Appearing in DICTA 2012 (Digital Image Computing: Techniques and Applications).

Generalized Exemplar-Based Full Pose Estimation from 2D Images without Correspondences

Damien Teney
University of Liège, Belgium
Damien.Teney@ulg.ac.be

Justus Piater
University of Innsbruck, Austria
Justus.Piater@uibk.ac.at

***Abstract*—This paper addresses the problem of full pose estimation of objects in 2D images, using registered 2D examples as training data. We present a general formulation of the problem, which departs from traditional approaches by not focusing on one specific type of image features. The proposed algorithm avoids relying on specific model-to-scene correspondences, allowing using similar-looking and generally unmatchable features. We effectively demonstrate this capability by applying the method to edge segments. Our algorithm uses successive histogram-based and probabilistic evaluations, which ultimately recover a complete description of the probability distribution of the pose of the object, in the 6 degree-of-freedom 3D pose space, thereby accounting for the inherent ambiguities in the 2D input data. Furthermore, we propose, in a rigorous framework, an efficient procedure for fusing multiple sources of evidence, such as multiple registered 2D views of the same scene. The proposed method is evaluated qualitatively and quantitatively on synthetic and real test images. It shows promising results under challenging conditions, including occlusions and heavy clutter, while being capable of handling objects with little texture and detail.**

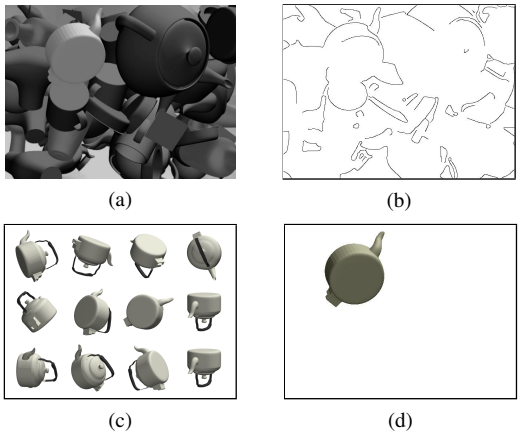


Fig. 1. Pose estimation in a single image, using 2D training examples; (a) test image; (b) edge map used as input; (c) sample training views; (d) rendering of a model of the object in the best pose found by the algorithm, note that the correct pose is recovered despite heavy clutter and missing observations.

I. Introduction and related work

Estimating the pose of a known object in a single 2D image is a fundamental problem in computer vision that has attracted a lot of attention over the years. The task is closely related to the problem of object recognition. However, state-of-the-art object recognition methods usually aims at identifying object *classes*, allowing small variability in appearance among different objects of the same class. We rather focus here on specific *instances* of objects, where such small changes in appearance are actually used as cues for determining the precise pose (3D position and orientation) of the object in a new scene.

The pose estimation task has many direct applications, such as robotic interaction and grasping, augmented reality, or the visual tracking of objects. Methods have been developed that make use of a 3D, explicit geometric model of the object of interest [1], [2], [3]. Those thus require precise a-priori knowledge of the 3D shape of the object, to be provided by external methods such as stereo vision or range sensors. In this paper, we rather present a 2D view-based, or exemplar-based method, which simply uses 2D views of the object as training data, in which the object appears in known poses. Those methods present the advantage of being easily trainable, *directly* using 2D visual data. Further motivation for the exemplar-based approach is brought by the human visual system, which was shown to exhibit properties of a view-based lookup function when recognizing objects, being robust to changes of about 20° around trained viewpoints [4]. Unfortunately, current, state-of-the-art methods following this approach present serious limitations, often relying on specific types of images features, or being suited to only particular types of objects, and are thus able to operate only under limited ranges of conditions. This led us to the reformulation of the problem in more general, probabilistic terms, and to the development of a novel method, that we will introduce after reviewing related work.

Early work in the field of exemplar-based methods used the appearance of the object as a whole. These so-called *appearance-based* methods [5], [6], [7] generally assumed a successful prior detection of the object in the test image and generally offered poor resistance to clutter and occlusions, or did not handle the full 6 degree-of-freedom pose space as needed in practical applications. More recent work, by contrast, focused on the use of individual, precisely located observations (such as *SIFT* features [8]) extracted in the 2D views of the object. These *feature-based* methods [9], [10] then rely on establishing matches, using their appearance, between observations in the test view and in the stored training examples. The limitations of this approach are obviously those of the extraction and matching of image features, which practically works best on texture-rich images, but can perform

poorly on scenes with mostly homogeneous surfaces or little detail.

The method proposed in this paper bridges a gap between the two approaches mentioned in the previous paragraph. It makes use of individual features extracted from the images, thereby offering the potential robustness of feature-based methods, e.g. against lighting changes, but does not rely on the matching of specific observations between the test and training views. Practically, this allows using similar-looking types of features. Although the method is generally applicable to different types of observations, we chose to demonstrate this key ability through the use of local edge segments. These correspond to points extracted in the images along the lines of maximum gradient, and they thus carry little appearance information individually. The result of our implementation is a pose estimation method readily trainable with 2D visual data, intrinsically robust to clutter and occlusions, and able to handle previously-problematic objects with little texture and detail.

