Evidential modeling for pose estimation

Fabio Cuzzolin
UCLA Vision Laboratory
University of California at Los Angeles
cuzzolin@cs.ucla.edu

Ruggero Frezza
Dipartimento di Ingegneria dell'Informazione
Università di Padova
frezza@dei.unipd.it

Abstract

Pose estimation involves reconstructing the configuration of a moving body from images sequences. In this paper we present a general framework for pose estimation of unknown objects based on Shafer's evidential reasoning. During learning an *evidential model* of the object is built, integrating different image features to improve both estimation robustness and precision. All the measurements coming from one or more views are expressed as *belief functions*, and combined through Dempster's rule. The best pose estimate at each time step is then extracted from the resulting belief function by probabilistic approximation. The choice of a sufficiently dense training set is a critical problem. Experimental results concerning a human tracking system are shown.

Keywords. Pose estimation, training set, feature-pose maps, belief functions, evidential model.

1 Introduction

Object tracking [8] is one of the most active field of computer vision, concerning the reconstruction of the actual pose of a moving object by processing the sequence of images taken during its motion. Most of the literature concerns *model-based* approaches [6]. In case of articulated objects composed by a number of rigid parts [2], a popular choice is to use a kinematic model, while when tracking general non-rigid objects one has to recur to flexible models whose parameters are no more in relation with the relative positions of rigid subsets of the body. Statistical estimation methods have been recently introduced, ranging from maximum a-posteriori (MAP) estimation, to applications of particle filtering or multiple hypothesis techniques. All those methods however contemplate the introduction of some kind of model of the moving body, in terms of observation likelihoods, volumetric models, or other.

Consider instead a scenario in which the configuration of an unknown object is desired, having *no a-priori information* about the nature itself of the body (i.e. rigid, articulated, deformable, etc.). As we show in this paper, the only way of doing inference on the object's pose is then building a connection between features and poses through a *learning* procedure, in which a finite approximation of the parameter space is acquired as a collection of poses assumed by the objects, while at the same time a number of features are extracted from one or more image sequences: the ground truth configuration values can be produced for instance by a motion capture system, as in [11].

We propose a feature integration approach based on the theory of evidence [12] in which the uncertainty is described through *belief functions* (b.f.), as they have no need for prior distributions, an attractive feature in a model-free estimation process. Model acquisition is based on a training session in which the object describes an exciting enough trajectory $\tilde{\mathcal{Q}}$ approximating the actual parameter space $\mathcal{Q}$. At the same time, the desired features are computed from the acquired images. Their ranges (*features spaces*) are then *discretized* by feeding a *hidden Markov model* (HMM) with the acquired samples, and the maps from these samples to the approximate configuration space $\tilde{\mathcal{Q}}$ are learned. These maps, together with discrete feature spaces and sample trajectory form the learned *evidential model* of the object: when the object evolves freely, the evidential model can be used to estimate its pose. The new features are transformed into belief functions, combined, and then a *pointwise estimate* of the body pose is extracted. We present experimental results concerning human body tracking using a motion capture system to provide the ground truth.

2 Pose estimation

In the pose estimation problem, an object (for instance a human body) moves in front of one or more cameras. We assume the *configuration* or *pose* of the moving object can be expressed as a point in some region $\mathcal{Q}$ of $\mathbb{R}^D$, called *parameter space*. In the case of a rigid body (for example, an airplane flying in the field of view of the cameras) the pose is simply the position and orientation of the object with respect to some fixed reference frame (for the plane, 3 coordinates giving its location and other 3 coordinates expressing its motion direction, $D = 6$).

If the body is *articulated* (composed by several rigid bodies) like a human arm or hand, its pose also describes its internal configuration. A common choice is to assume a kinematic model describing the angles between the rigid parts or links forming the object. For instance, Rehg and Kanade [10] used a kinematic model of the hand with $D = 27$ degrees of freedom (see Figure 1) representing the angles between each pair of links of the fingers.

The object evolution is observed through one or more cameras. Its pose is then estimated from the available images. However, as images themselves are too large and redundant too be useful, usually a number of salient measurements (*features*) are extracted from them to be used in the estimation process. Rehg, for instance, measured the projected location of the fingertips of the hand on the image plane.

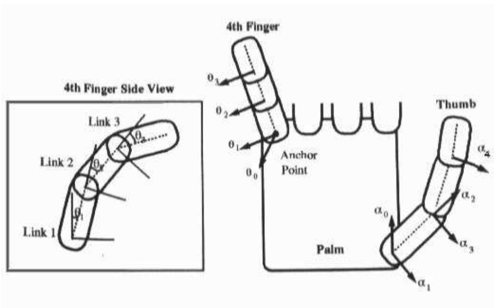


Figure 1: Rehg's kinematic model of the hand.

2.1 Pose estimation of unknown objects

Consider a situation in which we need to estimate the configuration of an unknown object, having no a-priori information about the nature of the body itself (i.e. rigid, articulated, deformable, etc.). The only way of doing inference on the object's pose is then *building a connection between features and poses.* This would allow us to make inference on the pose $q \in \mathcal{Q}$ given new evidence extracted from the images. In situations in which we have no knowledge of the analytic relations between image features and pose, this has to be done *directly from the data* by means of a learning strategy.

In a training session the object evolves, exploring its range of possible configurations, and a set of its poses is collected. This yields a finite approximation $\tilde{\mathcal{Q}}$ of the parameter space

$$\tilde{\mathcal{Q}} \doteq \{q_k, k = 1, ..., T\}. \quad (1)$$

At the same time a number N of features

$$\{y_i(k), k = 1, ..., T\}, \quad i = 1, ..., N \quad (2)$$

(represented as vectors of measurements) are extracted from the image sequences at each time instant k. As the images are collected by cameras at a finite, constant pace the number of feature measurements is finite.

To collect $\tilde{\mathcal{Q}}$ we need a source of "ground truth" telling us what the object pose is at each instant k of the training session. One possibility is using a motion capture system as in [11] or [7], for instance in the human body tracking problem. In particular, if we apply a number of reflective markers in fixed positions of the body, the system is able to provide (through some triangulation algorithm) the 3D locations of the markers during the motion in the training stage. Since we do not know the parameter space of the object, it is reasonable to adopt the markers' locations as body configuration. This way the motion capture system generates the desired sequence of poses (1).

