{x}_k)\ \ldots\ \mathbf{J}_{\boldsymbol{\eta}_\varphi}^{\mathrm{T}}(\mathbf{x}_k)\ \ldots\ \mathbf{J}_{\boldsymbol{\eta}_\Phi}^{\mathrm{T}}(\mathbf{x}_k)\Big]^{\mathrm{T}} \tag{23}$$

4. Physiological Modelling

In this section we address $\Phi = 7$ aspects of the human hand's physiology. Some of these are more easily understood than others, and the interested reader should refer to [5] or [13] for an in-depth description of the physiological aspects addressed in this paper. For each different aspect φ we consider the K_φ constraints relevant, and indicate the group of bones $\mathbf{B}_\varphi$ subject to this type of physiological limitation. We specify an appropriate measure of validity $\lambda_{\{\varphi,\kappa\}}$, give the bounds $\lambda^{\max}_{\{\varphi,\kappa\}}$ and $\lambda^{\min}_{\{\varphi,\kappa\}}$ within which this has to lie, and choose the standard deviation $\alpha_{\{\varphi,\kappa\}}$.

4.1. Inertia

The first and perhaps the most obvious aspect of the subject's physiology as a massive body yields a consideration of its inertia, leading to the imposition of continuity constraints on the dynamics of the articulated structure, which are modelled here through the *inertia* prior, $\varphi = 1$.

Considering the subject's skeleton as a first-order kinematic structure, the measure of validity for the inertia prior becomes the velocity of all moving parts of the skeleton at time k, which needs to be matched to that at time $k-1$. There is an overall spatial velocity $\dot{\mathbf{x}}_{\mathrm{s}}$ for the skeleton's root, and one angular velocity $\boldsymbol{\omega}_b$ for each moving bone. Continuity of the former is enforced by the additional constraint $\{\varphi = 1, \kappa = 0\}$, for which

$$\boldsymbol{\lambda}_{\{1,0\}}(\mathbf{x}_k) = \dot{\mathbf{x}}_{\mathrm{s}}(\mathbf{x}_k) \qquad \boldsymbol{\lambda}^{\max}_{\{1,0\}} = \boldsymbol{\lambda}^{\min}_{\{1,0\}} = \dot{\mathbf{x}}_{\mathrm{s}}(\mathbf{x}_{k-1}) \tag{24}$$

while continuity of the latter is enforced by defining

$$\boldsymbol{\lambda}_{\{1,\kappa\}}(\mathbf{x}_k) = \boldsymbol{\omega}_b(\mathbf{x}_k) \qquad \boldsymbol{\lambda}^{\max}_{\{1,\kappa\}} = \boldsymbol{\lambda}^{\min}_{\{1,\kappa\}} = \boldsymbol{\omega}_b(\mathbf{x}_{k-1}) \tag{25}$$

where $b = (\mathbf{B}_1)_\kappa$ and $\mathbf{B}_1 \in \mathbb{N}^{K_1}$ stores all the of the hand's phalanges, the thumb's metacarpal and the wrist (see Figure 2). Since the inertia of the subject's skeleton only gives a general indication about the current state, the inertia constraints were relaxed by setting an appropriately large value for the associated standard deviation. For this reason we have set $\alpha_{\{1,\kappa\}} = 20^{\mathrm{o}}$ per frame, and $\alpha_{\{1,0\}} = 5$cm per frame.

4.2. Flexion

The human hand's evolutionary design has imposed a prehensile nature to it which favours functional asymmetry: the hand may well access and manipulate objects situated on its ventral side, but is physically incapable of any other manipulation. As a consequence there are limits to the amount of flexion and extension that the hand's fingers can undergo. The role of the *flexion* prior, $\varphi = 2$, is to penalise any values of the current state that would otherwise allow the hand's fingers to hyperextend, or to hyperflex.

