A belief-theoretical approach to example-based pose estimation

Fabio Cuzzolin and Wenjuan Gong

Abstract

In example-based pose estimation, the configuration or “pose” of an evolving object is sought given visual evidence, having to rely uniquely on a set of examples. We assume here that, in a training stage, a number of feature measurements is extracted from the available images, while an “oracle” provides us with the true object pose at each instant. In this scenario, a sensible approach consists in learning maps from features to poses, using the information provided by the training set. In particular, multi-valued mappings linking feature values to set of training poses can be easily constructed. A probability measure on any feature space is then naturally mapped to a convex set of probabilities on the set of training poses, in a form of a “belief function”. Given a test image, its feature measurements translate into a collection of belief functions on the set of training poses, which when combined yield there an entire family of probability distributions. From the latter either a single central pose estimate or a set of extremal estimates can be computed, together with a measure of how reliable the estimate is.

We call this technique “Belief Modeling Regression” (BMR). We illustrate BMR’s performance in an application to human pose recovery, showing how it outperforms our implementation of both Relevant Vector Machine and Gaussian Process Regression. We discuss motivation and advantages of the proposed approach with respect to its competitors and outline an extension of this technique to tracking.

Index Terms

Example-based pose estimation, feature-pose maps, theory of evidence, belief functions.

I. INTRODUCTION

Pose estimation is a well studied problem in computer vision. Given an image sequence capturing the motion and evolution of an object of interest, the problem consists in estimating the
position and orientation of the object at each time instant, along with its internal configuration or pose. Such estimation is typically based on two pillars: the extraction of salient measurements or features from the available images and, when present, a model of the structure and kinematics of the moving body. Pose estimation is, among others, a fundamental ingredient of motion capture, i.e., the reconstruction of the motion of a person throughout a video sequence, usually for animation purposes in the movie industry or for medical analysis of posture and gait. Other major applications include human-computer interaction, image retrieval on the internet, robotics.

State of the art. Current methodologies for pose estimation can roughly be classified into “model-based”, “learning-based” and ‘example-based’” approaches. The former [1], [2] presuppose an explicitly known parametric body model: pose recovery is typically achieved by matching the pose variables to a forward rendered model based on the extracted features. Initialization is often difficult, and the pose optimization process can be subject to local minima [3]. In contrast, learning based approaches [14], [4], [5], [6] exploit the fact that typical (human) motions involve a far smaller set of poses than the kinematically possible ones, and learn a model that directly recovers pose estimates from observable image quantities. Such methods [7], [8], [9], [10] are appealing and generally faster, due to the lower dimensionality of the models employed, and typically provide a better predictive performance when the training set is comprehensive. On the other hand, they sometimes require heavy training to produce a decent predictive model, and the resulting description can lack generalization power.

Example-based methods, which explicitly store a set of training examples whose 3D poses are known, estimate pose by searching for training image(s) similar to the given input image and interpolating from their poses [5], [11]: they can then be used to initialize model-based methods in a “smart” way, as in the monitoring of an automobile driver’s head movements provided in [10]. No prior analytic structure of the pose space is incorporated in the estimation process, although the training data itself do amount to a rough approximation of the configuration space. Most of these methods share a common architecture. Vectors of feature measurements (such as moments of silhouette images [12], multi-scale edge direction histograms [13], distribution of shape contexts [14], and Harr-like features [15]) are extracted from each individual image. Indeed, the integration of multiple cues is crucial in pose estimation to increase both the resolution/accuracy of the estimation and its robustness [35], [36], [37], [38]. Then, the likely pose of the object is predicted by feeding this feature vector to a map from...
the features space to the pose space, which is learned from a training set of examples or a model whose parameters are learned from the training data, and whose purpose is to (globally or locally) represent the relationship between image and pose. This mapping, albeit unknown, is bound to be (in general) one-to-many: more than one object configuration can generate the same feature observation, because of occlusions, self-occlusions and the ambiguities induced by the perspective image projection model.

Since only limited information is provided to us in the training session, only an approximation of the true feature-pose mapping can be learned. The accuracy of the estimation depends on the forcibly limited size and distribution of the available examples, which are expensive and time consuming to collect. This has suggested in the past to consider a more constrained, activity-based setting to constrain the search space of possible poses.

In [12], for instance, an inverse mapping between image silhouette moments and 2D joint configurations is learned, for each cluster obtained by fitting a Gaussian mixture to 2D joint configurations via the EM algorithm. In [14] a Relevant Vector Machine (RVM) is used to infer human pose from the silhouette shape descriptor, while more recently an extension to mixtures of RVMs has been proposed by Thayananthan et al. [16]. In [17], a number of exemplar 2D views of the human body is stored; the locations of the body joints are manually marked and labeled. The input image is then matched via “shape context matching” to each stored view, and the locations of the body joints in the matched exemplar view are transferred to the test image. Other approaches include Local Weighted Regression [5], BoostMap [11], Bayesian Mixture of Experts [6] and Gaussian Process Regression (GPR) [27].

The accuracy of example-based approaches critically depends on the amount and representativeness of the training data. Queries can be potentially computationally expensive, and need to be performed quickly and accurately [5], [11]. In addition, example-based approaches often have problems when working in high dimensional configuration spaces, as it is difficult to collect enough examples to densely cover them.

A. Problem statement

In this paper we consider the following scenario:

- the available evidence comes in the form of a training set of images containing sample poses of an unspecified object;
• we only know that its configuration can be described by a vector $q \in Q \subset \mathbb{R}^D$ in a pose space $Q$ which is a subset of $\mathbb{R}^D$;
• a source of ground truth exists which provides for each training image $I_k$ the configuration $q_k$ of the object portrayed in the image;
• the location of the object within each training image is known, in the form of a bounding box containing the object of interest.

In the training session, the object explores its range of possible configurations, and a set of poses is collected to form a finite approximation $\tilde{Q}$ of the parameter space:

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

At the same time a number $N$ of distinct features are extracted from the available image(s), within the available bounding box:

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

In order to collect $\tilde{Q}$ we need a source of ground truth to tell us what pose the object is in at each instant $k$ of the training session. One option is to use a motion capture system, as it is done in [12] for the human body tracking problem. After applying a number of reflective markers in fixed positions of the moving object, the system is able to provide by triangulation the 3D locations of the markers throughout the training motion. Since we do not know the parameter space of the object, it is reasonable to use as body pose vector the collection of 3D locations of the markers’. Based on this evidence, in the testing stage:

• a supervised localization algorithm (trained in the training stage using the annotation provided in terms of bounding boxes, e.g. [26]) is employed to locate the object within each test image: image features are only extracted from within the resulting bounding box;
• such features are exploited to produce an estimate of the object’s configuration, together with a measure of how reliable this estimate is.

B. Contribution

In this paper we propose a regression framework for the example-based pose estimation problem as formulated above, based on the theory of belief functions [19], [20]. Belief functions are non-additive measures which admit a number of interpretations: i) as random sets, i.e.
probability distributions on the power sets of all subsets; ii) as convex sets of standard probability distributions (credal sets); iii) as objects induced by the application of a multi-valued map to a standard probability [20], [21]. Most relevant to us is the second interpretation, for a belief function on the pose space is equivalent to a \textit{set of linear constraints on the actual conditional pose distribution (given the features)}. Our Belief Modeling Regression (BMR) framework uses the finite amount of evidence provided in the training session to build, given a new feature value, a belief function on the set of training poses. Technically, this is done by learning a discrete mapping (a “refining”) between an approximation of the feature space obtained via EM and the set of training poses. According to interpretation iii), any test feature value, once encoded as a set of likelihoods, translates into a belief function on the set of training poses. This determines a convex sets of distributions there, which in turn generates an interval of pose estimates.

Multiple features are necessary to obtain decent accuracy in terms of pose estimation. All single-feature refinings are collected in an “evidential model” of the object: the information they carry is fused before estimating the object’s pose in the belief framework [20], [22], allowing a limited resolution for the individual features to translate into a relatively high estimation accuracy (in a similar way to tree-based classifiers [18] or boosting approaches in which weak features are combined to form a strong classifier). The size of the resulting convex set of probabilities reflects the amount of training information available: the larger and more densely distributed within the pose space the training set is, the narrower the resulting credal set. Both a point-wise estimate of the current pose and a measure of the accuracy of the resulting estimate can be extracted, for instance as a function of the size of this convex (credal) set. In alternative, a separate pose estimate can be computed for each vertex of the credal set, in a robust statistical fashion.

As we show in the last part of the paper, an evidential model essentially provides a constraint on the family of admissible feature-to-pose maps, in terms of smooth upper and lower bounds. All mappings (even discontinuous, or 1-many) within those smooth bounds are possible under the model. The width of this space of mappings reflects the uncertainty induced by the size and distribution of the available training set.

\textbf{C. Paper outline}

The paper is structured as follows. Firstly, the theory of belief functions is introduced in Section II, with a focus on evidence combination operators and the handling of evidence defined
on distinct but related domains. In Section III the different elements of our Belief Modeling Regression approach are described in detail. The learning of an “evidential model” of the body from the learning data, based on approximations of the unknown feature-to-pose maps, is described in Section III-A. In Section III-B Dirichlet belief functions are proposed to model the uncertainty due to scarce training data. From the belief estimate resulting from their conjunctive combination, either a pointwise estimate or a set of extremal estimates of the pose can be extracted. In III-C the computational complexity of learning and estimation algorithms is analyzed, while in Section III-D model assessment criteria are discussed. Section IV illustrates the performance of the Belief Modeling Regression (BMR) technique in an application to human pose recovery, showing how BMR outperforms our implementation of both Relevant Vector Machine and Gaussian Process Regression. Finally, Section V discusses motivation and advantages of the proposed approach in comparison with other competitors, analyzes approaches alternative to Dirichlet modeling for belief function inference, and outlines an extension of this technique to tracking, in which temporal consistency is achieved via the total belief theorem.

II. BELIEF CALCULUS

A. Semantics of belief functions

1) Belief functions induced by multi-valued mappings: suppose that we have a probability measure \( P \) for a question \( Q_1 \), and that there exists a one-to-many map \( \rho : \Omega \rightarrow 2^\Theta \) (a multi-valued mapping) from \( \Omega \) to \( \Theta \), the sets of possible answers to \( Q_1 \) and \( Q_2 \), respectively. If the answer to \( Q_1 \) is \( \omega \in \Omega \), then the answer to \( Q_2 \) is somewhere in \( A \) whenever \( \rho(\omega) \subseteq A \). As a number of seminal works by A. Dempster [20] have shown, the result of mapping a probability distribution via a multi-valued map is an object more general than a probability distribution: a belief function [21]. The “degree of belief” \( b(A) \) with which \( A \subseteq \Theta \) contains the answer to \( Q_2 \) (Figure 1-left) is then the total probability of all such answers \( \omega \): \( b(A) = P(\{\omega \in \Omega | \rho(\omega) \subseteq A\}) \).