Ideally, $\tilde{\mathcal{Q}}$ should be somehow "dense" in $\mathcal{Q}$: $\forall q \in \mathcal{Q}$ there should be a sample q_k such that $\|q - q_k\| < \epsilon$ for some ϵ small enough. Clearly as the actual shape of the parameter space is unknown such a condition is hard to impose. The correctness of the inference on a discrete approximation of the parameter space, rather than the space itself, is also a delicate matter, as we will see in Section 6.2.

3 Feature-pose maps

Model-free estimation is then to be based on feature-pose maps, that have to be learned from the training data. *Hidden Markov models* provide a method to build them automatically.

A hidden Markov model is a stochastic model in which the states $\{X_k\}$ form a *Markov chain*; the only observable quantity is a corrupted version Y_k of the state called *observation process*. We can associate the elements of the finite state space $\mathcal{X} = \{1, ..., n\}$ to coordinate versors $e_i = (0, .., 0, 1, 0, .., 0) \in \mathbb{R}^n$ and write the model as [9]

$$\begin{cases} X_{k+1} = AX_k + V_{k+1} \\ Y_{k+1} = CX_k + diag(W_{k+1})\Sigma X_k \end{cases}$$

where $\{V_{k+1}\}$ is a sequence of martingale increments and $\{W_{k+1}\}$ is a sequence of i.i.d. Gaussian noises $\mathcal{N}(0, 1)$. The HMM parameters will then be the *transition matrix* $A = (a_{ij}) = P(X_{k+1} = e_i | X_k = e_j)$, the matrix C collecting the *means of the state-output* *distributions* (being the j-th column C_j of C equal to $E[p(Y_{k+1}|X_k = e_j)]$), and the matrix Σ of the variances of the output distributions.

3.1 Approximate feature spaces

A, C and Σ can be estimated, given a sequence of observations, through the *Expectation-Maximization* (EM) algorithm. When applied to the sequence of feature vectors $\{y_i(k), k = 1, ..., T\}$ acquired in the training session, the EM algorithm provides a multimodal Gaussian approximation of the original feature space (the range $\mathcal{Y}$ of the unknown feature function $y : I \rightarrow \mathcal{Y}$, where I is the set of images). More precisely, a set of n_i Gaussian densities $\{\Gamma_i^j, j = 1, ..., n_i\}$ on the i-th feature space is set up, whose means and variances are collected in C and Σ, respectively. This is equivalent to an *implicit* partition

$$\Pi_{\mathcal{Y}_i} = \{\mathcal{Y}_i^1, ..., \mathcal{Y}_i^{n_i}\} \tag{3}$$

of the feature range driven by the training data clusters (see Figure 2), where $\mathcal{Y}_i^j = \{y \in \mathcal{Y}_i \; s.t. \; \Gamma_i^j(y) > \Gamma_i^l(y) \; \forall \, l \neq j\}\}$.

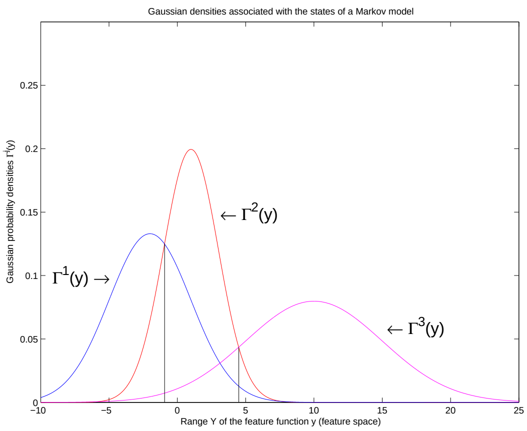


Figure 2: Implicit partition associated with a Markov model with three states. For each state a Gaussian density is set up on the range $\mathcal{Y}$ of the feature function (abscissa), detecting a partition of the range into three parts $\mathcal{Y}^1, \mathcal{Y}^2, \mathcal{Y}^3$ (separated by the vertical lines).

3.2 Learning feature-parameter maps

Therefore, for each feature i each sample pose q_k in the training set $\tilde{\mathcal{Q}} = \{q_k, k = 1, ..., T\}$ is then mapped to the element $\mathcal{Y}_i^j$ of the implicit partition (3) such that $y_i(k) \in \mathcal{Y}_i^j$ (i.e. $\Gamma_i^j(y_i(k)) > \Gamma_i^l(y_i(k))$ for all $l \neq j$)

$$\forall i, k \quad q_k \mapsto \mathcal{Y}_i^j : y_i(k) \in \mathcal{Y}_i^j.$$

On the other hand, each element $\mathcal{Y}_i^j$ of the discrete feature space is mapped to the set of training poses whose feature value fall in $\mathcal{Y}_i^j$,

$$\omega_i : \mathcal{Y}_i^j \mapsto \tilde{\mathcal{Q}}_i^j \doteq \{q_k \in \tilde{\mathcal{Q}} : y_i(k) \in \mathcal{Y}_i^j\}. \tag{4}$$

The collection (3) is a finite representation of the range of the i-th feature function: we call it *approximate* feature space. The application of the EM algorithm to all the feature sequences (2) then yields N maps (4) from each approximate feature space (3) to the approximate parameter space $\tilde{\mathcal{Q}}$.

4 Evidential modeling

Once feature-pose maps are learned from the training data they can be used to design a method for estimating the pose of the object of interest. As both local and global features cannot guarantee of stability and invariance to nuisance factors, extracting multiple cues can also be useful to solve ambiguities and increase the robustness of the overall system. Some researchers have actually posed the cue integration issue in a rigorous probabilistic setup [16]. However, in a Bayesian setting it is hard to design a model-independent feature extraction process, and a number of independence or incorrelation assumptions are needed to make the formulas tractable [14].

4.1 The theory of evidence

We propose an uncertainty description based on *theory of evidence* as, in contrast with the Bayesian formalism, *no a-priori distributions are needed* in an estimation process [12] as required in a model-free context.

Definition 1. *A* basic probability assignment *(b.p.a.) on a finite set (FOD) Θ is a function $m : 2^\Theta \rightarrow [0, 1]$ such that $m(\emptyset) = 0$, $m(A) \geq 0 \;\; \forall A \subset \Theta$, $\sum_{A \subset \Theta} m(A) = 1$.*

2^Θ denotes the power set of Θ. The elements of 2^Θ associated with non-zero values of m are called *focal elements* and their union $\mathcal{C}$ *core*. If a b.p.a. is introduced over an arbitrary FOD, then

Definition 2. *The* belief function *s associated with a basic probability assignment m is the unique function $s : 2^\Theta \rightarrow [0, 1]$ such that $s(A) = \sum_{B \subset A} m(B)$.*

In the theory of evidence finite probability functions are just peculiar belief functions called *Bayesian* b.f., such that $m(A) = 0 \;\; \forall \; A$ s.t. $|A| > 1$. Belief functions can be combined with no need of any *a-priori* distributions as in the Bayesian approach.