The flexion prior's measure of validity and its upper and lower bounds refer to the angle of flexion θ of each of the bones listed in $\mathbf{B}_2$. In other words, for every flexion constraint $\{2, \kappa\}$ we have,

$$\lambda_{\{2,\kappa\}}(\mathbf{x}_k) = \theta_b(\mathbf{x}_k) \qquad \lambda^{\max}_{\{2,\kappa\}} = \theta^{\max}_b \qquad \lambda^{\min}_{\{2,\kappa\}} = \theta^{\min}_b \tag{26}$$

where $b = (\mathbf{B}_2)_\kappa$, and $\mathbf{B}_2 \in \mathbb{N}^{K_2}$ stores all of the hand's phalanges and the thumb's metacarpal (see Figure 2). For most of these bones the upper bounds are defined as $\theta^{\max}_b = 90^{\mathrm{o}}$, this way allowing enough flexion for the formation of the fist, but not more. Similarly, the majority of the lower bounds are defined as $\theta^{\min}_b = -10^{\mathrm{o}}$, meaning that a limited degree of hyperextension is allowed. Finally, we set the standard deviation for each flexion constraint to $\alpha_{\{2,\kappa\}} = 5^{\mathrm{o}}$.

4.3. Abduction

Beyond the limitations in terms of flexion and extension, the hand's prehensile physiology is also responsible for a limitation in the amount of abduction and adduction (*i.e.* sideways rotation) that the fingers might undergo. These limitations are modelled by the *abduction* prior, $\varphi = 3$.

The abduction prior is only relevant to the four fingers' proximal phalanges and to the thumb's metacarpal; these bones are stored in $\mathbf{B}_3 \in \mathbb{N}^{K_3}$ and are shown in Figure 2.

The amount of abduction exercised by each $b \in \mathbf{B}_3$ is indicated by the abduction angles ϕ. We also define the upper and lower bounds of the abduction angles to be such that the amount of tolerable abduction is equal to the amount of tolerable adduction that the bones can exercise:

$$\lambda_{\{3,\kappa\}}(\mathbf{x}_k) = \phi_b(\mathbf{x}_k) \qquad \lambda^{\max}_{\{3,\kappa\}} = \phi^{\max}_b \qquad \lambda^{\min}_{\{3,\kappa\}} = -\phi^{\max}_b \tag{27}$$

for every bone $b = (\mathbf{B}_3)_\kappa$. The abduction bound $\phi^{\max}_b$ is *flexion*-dependent and is given in degrees as

$$\phi^{\max}_b(\theta_b) = \frac{\phi^{\max\min}_b - \phi^{\max\max}_b}{\theta^{\max}_b - \theta^{\min}_b}\left(\theta_b - \theta^{\min}_b\right) + \phi^{\max\max}_b \tag{28}$$

where $\phi^{\max\max}_b = \phi^{\max}_b(\theta^{\min}_b)$ and $\phi^{\max\min}_b = \phi^{\max}_b(\theta^{\max}_b)$ denote the maximum and minimum values respectively, of the abduction bound $\phi^{\max}_b$. With the flexion bounds $\theta^{\max}_b$ and $\theta^{\min}_b$ defined in Section 4.2, we set $\phi^{\max\max}_b = 40^{\mathrm{o}}$ and $\phi^{\max\min}_b = 0^{\mathrm{o}}$ for every proximal phalanx b. This has the effect of completely restricting a finger from *any* abduction if that finger has experienced its maximum allowed level of flexion, while preserving a generous degree of abductive ability for a hyperextended finger. Since the thumb's abduction bound demonstrates no dependence on the extent of flexion, the associated abduction constraint $\{3, 5\}$ is partially relaxed by placing $\phi^{\max\max}_{b=23} = \phi^{\max\min}_{b=23} = 50^{\mathrm{o}}$. Finally, the abduction standard deviation is set to $\alpha_{\{3,\kappa\}} = 5^{\mathrm{o}}$.