2) Belief functions as sum functions: a multi-valued mapping maps therefore a probability distribution \( P \) on \( \Omega \) to a distribution on the power set \( 2^\Theta = \{A \subseteq \Theta\} \) of the codomain \( \Theta \). This is a function \( m : 2^\Theta \rightarrow [0, 1] \) s.t. \( \sum_{A \subseteq \Theta} m(A) = 1 \), and is called basic probability assignment (b.p.a.) [19]. The belief function \( b : 2^\Theta \rightarrow [0, 1] \) induced by \( m \) has the form:

\[
b(A) = \sum_{B \subseteq A} m(B).
\]
The belief value (3) of $A$ is the sum of the masses of all its subsets. The domain $\Theta$ of a belief function is called “frame of discernment” (FOD), and the subsets of $\Theta$ associated with non-zero values of $m$ focal elements (f.e.s).

3) Belief functions as credal sets: a popular (albeit criticized) interpretation of belief functions sees each focal element $A$ with b.p.a. $m(A)$ as the indication of the existence of a mass $m(A)$ “floating” inside $A$, which can be assigned to any of its elements $x \in A$. This constrains the possible probability distributions on $\Theta$: a distribution “consistent” with $b$ is obtained by redistributing the mass $m(A)$ of each focal element $A$ to its singleton elements $x \in A$.

The credal set [29] of all and only the probabilities consistent with a belief function $b$ is $\mathcal{P}[b] = \{p \in \mathcal{P} : p(A) \geq b(A) \ \forall A \subseteq \Theta\}$, i.e., the set of probability measures whose values dominate that of $b$ on all events $A$. This is a polytope in the simplex of all probabilities we can define on $\Theta$. Its vertices are all the distributions $p^\pi$ induced by any permutation $\pi = \{x_{\pi(1)}, ..., x_{\pi(|\Theta|)}\}$ of the singletons of $\Theta$ of the form:

$$p^\pi[b](x_{\pi(i)}) = \sum_{A \ni x_{\pi(i)}; A \not\ni x_{\pi(j)} \ \forall j < i} m(A),$$

assigning to a singleton element put in position $\pi(i)$ by the permutation $\pi$ the mass of all focal elements containing it, but not containing any elements preceeding it in the permutation order.

B. Cue integration in belief calculus

The combination of pieces of information obtained from different sources, and represented as belief functions, is a central theme of belief calculus. When such belief functions are generated by “independent” sources of information they can be combined by means of Dempster’s rule of combination [39], or its variant the conjunctive rule of combination [22].

**Definition 1:** the conjunctive combination of two belief functions $b_1, b_2 : 2^\Theta \rightarrow [0, 1]$ is a new belief function $b_1 \Box b_2$ on the same FOD whose focal elements are all the possible intersections of focal elements of $b_1$ and $b_2$ respectively, and whose b.p.a. is given by:

$$m_{b_1 \Box b_2}(A) = \sum_{B \cap C = A} m_{b_1}(B) m_{b_2}(C).$$

Definition 1 can be extended to the combination of an arbitrary number of belief functions.

While it is axiomatically justifiable as the only combination rule which meets a number of sensible requirements such as least commitment, specialization, associativity and commutativity
the conjunctive combination also amounts to assuming that the sources of evidence to merge are both reliable and independent. In general, the current consensus is that different combination rules are to be employed under different assumptions [40]. However, it is difficult to decide in which situations the sources of information can indeed be considered independent: this is the case for features extracted from one or more views of the same object. An alternative point of view, supported by Shenoy, maintains instead that rather than employing a battery of combination rules whose applicability to a given problem is difficult to establish, we should adopt models which do meet the independence of sources assumption, as it happens in probability theory. We support this view here, and will test the adequacy of the assumption empirically.

C. Families of frames

In belief calculus a map between two FODs $\Theta$ and $\Omega$ of the form $\rho : 2^\Theta \to 2^\Omega$, $\rho(A) = \bigcup_{\theta \in A} \rho(\{\theta\})$, which maps a domain $\Theta$ to a disjoint partition of its codomain $\Omega$ ($\rho(\{\theta_1\}) \cap \rho(\{\theta_2\}) = \emptyset$ for all distinct $\theta_1, \theta_2 \in \Theta$) is called a refining. $\Omega$ is called a refinement of $\Theta$, while $\Theta$ is said a coarsening of $\Omega$. When a FOD $\Omega$ is a refinement for a collection $\Theta_1, ..., \Theta_N$ of others it is called a common refinement of them [19]. $\Theta_1, ..., \Theta_N$ are said to be independent [19] if

$$\rho_1(A_1) \cap \cdots \cap \rho_N(A_N) \neq \emptyset, \quad \forall \emptyset \neq A_i \subseteq \Theta_i, \forall i = 1, ..., N$$

where $\rho_i$ is the refining from $\Theta_i$ to $\bigotimes_i \Theta_i$. In such a case, their minimal refinement is nothing but their Cartesian product: $\bigotimes_i \Theta_i = \Theta_1 \times \cdots \times \Theta_N$.

A belief function $b'$ on $\Omega$, a refinement of $\Theta$, is called the vacuous extension of another belief function $b$ on $\Theta$ iff the focal elements of $b'$ are images (via $\rho$) of focal elements of $b$.

III. APPROACH: BELIEF MODELING REGRESSION

A. Model learning

1) Building feature-pose maps: consider an image feature $y$, whose values live in a feature space $\mathcal{Y}$, and let us denote by $\rho : \mathcal{Y} \to 2^\mathcal{Q}$ the unknown mapping from the feature space to the collection $2^\mathcal{Q} = \{Q \subseteq \mathcal{Q}\}$ of sets of object poses. We seek to learn from the training data an approximation $\tilde{\rho}$ of this unknown mapping, which is applicable to any feature value, and ideally produces only admissible object configurations. As evidence is limited, we can only constrain $\tilde{\rho}$ to have output in the space $\mathbb{R}^D$ the true pose space $\mathcal{Q}$ is embedded into.
We propose to obtain such a representation by applying EM clustering \cite{28} to the training data (1), (2), individually for each feature component.

Consider the $N$ sequences of feature values $\{y_i(k), k = 1, ..., T\}$, $i = 1, ..., N$, acquired during training. EM clustering can be applied to them to obtain a Mixture of Gaussians (MoG)

$$\left\{ \Gamma_i^j, j = 1, ..., n_i \right\}, \quad \Gamma_i^j \sim \mathcal{N}(\mu_i^j, \Sigma_i^j)$$

with $n_i$ Gaussian components, separately for each feature space (the range $\mathcal{Y}_i \subset \mathbb{R}^{d_i}$ of the unknown feature function $y_i : \mathcal{I} \rightarrow \mathcal{Y}_i$ on the set of all images $\mathcal{I}$). MoG models are often employed in bottom-up pose estimation, as their parameters can be speedily estimated via the EM algorithm \cite{28}. For instance, in \cite{6} several “experts” predictions are combined in a Gaussian mixture model. In \cite{25} conditional distributions are also assumed to be Gaussian mixtures.

Here we use the learnt MoG (7) to build a particle-based discrete approximation of the unknown feature pose mapping. The former induces an implicit partition

$$\Theta_i = \left\{ \mathcal{Y}_i^1, \ldots, \mathcal{Y}_i^{n_i} \right\}$$

of the $i$-th feature range, where $\mathcal{Y}_i^j = \left\{ y \in \mathcal{Y}_i \ s.t. \ \Gamma_i^j(y) > \Gamma_i^l(y) \ \forall \ l \neq j \right\}$ is the region of $\mathcal{Y}_i$ in which the $j$-th Gaussian component dominates all the others (Figure 1-right). We call (8) the $i$-th “approximate” feature space, though the purpose here is to model the feature-pose relation in an efficient way rather than to approximate the actual feature space.

In virtue of the fact that features are computed during training in synchronous with the true poses provided by the source of ground truth, each element $\mathcal{Y}_i^j$ of the approximate feature space

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Left: A probability measure $P$ on $\Omega$ induces a belief function $b$ on $\Theta$ through a multi-valued mapping $\rho$. Right: a Mixture of Gaussian learned via EM from the training features defines an implicit partition on the set of training poses $\tilde{Q}$. The Gaussian densities $\{\Gamma^j, j = 1, ..., n\}$ on the range $\mathcal{Y}$ of the feature function define a partition of $\mathcal{Y}$ into regions $\{\mathcal{Y}^j\}$. Each of those regions $\mathcal{Y}^j$ is in correspondence with the set $\tilde{Q}^j$ of sample poses $q_k$ whose feature value $y(q_k)$ falls inside $\mathcal{Y}^j$.}
\end{figure}
is associated with the set of training poses \( q_k \in Q_k \) whose \( i \)-th feature value falls in \( Y_{k}^i \) (Figure 1-right again):
\[
\rho_i : Y_{i}^i \mapsto Q_i = \left\{ q_k \in \tilde{Q} : y_i(k) \in Y_{k}^i \right\}.
\] (9)

Applying EM clustering separately to each training feature sequence (2) yields therefore both \( N \) approximate feature spaces \( \Theta_i = \{ Y_1^i, \cdots, Y_{n_i}^i \} \), \( i = 1, \cdots, N \), and \( N \) maps (9) from each of them to the approximate pose space (the set of training poses) \( \tilde{Q} \). The learned feature-pose maps (9) amount to constraints on the unknown feature pose maps \( \rho_i^* : Y \rightarrow 2^\Omega \), built on the evidence available in the specific regions covered by training feature/pose pairs (see Section V). Just as their unknown counterparts, they are inherently multi-valued, i.e., they map elements of each approximate feature space \( \Theta_i \) to sets of training poses. The number \( n_i \) of clusters can be estimated by cross-validation. Here we will set it to a fixed value for each feature space.

2) Continuous mapping via belief functions: the maps (9) only apply to partitions of the feature range, and cannot be used to map individual feature values as they are. The structure provided by the learned MoGs (7) can nevertheless be used to build universal mappings. Given the mixture (7), each new feature value \( y_i \) can be represented by its soft assignments
\[
y_i \mapsto \left[ \Gamma_1^i(y_i), \Gamma_2^i(y_i), \cdots, \Gamma_{n_i}^i(y_i) \right]
\] (10)
to each mixture component. The density values (10) constitute a vector of coordinates of the feature value in the feature range \( Y_i \); in this interpretation, the MoG approximation of \( Y_i \) provides an atlas of coordinate charts on the feature space itself. Rather than mapping \( y \) we can use (9) to map the associated coordinates (soft assignments) (10), extending the “particle”-like information on the shape of \( \rho_i^* \) given by a learnt refining (9) to map any test feature value.