Definition 3. *The* orthogonal sum *or* Dempster's sum *of two belief functions s_1, s_2 is a new belief func-*

tion $s_1 \oplus s_2$ whose focal elements are all the possible non-empty intersections between focal elements A_i, B_j of s_1 and s_2 respectively, and whose b.p.a. is given by

$$m(A) = \frac{\sum_{i,j:A_i \cap B_j = A} m_1(A_i) \cdot m_2(B_j)}{1 - \sum_{i,j:A_i \cap B_j = \emptyset} m_1(A_i) \cdot m_2(B_j)}. \quad (5)$$

The theory's major restriction (at least in its original formulation [1]) is the fact that belief functions are defined on finite domains. However, as the training set $\tilde{\mathcal{Q}}$ is finite, the use of belief functions is not problematic.

4.2 Evidential model

Dense training sets (as defined in Section 2.1) are good formal approximations of the unknown parameter space of the moving object. On their side, HMMs provide a method for finding a multi-modal approximation of the involved feature spaces, and an easy way of building maps from each feature space to the approximate parameter space. A unifying framework in which to arrange training sets, approximate feature spaces and feature-pose maps would be desirable. The theory of evidence is such a framework, as both approximate feature spaces (3) and training sets (1) can be thought of as FODs.

As the collection $\pi_i^{\tilde{\mathcal{Q}}} \doteq \{\tilde{\mathcal{Q}}_i^j, \ j = 1, ..., n_i\}$ forms a partition of the approximate parameter space $\tilde{\mathcal{Q}}$, in the evidential language the EM algorithm builds a *refining* ω_i (4) between each approximate feature space (3) and the training set $\tilde{\mathcal{Q}}$.

Definition 4. *Given two FODs* Θ *and* Ω*, a map* $\omega : 2^\Theta \rightarrow 2^\Omega$ *is called* refining *if it satisfies the following conditions: 1)* $\omega(\{\theta\}) \neq \emptyset \ \forall \theta \in \Theta$*; 2)* $\omega(\{\theta\}) \cap \omega(\{\theta'\}) = \emptyset$ *if* $\theta \neq \theta'$*; 3)* $\bigcup_{\theta \in \Theta} \omega(\{\theta\}) = \Omega$.

A refining ω maps Θ to a disjoint partition of Ω. Ω is called a *refinement* of Θ, while Θ is said a *coarsening* of Ω (see Figure 3-top)).

Summarizing, a refining ω_i allows us to make inferences in $\tilde{\mathcal{Q}}$ on the pose of the moving object, based on the evidence provided by a feature measurement living in Θ_i. However, as inferences based on a single feature are often inaccurate and unreliable, we select in the training stage a number of different features and build a refining for each of them.

$\tilde{\mathcal{Q}}$ then becomes the *common refinement* [12] of the collection of approximate feature spaces $\Theta_1, ..., \Theta_N$ (Figure 3-bottom), and is the place all the evidence provided by the different image measurements can be combined. This can be done by representing feature measurements as belief functions and projecting them on the approximate parameter space by means of the refinings learnt in the training stage, as we will see in Section 5.

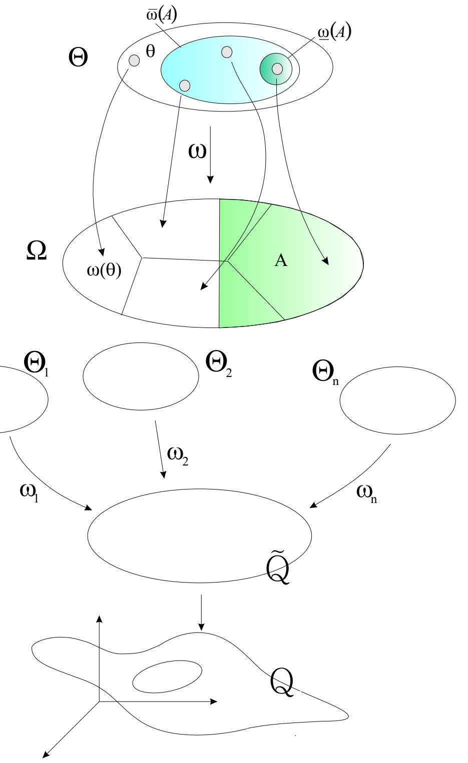


Figure 3: (top) A refining ω between two FODs Θ and Ω. Each element $\theta \in \Theta$ is associated with an element of a disjoint partition of Ω. Each subset $A \subset \Omega$ can be mapped to its inner $\underline{\omega}(A)$ or $\overline{\omega}(A)$ outer reduction as in figure. (bottom) Evidential model architecture.

$\tilde{\mathcal{Q}}, \Theta_1, ..., \Theta_N$ form a family of compatible FODs. This collection of FODs along with their refining maps is characteristic of both the unknown object to track and the chosen feature functions $y_1, ..., y_N$: we call it *evidential model*.

5 Pose estimation

5.1 Measurement functions

The evidential model of the object is built in the training stage, when an adequate training set of the object's motion is acquired (see Section 2.1). After that, it can be used to estimate the pose of the moving object, as new evidence becomes available in terms of image features.

When new images are acquired, at each time instant a pose estimate is computed from the visual features extracted from the images. Those continuous features

[1] In fact Shafer extended the theory to infinite frames of discernment in [13].

can be transformed into belief functions on the approximate feature space: $y_i \mapsto s_i : 2^{\Theta_i} \rightarrow [0,1]$. We call these b.f. *measurement functions* (m.f.).

This can be done in two steps. Given an input feature value $y_i \in \mathcal{Y}_i$ the Markov model $\mathcal{H}_i$ produces as output the set of *likelihoods* $\{\Gamma_i^j(y_i), j = 1, ..., n_i\}$ of this measurement with respect to each state e_j of the model (see Figure 2). Computing the measurement function $s_i(y_i)$ then reduces to building a belief function from this set of likelihoods. These in turn detect a *consonant* belief function (i.e. a b.f. whose focal elements $\mathcal{A}_1, ..., \mathcal{A}_n$ are *nested*, [12]) through the following expression, due to Shafer:

$$s_i(A) = 1 - \frac{\max_{\theta_i^j \in \bar{A}} \Gamma_j(y_i)}{\max_{\theta_i^j \in \Theta} \Gamma_j(y_i)}.$$

The consonant measurement functions have then to be projected to the common refinement $\tilde{\mathcal{Q}}$ to be combined, in order to infer the corresponding pose $\hat{q}$ of the object.