4.4. Intradigital Correlation

Muscles responsible for phalangeal flexion in a particular finger form tendinous synapses with more than one of that finger's phalanges, meaning that the relation between muscle contraction and phalangeal flexion is not a one-to-one mapping. Even more significant is the intricate interconnection formed between the tendons of the *flexor digitorum profundus* and the *flexor digitorum superficialis* muscles, which are the prime movers for the flexion of the distal and intermediate phalanges respectively. The *intradigital correlation* prior, $\varphi = 4$, is responsible for modelling such interdependencies.

The $K_4 = 5$ bones affected by the intradigital correlation constraints are the distal phalanges of the four fingers and the proximal phalanx of the thumb. These are illustrated in Figure 2 and are listed in $\mathbf{B}_4 \in \mathbb{N}^{K_4}$. The measure of validity is once again the amount of flexion that these bones undergo according to the current state, but in this case the upper and lower bounds are fixed to a target value specified by the flexion of the corresponding parent-bone. In other words, for every bone $b = (\mathbf{B}_4)_\kappa$,

$$\lambda_{\{4,\kappa\}}(\mathbf{x}_k) = \theta_b(\mathbf{x}_k) \qquad \lambda^{\max}_{\{4,\kappa\}} = \lambda^{\min}_{\{4,\kappa\}} = \mu_b\, \theta_{\gamma(b)} \tag{29}$$

where $\gamma(b)$ is the parent of bone b and typically $\mu_b = \frac{2}{3}$.

The intradigital correlation standard deviation is set to $\alpha_{\{4,\kappa\}} = 5^{\mathrm{o}}$ for all constraints $\{4, \kappa\}$ with the sole exception of the weaker correlation in the flexion of the thumb's proximal phalanx, for which we set $\alpha_{\{4,5\}} = 30^{\mathrm{o}}$.

4.5. Transdigital Correlation

Further to the intradigital correlation in flexion discussed above, the *palmar ligament* of the hand shared by the four fingers causes part of the flexion of a proximal phalanx to be transmitted across neighbouring fingers. In addition, the *extensor digitorum* muscle, which is active during digital extension (as the prime mover) *and* during flexion (as the antagonist) forms tendinous synapses with more than one finger. These produce correlation in the flexion between phalanges across neighbouring fingers, which we model here with the *transdigital correlation* prior, $\varphi = 5$.

There are $K_5 = 7$ transdigital correlation constraints in our model and each constraint $\{5, \kappa\}$ is associated with a unique pair of phalanges $\{b_1, b_2\}$, which is stored in $\mathbf{B}_5 \in \mathbb{N}^{K_5 \times 2}$ (see Figure 2). We define the measure of validity $\lambda_{\{5,\kappa\}}$ as the *relative* amount of flexion $\Delta\theta_{\{b_1,b_2\}} > 0$ that separates the two bones. Hence, for every pair of bones $b_1 = (\mathbf{B}_5)_{\kappa,1}$ and $b_2 = (\mathbf{B}_5)_{\kappa,2}$,

$$\lambda_{\{5,\kappa\}}(\mathbf{x}_k) = \Delta\theta_{\{b_1,b_2\}}(\mathbf{x}_k) \qquad \lambda^{\max}_{\{5,\kappa\}} = \Delta\theta^{\max}_{\{b_1,b_2\}} \tag{30}$$

where $\Delta\theta^{\max}_{\{b_1,b_2\}}$ is known *a priori* and represents the upper bound on the relative flexion between bones b_1 and b_2. Note that the value of $\Delta\theta^{\max}_{\{b_1,b_2\}}$ may depend on *which* of b_1 and b_2 is causing its counterpart to flex. Due to the bigger variation in transdigital correlation across subjects, it was decided to partially relax these constraints by allowing for a bigger standard deviation $\alpha_{\{5,\kappa\}} = 10^{\mathrm{o}}$.