By normalizing (10), each (test) feature value is associated with a probability distribution on the “approximate” feature space \( \Theta_i \). By comparing Figures 1-left and 1-right, it is clear that the maps (9) are multi-valued mappings linking the question \( Q_1 \) “to which Gaussian component of the MoG (7) does the new feature value \( y \) belong” to the question \( Q_2 \) “what is the object pose whose observed feature value is \( y \)”. From Section II-A.1, it follows that the probability distribution associated with any feature value induces a belief function on the (approximate) pose range \( \tilde{Q} \). Overall, the learnt universal feature-pose mapping is a cascade of soft assignment and refining-based multi-valued mapping:
\[
y_i \in Y_i \overset{(10)}{\mapsto} \left[ \Gamma_1^i(y_i), \Gamma_2^i(y_i), \cdots, \Gamma_{n_i}^i(y_i) \right] \mapsto p_i = [p_i(\mathcal{Y}_1^i), \cdots, p_i(\mathcal{Y}_{n_i}^i)] \overset{(9)}{\mapsto} b_i : 2^{\tilde{Q}} \rightarrow [0,1]
\] (11)
where \( p_i(\mathcal{Y}_i^j) = \frac{\Gamma_j(y_i)}{\sum_k \Gamma_k(y_i)} \), associating any test feature value with a belief function (a specific convex set of probability distributions) on the set of training poses \( \tilde{Q} \).

3) **Training algorithm:** to summarize, in the training stage the body moves in front of the camera(s), exploring its configuration space, while a sequence of training poses \( \tilde{Q} = \{q_k, k = 1, ..., T\} \) is provided by a source of ground truth (for instance a motion capture system, Section I-A). The sample images are annotated by a bounding box indicating the location of the object within each image. At the same time:

1. for each time instant \( k \), a number of feature values are computed from the region of interest of each available image: \( \{y_i(k), k = 1, ..., T\}, i = 1, ..., N; \)
2. EM clustering is applied to each feature sequence \( \{y_i(k), k = 1, ..., T\} \) (after setting the number of clusters \( n_i \)), yielding:
   2.a \( N \) approximate feature spaces \( \Theta_i = \{\mathcal{Y}_i^j, j = 1, ..., n_i\} \), i.e., the implicit partitions of the feature ranges \( \mathcal{Y}_i \) associated with the EM clusters (Section III-A.1);
   2.b \( N \) maps (9) \( \rho_i: \mathcal{Y}_i^j \in \Theta_i \mapsto \tilde{Q}_i^j \equiv \{q_k \in \tilde{Q}: y_i(k) \in \mathcal{Y}_i^j\} \) mapping EM feature clusters to sets of sample training poses in the approximate pose space \( \tilde{Q} \).

As the applications (9) map approximate feature spaces to disjoint partitions of the approximate pose space \( \tilde{Q} \) they are refining, and \( \tilde{Q} \) is a common refinement for the collection of approximate feature spaces \( \Theta_1, ..., \Theta_N \). The collection of FODs \( \tilde{Q}, \Theta_1, ..., \Theta_N \) along with their refinings \( \rho_1, ..., \rho_N \) is characteristic of both the object to track, the chosen features functions \( y_i \), and the actual training data: we call it the **evidential model** (Figure 2) of the object.

**B. Estimation**

Once an evidential model has been learned from the available training set, it can be used to provide robust estimates of the pose of the moving object when new evidence becomes available.

1) **Dirichlet belief function modeling of soft assignments:** when one or more new test images are acquired, new visual features \( y_1, ..., y_N \) are extracted. Such feature values can be mapped by the learnt universal mappings (11) to a collection of belief functions \( b_1, ..., b_N \) on the set of training poses \( \tilde{Q} \). From Section II-A.3, each \( b_i \) corresponds to a convex set of probability distributions, whose width encodes the uncertainty on the pose value due to the uncertainty on the analytical form of the true, unknown feature-pose map \( \rho_i^* \).
In addition, belief functions allow to take into account the scarcity of the training samples, by introducing uncertainty on the soft assignment (10) itself. This can be done by assigning some mass \( m(\Theta_i) \) to the whole approximate feature space, prior to applying the refining \( \rho_i \). This encodes the fact that there are other samples out there which, if available, would alter the shape of the MoG approximation of \( \mathcal{Y}_i \) in unpredictable ways. Namely, we map the soft assignment (10) to a Dirichlet belief function [41], with basic probability assignment:

\[
m_i : 2^{\Theta_i} \rightarrow [0, 1], \quad m_i(\mathcal{Y}_i^j) = \frac{\Gamma_i^j(y_i)}{\sum k \Gamma_i^k(y_i)} (1 - m_i(\Theta_i)).
\] (12)

The b.p.a. (12) "discounts" the probability distribution obtained by simply normalizing the likelihoods (10) by assigning some mass \( m_i(\Theta_i) \) to the entire FOD \( \Theta_i \).

As we need to discount the limited accuracy achieved by using as coordinates in \( \mathcal{Y}_i \) those derived by the MoG representation \( \Theta_i \), a plausible choice is \( m_i(\Theta_i) = \frac{1}{n_i} \). When \( n_i \rightarrow \infty \) the discount factor tends to zero, and the approximate feature space converges (in theory) to the real thing. In addition, as \( n_i \) cannot be greater than the number of training pairs \( T \), such a discounting factor also takes into account the limited number of training samples.

Fig. 2. Evidential model. The EM clustering of each feature set collected in the training stage yields an approximate feature space \( \Theta_i = \{\mathcal{Y}_i^j, j = 1, ..., n_i\} \). Refining maps \( \rho_i \) between each approximate feature space and \( \tilde{Q} = \{q_1, ..., q_T\} \) (the training approximation of the unknown pose space \( Q \)) are learned, allowing the fusion on \( \tilde{Q} \) of the evidence gathered on \( \Theta_1, ..., \Theta_N \).

2) Belief estimate: the measurement belief functions \( \{b_i : 2^{\Theta_i} \rightarrow [0, 1], i = 1, ..., N\} \) inferred from the test feature values \( y_1, ..., y_N \) via (12) are then mapped to belief functions \( \{b'_i : 2^{\tilde{Q}} \rightarrow \} \)
\[ \forall A \subset \tilde{Q} \] on the approximate pose space \( \tilde{Q} \) by vacuous extension: \( \forall A \subset \tilde{Q} \)

\[
m'_i(A) = \begin{cases} 
m_i(A_i) & \exists A_i \subset \Theta_i \text{ s.t. } A = \rho_i(A_i); \\ 0 & \text{otherwise.} \end{cases}
\]  

(13)

The resulting b.f.s on \( \tilde{Q} \) are combined by conjunctive combination (5). The result is a belief function \( \hat{b} = b'_1 \cap \cdots \cap b'_N \) on \( \tilde{Q} \) which is right to call the belief estimate of the object pose.

3) Example: it is important to understand how sophisticated a description of the object’s pose a belief function is, as opposed to any estimate in the form of a “precise” probability distribution, even complex multi-modal or particle-based descriptions based on extracting samples using Monte-Carlo methods, as a belief estimate \( \hat{b} \) of the pose represents an entire convex collection of probabilities (credal set) on the approximate pose space (see Section II-A.3). Suppose that the approximate pose space contains just three samples: \( \tilde{Q} = \{q_1, q_2, q_3\} \). Suppose also that the evidence combination process has produced a belief estimate \( \hat{b} \) with b.p.a.:

\[
\hat{m}(\{q_1, q_2\}) = \frac{1}{3}, \quad \hat{m}(\{q_3\}) = \frac{1}{6}, \quad \hat{m}(\{q_1, q_2, q_3\}) = \frac{1}{2}.
\]  

(14)

By (4), the vertices of \( P[\hat{b}] \) are those probabilities generated by reassigning the mass of each focal element to any one of its singletons: there are \( \prod_k |A_k| \) such possible choices, where \( \{A_k\} \)

---

Fig. 3. The convex set of probability distributions \( P[\hat{b}] \) (in red) associated with the belief function \( \hat{b} \) (14) on the approximate parameter space \( \tilde{Q} = \{q_1, q_2, q_3\} \), displayed on the triangle of all probability distributions on \( \tilde{Q} \). The pignistic approximation \( BetP[\hat{b}] \) (the center of mass of \( P[\hat{b}] \), in blue) is also shown.
is the list of focal elements of \( \hat{b} \). As (14) has 3 focal events of size 1, 2 and 3, the corresponding credal set \( \mathcal{P}[\hat{b}] \) will be the convex closure of \( 1 \cdot 2 \cdot 3 = 6 \) probability distributions, namely:

\[
p_1 = [\frac{5}{6}, 0, \frac{1}{6}], \quad p_2 = [\frac{1}{2}, \frac{1}{3}, \frac{1}{6}], \quad p_3 = [\frac{1}{3}, \frac{1}{2}, \frac{1}{6}], \quad p_4 = [0, \frac{5}{6}, \frac{1}{6}], \quad p_5 = [0, \frac{1}{3}, \frac{2}{3}], \quad p_6 = [\frac{1}{3}, 0, \frac{2}{3}].
\]

Figure 3 shows the credal set (a polygon) associated with the belief estimate (14) in the simplex of all probability distributions on \( \tilde{Q} \) (a triangle in this case), in which each probability distribution is represented as a point of \( \mathbb{R}^3 \). The larger the polygon, the greater the uncertainty of the estimate. In the figure \( \mathcal{P}[\hat{b}] \) covers almost all the probability simplex, displaying a large degree of imprecision of the estimate due to lack of evidence.

4) Computing expected pose estimates: point-wise information on the object’s pose can be extracted from \( \hat{b} \) in two different ways.

a) Extracting a set of extremal point-wise estimates: each of the vertices (4) of the credal set associated with the belief estimate \( \hat{b} \) is a probability distribution on the approximate pose space \( \tilde{Q} \). We can then compute the associated expected pose as:

\[
\hat{q} = \sum_{k=1}^{T} p(q_k)q_k.
\]

The set of such “extremal” estimates describes therefore an entire polytope of expected pose values in the object’s pose space \( Q \). In the example, the expected poses for the vertices \( p_1, p_4, p_5, p_6 \) of \( \mathcal{P}[\hat{b}] \) are: \( \hat{q}[p_1] = \frac{5}{6}q_1 + \frac{1}{6}q_3, \hat{q}[p_4] = \frac{5}{6}q_2 + \frac{1}{6}q_3, \hat{q}[p_5] = \frac{1}{3}q_2 + \frac{2}{3}q_3, \hat{q}[p_6] = \frac{1}{3}q_1 + \frac{2}{3}q_3. \)

b) Extracting a point-wise estimate: an alternative way to extract a single pose estimate \( \hat{q} \) from the belief estimate consists in approximating \( \hat{b} \) with a probability \( \hat{p} \) on \( \tilde{Q} \), and then computing its mean value as above. The problem has been indeed extensively studied. In particular, Smets’ pignistic function [42]

\[
\text{BetP}[\hat{b}](x) = \sum_{A \supseteq x} \frac{m_b(A)}{|A|} \quad \forall x \in \Theta
\]

has been proposed within the framework of the Transferable Belief Model as the unique transformation which meets a number of rationality principles. Geometrically, it is nothing but the barycenter of the convex set of probabilities \( \mathcal{P}[\hat{b}] \) associated with \( \hat{b} \) (see Figure 3). As such, it is more compatible with the interpretation of belief functions as credal sets of probabilities. Though other transforms such as the “relative plausibility of singletons” [43], [44], [45] and the “intersection probability” [46] have been proposed, empirically the performances of the different
pointwise transformations in the human pose tests presented here have been proven comparable. Therefore, in the following we will simply adopt the pignistic transform.