5.2 Projections of measurement functions

In model-free estimation it is hard to formulate analytic relations between distinct features as they may concern completely unrelated aspects of the images. The evidential model, instead, provides a formal description of these relations in terms of refinings.

Definition 5. *The maps* $\underline{\omega} : 2^\Omega \rightarrow 2^\Theta$, $\underline{\omega}(A) = \{\theta \in \Theta | \omega(\{\theta\}) \subset A\}$ *and* $\bar{\omega} : 2^\Omega \rightarrow 2^\Theta$, $\bar{\omega}(A) = \{\theta \in \Theta | \omega(\{\theta\}) \cap A \neq \emptyset\}$ *are called* inner *and* outer reduction *associated with the refining* ω.

Roughly speaking, $\underline{\omega}(A)$ is the largest subset of Θ that implies A, while $\bar{\omega}(A)$ is the smallest subset of Θ implied by A (see Figure 3-top again).
If Ω is a refinement of Θ with refining ω then a b.f. $s' : 2^\Omega \rightarrow [0,1]$ on Ω is called the *vacuous extension* of a second b.f. $s : 2^\Theta \rightarrow [0,1]$ on Θ iff $s(A) = s'(\omega(A))$ for all $A \subset \Theta$, and its focal elements are exactly the images of focal elements of s: $m'(B) \neq 0$ iff $B = \omega(A)$ for some A s.t. $m(A) \neq 0$. Every m.f. generates a collection of "sister" functions defined on the other FODs, called *projections* of the m.f. onto the other spaces of the family.

Theorem 1. *The projection of a m.f.* s_i *defined over a FOD* Θ_i *onto another discrete feature space* Θ_j *has the following expression:* $\pi_j[s_i] : 2^{\Theta_j} \rightarrow [0,1]$, $\pi_j[s_i] = s_i \circ \underline{\omega}_i \circ \omega_j$. *Its basic probability assignment can be computed from* m_i *as* $\pi_j[m_i](A) = \sum_{B \subset \Theta_i : \bar{\omega}_j(\omega_i(B)) = A} m_i(B)$.

Proof. The vacuous extension s' of a belief function $s : 2^\Theta \rightarrow [0,1]$ to a refinement Ω, $\omega : 2^\Theta \rightarrow 2^\Omega$ has a b.p.a. m such that $m(A) = m'(\omega(A)) = m' \circ \omega$ and can be expressed as $s' = s \circ \underline{\omega}$ ([12], Theorem 7.4). The projection $\pi_j[s_i]$ of a b.f. $s_i : 2^{\Theta_i} \rightarrow [0,1]$ onto another discrete feature space Θ_j is then the restriction to Θ_j of the vacuous extension of s_i to $\tilde{\mathcal{Q}}$, so that $\pi_j[s_i] = s'_j \circ \omega_j = s_i \circ \underline{\omega}_i \circ \omega_j$. Its basic probability assignment is $\pi_j[m_i](A) = \sum_{C \subset \tilde{\mathcal{Q}} : \bar{\theta}_j(C) = A} m'_i(C)$ which is equal to

$$\sum_{B \subset \Theta_i : \bar{\omega}_j(\omega_i(B)) = A} m'_i(\omega_i(B)) = \sum_{B \subset \Theta_i : \bar{\omega}_j(\omega_i(B)) = A} m_i(B)$$

for $m'_i(C) \neq 0$ iff C is the image of a f.e. B of s_i. □

As a consequence, the set of focal elements of $\pi_j[s_i]$ is $\mathcal{E}_{\pi_j[s_i]} = \{A \subset \Theta_j | A = \bar{\omega_j}(\omega_i(B)),\ B \in \mathcal{E}_{s_i}\}$ and its core $\mathcal{C}_{\pi_j[s_i]} = \bar{\omega_j}(\omega_i(\mathcal{C}_{s_i}))$.

Theorem 2. *The projection of a consonant belief function is still consonant, even if it has not in general the same number of focal elements.*

Proof. Let us call $\mathcal{A}_1 \subseteq \cdots \subseteq \mathcal{A}_K$ the nested focal elements of s: obviously its extension s' to $\tilde{\mathcal{Q}}$ has as focal elements $\omega_i(\mathcal{A}_1), ..., \omega_i(\mathcal{A}_K)$. They are all distinct subsets of $\tilde{\mathcal{Q}}$ for ω is a refining. They are also nested so that s' is still a consonant b.f. over $\tilde{\mathcal{Q}}$.
Now, the focal elements of $\pi_j[s_i]$ are the restrictions of those subsets to Θ_j:

$$\bar{\omega}_j(\omega_i(\mathcal{A}_k)) = \{\theta \in \Theta_j | \omega_j(\{\theta\}) \cap \omega_i(\mathcal{A}_k) \neq \emptyset\}.$$

Since they are nested, if $\omega_j(\{\theta\}) \cap \omega_i(\mathcal{A}_k) \neq \emptyset$ then $\omega_j(\{\theta\}) \cap \omega_i(\mathcal{A}_l) \neq \emptyset$, $l \geq k$ so that $\bar{\omega}_j(\omega_i(\mathcal{A}_k)) \subset \bar{\omega}_j(\omega_i(\mathcal{A}_l))$, $l \geq k$ i.e. $\pi_j[s_i]$ is consonant. However, it is not generally true that $\forall k, l\ \exists \theta \in \Theta_j$ s.t. $\theta \in \bar{\omega}_j(\omega_i(\mathcal{A}_l))$ and $\theta \notin \bar{\omega}_j(\omega_i(\mathcal{A}_k))$: it is actually easy to find a counterexample. Hence in the chain $\bar{\omega}_j(\omega_i(\mathcal{A}_1)) \subset \cdots \subset \bar{\omega}_j(\omega_i(\mathcal{A}_K))$ some equalities may appear. □

Theorem 2 shows how consonant belief functions are then *invariant under projection*, and hence a formally correct representation of feature measurements.