4.6. Rigidity

The role of the hand as the most dexterous part of the human body suggests that a considerable degree of interaction between fingers should be expected to take part in any of the subject's gestures, despite the significant reduction in their mobility due to the physiological constraints that have already been visited. It is during such interactions that the silhouettes of different fingers in the visual data merge, causing the information content in the observations to deteriorate. Under such circumstances the tracking system is more vulnerable to a misfit, meaning that the ML state is now closely surrounded by local maxima in the likelihood. Because the earliest stages of a misfit often involve the mutual intersection of neighbouring fingers of the skeleton, the main motivation in formulating the *rigidity* prior, $\varphi = 6$, is to increase robustness against such misfits.

Before modelling the hand's rigidity it is worth noting that many of the bones are guaranteed *not* to intersect each

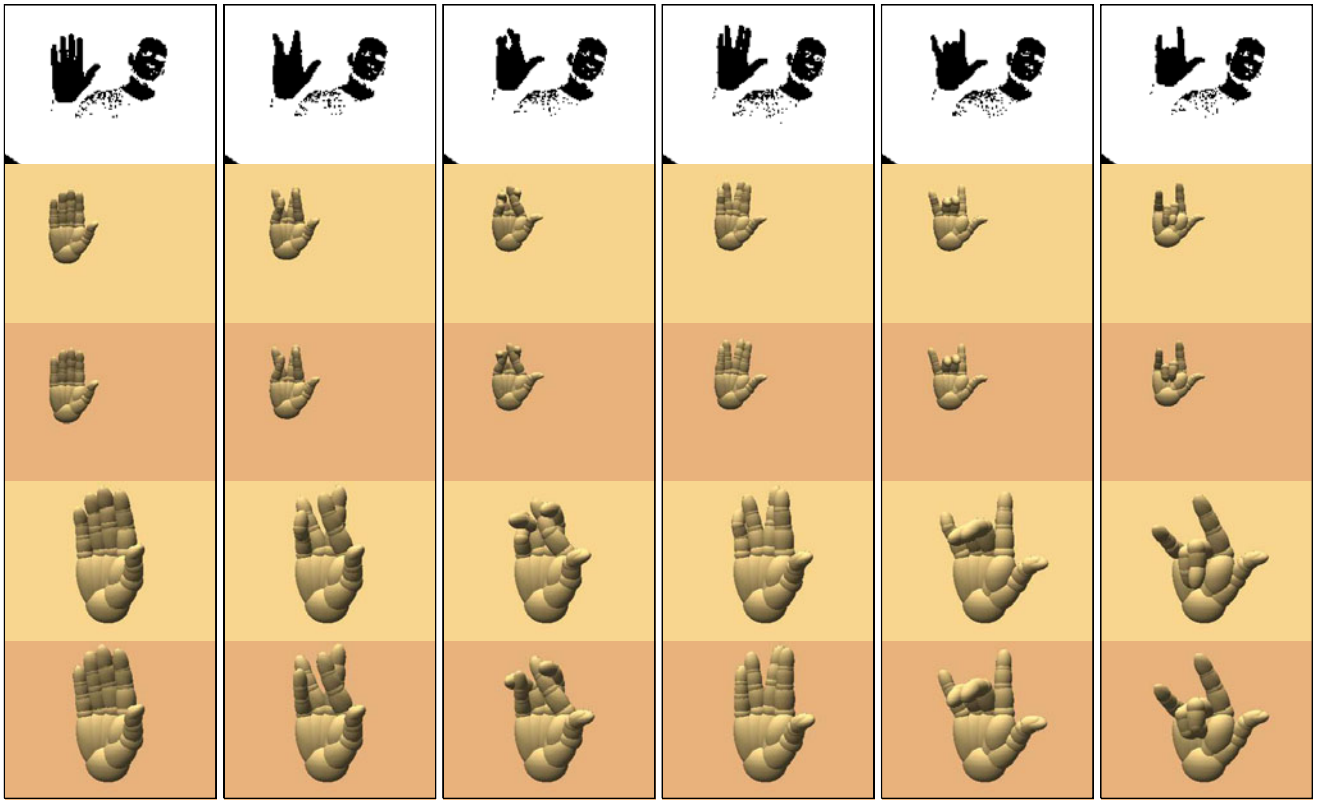

Figure 3. Three-camera experiment on a sequence with a significant degree of self-occlusion. Here we show the data as recorded by one of the cameras (1st row) and a reconstruction of the hand's pose, before application of the physiological prior (lighter background) and after (darker background), from a view that matches that of the data (2nd and 3rd rows) and from an arbitrarily chosen view (4th and 5th rows). The extracted MAP motion (darker background) is more robust against problems related to intersection of the fingers (3rd snapshot), phalangeal hyperextension (2nd and 4th snapshots), and intradigital correlation (6th snapshot).

other thanks to the activity of other priors that have been visited already — *e.g.* the flexion prior guarantees no intersection between phalanges of the same finger, and also prevents intersection of the carpals and metacarpals. Hence, the total number of unique pairs of bones that need to be addressed for the formation of the rigidity prior reduces to $K_6 = 42$; all these pairs are listed in $\mathbf{B}_6 \in \mathbb{N}^{K_6 \times 2}$.