The identification of the object of interest in a new image ressembles the traditional problem of object recognition and localization. A number of successful methods have been developed that specifically make use of edges as image features. The classical measures of chamfer distance [11] and Hausdorff distance [12] evaluate the fit of a template over a test image; their initial formulations were refined in different ways to provide practical algorithms capable of finding such a template (a training image of the object of interest) in a cluttered scene [13], [14]. One key addition proved to be the use of the orientation of the edges, as we also do in the proposed method. Other state-of-the-art methods include the work of Ferrari *et al*. In [15], they use descriptors of simple edge groupings to train an SVM classifier, capable of recognizing object classes, then using a traditionnal sliding window over the test image. In [16], they focus on the learning of shape models from unsegmented training views, and then use a soft-matching procedure of those shapes to recognize objects in new images. The purpose of those two methods is however to specifically handle intra-class variations of appearance. The work presented in this paper differs from the cited methods in 3 important ways: (i) we present a generally-applicable method not bound to one specific type of image features; it offers the flexiblity to use additional characteristics (e.g. edge curvature) or other features (e.g. interest points); (ii) we do not seek to identify objects or object classes, but rather to determine their pose, using the small changes of appearance as clues to this end; (iii) we go beyond a simple localization in the image (e.g. as 2D bounding box), as we directly consider the full 6-degree-of-freedom pose of the object in the 3D space, of which we recover a probability distribution, and not a single maximum.

The method proposed in this paper is based on a probabilistic representation of both the test and the training data. Such a representation has been used in the slightly different context of pose estimation using 3D models and observations [17], [3], and this work can be seen as their extension to the case of 2D data. In addition to modelling the uncertainty inherent in the input data, the probabilistic approach leads to the definition of the pose of the observed object as a probability distribution in the 3D pose space, of which we want to identify the peaks. This is justified by the uncertainty in the pose estimation problem arising from the 3D-to-2D projection ambiguities. Intuitively, a given 2D view may often be explained by several 3D poses of the object of interest, and we are generally interested in recovering *all* these potentially correct results. Our probabilistic approach, as will be demonstrated, is able to address this objective. Another contribution of this paper is the introduction of successive histogram- and probabilistic-based evaluations that seek to identify *all* significant modes in the distribution of interest. The aforementioned references, which had to deal with less complex distributions, employed approximations such as Monte Carlo methods [18], which generally recovered only a unique solution. This would have been insufficient in the present case, due to the particular ambiguities mentioned above.

Finally, we propose an efficient method for fusing multiple sources of evidence in the same probabilistic framework. This information may be available e.g. through multiple 2D views of the scene, observed under different viewpoints, but the same principle can also serve to jointly handle multiple types of features extracted from a same image. Viksten *et al.* [10] proposed another method for combining such multiple sources of information through a simple clustering step on top of several instances of existing methods. This however lacks the genericity offered by the rigorous approach proposed here. We make full use of the probabilistic nature of the problem, combining the different sources of information in a Markov random field, on which inference is performed using non-parametric belief propagation. The power of the technique is demonstrated through the use of two 2D views of the same scene, thereby increasing the accuracy of the pose estimation process. A comparable approach was used by Toshev *et al.* [19] for tracking of the pose of an object over time in a video. Other methods for handling multiple views with a 2D pose estimation method have been proposed [1], [20], but with the underlying process based on feature matching, as opposed to the more generic approach proposed here.

II. Probabilistic representation of pose and appearance

In this section, we introduce a rigorous formulation of the pose estimation problem, using a probabilistic representation of the input data. As mentioned above, the proposed method is not specific to one particular type of image features, but the general formulation is illustrated with local edge segments. Those correspond to points extracted from the images along the lines of maximum gradient (see Section V).

A. Representation of test data

Let us first consider the test data, which consists of a single 2D image, from which we extract features x_i. They form the set of *observations* $\mathcal{O} = \{x_i\}_{i=1}^{N}$, where $x_i \in \mathcal{A}$, the space on which is defined the appearance of our observations. In the

case of local edge segments, an observation is characterized by its 2D position in the image, and by its orientation (without direction, i.e. an element on the semicircle). Therefore, we have $\mathcal{A} = \mathbb{R}^2 \times S_1^+$. Considering another case where each observation would be a texture patch extracted around an interest point, the appearance space $\mathcal{A}$ would then encompass the position of that point, and a description of the texture itself.

As proposed in [3], such a set of observations can be used to define a continuous probability density ϕ on $\mathcal{A}$. This distribution is defined in a non-parametric fashion, using Kernel Density Estimation (KDE), directly using the elements of $\mathcal{O}$ as supporting particles. The probability density function of ϕ is then given by

$$\phi(x) \;=\; \frac{1}{N} \sum_{x_i \in \mathcal{O}} K_1(x_i, x) \;, \tag{1}$$

where $x \in \mathcal{A}$, and $K_1(\cdot,\cdot)$ a kernel function on $\mathcal{A}$. This formulation allows modelling the uncertainty that may be present in the observations, e.g. due to image noise or to other artifacts occurring during image formation and processing. The kernels used will depend on the appearance space considered [3]. In our application, using edge segments, we found that using kernels allowing only a small deviation on the position and on the orientation was sufficient, as our edge detection algorithm could provide results of good accuracy (see Section V). In practice, the narrow bandwidth of the chosen kernels implies that sampling from $\phi(x)$ amounts to selecting random points x_i from $\mathcal{O}$, with only small variations (see Fig. 2b).