5.3 Pointwise estimation

Once the measurement functions $s_1, ..., s_N$ are projected onto $\tilde{\mathcal{Q}}$, they can be combined using Dempster's rule. In absence of conflict the orthogonal sum of the m.f. finally yields a belief function $\hat{s}$ on $\tilde{\mathcal{Q}}$. The easier way to extract a single pose estimate $\hat{q}$ from the belief estimate is then approximating $\hat{s}$ with a finite probability $\hat{p}$ on $\tilde{\mathcal{Q}}$, and then compute its mean value

$$\hat{q} = \sum_{k=1}^{T} \hat{p}(q(k)) \cdot q(k).$$

The problem of approximating a belief function with a finite probability has been widely studied [3, 18]. Natural approximations of a belief function s are the *relative plausibility of singletons* $\hat{p} = \tilde{P}^*_s$ [17]

$$\tilde{P}^*_s(\theta) = \frac{P^*_s(\{\theta\})}{\sum_{\theta' \in \Theta} P^*_s(\{\theta'\})}$$

where $P^*_s : 2^\Theta \rightarrow [0,1]$ is the associated plausibility function, or the pignistic function [15]. Under weak local linearity assumptions on the parameter space, this approximation produces good results.

Several fast implementation of Dempster's rule exist for restricted situations (see [1, 5] among the others). However, consonant b.f. can only have n focal elements (being nested, Section 5.1), reducing the computational complexity of their sum to $O(n^2)$ instead of $O(2^{2n})$.

6 Analysis of the evidential model

6.1 Quantization and ambiguities

We have seen that learning a model directly from the data necessarily implies some sort of approximation of the involved spaces in terms of samples. However, this has non-trivial consequences on the performance of the resulting estimation. In particular, the degree of discretization of feature spaces and training set have a direct influence on the presence of ambiguities during estimation.

Again, the evidential formalism is powerful enough to give a formal picture of the ambiguity issue, and help the assessment of the adequate level of quantization. We have seen how $\tilde{\mathcal{Q}}$ is a common refinement of the approximate feature spaces. However, a collection of compatible FODs has many common refinements. One of these is particularly simple.

Definition 6. *Consider a collection of compatible FODs $\Theta_1, ..., \Theta_N$. If another FOD Θ is such that: 1) $\forall i \; \exists \omega_i : 2^{\Theta_i} \rightarrow 2^\Theta$ refining; 2) $\forall \theta \in \Theta \; \exists \; \theta_i \in \Theta_i$ for all $i = 1, ..., N$ such that $\{\theta\} = \omega_1(\{\theta_1\}) \cap ... \cap \omega_N(\{\theta_N\})$, then the unique [12] FOD $\Theta \doteq \Theta_1 \otimes \cdots \otimes \Theta_N$ is called* minimal refinement *of the collection $\Theta_1, ..., \Theta_N$.*

$\Theta_1, ..., \Theta_N$ are said to be *independent* if $\omega_1(A_1) \cap \cdots \cap \omega_N(A_N) \neq \emptyset$ whenever $\emptyset \neq A_i \subset \Theta_i \; \forall i$.

Now, imagine that the N feature functions assume the discrete values $\theta^1, ..., \theta^N$ respectively. Each piece of evidence implies an object pose inside the subset $\omega_i(\theta^i)$ of the training set, so that from $\theta^1, ..., \theta^N$ it can be inferred that the estimated pose falls inside the set

$$\omega_1(\theta^1) \cap ... \cap \omega_N(\theta^N). \qquad (6)$$

In other words, configurations in the same intersection of the form (6) are indistinguishable. But now, those subsets *are* the elements of the minimal refinement, so that

Theorem 3. *Each pair of sample poses of the training set can be distinguished under the evidential model iff $\tilde{\mathcal{Q}}$ is the minimal refinement of the feature spaces.*

In this case, for each training pose q_k there exists a combination of feature values $\theta_1, ..., \theta_N$ s.t. q_k is perfectly resolved: $q_k = \omega_1(\theta^1) \cap ... \cap \omega_N(\theta^N)$.

Given a training set it is hence desirable to choose the number of states of the N HMMs associated with each feature space in order to bring the minimal refinement $\Theta_1 \otimes \cdots \otimes \Theta_N$ as close as possible to $\tilde{\mathcal{Q}}$. Naturally, since the "natural" number of states is determined by the clusters actually formed by the feature data acquired in the training session, sometimes the addition of new features becomes necessary to resolve the ambiguities.

6.2 Conflict management

Dempster's combination of generic b.f. is guaranteed only if $\Theta_1, ..., \Theta_n$ are *independent* [4]. This could seem a serious obstacle to evidence combination in the evidential model. However, measurement functions are not generic b.f., but they are peculiar b.f. built by means of HMMs as seen in Section 5.1.

Let us first suppose one of the sample poses q_k is presented to the model, with feature values $y_1(k), ..., y_N(k)$. In the worst case, for each feature i there will be a single cluster j_i with $\Gamma^{j_i}_i(y_i) >> \Gamma^j_i(y_i)$ $\forall l \neq j_i$. The algorithm of Section 5.1 will then yield a set of N Bayesian b.f. with $m_i(\theta^{j_i}) = 1$, $m_i(A) = 0$ for $A \neq \{\theta^{j_i}\}$. Now, as $q(k) \in \omega_i(\theta^{j_i})$ for all i (being the sample $q(k)$ attributed to the state θ^{j_i} of $\mathcal{H}_i$), $\omega_1(\theta^{j_1}) \cap \cdots \cap \omega_N(\theta^{j_N}) \supset q(k) \neq \emptyset$ and a belief estimate $\hat{s}(k)$ of the pose always exists. Hence, a conflict is possible only if $q \notin \tilde{\mathcal{Q}}$. Consider then two scalar features as in Figure 4, with ranges partitioned in 2 clusters delimited by the horizontal lines and a training set composed by the 6 points marked on the absciss.

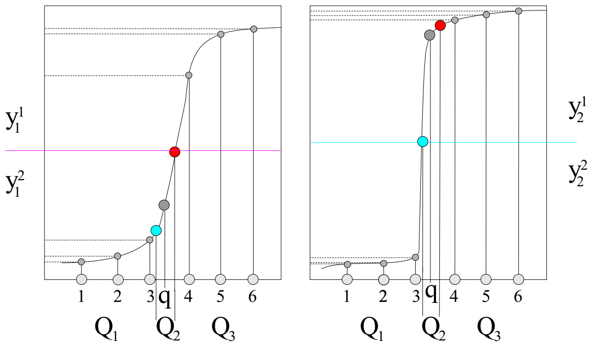


Figure 4: Conflict in the evidential model.