In order to assess whether bones b_1 and b_2 have intersected each other, we define the corresponding measure of validity $\lambda_{\{6,\kappa\}}$ to be associated with the shortest interphalangeal distance $\delta_{\{b_1,b_2\}}$. In fact, it is easiest to work with the *squared* shortest interphalangeal distance, because this makes calculations easier. In any case, the lowest bound for this distance relates to the radii ρ_{b_i} of these bones, which are assumed to be known *a priori*. Hence, for every pair of bones $b_1 = (\mathbf{B}_6)_{\kappa,1}$ and $b_2 = (\mathbf{B}_6)_{\kappa,2}$,

$$\lambda_{\{6,\kappa\}}(\mathbf{x}_k) = \delta^2_{\{b_1,b_2\}}(\mathbf{x}_k) \qquad \lambda^{\text{min}}_{\{6,\kappa\}} = (\rho_{b_1} + \rho_{b_2})^2 \quad (31)$$

The standard deviation for each rigidity constraint is finally set to correspond to half of the minimum interphalangeal distance permissible *i.e.* $\alpha_{\{6,\kappa\}} = \left[\frac{1}{2}(\rho_{b_1} + \rho_{b_2})\right]^2$.

4.7. Friction

Further to the implications of transdigital interaction with regards to the rigidity of the fingers, the nature of the skin as a rough surface produces frictional forces which tend to reduce the amount of transdigital slip. In other words, the skin causes motion to be transmitted from one finger to another as the fingers come to contact during gestures, and particularly during the formation of the fist. The *friction* prior, $\varphi = 7$, accounts for the skin's 'sticky' nature.

In order to form the friction prior we only take into consideration the four fingers' distal and intermediate phalanges, which would otherwise undertake most of the transdigital slip for the majority of the gestures anticipated. The $K_7 = 6$ pairs of bones subject to the friction constraints are shown in Figure 2 and are listed in $\mathbf{B}_7 \in \mathbb{N}^{K_7 \times 2}$.