B. Representation of training data

The training data is composed of a number of pre-segmented 2D images, in which the object of interest appears in known poses. Each of those images is processed, in a similar way as the test image, to extract image features. Each observation x_i is then associated with the pose w_i of the image it was extracted from, thereby forming a set of *pose/appearance pairs* $\mathcal{T} = \{(w_i, x_i)\}_{i=1}^{M}$, where $x_i \in \mathcal{A}$, the appearance space of our observations, and $w_i \in \mathrm{SE}(3)$, the space of 3D poses. Similarly to the observations, these points are used to support a KDE, therefore defining a probability distribution on the joint pose/appearance space. This distribution, called ψ, represents the probability of observing an image feature of a given appearance when the object is in a given pose. Formally, ψ is defined by its density function

$$\psi(w, x) \;=\; \frac{1}{N} \sum_{(w_i, x_i) \in \mathcal{T}} K_2\big((w_i, x_i),\, (w, x)\big) \;, \tag{2}$$

where $w \in \mathrm{SE}(3)$, $x \in \mathcal{A}$ and $K_2(\cdot,\cdot)$ is a kernel function on $\mathrm{SE}(3) \times \mathcal{A}$. The use of kernels on the training data can be seen here as a smoothing over the available training points, effectively yielding a continuous distribution and allowing us to interpolate, to some extent, the value of ψ over regions not covered by the training data. Practical details on the use of kernels in $\mathrm{SE}(3)$ are discussed e.g. by Detry and Piater [18].

In addition to the training data, a number of possible transformations in the pose/appearance space are usually known. For example, under orthographic projection[1], the camera intrinsic parameters dictate how a translation (in pose space) parallel to the camera image plane relates to a translation of the observations in the image (in appearance space). In our case, with edge segments, we chose to hard-code three such transformations, namely the translation and rotation in the image plane, and the change of depth along the camera projection rays which give identical projections on the image plane. Formally, we represent these transformations via a single function f, parameterized by a vector of parameters $p \in \mathcal{P}$, such that

$$f\big((w, x),\; p\big) \;=\; (w', x') \tag{3}$$

with (w, x) and (w', x') being pose/appearance pairs, equivalent through the hard-coded transformations under the parameters p. Those transformations allow us to extend our definition of ψ to larger regions of the pose/appearance space than with the training points alone. To that effect, we substitute $\mathcal{T}'$ for $\mathcal{T}$ in Eq. 2, where

$$\mathcal{T}' \;=\; \mathcal{T} \;\cup\; \{\quad (w', x') : \exists\, (w, x) \in \mathcal{T},\; p \in \mathcal{P} : \\ f\big((w, x),\; p\big) = (w', x') \qquad \}. \tag{4}$$

This *augmented* training set $\mathcal{T}'$ complements $\mathcal{T}$ with all transformations of its elements that can be obtained using f. As we will see in Section III however, our implementation does not require an explicit representation of $\mathcal{T}'$, and, in practice, only a small subset of its elements will have to be identified.

For practical purposes, we remark that the definition of ψ (Eq. 2) presents the problem of making its value dependent on the density of training examples in the corresponding region. For example, including two identical views of the object in the training data, in the same pose, would simply double the density of ψ in the corresponding regions, which is not desirable. This effect is alleviated by using the maximum value of the neighbouring kernels (see Fig. 2c) instead of a summation over their values. This leads to the alternative definition

$$\psi(w, x) \;=\; \frac{1}{C} \max_{(w_i, x_i) \in \mathcal{T}'} K_2\big((w_i, x_i),\, (w, x)\big) \;, \tag{5}$$

where C is a normalization constant.

C. Probability distribution of 3D pose

The probabilistic representations of the test and the training data, given respectively as ϕ and ψ, are now used together to model the pose of the object in the test image. The pose is modelled as a random variable $W \in \mathrm{SE}(3)$, and its distribution is simply given by

$$p(w) \;=\; \int_{\mathcal{A}} \psi(w, x)\; \phi(x)\; \mathrm{d}x \;. \tag{6}$$

[1]Our implementation of the method assumes an orthographic or near-orthographic projection, which in practice is easily satisfied with a camera of sufficient focal length relative to the scene depth (see Section V).

This expression, in effect, measures the compatibility of a pose w with the whole distribution of features observed in the image. Another interpretation is to see it as the cross-correlation of the distribution ϕ of observations in the test image with the distribution $\psi(w,\cdot)$ of training points at a given pose. Note that this formulation of $p(w)$ is similar to that proposed in [18], [3] for the use of 3D models and observations.

III. Pose inference

This section presents a practical method for solving the pose estimation problem as formulated in Section II. The method is based on two key observations, presented below, which allow an approximate evaluation of $p(w)$.

First, the value of the integral in Eq. 6 can be approximated using Monte Carlo integration [21], [18]. This method, which involves a random exploration of the integration domain, gives

$$p(w) \;\approx\; \frac{1}{n}\sum_{i}^{n}\psi(w,x_i) \quad \text{where} \quad x_i \sim \phi(x)\;. \qquad (7)$$

The evaluation of $p(w)$ (see Fig. 2a–d) thus amounts to successive evaluations of $\psi(w,x_i)$ for different values of x_i, drawn from the distribution of observations in the test image (ϕ).