The refining maps are $\omega_1(\mathcal{Y}_1^1) = \{4,5,6\}$, $\omega_1(\mathcal{Y}_1^2) = \{1,2,3\}$ and $\omega_2(\mathcal{Y}_2^1) = \{4,5,6\}$, $\omega_2(\mathcal{Y}_2^2) = \{1,2,3\}$ respectively. But now, for each of the poses q in the highlighted region $\mathcal{Q}_2$, $\omega_1(y_1(q)) = \{1,2,3\}$ while $\omega_2(y_2(q)) = \{4,5,6\}$ so that $\omega_1(y_1(q)) \cap \omega_2(y_2(q)) = \emptyset$ and there is no combination.

Formally, each implicit partition of a feature range $\mathcal{Y}_i$ is naturally associated with a partition of the parameter space $\mathcal{Q}$ through the inverse feature map

$$\{\mathcal{Y}_i^1, ..., \mathcal{Y}_i^{n_i}\} \overset{y_i^{-1}}{\mapsto} \Pi_{\mathcal{Q}}^i = \{y_i^{-1}(\mathcal{Y}_i^1), ..., y_i^{-1}(\mathcal{Y}_i^{n_i})\}. \quad (7)$$

On the other side, there exists a refining ω_i' between Θ_i and $\mathcal{Y}_i$, namely $\omega_i'(\theta_i^j) = \mathcal{Y}_i^j$, so that each feature value $\mathbf{y}_i$ can be mapped to a set of training poses in $\tilde{\mathcal{Q}}$, $\mathbf{y}_i \mapsto \omega_i(\bar{\omega}_i'(\mathbf{y}_i)) = \{q_k \in \tilde{\mathcal{Q}} : y_i(q_k) \in \mathcal{Y}_i^{j_{\mathbf{y}_i}}\}$ where $\bar{\omega}_i'$ is the reduction of ω_i' and $\mathcal{Y}_i^{j_{\mathbf{y}_i}}$ is the region of $\mathcal{Y}_i$ containing $\mathbf{y}_i$. Hence, each configuration q has a set of *representatives* in $\tilde{\mathcal{Q}}$,

$$\omega_1(\bar{\omega}_1'(y_1(q))) \cap \cdots \cap \omega_N(\bar{\omega}_N'(y_N(q))). \quad (8)$$

As the above example explains,

Proposition 1. *There is conflict between measurement functions iff the underlying object pose q has no representatives (8) in the training set.*

The intersection of the partitions $\Pi_{\mathcal{Q}}^i$ of the parameter space associated with the quantized feature ranges (Equation (7)) finally detects another partition of $\mathcal{Q}$,

$$\Pi_{\mathcal{Q}} = \{y_1^{-1}(\mathcal{Y}_1^{j_1}) \cap \cdots \cap y_N^{-1}(\mathcal{Y}_N^{j_N})\}$$

$\forall\ j_1 = 1, ..., n_1,\ \cdots,\ j_N = 1, ..., n_N$. The non-represented region of $\mathcal{Q}$ is then the set of the elements of $\Pi_{\mathcal{Q}}$ which do not contain any sample pose (as the marked region $\mathcal{Q}_2$ in Figure 4). This suggests that

Proposition 2. *The conflict is impossible in the evidential model iff* $\Pi_{\mathcal{Q}} \sim \Theta_1 \otimes \cdots \otimes \Theta_N$.

Adding new samples in each of the non-represented elements of $\Pi_{\mathcal{Q}}$ (like $q \in \mathcal{Q}_2$ in Figure 4) would ideally solve the problem. Unfortunately, as we have no a-priori knowledge about the parameter space, the best we can do is updating the evidential model each time a conflict takes place, progressively reducing the chance of conflict.

7 Algorithms

Let us now summarize the learning and estimation algorithms we introduced in Sections 3, 4 and 5.

7.1 Model learning

In the training stage an evidential model of the object is built. The body moves in front of the camera(s), exploring its parameter space, while a sequence of training poses $\tilde{\mathcal{Q}} = \{q_k, k = 1, ..., T\}$ is provided by a source of ground truth (for instance a motion capture systems, Section 2.1). At the same time:

1. we compute a number of feature measurements y_i, $i = 1, ..., N$ from all the images;

2. we process every feature sequence $\{y_i(k), k = 1, ..., T\}$ by means of a HMM $\mathcal{H}_i$ with n_i states, obtaining

   2.a the approximate feature space Θ_i, i.e. the implicit partition of the feature range associated with the states of the HMM (Section 3.1);

   2.b the N refinings ω_i (4) between each feature space Θ_i and the approximate parameter space $\tilde{\mathcal{Q}}$.

Even after choosing the desired feature functions, the resulting evidential model still depends on the dimensions $\{n_i,\ i = 1, ..., N\}$ of the approximate feature spaces (the number of states of the HMMs). The "best" value of n_i can be estimated by analyzing the clusters formed by the training data, or alternatively learned by measuring the estimation error as n_i increases.

7.2 Pose estimation

Given an evidential model of the moving body with N features, estimating the object's pose reduces to the following steps: Given one or more images at time t

1. the N feature values are extracted from the acquired image(s);

2. the likelihoods $\{\Gamma_i^j(y_i(t)), j = 1, ..., n_i\}$ of each feature value $y_i(t)$ associated with the corresponding HMM $\mathcal{H}_i$ are computed (Section 3.1);

3. for each feature i a measurement belief function $s_i(t)$ is built from these likelihoods (Section 5.1);

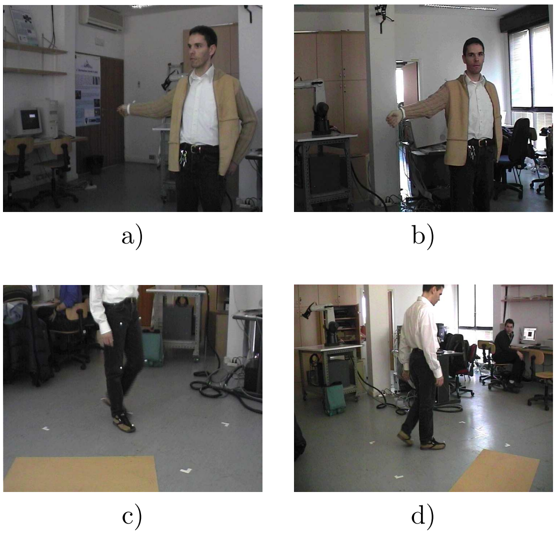
4. all the m.f. $\{s_i(t), i = 1, ..., N\}$ are projected onto the approximate parameter space $\tilde{\mathcal{Q}}$ by vacuous extension (Section 5.2);

5. in absence of conflict their orthogonal sum $\hat{s}(t) \doteq s_1(t) \oplus ... \oplus s_N(t)$ is obtained through Dempster's rule; otherwise the new pose is added to $\tilde{\mathcal{Q}}$, and the model is updated;

6. the pointwise estimate $\hat{q}(t)$ of the object configuration is obtained from $\hat{s}(t)$ after Bayesian approximation (Section 5.3).