Since the purpose of the friction prior is to limit the amount of transdigital slip, we associate the measure of validity with the (squared) shortest interphalangeal distance, which, in the case of two fingers coming to contact, needs to be maintained roughly fixed. We define the measure of validity relevant to the friction between two phalanges

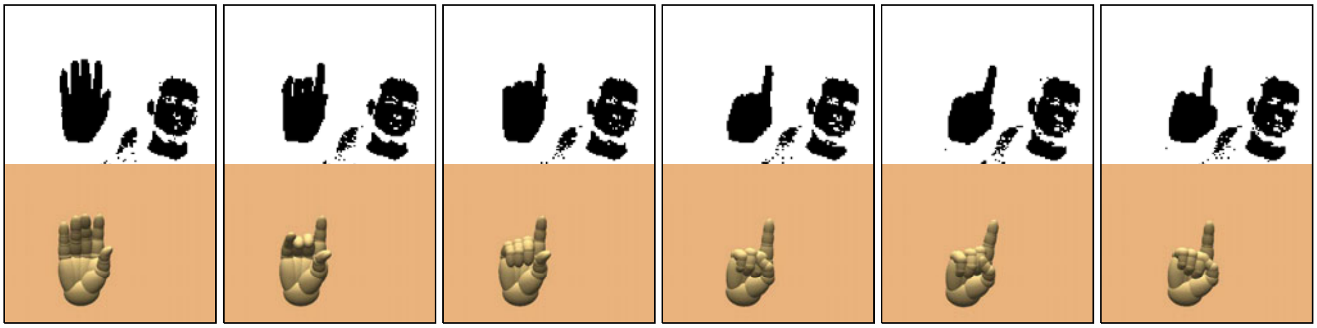

Figure 4. Three-camera MAP tracking on a sequence with an extended degree of self-occlusion. Despite the ambiguity in the silhouettes, the extracted motion maintains a high level of precision. Our tracking system can follow small and subtle changes in the subject's state, even when the degree of self-occlusion is at its highest (*e.g.* the partial movement of the fully flexed fingers in the 5th and 6th snapshots).

$b_1 = (\mathbf{B}_7)_{\kappa,1}$ and $b_2 = (\mathbf{B}_7)_{\kappa,2}$ as

$$\lambda_{\{7,\kappa\}}(\mathbf{x}_k) = \begin{cases} \delta^2_{\{b_1,b_2\}}(\mathbf{x}_k) & \text{if} \quad a_3 \leq \delta_{\{b_1,b_2\}} \leq a_4 \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

where $a_3 = 0.9(\rho_{b_1} + \rho_{b_2})$ and $a_4 = 1.1(\rho_{b_1} + \rho_{b_2})$. Hence, if the two phalanges are not in contact, $\delta_{\{b_1,b_2\}} > a_4$ and the friction prior is inactive. The friction prior is also inactive when $\delta_{\{b_1,b_2\}} < a_3$, leaving violations due to phalangeal intersection to be handled by the rigidity prior. In a similar fashion, the target value for $\lambda_{\{7,\kappa\}}$ is defined as

$$\lambda^{\max}_{\{7,\kappa\}} = \lambda^{\min}_{\{7,\kappa\}} = \begin{cases} a_3^2 & \text{if} \quad a_3 \leq \delta_{\{b_1,b_2\}} \leq a_4 \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

and we set the standard deviation for friction to half of the target value specified above, *i.e.* $\alpha_{\{7,\kappa\}} = \left(\frac{1}{2}a_3\right)^2$.

5. Results

Beyond the computations relevant to the models of Sections 4.1 to 4.7 we used equations (12) to (20) to form $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{Z}}$. We then performed 5 iterations of equation (21) to converge to the MAP estimate for the state at any time k in a sequence. The initial state $\hat{\mathbf{x}}_1^0$ for the first frame was taken to be the hand's position of extension, meaning that only the 6 state parameters responsible for global position and orientation were manually initialised.

Figure 3 makes a comparison of the results taken from an experiment based on real data captured with 3 synchronised, calibrated cameras running at 25Hz with pixel resolution 120×160 each. Due to the activity of the physiological prior, the extracted motion is natural and convincing, since the subject's physiology is respected — something which is not true for the results taken from the purely data-fitting system. Problems encountered during ML tracking relevant to phalangeal hyperextension, phalangeal intersection, and transdigital correlation are now solved without compromising the system's generality in tracking unanticipated gestures.