5) Handling of conflict: the mass assigned to the empty set by the conjunctive combination (5) represents the amount of evidence which is in contradiction, as it is assigned by the input b.f.s to contradictory (disjoint) focal elements. In our pose estimation scenario, conflict can arise when combining feature evidence via $b'_1 \bigcap \cdots \bigcap b'_N$ for basically two reasons: 1- the object is localized in an imprecise way (due to limitations of the trained detector), so that background features conflicting with the foreground information are also extracted; 2- occlusions are present, generating conflict for similar reasons. The critical case is when all focal elements of one belief function have empty intersection with those of the other b.f.s to combine: all the mass is assigned to $\emptyset$, and no estimation is possible.

When modeling the scarcity of training pairs via Dirichlet belief functions (Section III-B.1), however, this extreme scenario never materializes, as each individual b.f. has $\Theta_i$ as a focal element. In case of disagreement, then, some mass is always assigned to the focal elements of the remaining belief functions. As we argue in Section V-C, combining Dirichlet belief functions amounts to say that all the partial combinations of feature evidence are given some credit. Maybe, this is the reasoning, only a subset of features is telling the truth [47]. Under the assumption that most features come from the foreground, this brings robustness to localization errors and presence of occlusions. In the following, therefore, we do not employ any explicit conflict resolution mechanism.

6) Pose estimation algorithm: let us summarize the whole pose estimation procedure. Given an evidential model of the moving body with $N$ feature spaces, and given at time $t$ one or more test images, possibly coming from different cameras:

1. the object detector learned during training is applied to the test image(s), returning for each of them a bounding box roughly containing the object of interest;
2. the $N$ feature values are extracted from the resulting bounding boxes, as during training;
3. the likelihoods $\{\Gamma^j_i(y_i(t)), j = 1, ..., n_i\}$ of each feature value $y_i(t)$ with respect to the appropriate learned Mixture of Gaussian distribution on $\mathcal{Y}_i$ are computed (10);
4. for each feature $i = 1, ..., N$, a separate b.f. $b_i(t) : 2^{\Theta_i} \rightarrow [0, 1]$ on the appropriate feature space $\Theta_i$ is built from the set of likelihoods $\{\Gamma^j_i(y_i(t)), j = 1, ..., n_i\}$ as in Section III-B.1;
5. all the resulting b.f.s $\{b_i(t) : 2^{\Theta_i} \rightarrow [0, 1], i = 1, ..., N\}$ are projected onto $\tilde{Q}$ by vacuous
extension (13), yielding a set of belief functions \( \{b'_i : 2^{\tilde{Q}} \to [0, 1], i = 1, ..., N \} \) on \( \tilde{Q} \);

6. their conjunctive combination \( \hat{b}(t) = b'_1(t) \cap \cdots \cap b'_{N}(t) \) is computed via (5);

7a. the pignistic transform (16) is applied to \( \hat{b}(t) \), yielding a distribution on \( \tilde{Q} \) from which an expected pose estimate \( \hat{q}(t) \) is obtained by (15);

7b. the vertices (4) of the convex set of probabilities \( P[\hat{b}(t)] \) associated with the current belief estimate \( \hat{b}(t) \) are computed, and a mean pose estimate (15) obtained for each one of them.

C. Computational cost

1) Learning: EM’s computational cost is easy to assess, as the algorithm usually takes a constant number of steps to converge, \( c \sim 5 - 10 \), while at each step the whole observation sequence of length \( T \) is processed, yielding \( O(cNnT) \) (where again \( N \) is the number of features, \( n \) the average number of EM clusters, \( T \) the number of samples collected in the training stage). This is quite acceptable for real-world applications since this has to be done just once in the training session. In the experiments of Section IV the whole learning procedure in Matlab required some 17.5 seconds for each execution of the EM on a rather old Athlon 2.2 GHz processor with \( N = 5 \) features, \( n_i = n = 5 \) states for each feature space, and \( T = 1726 \).

2) Estimation: although the conjunctive combination (5) is exponential in complexity if naively implemented, fast implementations of \( \cap \) exist, under additional constraints [48]. Numerous approximation schemes have been proposed, based on Monte-Carlo techniques [49]. Furthermore, the particular form of the belief functions we use in the estimation process needs to be taken into account. Dirichlet b.f.s (12) have \( n_i + 1 \) non-zero focal elements, reducing the computational complexity of their pairwise combination from \( O(2^{2n}) \) (associated with the mass multiplication of all possible \( 2^n \) focal elements of the first b.f. and all the focal elements of the second b.f.) to \( O(n^2) \). The other steps of the algorithm have a computational cost which is negligible with respect to belief combination.

D. Assessing evidential models

A number of aspects of the evidential model architecture are strictly related to fundamental questions of the example-based pose estimation problem: 1) whether the model is self-consistent, i.e., whether it produces the correct ground truth pose values when presented with the training feature data; 2) what resolutions \( \{n_i, i = 1, ..., N\} \) of the feature’s MoG representations are
adequate to guarantee a sufficient accuracy of the learned feature-pose mapping, and through the latter of the estimation process itself; 3) whether the training set of poses \( \tilde{Q} \) is indeed a proper approximation of the unknown parameter space \( Q \) (see Figure 2).

As it turns out, those issues are related to discussing, respectively: 1) whether \( \tilde{Q} \) is the “minimal refinement” of the approximate feature spaces \( \Theta_i \); 2) whether the selected features space are independent, in a very precise way; 3) whether a flag can be derived to indicate the need to update the evidential model by adding more training poses.

1) Model consistency and \( \tilde{Q} \) as minimal refinement: in order for the model to return the correct ground truth pose when presented with a set of training feature values \( \{y_i(k), i = 1, ..., N\} \) it is necessary that each sample in the training set \( \tilde{Q} \) be characterized by a distinct set of feature MoG components, i.e., no two training poses \( q_1, q_2 \) may have feature components falling in the same cluster for each approximate feature space: \( \not\exists q_1, q_2 \) s.t. \( y_i(q_1), y_i(q_2) \in \mathcal{Y}_i \) \( \forall i \) for the same \( j_1, ..., j_N \).

Imagine that the \( N \) feature vector components \( y_1, ..., y_N \) generated by a test image are such that: \( y_1 \in \mathcal{Y}_i^{j_1}, ..., y_N \in \mathcal{Y}_i^{j_N} \). Each piece of evidence \( y_i \in \mathcal{Y}_i^{j_i} \) implies that the object’s pose lies within the subset \( \rho_i(\mathcal{Y}_i^{j_i}) \) of the training set \( \tilde{Q} \). The estimated pose must then fall inside the set:

\[
\rho_1(\mathcal{Y}_1^{j_1}) \cap \cdots \cap \rho_N(\mathcal{Y}_N^{j_N}) \subset \tilde{Q},
\]

Sample object poses in the same intersection of the above form are indistinguishable under the given evidential model. The collection of all the non-empty intersections of the form (17) is called the minimal refinement \( \Theta_1 \times \cdots \times \Theta_N \) of the FODs \( \Theta_1, ..., \Theta_N \). It follows that:

**Theorem 1:** Any two poses of the training set can be distinguished under the evidential model iff \( \tilde{Q} \) is the minimal refinement of \( \Theta_1, ..., \Theta_N \).

**Proof:** \( \Rightarrow \): if any two sample poses can be distinguished under the model, i.e., for all \( k, k' \)

\[
q_{k'} \not\in \rho_1(\mathcal{Y}_1^{j_1}) \cap \cdots \cap \rho_N(\mathcal{Y}_N^{j_N}) \ni q_k;
\]

it follows that each intersection of the form (17) cannot contain more than one sample pose, otherwise there would exist a pair violating (18) (note that the intersection can instead be empty).

Furthermore, each sample pose \( q_k \) falls within such an intersection, the one associated with the visual words \( \mathcal{Y}_1^{j_1}, ..., \mathcal{Y}_N^{j_N} \) s.t. \( y_1(q_k) \in \mathcal{Y}_1^{j_1}, ..., y_N(q_k) \in \mathcal{Y}_N^{j_N} \). Hence, the minimal refinement of \( \Theta_1, ..., \Theta_N \) has as elements (17) all and only the individual sample poses (elements of \( \tilde{Q} \)):
therefore, \( \tilde{Q} = \Theta_1 \otimes \cdots \otimes \Theta_N \).

\[ \iff \text{if } \tilde{Q} \text{ is the minimal refinement of } \Theta_1, \ldots, \Theta_N \text{ then for all } q_k \in \tilde{Q} \text{ we have that } \{ q_k \} = \rho_1(Y_{i_1}^{j_1}) \cap \cdots \cap \rho_N(Y_{i_N}^{j_N}) \text{ holds for some unique selection of feature components } Y_{i_1}^{j_1}, \ldots, Y_{i_N}^{j_N}, \text{ distinct for each training pose. Any two different sample poses belong therefore to different intersections of the form (17), i.e., they can be distinguished under the model.} \]

The “self-consistency” of the model can then be measured by the ratio between the cardinality of the minimal refinement of \( \Theta_1, \ldots, \Theta_N \), and that of the actual approximate parameter space \( \tilde{Q} \):

\[ \frac{1}{T} \leq \frac{|\otimes_i \Theta_i|}{|\tilde{Q}|} \leq 1. \]

It is hence desirable, in the training stage, to select a collection of features which brings the minimal refinement \( \Theta_1 \otimes \cdots \otimes \Theta_N \) as close as possible to \( \tilde{Q} \): sometimes the addition of new features will be desirable in order to resolve the ambiguities.

2) Complementarity of features and model quantization: when the approximate feature spaces \( \Theta_i \) are independent (Equation (6)) for each combination of feature clusters \( Y_{i_1}^{j_1}, \ldots, Y_{i_N}^{j_N} \) there exists a unique sample pose \( q_k \) characterized by feature values in those clusters: \( \{ q_k \} = \rho_1(Y_{i_1}^{j_1}) \cap \cdots \cap \rho_N(Y_{i_N}^{j_N}) \). Different cues carry then complementary pieces of information about the object’s pose: to resolve an individual sample pose \( q_k \) you need to measure all its feature values.

When the approximate feature spaces are not independent, instead, two opposite facts may take place: while in some cases fewer than \( N \) feature values may be enough to resolve some training poses, in general each combination of feature values will yield a whole set of training poses.