8 Experimental results

We have tested our evidential modeling technique in a challenging setup, involving the pose estimation of human arms and legs from two well separated views. To collect the necessary ground truth we used a marker-based motion capture system [11, 7] built by E-motion. The person was filmed by two DV cameras (Figure 5). In the first experiment we asked him to make his arm following a trajectory which could hopefully cover the configuration space of the arm itself, keeping his wrist blocked and standing on a fixed spot on the floor to reduce the dimensionality of the space (2 d.o.f.s for the shoulder and 2 for the elbow). In the second experiment we tracked the legs, assuming that the person was walking normally on the floor, and collected the training set by sampling a random walk on a small part of the floor.


a)

b)

c)

d)

Figure 5: Human bodypart tracking. a),b) Sample images of a person moving his right arm. c),d) Sample images of the person walking inside a rectangle drawn on the lab floor.

We adopted color segmentation to separate the moving arm from a non-static background through a colorimetric analysis of the body of interest (Figure 6). We then found the bounding box around the the silhouette of the segmented upper torso or legs, building a feature vector as the collection max(row), min(row), max(col), min(col) of the row and column indexes defining the box. For the arm experiment we placed four markers on shoulder, elbow, wrist and hand, and used the resulting 12-component vector of their 3D positions as body pose. For the leg experiment we placed 6 markers on hips, knees and feet. The length of the training sequences was 1726 frames for the arm and 1952 frames for the legs. We then built for the arm experiment *three* different evidential models: one using the left view only, with $N = 2$ feature spaces and a $n_i = 5$ states for both features; a second model for the right view only, with $N = 3$ and $n_i = const = 5$; an overall model in which all the features from both view were integrated. For the leg experiment we instead built two models with $N = 6$ and $n = 4$ or $n = 5$ respectively to investigate the influence of the quantization level. We acquired a testing sequence for each experiment and compared the results with the ground truth provided by the E-motion system. The single-view estimates were rather mediocre. Accordingly, the minimal re-

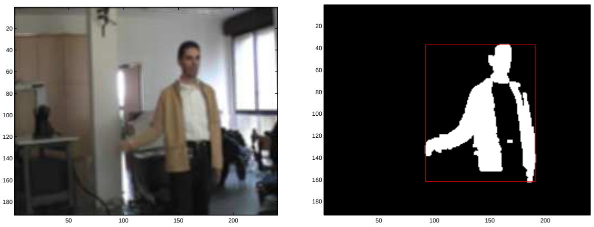


Figure 6: Feature extraction process. Left: original image. Right: the object of interest is color segmented, and the bounding box around the object of interest is detected.

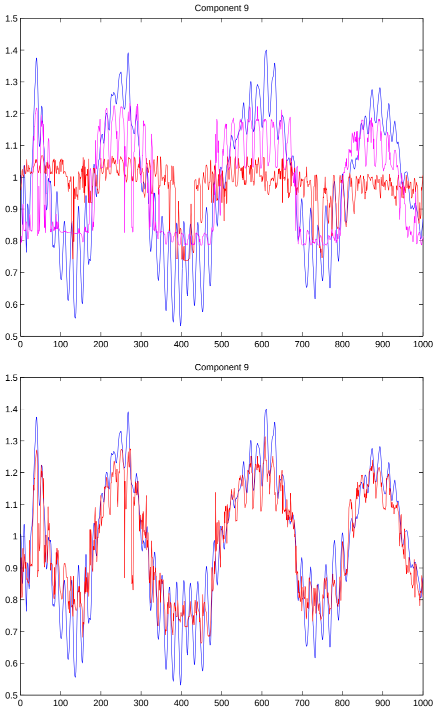


Figure 7: Left: pose estimates of the ninth component of the configuration vector (Y coordinate of the hand marker) produced by the left (red) and right (magenta) model compared with the ground truth (blue), plotted against time. Right: the estimate sequence (red) yielded by the overall model compared with the ground truth.

finements $\bigotimes \Theta_i$ for the left and right models were of size 22 and 80 respectively (the larger the size, the finer the resolution of the model).

As Figure 7 shows, the estimates obtained exploiting both views are clearly better than a simple selection of the best partial estimate at each instant, as the ambiguity analysis confirms. $\bigotimes \Theta_i$ for the overall model had size 372 and the sizes of the five largest undistinguished regions of $\mathcal{Q}$ were 69,53,43,38,28 respectively, with 139 samples perfectly resolved!
We also computed the estimation error by measuring the Euclidean distance between the real and estimated 3D location of each marker over the whole testing sequence. The average estimation errors are 17.3, 7.95, 13.03, and 2.7 centimeters respectively for the four markers. Figure 9 shows instead the performance of the estimation algorithm in the leg experiment, for a 200-frame-long test sequence. Here the results are a bit less impressive but still very good, mainly due to the difficulty of segmenting a pair of black pants against a dark background (see Figure 5c). This is confirmed by the numerical results. The average estimation errors are 25.41, 19.29, 21.84, 19.88, 23.00, and 22.71 centimeters respectively. Consider that the cameras were located at a distance of about three meters. No significant conflict problem was reported.

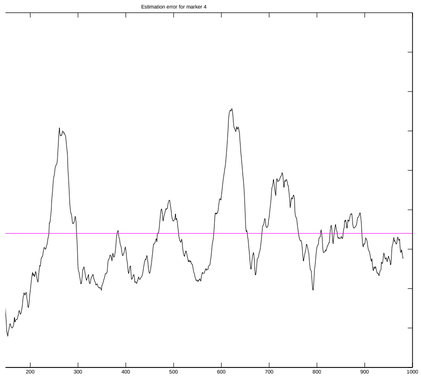


Figure 8: Estimation errors for markers 4 in the arm experiment. The horizontal line indicates the average value.