Figures 4 and 5 show two further three-camera experiments involving more challenging gestures that the data-fitting scenario failed to track robustly. Despite the extended degree of self-occlusion (which reaches its maximum with the formation of the fully clenched fist) and despite the ambiguous nature of the silhouette data, it is clear that use of the improper physiological prior provides adequate assistance to the gradient-based system for extracting a motion that is natural, precise and in agreement with the visual data. All the experiments outlined here as well as many additional ones can be found on the paper's website, http://www-sigproc.eng.cam.ac.uk/~pk228.

Calculation of the physiological model, bounds and Jacobian was implemented entirely in Matlab and took around 0.05s, subject to the number of physiological constraints violated. Along with the calculations relevant to the measurement model, processing of 5 iterations of the Gauss-Newton algorithm took less than 0.5s for each 3-camera frameset on a 1.8GHz, 2GB RAM Intel Centrino Duo processor running on Windows Vista. From an extensive evaluation of our system using data acquired with three cameras, it was found that gestures of various styles and degrees of complication were followed well, assuming a reasonable state initialisation (involving only 6 parameters), and well-segmented silhouette observations.

6. Conclusions

We have presented a gradient-based system for extracting the MAP state trajectory of the human hand (using 26 degrees of freedom), based on low-level and very often ambiguous visual information. At the core of our system's robustness lies an in-depth examination of seven aspects of the hand's physiology, some of which have never been addressed in the vision literature before. We have shown how the information arising from the physiology is combined with the measurements, leading to a system that can track any kind of unanticipated motion in near real-time, even in the presence of severe self-occlusion.

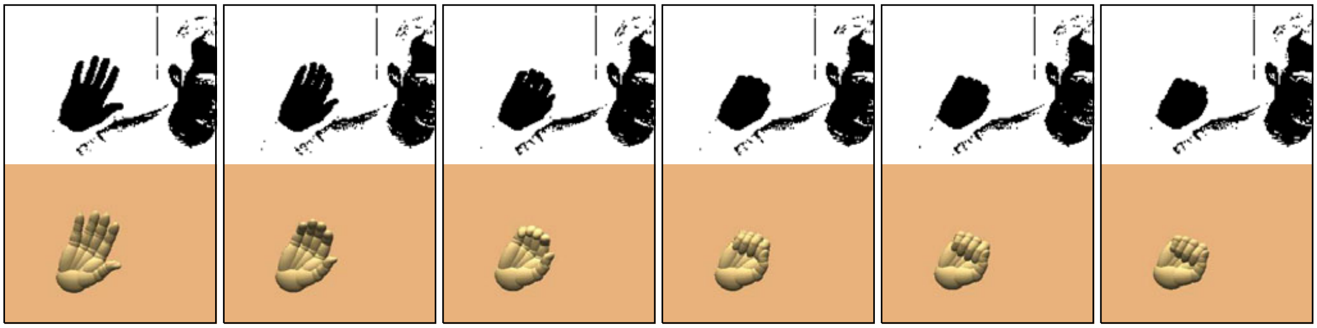

Figure 5. Three-camera MAP tracking on a sequence with a severe degree of self-occlusion. The ambiguity of the data reaches its maximum as the hand closes to a fist. Use of the physiological prior makes tracking possible and the extracted motion is natural and precise.

An important part of our contribution is that our physiology model may be used alongside *any* measurement model, whatever the choice of features used for tracking. Furthermore, physiological limitations affect the kinematic capability of *every* articulated body encountered in nature. Although Sections 4.1 to 4.7 are specific to the physiology of the human hand, the method with which the physiology-specific constraints are incorporated into the data-fitting tracking system is a generic one, and may thus be used regardless of the subject's underlying physiology.