Importantly, and this is our second key observation, each of these evaluations of $\psi(w,x_i)$ only requires a small number of elements of the *augmented* training set $\mathcal{T}'$. For a fixed x_i, using the hard-coded transformations (in-plane rotation and translation), any original training pair $(w,x) \in \mathcal{T}$ can be transformed into a pair $(w',x_i) \in \mathcal{T}'$ of appearance x_i. Those pairs will have the strongest influence on the value of $\psi(w,x_i)$ (Eq. 5), and its evaluation can therefore be limited in practice to the use of those pairs, which formally correspond to the following subset of $\mathcal{T}'$:

$$\{\quad (w',x_i) : \exists\,(w,x)\in\mathcal{T},\; p\in\mathcal{P} : \\ f\big((w,x),\,p\big) = (w',x_i) \qquad \} \;\subset\; \mathcal{T}' \qquad (8)$$

The practical consequence of this property is that an explicit and complete representation of $\mathcal{T}'$ is not required, and that only a fraction of its elements have to be identified.

A. Exhaustive search algorithm

The two properties we just presented make the evaluation of $p(w)$ possible for any pose w. Various methods can then in principle be used to identify the main modes of this distribution, such as a Monte Carlo-type search as proposed in [18], [3]. However, the purpose of such methods is generally to identify the global maximum of the density. As argued above, the particular ambiguities in the 2D input data are likely to induce a very complex distribution, potentially presenting multiple weak modes that we wish to identify. We therefore devised an algorithm to exhaustively explore the relevant parts of the 3D pose space. This task is particularly challenging [22] due to the high dimensionality of SE(3). We propose a two-stage process that first relies on a histogram-based approximation, in order to pre-select regions of interest in SE(3). This serves to discard those bins of the histogram that correspond to areas of low density, dramatically reducing the amount of data used at the second stage. It is then possible to perform a full-scale kernel-based evaluation of the density (Eq. 5,7), limited to the pre-selected regions of the pose space. The algorithm returns a set of poses $\mathcal{R}$ where the density exceeds a certain threshold.

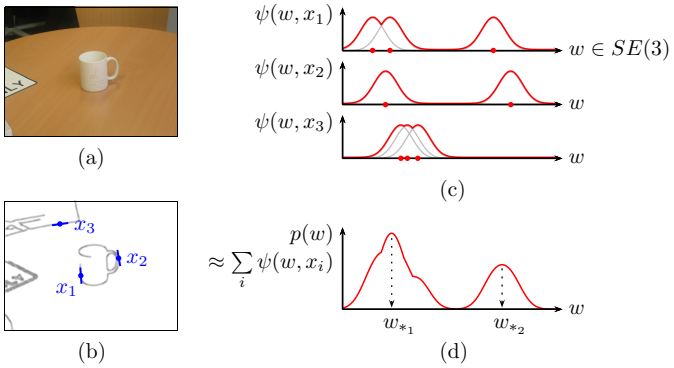


Fig. 2. Proposed method for pose estimation, using edge segments as image features. (a) Test image; (b) distribution of edges in test image, denoted $\phi(x)$ in the text, and three samples x_{1-3} (*blue oriented points*) of that distribution; (c) distribution (*red curve*) of poses compatible with each observation x_i (Eq. 5), made up of individual kernels (*gray curves*) supported by a small subset of poses $w' \in \mathcal{T}'$ (Eq. 8, *red dots*); (d) distribution of poses compatible with all observations x_i (Eq. 7) and local maxima w_{*_i}, as recovered by our method.

The computational complexity of this algorithm is proportional to the number of training points (M, Section II-B), multiplied by the number of samples used from the observations (n, in Eq. 7), itself chosen as a fraction of the total number of observations (N, Section II-A). Note also that it is not mandatory to process all possible combinations of observation samples and training points, but a stochastic approach can rather make use of the probabilistic representation of the training data ϕ, and use a limited number of random samples thereof. This scheme was previously used in the related problem of 3D models and observations [17], [3].

B. Post processing of pose estimates

As a post-processing step, one may want to identify the actual peaks of each mode. This could be accomplished by a traditional gradient-ascent method, such as mean shift [23]. In our case, this procedure would be costly due to the complexity of the pose space. Fortunately, in practice, the proposed algorithm usually returns poses in the close neighbourhood of the actual peaks. A simple *non-maximum suppression* step therefore proves sufficient. In this method, an element is discarded if it lies in the close neighbourhood of an element of greater density, the neighbourhood being defined by a fixed radius in the pose space. This procedure, efficiently implemented by processing the poses of $\mathcal{R}$ in order of decreasing density, therefore selects the poses that are the closest to the peaks of the distribution (Fig. 2d).

IV. Extension to multiple sources of evidence

The method presented above uses a single source of information as input data, i.e. a single 2D image, to evaluate the most probable poses of the object. However, it is sometimes desirable to use several sources of information to disambiguate the result, or make it more precise. Such extra information could be available, e.g. as multiple images of the same scene, observed under different viewpoints, or as several types of image features, extracted from one same image. This section proposes a rigorous method for fusing the results produced by each different cue, thereby determining globally consistent poses. The method is presented in the concrete context of multiple views, but it directly extends to other scenarios, e.g. with multiple types of image features.