8.1 Other learning-based approaches

A comparison with the performance of other methods only makes sense for data-driven model building methods. In fact, a small number of other researchers have proposed an estimation framework based a learning process. Howe *et al.* [7] proposed to use some a-priori knowledge about human motion learned from training data to resolve the ambiguities of monocular tracking. They modeled the probabilities of short motions called *snippets* as a mixture of Gaussian density, and adopted EM to find the weight of each cluster. However, the system was tested on short clips only, as the authors admitted it would not be reliable for significantly long sequences. Here instead we have tested our algorithms over rather long sequences.
Rosales and Sclaroff [11] proposed instead a pose estimation technique in which the object pose was expressed as a 2D marker configuration. In the training stage, a motion capture system provided the 3D marker positions of the given object in the motion sequences, whose 2D projections formed the ground truth of the learning process. A 3D model of the object was then used to generate synthetic silhouette views, from which some visual features were extracted. The training set of projected 2D markers was clustered using the EM algorithm, and for each cluster a multi-layer perceptron was trained to map features to 2D marker positions, analogously to what we have shown is Section 3.2.
The training data was more extensive, consisting on 5 sequences of some 200 frames each. The method was applied to synthetic silhouettes obtained through a 3D model, while the results seemed to degrade a little when real video data were considered.

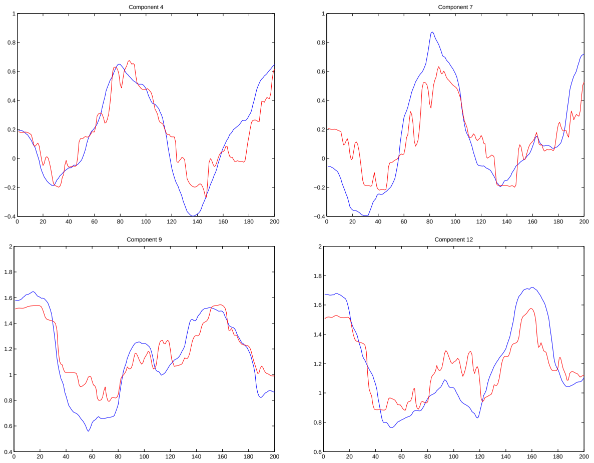


Figure 9: Performance of the two-view pose estimation for the leg experiment on a testing sequence of length 200 with $N = 6$ and $n_i = n = 5$. The estimates for some representative components of the pose vector (components 4,7,9, and 12) are shown in red, against the ground truth values (in blue).

9 Conclusions

In this paper we presented a bottom-up technique for pose estimation of arbitrary unknown objects (articulated, flexible, etc.) based on learning a direct mapping between features and poses in a learning stage, and combining the available data in the framework of the evidential reasoning. We showed how to use the EM algorithm for HMMs to build multi-Gaussian approximations of the feature spaces and direct feature-pose maps.

We adopted the evidential language to formalize and implement the information fusion, as the ToE makes no use of a-priori distributions, a very attractive feature in the context of model-free estimation. The ToE also provides tools to represent feature-pose maps through the notion of refining, and analyze incorrect inferences due to coarse approximations or ambiguities in the learnt model.

We chose to represent features as belief functions, also because of their relative simplicity. However, the use of Dempster's rule as fusion mechanism is implicitly based on the assumption that the pieces of evidence carried by the features are independent. This is reasonable in a model-free situation in which no a-priori knowledge of their relationships is available, and is often assumed in the literature. The approach has been tested in a challenging setup, concerning real video sequences with realistic non-static backgrounds. The experimental results support this assumption.

It would be interesting to understand what happens when we represent the uncertainty by means of some other class of imprecise probabilities, such as credal sets, coherent previsions, or the imprecise Dirichlet model. However, this well deserves a paper of its own, and will be the object of further research in the near future.

References

[1] M. Bauer, *Approximation algorithms and decision making in the Dempster-Shafer theory of evidence–an empirical study*, International Journal of Approximate Reasoning **17** (1997), 217–237.

[2] C. Bregler and J. Malik, *Tracking people with twists and exponential maps*, Proceedings of CVPR'98, Santa Barbara, CA, 1998.

[3] B. R. Cobb and P. P. Shenoy, *A comparison of methods for transforming belief function models to probability models*, Proceedings of ECSQARU'2003, Aalborg, Denmark, July 2003, pp. 255–266.

[4] F. Cuzzolin, *Algebraic structure of the families of compatible frames of discernment*, accepted for a Special Issue of the Annals of Mathematics and Artificial Intelligence (2005).

[5] T. Denoeux and A. Ben Yaghlane, *Approximating the combination of belief functions using the fast moebius transform in a coarsened frame*, International Journal of Approximate Reasoning **31(1-2)** (October 2002), 77–101.

[6] D. M. Gavrila and L. S. Davis, *3D model-based tracking of humans in action: A multi-view approach*, Proceedings of CVPR'96, San Francisco, CA, 18-20 June 1996, pp. 73–80.

[7] N. Howe, M. Leventon, and W. Freeman, *Bayesian reconstruction of 3D human motion from single-camera video*, Neural Information Processing Systems, Denver, Colorado, 1999.

[8] T. Moeslund and E. Granum, *A survey of computer vision-based human motion capture*, Image and Vision Computing **81** (2001), 231–268.

[9] L. Aggoun R. Elliot and J. Moore, *Hidden Markov models: estimation and control*, 1995.

[10] J. M. Rehg and T. Kanade, *Model-based tracking of self-occluding articulated objects*, Proceedings of ICCV'95, 20-23 June 1995, pp. 618–623.

[11] R. Rosales and S. Sclaroff, *Learning and synthesizing human body motion and posture*, Fourth Int. Conf. on Automatic Face and Gesture Recognition, Grenoble, France, March 2000.

[12] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, 1976.

[13] Glenn Shafer, *Allocations of probability*, Annals of Probability **7:5** (1979), 827–839.

[14] H. Sidenbladh and M.J. Black, *Learning the statistics of people in images and video*, IJCV **54** (2003), 189–209.

[15] Philippe Smets, *Belief functions*, Non-Standard Logics for Automated Reasoning (Ph. Smets, A. Mamdani, D. Dubois, and H. Prade, eds.), Academic Press, London, 1988, pp. 253–286.

[16] C. Sminchisescu and B. Triggs, *Covariance scaled sampling for monocular 3D body tracking*, Proceedings of CVPR'01, Hawaii, December 2001.

[17] F. Voorbraak, *A computationally efficient approximation of Dempster-Shafer theory*, International Journal on Man-Machine Studies **30** (1989), 525–536.

[18] A. Ben Yaghlane, T. Denoeux, and K. Mellouli, *Coarsening approximations of belief functions*, Proceedings of ECSQARU'2001 (S. Benferhat and P. Besnard, eds.), 2001, pp. 362–373.