If the model is self-consistent (\( |\tilde{Q}| = |\otimes_i \Theta_i| \)) and the chosen features are complementary (i.e., they are such that \( |\otimes_i \Theta_i| = \prod_i |\Theta_i| \)), we have that \( T = |\tilde{Q}| \sim n_1 \times \cdots \times n_N \): assuming \( n_i = \text{const} = n \) this means \( n \sim \sqrt{T} \). Given a realistic sampling of the parameter space with \( T = 20000 \), the use of \( N = 9 \) complementary features allows to require no more than \( \sqrt{20000} \sim 3 \) MoG components for each feature space in order to assure a decent accuracy of the estimate.

This shows the clear advantage of encoding feature-pose maps separately: as long as the chosen features are uncorrelated, a relatively coarse MoG representation for each feature space allows us to achieve a decent resolution in terms of pose estimates, in analogy to what proposed in [15] or [18], where trees of classifiers are used for face pose estimation.

3) On conflict and the relation between approximate and actual pose space: ideally, the set \( \tilde{Q} \) of training poses, as an approximation of the actual pose space \( Q \), should be somehow “dense” in \( Q \): \( \forall q \in Q \) there should be a sample \( q_k \) such that \( \|q - q_k\| < \epsilon \) for some \( \epsilon \) small enough.
Clearly, such a condition is hard to impose. The distribution of the training poses within $Q$ has nevertheless a number of consequences on the estimation process:

1) as the true pose space $Q$ is typically non-linear, while the pose estimate is a linear combination of sample poses (see Section III-B.2), the pointwise estimate can be non-admissible (fall outside $Q$). This can be fixed by trying to make the feature spaces independent, as in that case every sample pose $q_k$ is characterized by a different combination of feature clusters: $\{q_k\} = \rho_1(Y_{j_1}) \cap \cdots \cap \rho_N(Y_{j_N})$. As a consequence, any set of test feature values $y_1 \in Y_{j_1}, \ldots, y_N \in Y_{j_N}$ generates a belief estimate in which a single sample pose $q_k$ is dominant: its credal set (III-B.4.a) is of limited extension around a single sample pose, and the risk of non-admissibility reduced.

2) there can exist regions of $Q$ characterized by combinations of feature clusters which are not in the model: $\exists q \in Q : \forall Y_{j_1} \in \Theta_1, \ldots, Y_{j_N} \in \Theta_N \quad q \notin \rho_1(Y_{j_1}) \cap \cdots \cap \rho_N(Y_{j_N})$. This generates high level of conflict $m(\emptyset)$ in the conjunctive combination (5) (although combination is always guaranteed for Dirichlet belief functions, see above), a flag of the inadequacy of the current version of the evidential model. This calls, whenever new ground truth information can be provided, for an update of the model by incorporating the sample poses generating the problem.

IV. EXPERIMENTAL RESULTS

We tested our Belief Modeling Regression technique in a rather challenging setup, involving the pose estimation of human arms and legs from two well separated views. While the bottom line of the evidential approach is doing the best we can with the available examples, regardless the dimensionality of the pose space, and without having at our disposal prior information on the object at hand, we ran test on articulated objects (one arm and a pair of legs) with a reasonably limited number of degrees of freedom to show what can be achieved in such a case, and show how the technique outperforms competitors such as Relevant Vector Machines and GPR.

A. Setup: two human pose estimation experiments

To collect the necessary ground truth we used a marker-based motion capture system [25], [23] built by E-motion, a Milan firm. The number of markers used was 3 for the arm (yielding a pose space $Q \subset \mathbb{R}^9$, using as pose components the 3D coordinates of the marker), and 6 for the pair of legs ($Q \subset \mathbb{R}^{18}$). The person was filmed by two uncalibrated DV cameras (Figure 4). In the training stage of the first experiment we asked the subject to make his arm follow a trajectory
(approximately) covering the pose space of the arm itself, keeping his wrist locked and standing on a fixed spot on the floor to limit the intrinsic dimensionality of the pose space (resulting in 2 d.o.f.s for the shoulder and 3 for the elbow). In the second experiment we tracked the subject’s legs, assuming that the person was walking normally on the floor, and collected a training set by sampling a random walk on a small section of the floor. This is similar to what is done in other works, where the set of examples are taken for a particular family of motions/trajectories, normally associated with action categories such as the walking gait. The length of the training sequences was 1726 frames for the arm and 1952 frames for the legs.

While the number of degrees of freedom was limited by constraining the articulated object (person) to performing motions of a specific class (walking versus brandishing an arm), the tests are sufficiently complex to allow us to illustrate the traits of the BMR approach to pose estimation. In addition, in both experiments the background was highly non-static, with people coming in and out the scene and flickering monitors; the object of interest would self-occlude itself a number of times on at least one of the two views (e.g. when one leg occludes the other when seen from the left camera), making the experimental setup quite realistic.

B. Automatic annotation of training images

Under the assumptions listed in Section I-A, in the training stage the images ought to be annotated by a bounding box, which provides a rough localization of the unknown object.

Fig. 4. Two human body-part pose estimation experiments. Left: training images of a person standing still and moving his right arm. Right: training images of the person walking inside a rectangle on the floor.

To simulate this annotation process, and isolate the performance of the proposed example based estimation approach from that of the object detector employed, in these tests we used color-based segmentation to separate the object of interest from the non-static background, implemented via a colorimetric analysis of the body of interest (Figure 5): pixels were clustered in the RGB space; the cluster associated with the yellow sweater (in the arm experiment) or the black pants (legs
one) was detected, and pixels in that cluster assigned to the foreground; the minimal bounding box containing the silhouette of the segmented foreground pixels was finally detected. Note that this is just a way of automatically generate, rather than manually construct, the bounding box annotation required in the assumptions of the initial scenario: the notion that no a-priori information on the object of interest needs to be used still holds.

C. Feature extraction and modeling

For these tests we decided to build an extremely simple feature vector for each image directly from the bounding box, as the collection \( \max(\text{row}), \min(\text{row}), \max(\text{col}), \min(\text{col}) \) of the row and column indexes defining the box (Figure 5). As two views were available at all times, at each time instant two feature vectors of dimension 4 were computed from the two views.

In the arm experiment we build \textit{three} different evidential models from these vectors: one using \( N = 2 \) features (\( \max(\text{row}) \) and \( \max(\text{col}) \)) from the left view only, and a Mixture of Gaussians with \( n_i = n = 5 \) components for both feature spaces; a second model for the right view only, with \( N = 3 \) feature spaces (associated with the components \( \max(\text{row}), \min(\text{col}) \) and \( \max(\text{col}) \)) and \( n_i = n = 5 \) MoG components for each feature space; an overall model in which both the 2 features from the left view and the 3 features from the right one were considered, yielding a model with \( N = 5 \) feature spaces with the same MoG representation.

In the leg experiment, instead, we built two models with \( N = 6 \) feature spaces (the \( \max(\text{row}), \min(\text{col}) \) and \( \max(\text{col}) \) feature components from both views), but characterized by a different
number of Gaussian components \( n = 4 \) or \( n = 5 \), respectively) to test the influence of the quantization level on the quality of the mapping and therefore of the estimates.

D. Performance

To measure the accuracy of the estimates produced by the different evidential models, we acquired a testing sequence for each of the two experiments and compared the results with the ground truth provided by the motion capture equipment.

![Fig. 6. Top left: pose estimates of component 9 of the pose vector (Y coordinate of the hand marker) produced by the left (red) and right (magenta) model compared to the ground truth (blue), plotted against time. Top right: the sequence of pose estimates yielded by the overall model which uses features computed in both left and right images is plotted in (solid) red against the ground truth in (dashed) blue. Bottom: performance of the overall model on components 1 (left) and 6 (right) of the pose vector, for the first 400 frames of the test sequence. The pignistic function is here used to compute the pointwise estimates.](image)

1) Arm experiment: in the arm experiment the test sequence was 1000 frames long. Pointwise pose estimates were extracted from belief estimates via pignistic transform (16). As the anecdotal
evidence of Figure 6-top left indicates, the estimates of the single-view models were of rather poor quality. Indeed, recalling the discussion of Section III-D.1, the minimal refinements $\bigotimes \Theta_i$ for the left-view and the right-view models were of size 22 and 80 respectively, signalling a poor model resolution. In opposition, the estimates obtained by exploiting image evidence from both views (Figure 6-top right) were clearly better than a simple selection of the best partial estimate at each instant. This was confirmed by a minimal refinement $\bigotimes \Theta_i$ for the overall model with cardinality equal to 372 (the $N = 5$ features encoded by a MoG with $n = 5$ components were enough to resolve 372 of the 1700+ sample poses), with 139 sample poses individually resolved by some particular combination of the $N = 5$ feature values. Figure 6-bottom illustrates analogous results for components 1 and 6 of the pose vector.

We also measured the Euclidean distance between real and expected 3D locations of each marker over the whole testing sequence. For the arm experiment, the average estimation errors were 17.3, 7.95, 13.03, and 2.7 centimeters for the markers “hand”, “wrist”, “elbow” and “shoulder”, respectively. As during testing the features were extracted from the estimated foreground, and no significant occlusions were present, the conflict between the different feature components was negligible throughout the test sequence.

2) Lower and upper estimates associated with the credal estimate: as each belief estimate $\hat{b}$ amounts to a convex set $P[\hat{b}]$ of probability distributions on $\tilde{Q}$, an expected pose estimate can be computed for each of its vertices (4). The BMR approach can therefore provide a “robust” pose estimate, for instance by computing for each instant $t$ the maximal and minimal expected value (over the vertices of $P[\hat{b}]$) of each component of the pose vector.

Figure 7 plots these upper and lower bounds to the expected pose values in the arm experiment, for three different components of the pose vector, over three subsequences of the test sequence. As it can be clearly observed, even for the rather poor (feature-wise) evidential model built here, most of the time the true pose falls within the provided interval of expected pose estimates. Quantitatively, the percentage of test frames in which this happens for the twelve pose components is 49.25%, 44.92%, 49.33%, 50.50%, 48.50%, 48.33%, 49.17%, 54.42%, 49.67%, 51.50%, 39.33% and 43.50%, respectively. We can also measure the average Euclidean distance between the true pose estimate and the boundary of the interval of expected poses, for the four markers and along the entire test sequence: we obtain average 3D distances of 7.84cm, 3.85cm, 5.78cm and 2.07cm for the four markers, respectively, which give a better indication of the
robustness of the evidential approach than the errors associated with the central expected pose estimate associated with the pignistic function (see Figure 12-right for a comparison).

Note that in these tests the pose estimate interval was computed using just a subset of the
3) Comparison with GP and RVM regression: it is interesting to compare the performance of our technique with that of two classical regression approaches: Gaussian Process Regression [31], [32] and Relevant Vector Machines (RVMs) [50]. The latter are used to build feature-pose maps in, for instance, [14] and [16]. Figure 8 shows the estimates produced by a RVM on the same test sequences and components of Figure 6. It is clear from a visual comparison of Figures 8 and 6 that our evidential model significantly outperforms a standard RVM implementation. Quantitatively, the average Euclidean distances between real and estimated 3D location of each marker over the whole arm testing sequence were, in the RVM tests, 31.2, 13.6, 23.0, and 4.5 centimeters for the markers “hand”, “wrist”, “elbow” and “shoulder”, respectively.