We represent by the random variable $W \in \mathrm{SE}(3)$ the pose of the object, and by $X_i \in \mathcal{A}$ the distribution of observations in the ith view. The dependency between these random variables can be represented by a pairwise Markov random field [24], [17], organized in a tree structure, W being the root node (see Fig. 3). The *compatibility potential functions* parameterizing the relationship between W and a X_i are called ψ_i. These are identical to the ψ introduced in Section II, apart from now taking into account the actual viewpoint of the corresponding view. Each node X_i is moreover connected to its corresponding observed variable, Y_i , their relationship being parameterized by ϕ_i, defined similarly to the ϕ of Section II. To determine the marginal density of the top node W, inference on such a graphical model can be performed using Non-parametric Belief Propagation (NBP), as proposed in [24]. The application of the NBP algorithm on a model as simple as that considered here allows many simplifications. In particular, the distribution of W is simply given by

$$p(w) \;=\; \prod_{i=1}^{q} m_i(w) \;, \tag{9}$$

with a *message* $m_i(w)$, conceptually sent from a node X_i to the root node W (see Fig. 3), and expressing its *belief* about the state of W, being defined as

$$m_i(w) \;=\; \int_{\mathcal{A}} \psi_i(w,x)\,\phi_i(x)\,\mathrm{d}x \;. \tag{10}$$

Note that this definition of $m_i(w)$ is identical to Eq. 6, but is now indexed on the source of the observations. Practically, each $m_i(w)$ can be independently evaluated, using the method of Section II-C. This method returns a set of poses in the most dense regions of SE(3), which can directly be used to represent the distribution $m_i(w)$ in a non-parametric fashion, using KDE, weighting each of them with its evaluated probability density. Fusing the results from all sources of evidence, via Eq. 9, then amounts to computing the product, or intersection, of all of these non-parametric representations of densities on SE(3). In practice, the representation of each $m_i(w)$ is usually quite compact, and the evaluation of $p(w)$ for a given w can thus be performed at a reasonable computational cost. We therefore identify the maxima of $p(w)$ with a Markov Chain Monte Carlo (MCMC) type search, using a simple random walk scheme [18]. This local optimization process is performed from several different starting points, selected from the supporting particles of the $m_i(w)$. Using this process, the output of the algorithm is finally the set of poses corresponding to the local maxima of $p(w)$ as defined by Eq. 9, i.e. the poses globally the most consistent with all available sources of evidence.

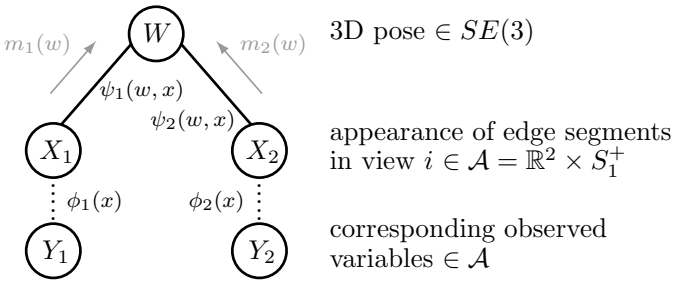


Fig. 3. Markov tree representing the integration of multiple source of evidence, in this case 2 views from which edge segments are extracted. The messages $m_i(w)$ represent the belief about the state of W sent from each node X_i; they are fused to determine the values of W globally consistent with the two views.

V. Evaluation

The evaluation of a pose estimation method is not trivial, due to the difficulty of obtaining the ground truth 3D poses themselves, especially in realistic scenes. We considered using various existing datasets, as reviewed below, but finally decided to produce new datasets, with synthetic images and thus known ground truth, which allowed performing a rigorous quantitative evaluation. Practically, the image features were extracted using the well-known Canny edge detector, followed by a smoothing and subsampling step to reduce the noise in the observations (Fig. 1b). All images used were 640×480 pixel grayscale images, and all the parameters of the algorithm were set to identical values for all the tests (with both synthetic and real images).

Among candidates public datasets, we considered the *ETHZ Shape dataset* [15], which features shape-based object classes in various cluttered scenes. It is however specifically targeted at class recognition algorithms, designed to handle variations in shape, as opposed to our method, which actually uses those slight changes in appearance as clues for estimating the 3D pose of the object. The dataset does not include any suitable training data or any ground truth for 3D pose. The *NORB dataset* [25] is made up of images of toy objects in different poses, and of artificial compositions of such images proposed as cluttered scenes. In addition to being evaluated only with class recognition methods (as far as we are aware), the very-low-resolution images prevent any reliable use of edge features, as our method requires. The *RGB-D dataset* [26] is made up of household objects on a turntable, viewed at 3 different elevations, thus in a fairly limited range of poses. We also argue that the basic evaluation methodology proposed for those sequences, which is basically to use every

other image for training and test alternatively, in the absence of clutter and object translations, is overly simplistic and of limited diagnostic value. The capture setup (e.g. constant-speed turntable) is also acknowledgedly imprecise and ruled out this dataset as an interesting candidate for a rigorous evaluation.