The system presented in this paper can be extended to take into consideration further aspects of the hand's physiology such as the transdigital correlation in abduction, the opposition of the thumb to the rest of the fingers *e.g.* via the conditional relaxation of the intradigital correlation prior, or perhaps an angular extension to the friction model. Finally, we also aim to reduce the system's processing time for future real-time tracking.

References

[1] M. Bray, E. Koller-Meier and L. Van Gool, "Smart particle filtering for high-dimensional tracking", *Computer Vision and Image Understanding (CVIU)*, vol. 106, pp. 116–129, April 2007.

[2] G. Buzsáki, *Rhythms of the Brain*, Oxford University Press Inc., first ed., October 2006.

[3] F. Caillette and T. Howard, "Real-time markerless human body tracking with multi-view 3-D voxel reconstruction", *Proc. British Machine Vision Conference*, vol. 2, pp. 597–606, London, UK, September 6-9 2004.

[4] G. Dewaele, F. Devernay and R. Horaud, "Hand motion from 3D point trajectories and a smooth surface model", *Proc. European Conference on Computer Vision*, vol. 1, pp. 495–507, Prague, Czech Republic, May 11-14 2004.

[5] H. Gray, *Gray's Anatomy*, Running Press, unabridged ed., May 1974.

[6] K. Grochow, S. L. Martin, A. Hertzmann and Z. Popović, "Style-based inverse kinematics", *ACM Trans. on Graphics*, vol. 23(3), pp. 522–531, August 8-12 2004.

[7] P. Kaimakis and J. Lasenby, "Gradient-based hand tracking using silhouette data", *Proc. International Symposium on Visual Computing*, vol. 1, pp. 24–35, Lake Tahoe, NV/CA, USA, November 26-28 2007.

[8] B. Michoud, E. Guillou, H. Briceño and S. Bouakaz, "Real-time marker-free motion capture from multiple cameras", *Proc. International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, October 14-20 2007.

[9] I. Mikić, M. M. Trivedi, E. Hunter and P. C. Cosman, "Articulated body posture estimation from multi-camera voxel data", *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 455–462, Kauai, HI, USA, December 8-14 2001.

[10] R. Navaratnam, A. W. Fitzgibbon and R. Cipolla, "The joint manifold model for semi-supervised multi-valued regression", *Proc. ICCV*, Rio de Janeiro, Brazil, October 2007.

[11] R. Plänkers and P. Fua, "Tracking and modeling people in video sequences", *CVIU*, vol. 81(3), March 2001.

[12] M. Ringer, T. Drummond and J. Lasenby, "Using occlusions to aid position estimation for visual motion capture", *Proc. CVPR*, vol. 2, pp. 464–469, December 8-14 2001.

[13] R. S. Snell, *Clinical Anatomy for Medical Students*, Lippincott Williams & Wilkins, sixth ed., February 2000.

[14] B. Stenger, A. Thayananthan, P. H. S. Torr and R. Cipolla, "Model-based hand tracking using a hierarchical Bayesian filter", *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28(9), pp. 1372–1384, September 2006.

[15] E. B. Sudderth, M. I. Mandel, W. T. Freeman and A. S. Willsky, "Visual hand tracking using nonparametric belief propagation", *Proc. CVPR Workshop on Articulated and Nonrigid Motion*, vol. 12, p. 189, Washington, DC, USA, July 2004.

[16] Y. Wu, J. Lin and T. S. Huang, "Analyzing and capturing articulated hand motion in image sequences", *Trans. PAMI*, vol. 27(12), pp. 1910–1922, December 2005.

[17] M. P. Young and J. W. Scannell, "Brain structure-function relationships: Advances from neuroinformatics", *Phil. Trans. Royal Society of London. Biological Sciences*, vol. 355, pp. 3–6, 2000.

[18] J. Ziegler, K. Nickel and R. Stiefelhagen, "Tracking of the articulated upper body on multi-view stereo image sequences", *Proc. CVPR*, vol. 1, pp. 774–781, June 17-22 2006.