Figure 9 shows instead the estimates produced by Gaussian Process Regression for the same experimental setting of Figures 8 and 6. A visual inspection of Figures 9 and 6 shows a rather comparable performance with that of the BMR approach, although the partial models obtained...
from left and right view features only seem to perform relatively poorly. Quantitatively, however, the average Euclidean distances between real and estimated 3D location of each marker over the whole arm testing sequence were, in the GPR tests, 25.0, 10.6, 18.6, and 7.0 centimeters for the markers “hand”, “wrist”, “elbow” and “shoulder”, respectively, showing how our belief-theoretical approach outperforms this competitor as well.

Figure 10 plots the confidence intervals of the estimates produced by Gaussian Process Regression for the same test sequences of Figure 7. A confidence level of 95% (corresponding to an interval of two standard deviations) is used. It should be clear, however, the difference between the confidence band (shown in Figure 10) associated with a single Gaussian distribution on the outputs (poses) (such as the prediction function \( p(q|y, \tilde{Q}, \tilde{y}) \) of a GPR) which is characterized by a single mean estimate and a (co)-variance, and the interval of expected (mean) poses associated with a belief estimate (which amounts to entire family of probability distributions) shown in Figure 7. This is the consequence of the second-order uncertainty encoded by belief functions, as opposed to single classical probability distributions. Indeed, for each vertex of the credal estimate produced by BMR we could also compute (besides an expectation) a covariance and a confidence band: the cumulated confidence bands for all Probability Distribution Functions.
(PDFs) in the credal estimate would be a fairer comparison for the single confidence band depicted in Figure 10, and better illustrate the approach’s robustness.

Fig. 10. Confidence intervals (two standard deviations) associated with the GPR estimates (solid red) and the ground truth (in blue) for the same test sequences of Figure 7.

4) Testing models of different resolutions in the legs experiment: Figure 11 shows instead BMR’s performance in the leg experiment, for a 200-frame-long test sequence. Again, the pignistic transform was adopted to extract a pointwise pose estimate at each time instant. The estimates generated by two models with the same number of feature spaces ($N = 6$) but different number of MoG components per feature space ($n = 5$, red; $n = 4$, magenta) are shown, to analyze the effect of quantization on the model. The results were a bit less impressive but still good, mainly due to the difficulty of automatically segmenting a pair of black pants against a dark background (see Figure 4-right). Again, this cannot be considered an issue of the BMR approach, as annotation is supposed to be given in the training stage. A quantitative assessment returned average estimation errors (for the pignistic expected pose estimate and the model with $n = 5$) of 25.41, 19.29, 21.84, 19.88, 23.00, and 22.71 centimeters, respectively, for the six markers (located on thigh, knee and toe for each leg). Consider that the cameras were located at
a distance of about three meters. No significant differences in accuracy could be observed when reducing the number of MoG components to 4. As in the arm experiment, no significant conflict was reported. In this sense these tests did not allow us to illustrate the ability of the evidential approach to detect foreground features in the case of occlusions or imprecise localization: more challenging tests will need to be run in the near future.

Fig. 11. Performance of two versions of the two-view evidential model with $N = 6$ feature spaces, in the leg experiment, on a test sequence of length 200. The pignistic expected pose is computed for a number of MoG components equal to $n_i = n = 5$ for each feature space (red), and a model with $n_i = n = 4$ (magenta), and plotted versus the ground truth (blue). The estimates for components 4, 7, 9 and 12 of the 18-dimensional pose vector (the 3D coordinates of each of the 6 markers) are shown.

5) When ground truth is not available: visual estimates: when ground truth is not available in the training stage, the pignistic probability $\hat{p}$ on $\hat{Q}$ extracted from the belief estimate $\hat{b}$ can be used to render, given a test image, a visual estimate in terms of the weighted sum of sample images: $\hat{I} = \sum_{k=1,\ldots,T} \hat{p}(q_k) \cdot I(k)$. Figure 12 compares the results of this visual estimate with
the corresponding, real, test image. The accuracy of this visual reconstruction can be easily appreciated. Some fuzzyness is present, due to the fact that the visual estimate is the extrapolation of possibly many sample images, and expresses the degree to which the estimate is inaccurate.

Fig. 12. Left: visual comparison between a real test image of the arm experiment (top) and the corresponding visual reconstruction (bottom) \( \hat{I} = \sum_{k=1, \ldots, T} \hat{p}(q_k) \cdot I(k) \). Right: mean estimation errors for each of the four markers throughout the test sequence delivered by the multiple feature space model versus those produced by a number of single (vectorial) feature space model. When each scalar feature component is considered separately and combined by conjunctive rule (solid red), instead of being piled up in a single observation vector, the performance is significantly superior. The dashed black line corresponds to a single feature space with \( n = 30 \) Gaussian components. Solid black lines are associated with a quantization of \( n = 20, 10, 7, 5 \) and 3, respectively. In magenta the average 3D distance from the interval of expected poses (Section IV-D.2).

6) Conjunctive combination versus vectorization: finally, it is interesting to assess the advantage of combining a number of separate belief functions for each component of the feature vector, rather than piling up all the features in a single observation vector. As a term of comparison, therefore, we applied the same estimation scheme to a single feature space, composed by whole feature vectors, rather than the collection of spaces associated with individual feature components. We applied EM to the set of training feature vectors, with a varying number \( n \) of MoG clusters. Figure 12-right plots the different average estimation errors for the four markers in the arm experiment along the whole testing sequence of length 1000, as produced by the two-view, multiple feature space evidential model versus the single feature space one generated by applying EM to whole feature vectors. The pignistic function was again used to compute the point-wise expected estimate. The solid red line represents the performance of the multiple feature space
model, versus a number of black lines associated with single feature space models with a number of MoG clusters $n$ equal to 3, 5, 7, 10, 20 and 30, respectively. In a way, these tests compare the efficacy of the conjunctive combination of belief functions to that of vectorization as a data fusion mechanism. Not only the former proves to be superior, but the plot suggests that, after a certain threshold, increasing the number of MoG components does not improve estimation performance anymore. Figure 13 visually compares the quality of the estimates for two components (2 and 4) of the pose vector and a 100-frame long sub-sequence of the testing sequence. Even though (in principle) there is no reason why quantizing a single, vectorial feature space should yield poor performances, in practice it is impossible to learn the parameters of a Mixture of Gaussians with a number of states comparable to the product $n_1 \cdot \ldots \cdot n_N$ of the number of clusters of the $N$ separate feature spaces. The EM algorithm is unable to converge: the best we can get to is a few dozen states, a number insufficient to guarantee an adequate estimation accuracy.

![Figure 13](image)

**Fig. 13.** Visual comparison between the estimates yielded by the belief combination of scalar features (solid red line) and those produced by a single (vectorial) feature space with $n = 20$ (dotted black) and $n = 30$ (dashed black) Gaussian components, versus the provided ground truth (solid blue), in the arm experiment. A 100-frame long interval of the testing sequence is considered; results for components 2 and 4 of the pose vector are shown.

V. DISCUSSION

We wish to conclude by discussing the methodological justification of the proposed regression framework, in the light of the problem to solve and in comparison with similar Bayesian approaches, in particular Gaussian Process regression.
A. Justification

In the scenario of Section I-A, any regression method we design needs to represent the feature-to-pose mapping \( y \mapsto q \), which is unknown. Consider first the case of a single feature function.

**Naive interpolation.** The training data \( \{ \tilde{Q}, \tilde{Y} \} \) already provides us with a first, rough approximation of the unknown mapping. A naive regression approach may, for instance, apply to any test feature value \( y \) a simple linear interpolator \( y \mapsto \sum_k w_k q_k \) with coefficients \( w_k \) function of some distance \( d(y, y_k) \) between \( y \) and each training feature \( y_k \). We obtain a one-to-one, piecewise linear map (see Figure 14-left) which, when the training samples are dense in the unknown pose space \( Q \), deliver a decent approximation of the (also unknown) feature-to-pose mapping.

However, such a naive interpolator does not allow us to express any uncertainty due to lack of training information. Also, although the source of ground truth provides a single pose value \( q_k \) for each sample feature value \( y_k \), occlusions, self-occlusions and projection ambiguities mean that each observed feature value \( y \) (including the sample feature values \( y_k \)) can be generated by a continuum \( Q(y) \) of admissible poses. In particular, this is true for the extremely simple bounding box features implemented in Section IV-C.

When presented with a training feature value \( y_k \) during testing, our naive interpolator associates it with the corresponding training pose \( q_k \), which is in fact only one of a continuous set of poses \( Q(y_k) \) that could have generated that particular feature value.

**Intervals of pose estimates in the cluster-refining framework.** The most interesting and realistic situation is that in which the training samples are sparse in \( Q \). Even assuming that the sought map is one-to-one (which is not), any regressor will be uncertain about the value of the pose far from the available samples. What we need, is to be able to: 1- express the uncertainty on the value of the map far from the samples, but also 2- express the fact that to any \( y \) may correspond an entire set \( Q(y) \) of poses. The evidential modeling framework of Section III-A.1 addresses, to some extent, both these questions, as: i) it provides an interval of (rather than pointwise) estimates in the regions of \( Q \) covered by samples (clusters); ii) it describes the uncertainty on the value of the pose far from the samples. Let us see how.

In the former regions, the interval of possible poses is estimated on the basis of the interval of sample pose values in the cluster \( \tilde{Q}_k \): \([q_{\text{min}}, q_{\text{max}}]\), where \( q_{\text{min}} = \min_{q_k \in \tilde{Q}_k} q_k \) and \( q_{\text{max}} = \max_{q_k \in \tilde{Q}_k} q_k \). The rationale is that if close feature values \( y_k \) yield different poses \( q_k \),s, this may
Fig. 14. Left: lower and upper bounds to the pose estimate generated by a single-feature evidential model. We picked component $c = 2$ of the sample poses $q_1$, $q_2$ and $q_4$ of the training sequence of the arm experiment of Section IV, and built a single-feature evidential model using as training feature values $y_1 = 23$, $y_2 = 38$ and $y_4 = 86$ and $n = 2$ EM clusters (corresponding to $\{q_1, q_2\}$ and to $\{q_4\}$). A simple linear interpolator (in green) yields a 1-1, piecewise feature-to-pose map which does a decent job when the samples are dense in $\mathcal{Q}$. Using Bayesian belief functions to encode feature values, the uncertainty on the feature value in each cluster is smoothly propagated to the entire range of feature values, obtaining the solid blue lower and upper bounds. Using Dirichlet belief functions delivers wider, more cautious bounds (dashed blue). Right: the conjunctive combination (5) of multiple features generates a framework with much expressive power in terms of the family of mappings modeled. Here the complex, but still smooth, shape of the lower and upper bounds generated by an evidential model with two feature spaces in the same toy experiment is shown.

signal what we called “inherent” ambiguity (in that region of $\mathcal{Q} \times \mathcal{Y}$ each feature value $y$ may be generated by a wide interval $\mathcal{Q}(y)$ of admissible poses): see zone 1 in Figure 14-left.