A. Quantitative evaluation on synthetic images

The synthetic datasets were produced with manually designed 3D models and rendered with ray tracing software. The training examples (Fig. 1c) correspond to different views of the object of interest on a uniform background; the poses of the object in the training set are chosen uniformly in the orientation space. The amount of clutter in a test image is measured as the ratio of the number of observations *not* belonging to the object of interest over the total number of observations in the image. For example, a clutter ratio of 0% corresponds to absence of clutter, whereas a clutter ratio of 80% means that about $4/5$ of the observations are actually noise. We measure the success rate as the ratio of experiments that returned a *correct* pose in the first k results (the algorithm returns a list of poses sorted by decreasing probability density). This aspect is important, as the ambiguities the 2D input data often prevent one from distinguishing between different 3D poses that have very similar appearances on the image plane. The threshold for considering a pose as *correct* was set in accordance to the typical dimensions of a scene: considering our objects are of a size of 100–200 mm and distant from the camera of 1000–2000 mm, this threshold was set to a translation error of 20 mm parallel to the image plane (XY), 100 mm in depth (Z), and a maximum rotation error of $20°$. The greater tolerance on the Z translation is justified by the fact that the use of a single 2D image makes the determination of depth very difficult. Note however that this error threshold remains a small fraction of the actual depth of the scenes. Using these conventions, the success rates of the algorithm for various conditions are reported in Fig. 5a. Please also note that relaxing the threshold discussed above does not necessarily lead to better quantitative results, as we also report, in Fig. 5c, the mean error of the first *correct* result returned by each run of the algorithm. The reported average numbers were computed over 30 runs of the algorithm for each of the 6 objects considered (Fig. 5b), each scene being generated at random, with clutter made up of different objects disposed randomly in the background. The measure of the error in orientation for the cylindrically symmetric objects (e.g. the bottle) naturally takes only their relevant degrees of freedom into account.

Systematic test cases including occlusions are hard to design, as the amount of occlusion is difficult to quantify: masking one half or the other of an object can have dramatically different effects due to different levels of detail. We are however confident in the ability of the system to cope with significant occlusion, since this is actually simulated by a common large fraction of missing observations (Fig. 4), due to background clutter preventing a good extraction of edges.

The algorithm presents very good success rates under common amounts of clutter (Fig. 5a). This success rate even remains acceptable as the amount clutter is raised to very challenging values (Fig. 4). Increasing the number of training views for each object was not found to have a significant impact on the success rate, but increased the accuracy of the results (Fig. 5c). Similarly, the amount of clutter did not have a significant influence on the precision of the results (Fig. 5c), but only makes harder the identification of the modes of the distribution. In general, the erroneous results can be attributed to two sources (see Fig. 4, last row). First, the edge segments we restrict ourselves to cannot always be extracted consistently. For example, in an image of the kettle, if the edges of the handle are extracted on one of its sides but not on the other, this side may be "matched" with any of the two sides of a training view, potentially leading to a large error on the orientation of the recovered pose – despite both being globally good matches with the 2D input view. Second, using the 2D projections of any 3D object introduces inevitable ambiguities. For example, it may be very difficult to differentiate between a cylindrical object pointing away and towards the camera (Fig. 4, bottom left); this effect is particularly true for our objects consisting of mostly homogeneous surfaces.

We used a similar protocol to evaluate the use of multiple views of a same scene, as proposed in Section IV. In those experiments, we used, instead of a single 2D image, 2 images of the scene from viewpoints spaced by $45°$. Such a wide baseline is generally too large to be handled by traditional stereo methods, and thus demonstrates one of the interests of our approach. The success rate was generally not noticeably affected by the use of two views over one, but the error was almost always substantially decreased, as reported in Fig. 5c. Using a second view helps the algorithm disambiguating between the different possible orientations of the object, and also provides much better clues for determining the actual depth of the scene (Z translation).

B. Real test images

The method was evaluated on real test images. For practical reasons, we relied here again on computer-generated images as training data. We used 128 training views of each object, that were produced as explained above (Fig. 1c), through ray tracing with manually-designed 3D models. In a realistic application, such images are to be acquired, e.g. by a robotic agent taking pictures of the real object under various viewpoints [27]. This alternative option was chosen purely for practical reasons, but added an additional challenge as the models used for generating the training images inevitably did not match the real objects perfectly. The test images were taken with a handheld camera at about 1000–2000 mm from the scenes.

We performed many experiments on typical household scenes with common objects. We purposely chose objects presenting large homogeneous surfaces with little texture and details, on which classical feature-based pose estimation would likely fail. We present, in Fig. 6, typical results of both successful and failed experiments. As the ground truth

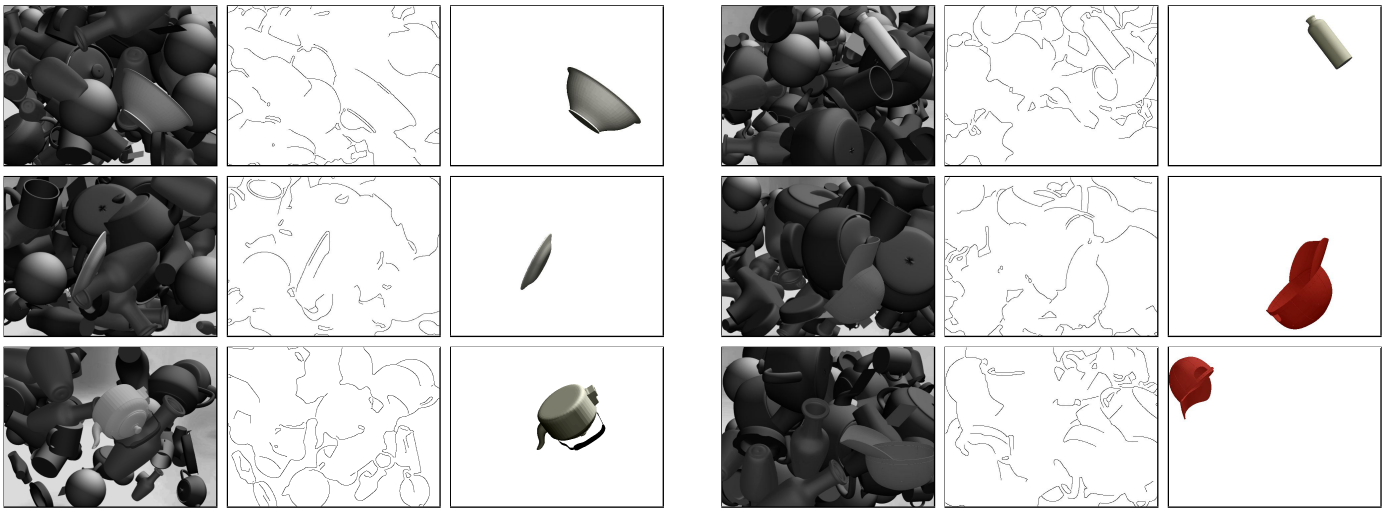
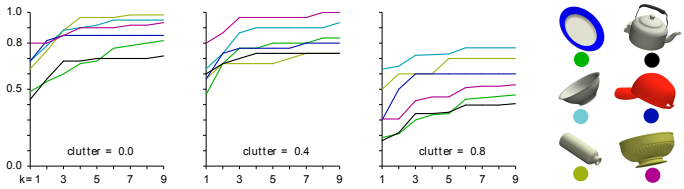


Fig. 4. Sample results of our quantitative evaluation (for each: test image, edge map used as input, rendering of object model in the first pose proposed by the algorithm); these tests used a single test view, 128 training views per object, and clutter=80%. The last row shows typical incorrect results: although a close match is found with the given edges, the 3D pose is incorrect.



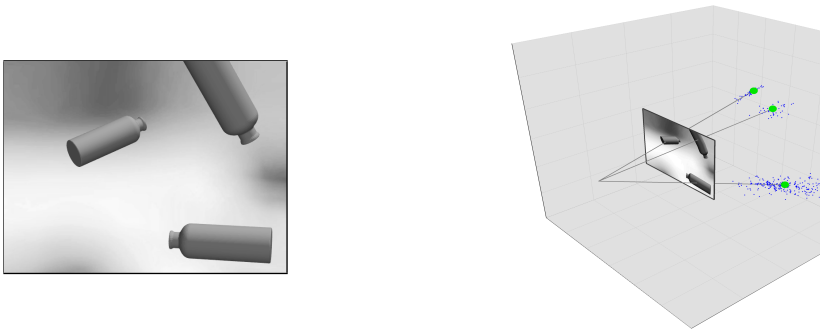
(a) (b)

N. of train. examples	**32**	**128**	**512**	**128, stereo obs.**
Translation (mm)	44.1	30.5	31.1	14.2
XY only	4.6	3.1	2.1	N/A
Z (depth) only	43.2	30.1	30.9	N/A
Orientation (°)	11.6	8.3	7.7	8.1
Rank (k)	2.9	1.8	1.3	1.2

(c)

Fig. 5. Quantitative results on synthetic images (see text for details); (a) for each object, success rate of having a correct result among the first k ones (128 training examples), in scenes of no/medium/heavy clutter; (b) test objects used; (c) average error of first correct result.

Fig. 6. Sample results on real images (similar conditions as Fig. 4); for visualization, we render, in yellow, the outline of artificial models set in the first pose found. The last two images show common failures, typically due to uncertainty in the limited input data used (edges): the mitten identified in background clutter, and the rim of the plate matched onto its shadow.



(a) (b)

Fig. 7. Recovery of probability distribution of 3D pose; (a) input image; (b) plot of 3D position as a non-parametric description (blue points), and local maxima (green points). Each occurrence of the object in the image correctly generates one mode in the distribution.

pose is not available, measuring the errors is not possible. Instead, we visualize the results by rendering, onto the input images, synthetic models of the objects in the poses found by the algorithm. One can observe good matches with the input images, demonstrating the good use made of the 2D information available. As discussed before, the use of 2D observations, especially edge segments alone, often makes it hard to distinguish between different poses that may appear similar in one image. The first pose returned by the algorithm may thus correspond to an erroneous result, but the correct result will often be found in the other poses proposed by the algorithm (identified with slightly lower probability). The actual disambiguation is thus to be left to the end application, which may best make use of this uncertainty in the results.

C. Retrieval of full pose distribution

One key capability that we propose is to recover a *distribution* of 3D poses, rather than a single optimum. We illustrate this in Fig. 7: the pose of a bottle is evaluated in an image containing several occurrences of the object. The distribution is recovered in a non-parametric fashion as a collection of particles, of which we plot the 3D position. One mode is correctly identified for each occurrence of the object, the main uncertainty remaining unsurprisingly in the depth dimension, extending along the camera projection axis.