Far from the clusters (zone 2), this interval uncertainty is propagated via Equation (10), by assigning a total weight $\sum_k w_k = \Gamma_j(y)/Z$ to the ensemble of samples of each cluster (equivalently, a set of PDFs, a belief function, on the collection $\tilde{\mathcal{Q}}$ of sample poses). Given an interpolator function $\hat{\mathcal{q}} = \mathcal{I}(\{p_k, q_k\}, y)$ mapping a feature value $y$ to a pose vector $\hat{q}$ (given a certain probability distribution $\{p_k\}$ on the training poses $\{q_k\}$), this translates into an interval of possible poses for each test feature $y$.

The overall effect is a “robustification” of the estimates produced by the chosen interpolator function $\mathcal{I}$ (observe Figure 14, zone 2 for the case of the expectation interpolator of Equation (15)). Still, isolated values which form clusters on their own are taken at face value (zone 3, as in GPR): this is an undesirable but unlikely result of EM clustering, if the number of clusters $n$ is much smaller than the number of training poses $T$.

Expressive power in terms of a family of mapping. In a single-feature evidential model, then,
the learned refining does not constitute an approximation of the true feature-pose map under the model, but determines a constraint on the latter associated with a whole family of feature-pose mappings compatible with the given training observations. Such compatible maps are those and only which would generate the learned refinings given the same training data, and form an $\infty$-dimensional family bounded by an upper and a lower admissible feature-to-pose function. We can prove that these lower and upper mappings are smooth, due to the smoothness of the Gaussian likelihoods $\Gamma$. For sake of simplicity we consider a single feature model, with $m(\Theta) = 0$.

**Theorem 2:** Suppose the interpolator function (15) is used to infer a pose estimate $\hat{q}(y)$ from a feature value $y$, given a probability distribution $\{p_k, k = 1, ..., T\}$ on the set of training poses $\hat{Q} = \{q_k, k = 1, ..., T\}$. Then, for each component $q^c$ of the pose vector $q$, both the upper bound $\sup \hat{q}^c(y)$ and the lower bound $\inf \hat{q}^c(y)$ to the pose estimates under a single-feature evidential model for all possible test feature values $y \in Y$ are smooth functions of $y$.

**Proof:** we prove the statement for the upper bound. A dual proof can be easily derived for the lower bound. The former quantity reads as:

$$\sup_{p \in P[\hat{b}(y)]} \sum_{k=1}^{T} p_k(y)q_k^c,$$

where $\hat{b}(y)$ is the belief estimate generated by a test feature value $y$, and $P[\hat{b}(y)]$ is the corresponding credal set. Since we consider a single feature model with $m(\Theta) = 0$, $\hat{b}(y)$ has $n$ focal elements $\hat{Q}^1, ..., \hat{Q}^n$ with mass $m(\hat{Q}^j) = \Gamma^j(y)/Z$, with $Z$ a normalization factor. Each is the image of a EM cluster in the feature space; together they form a disjoint partition of $\hat{Q}$, so that $\sum_{q_k \in \hat{Q}^j} p_k(y) = m(\hat{Q}^j) = \frac{\Gamma^j(y)}{Z}$. Therefore, we can decompose the upper bound as

$$\sup \hat{q}^c(y) = \sup \sum_{q_k \in \hat{Q}^j} p_k(y)q_k^c = \sup \left( \sum_{j=1}^{n} \sum_{q_k \in \hat{Q}^j} p_k(y)q_k^c \right) = \sum_{j=1}^{n} \sup \left( \sum_{q_k \in \hat{Q}^j} p_k(y)q_k^c \right).$$

But $\sup \left( \sum_{q_k \in \hat{Q}^j} p_k(y)q_k^c \right) = \frac{\Gamma^j(y)}{Z} \sup_{q_k \in \hat{Q}^j} q_k^c$, for the sup is obtained by assigning all mass $\frac{\Gamma^j(y)}{Z}$ to the sample with the largest pose component value. The quantity $\sup_{q_k \in \hat{Q}^j} q_k^c$ does not depend on the test feature value $y$, but is a function of the samples in the considered cluster $j$. Therefore,

$$\sup \hat{q}^c(y) = \frac{1}{Z} \sum_{j=1}^{n} \Gamma^j(y) \sup_{q_k \in \hat{Q}^j} q_k^c$$

is a smooth function, the linear combination of the smooth functions $\Gamma^j(y)$ with coefficients $\sup_{q_k \in \hat{Q}^j} q_k^c$. 

January 17, 2014

DRAFT
The lower and upper bounds are depicted as solid blue lines in the example of Figure 14-left. Within those smooth bounds, any one-to-many mapping is admissible, even discontinuous ones: a quite realistic situation, for the actual pose space $Q$ can have holes composed by non-admissible poses, which generate discontinuities in the feature-pose map.

The width of this family of mappings is a function of the number $n$ of EM clusters:

$$\sup \hat{q}(y) - \inf \hat{q}(y) = \frac{1}{Z} \sum_{j=1}^{n} \Gamma_j(y) \left( \sup_{q_k \in \tilde{Q}^j} q_k - \inf_{q_k \in \tilde{Q}^j} q_k \right).$$

A low $n$ amounts to a cautious approach in which training feature values are not "trusted", and the inherent ambiguity (number of training samples in $Q(y)$) is higher. Many clusters ($n \to |\tilde{Q}|$) indicate that we much trust the one-to-one mapping provided by the interpolator over the samples.

**Effect of $m_i(\Theta_i)$ and sparsity of samples.** An additional element in our regression framework is constituted by the mass $m_i(\Theta_i)$ assigned to the whole approximate feature space $\Theta_i$ (Section III-B.1). The effect on the family of admissible maps is to further expand the band of estimates, depending on the number of available training samples (as $m_i(\Theta_i) = 1/T$). In Figure 14-left the expanded upper and lower bounds due to the use of Dirichlet belief functions are depicted as dashed blue lines. It can be appreciated how these are still smooth functions of $y$.

**Choice of an interpolation operator.** The shape of the family of mappings represented by a single-feature evidential model (i.e., of its lower and upper bounds) is also determined by the choice of the interpolation operator $\hat{q} = I(\{p_k, q_k\}, y)$. In Section III-B the interpolator function was the expectation $\hat{q} = \sum_k p_k q_k$, but other choices are of course possible. Different interpolators generate different families of feature-pose mappings.

**Fusion of individual features.** An additional layer of sophistication is introduced by the combination of distinct features via the conjunctive combination of the associated belief functions (Section II-B). This produces a rather complex families of compatible feature-to-pose mappings. Figure 14-right illustrate the shape of the lower and upper bounds generated by an evidential model with $N = 2$ feature spaces. It can be noted that, despite their more complex shape, these bounds are still smooth.

**B. Differences and similarities with Gaussian Process Regression**

It can be interesting to compare the behavior of BMR with that of the classical Gaussian Process Regression (GPR) [27]. The latter assumes that any finite set of observations are drawn
from a multivariate Gaussian distribution. According to [27], a Gaussian process is defined as “a collection of random variables, any finite number of which have (consistent) joint Gaussian distribution”. It is then completely specified by a mean $m(s)$ and a covariance $k(s, s')$ function over the samples’ (observations) domain. It can be seen as a distribution over functions: $\zeta(s) \sim \mathcal{GP}(m(s), k(s, s'))$, where $m(s) = E[\zeta(s)]$, and $k(s, s') = E[(\zeta(s) - m(s))(\zeta(s') - m(s'))]$. If the covariance function depends on a set of hyperparameters, given a training set of noisy observations $\{(s_k = y_k, \zeta_k = q_k)\}_{k=1,...,T}$, and assuming the prediction noise to be Gaussian, we can find the optimal hyperparameters of the Gaussian Process $\mathcal{GP}$ which best fits the data by maximizing the log marginal likelihood (see [27] for details). With the optimal hyperparameters, we obtain a Gaussian prediction distribution in the space of targets (poses):

$$
\mathcal{N}(k(s^*, s)^T[K + \sigma^2_{\text{noise}}I]^{-1}\Psi', k(s^*, s^*) + \sigma^2_{\text{noise}} - k(s^*, s)^T[K + \sigma^2_{\text{noise}}I]^{-1}k(s^*, s)),
$$

where $K$ is the covariance matrix calculated from the training image features $s$ and $\sigma^2_{\text{noise}}$ is the covariance of the Gaussian noise. This is equivalent to having an entire family of regression models, all of which agree with the sample observations.

Both GP and BM Regression model a family of feature-to-pose mappings, albeit of a rather different nature. In Gaussian Process Regression, mappings are one-to-one, and a Gaussian Process amounts to a probability distribution over the set of mappings. The form of the family of mappings actually modeled is determined by the choice of a covariance function, which also determines a number of characteristics of the mappings such as periodicity, continuity, etcetera. After conditioning a Gaussian Process by the training data, we obtain a prediction function (Equation (19)) on $Q$ which comply with Gaussian distributions given a test observation and the trained model parameters. The predicted mean and variance vary according to the test observations. In particular the training samples are assumed correct and trustworthy: as a result, the posterior GP has zero uncertainty there.

In opposition, Belief Modeling Regression produces a random set, an entire convex set of discrete but arbitrary PDFs, but on the set of sample poses $\hat{Q}$, rather than on $Q$. As we have seen, given an interpolation function this random set corresponds to a constrained family of mappings, rather than a distribution over the possible maps as in GPR. The resulting mappings are arbitrary and one-to-many, as long as they meet the upper and lower constraints, or, equivalently, as long as they generate the learned refinings under the training data. The shape of the family of mappings
does depend on the chosen interpolation operator, while its width is a function of the number of clusters \( n \) and the mass of the whole feature space \( \Theta \). A trait of BMR is that uncertainty is present even in correspondence of sample feature values (see above).

Different is the treatment of the uncertainty induced by the scarcity of samples (i.e., far from the samples). In GPR the standard deviation of the prediction function is influenced by both the type of prior GP selected and the distance from the samples. In BMR the width of the interval of pose estimates is influenced by both the number \( n_i \) of EM feature clusters, and the mass \( m_i(\Theta_i) \) Dirichlet belief functions assign to the whole (approximate) feature space.