VI. Conclusions and future work

We presented a novel method for exemplar-based pose estimation in single images. Relying on a general, probabilistic formulation of the problem, the method avoids establishing specific correspondences between training and test views, thus allowing similar-looking types of images features. The pose of the object is treated as a probability density over the 3D pose space, from which we identify the different modes, thereby accounting for the ambiguities of 2D input data. We also proposed an elegant way of fusing evidence from multiple sources, such as several views of the same scene, or different types of image features. A first validation of the overall approach showed promising results. Further developments will mainly focus on the use of other types of image features within this framework, extending its applicability further to more types of scenes, objects and imaging conditions.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 (Specific Programme Cooperation, Theme 3, Information and Communication Technologies) under grant agreement no. 270273, Xperience. Damien Teney is supported by a research fellowship of the Belgian National Fund for Scientific Research.

References

[1] A. Collet and S. S. Srinivasa, "Efficient multi-view object recognition and full pose estimation," in *IEEE Int. Conf. on Rob. and Autom.*, 2010, pp. 2050–2055.

[2] I. Gordon and D. G. Lowe, "What and where: 3D object recognition with accurate pose," in *Toward Category-Level Object Recognition*, 2006, pp. 67–82.

[3] D. Teney and J. Piater, "Probabilistic Object Models for Pose Estimation in 2D Images," in *DAGM*, ser. LNCS, vol. 6835. Springer, 2011, pp. 336–345.

[4] S. Edelmann and H. Bülthoff, "Modeling human visual object recognition," in *Int. Joint Conf. on Neural Networks*, 1992, pp. 37–42.

[5] S. Ekvall, F. Hoffmann, and D. Kragic, "Object recognition and pose estimation for robotic manipulation using color cooccurrence histograms," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2003.

[6] P. Mittrapiyanuruk, G. N. DeSouza, and A. C. Kak, "Calculating the 3D-pose of rigid-objects using active appearance models," in *IEEE Int. Conf. on Rob. and Autom.*, 2004, pp. 5147–5152.

[7] A. R. Pope and D. G. Lowe, "Probabilistic models of appearance for 3D object recognition," 2000.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints." *Int. J. Comp. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[9] R. Soderberg, K. Nordberg, and G. Granlund, "An invariant and compact representation for unrestricted pose estimation," in *Patt. Rec. and Im. Anal.*, J. Marques, N. Prez de la Blanca, and P. Pina, Eds. Springer, 2005, vol. 3522, pp. 489–500.

[10] F. Viksten, P.-E. Forssén, B. Johansson, and A. Moe, "Comparison of local image descriptors for full 6 degree-of-freedom pose estimation," in *IEEE Int. Conf. on Rob. and Autom.*, 2009, pp. 2779–2786.

[11] G. Borgefors, "Hierarchical chamfer matching: A parametric edge matching algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 6, pp. 849–865, 1988.

[12] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, 1993.

[13] M.-Y. Liu, O. Tuzel, A. Veeraraghavan, and R. Chellappa, "Fast directional chamfer matching," in *Conf. on Comp. Vis. and Patt. Rec.*, 2010, pp. 1696–1703.

[14] C. F. Olson and D. P. Huttenlocher, "Automatic target recognition by matching oriented edge pixels," *IEEE Trans. Im. Proc.*, pp. 103–113, 1997.

[15] V. Ferrari, T. Tuytelaars, and L. Van Gool, "Object detection by contour segment networks," in *European Conf. on Comp. Vis.*, 2006.

[16] V. Ferrari, F. Jurie, and C. Schmid, "From images to shape models for object detection," *Int. J. Comp. Vis.*, vol. 87, no. 3, pp. 284–303, 2010.

[17] R. Detry, N. Pugeault, and J. Piater, "A probabilistic framework for 3D visual object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1790–1803, 2009.

[18] R. Detry and J. Piater, "Continuous surface-point distributions for 3D object pose estimation and recognition," in *Asian Conf. on Comp. Vis.*, 2010.

[19] A. Toshev, A. Makadia, and K. Daniilidis, "Shape-based object recognition in videos using 3d synthetic object models," in *Conf. on Comp. Vis. and Patt. Rec.*, 2009, pp. 288–295.

[20] F. Viksten, R. Soderberg, K. Nordberg, and C. Perwass, "Increasing pose estimation performance using multi-cue integration," in *IEEE Int. Conf. on Rob. and Autom.*, 2006, pp. 3760–3767.

[21] R. Caflisch, "Monte carlo and quasi-monte carlo methods," *Acta Numerica*, vol. 7, pp. 1–49, 1998.

[22] R. C. Nelson and A. Selinger, "Large-scale tests of a keyed, appearance-based 3D object recognition system," *Vis. Res.*, vol. 38, pp. 38–15, 1998.

[23] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790 –799, 1995.

[24] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky, "Nonparametric belief propagation," in *Comp. Vis. and Patt. Rec.*, 2003.

[25] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Conf. on Comp. Vis. and Patt. Rec.*, 2004, pp. 97–104.

[26] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *IEEE Int. Conf. on Rob. and Autom.*, 2011.

[27] K. Welke, J. Issac, D. Schiebener, T. Asfour, and R. Dillmann, "Autonomous acquisition of visual multi-view object representations for object recognition on a humanoid robot," in *IEEE Int. Conf. on Rob. and Autom.*, 2010, pp. 2012–2019.