### C. Different inference mechanisms

Dirichlet belief functions are not the only possible way of representing as a belief function evidence in the form of a set of likelihoods. Another option is to normalize the likelihoods \( (7) \) the MoG generates, obtaining a probability (or Bayesian b.f.) on \( \Theta_i = \{ \gamma^1_i, \cdots, \gamma^m_i \} \):

\[
m_i(\gamma^j_i) = \frac{\Gamma^j_i(y_i)}{\sum_k \Gamma^k_i(y_i)}.
\]  

(20)

Alternatively, the likelihood values can be used to build a consonant belief function, i.e., a b.f. whose focal elements \( A_1 \subset \cdots \subset A_m \) are nested, as in [19]:

\[
b_i(A) = 1 - \max_{j: \gamma^j_i \in A^c} \frac{\Gamma^j_i(y_i)}{\max_j \Gamma^j_i(y_i)}.
\]  

(21)

The three different Bayesian (20), consonant (21), and Dirichlet (12) inference algorithms seem to produce comparable results in terms of pointwise estimates, at least under this experimental setting characterized by low conflict. Significant differences emerge, however, if we investigate the nature of the belief estimate the different inference techniques generate.

In the Bayesian case, as the belief functions on the individual feature spaces \( \Theta_i \) have disjoint (singleton) focal elements, their projection onto \( \tilde{Q} \) also has disjoint focal elements. The conjunctive combination of all such b.f.s yields again a belief estimate \( \hat{b} \) whose focal elements are disjoint (Figure 15-left). This means that a region of the pose space is supported by \( \hat{b} \) only to the extent by which it is supported by \textit{all} the individual features. If the belief functions built on the feature spaces are Dirichlet, their projections onto \( \tilde{Q} \) all have the whole \( \tilde{Q} \) as a focal element. Therefore, their conjunctive combination (5) will have as f.e.s not only all the intersections of the form \( \rho_1(A_1) \cap \cdots \cap \rho_N(A_N) \) for all possible selections of a single focal element \( A_i \) for each
measurement function \( b_i \), but also all the intersections \( \rho_{i_1}(A_{i_1}) \cap \cdots \cap \rho_{i_m}(A_{i_m}) \) (where \( i_1, \ldots, i_m \) is any subset of features), and the whole approximate pose space \( \tilde{Q} \) (Figure 15-middle). This is equivalent to say that all the partial combinations of feature evidence are given some credit: maybe, this is the reasoning, only a subset of features is telling the truth. When there exists conflict among different feature models, this amounts to a cautious approach in which the most consensual group of features is given support, the more so whenever the remaining features are highly discounted as less reliable (\( m_i(\Theta_i) \) is high). In the consonant case, the conjunctive combination of single-feature belief functions yields a belief estimate \( \hat{b} \) whose focal elements also form chain of nested sets of poses: one can say that the belief estimate is “multi-modal”, with a focus on a few regions of the (approximate) pose space (Figure 15-right).

It is interesting to compare these three approaches by looking at the associated credal sets as well. The size of the credal set represented by a belief estimate is a function of two distinct sources of uncertainty: that associated with the belief function \( b_i \) we build on each approximate feature space \( \Theta_i \), and the multi-valued mapping from features to poses. Even when the former is a probability (Bayesian case), the multi-valued mapping still induces a belief function on \( \tilde{Q} \). Consider a restricted, toy model obtained from just the first four training poses \( q_1, q_2, q_3 \) and \( q_4 \) in the arm experiment. This way, \( \tilde{Q} \) has size 4, and the probability simplex there (see Figure 3) is 3-dimensional and can be visualized. Figure 16 depicts the credal sets generated by the three inference mechanisms (under the restricted model) in correspondence of frame 13 of the test feature sequence of the arm experiment. One can note how the credal set in the Bayesian case is
Fig. 16. Left: belief estimate, represented by a credal set in the simplex of all probability distributions on $\tilde{Q}$, generated by using Bayesian measurement belief functions for frame $k = 13$ of the arm experiment, for a model learned from the first 4 sample poses only. Middle: belief estimate for the same frame, generated using consonant measurement functions. Right: belief estimate produced via Dirichlet measurement functions.

the narrower (in fact in this example it reduces to a single point, although not in general), while it is the widest in the Dirichlet case. The latter amounts therefore to a more cautious approach to estimation yielding a wider uncertainty band around the central expected value (which is the one we adopt in the tests, Figure 7). In the former case, instead, all the uncertainty in the belief estimate comes from the multi-valued nature of refining maps.

While the size and shape of the credal set varies, the pignistic probability (magenta star) is pretty close in the Bayesian and Dirichlet cases: the empirical evidence seems to suggest that, when conflict is limited, pointwise estimates in the three cases are fairly close, while differing in the attached degree of uncertainty.

D. Towards evidential tracking

To conclude, we outline feasible options for extending the proposed belief-theoretical approach to fully fledged tracking, in which the temporal information provided by the time series of feature is exploited to help the estimation of the pose. Suppose you have a belief estimate $\hat{b}(t)$ of the pose
at time $t$, and a fresh set of features at time $t+1$. The simplest way of ensuring the temporal consistence of the estimated pose is to combine the current estimate $\hat{b}(t)$ with the evidence provided by the new feature values. Namely, the latter will induce a belief estimate $\hat{b}(t+1)$ via the algorithm of Section III-B.6; this has then to be combined with the old estimate by conjunctive combination, yielding an overall, “smoothed” version of the estimate: $\hat{b}(t) \land \hat{b}(t+1)$.

This approach, however, can easily lead to a drifting of the estimates, as no motion model whatsoever is employed. In addition, it can be argued that in this way features at time $t$ condition the estimates at time $t+1$ just as feature at time $t+1$ do, which is wrong in principle.

The use of a motion model encoding the dynamics of the object to track is more sensible: however, if this model were to be a-priori we would violate the assumptions of the example-based scenario. The way to go is learning a motion model from the training set, in the same way as we learn feature-pose maps from it. Assuming that the temporal dependency satisfies a Markovian-like condition, i.e., that the pose at time $t+1$ only depends on the pose at time $t$, the following framework can be formulated.

**Learning a motion model from the training set.** Consider a frame of discernment $\tilde{Q} = \tilde{Q} \times \tilde{Q}$ whose elements $(q_k, q'_k)$ can be interpreted as transitions $q_k \rightarrow q'_k$ from sample pose $q_k$ to sample pose $q'_k$. This frame is trivially partitioned into $T$ disjoint subsets $\tilde{Q} = \tilde{Q}_1 \cup \cdots \cup \tilde{Q}_T$, each of them $\tilde{Q}_k$ associated with a sample pose $q_k$, and collecting all possible $q_k \rightarrow q'_k$ transitions originating from $q_k$. We can then mine the information carried by the training set, and infer for each element of this partition a belief function with the following mass assignment: $m_k(q_k \rightarrow q'_k) = (1 - \epsilon) \cdot \frac{\# \text{transitions from } k \text{ to } k'}{\# \text{times } q_k \text{ appears in } \tilde{Q}}$, $m_k(\tilde{Q}_k) = \epsilon$. The discounting factor $\epsilon$ is a measure of how well the motion model learned from the training set approximates the true, unknown model of the object’s dynamics: in the ideal case $m_k(\tilde{Q}_k) = 0$. As this is achieved only by collecting an infinite number of samples, $\epsilon = \frac{1}{\# \text{times } q_k \text{ appears in } \tilde{Q}}$ is a reasonable albeit imperfect choice for such a factor. The training set also provides a-priori information on the sample poses themselves (i.e., on the elements $\tilde{Q}_k$ of the above disjoint partition of $\tilde{Q}$), which can be encoded as a probability distribution on the partition itself: $m_0(\tilde{Q}_k) = \frac{\# \text{times } q_k \text{ appears in } \tilde{Q}}{T}$.

**The total belief theorem.** Given the above belief functions with mass $m_1, \ldots, m_T$ defined on the individual elements $\tilde{Q}_1, \ldots, \tilde{Q}_T$ of the partition, and the a-priori distribution $m_0$ on the latter, we need to obtain a single belief function on the transition frame $\tilde{Q}$. This amounts to solving the total belief theorem [51], formulated as follows. Consider a set $\Theta$ and a disjoint partition $\Omega$. 

January 17, 2014

DRAFT
Fig. 17. Left: the total belief problem is the generalization of the classical total probability theorem to belief functions, as finite random sets. A total belief function \( b \) on \( \Theta \) is sought whose restriction to a partition \( \Omega \) of \( \Theta \) is equal to another (a-priori) b.f. \( b_0 \), and its conditional version with respect to each element \( \Theta_i \) of the partition \( \Omega \) coincides with a given (conditional) belief function \( b_i \). Right: diagram of the proposed evidential tracking process.

of \( \Theta \). Suppose a (conditional) belief function is defined on each element of the partition, and a (a-priori) b.f. is given on the partition \( \Omega \) itself. We seek a total belief function on the whole of \( \Theta \) whose restriction to \( \Omega \) coincides with the a-priori, and whose conditional versions (in our case under conjunctive combination) coincide with the given ones for all the elements \( \Theta_i \) of the partition \( \Omega \) of \( \Theta \) (see Figure 17-left).

**Tracking process.** Once the total belief function on \( \tilde{Q} \) which represents the learned motion model is obtained, it can be combined with the current pose estimate \( \hat{q}(t) \), which is defined on \( \tilde{Q} \), on the joint estimation space \( \tilde{Q} \times \tilde{Q} \), the Cartesian product of the two (Figure 17-right). The resulting belief function can be re-projected back to the approximate pose space \( \tilde{Q} \), where it represents the predicted pose given the pose at time \( t \) and the learned motion model. Finally, the latter is combined with the belief function resulting from the current feature measurements at time \( t + 1 \), to yield a belief estimate of the pose \( \hat{q}(t + 1) \) which incorporates both the current feature evidence and the predictions based on the motion model learned in the training stage.

We will thoroughly study and implement this natural extension to object tracking of the presented belief-theoretical approach to pose estimation in the near future.
VI. CONCLUSIONS

In this paper we presented a novel approach to example-based pose estimation, in which the available evidence comes in the form of a training set of images containing sample poses of an unspecified object, whose location within those images is provided. Ground truth is available in the training stage in the form of the configurations of these sample poses. An evidential model of the object is learned from the training data, under weak likelihood models built separately for each feature, and is exploited to estimate the pose of the object in any test image. Framing the problem within belief calculus is natural as feature-pose maps induce belief functions in the pose space, and it allows to exploit the available, limited evidence without additional assumptions, with the goal of producing the most sensible possible estimate with an attached degree of reliability. The approach was tested in a fairly challenging human pose recovery setup and shown to outperform popular competitors, demonstrating its potential even in the presence of poor feature representations. These results open a number of interesting directions: a proper empirical testing of object localization algorithms in conjunction with the proposed Belief Modeling Regression approach; an efficient conflict resolution mechanism able to discriminate as much as possible foreground from background features; the testing of the framework in more higher-dimensional pose ranges; the full development of the outlined evidential tracking approach.

REFERENCES

[41] A. Jsang and S. Pope, “Normalising the consensus operator for belief fusion,” *AIJCAI’06*.


