UNIVERSITAT POLITÈCNICA DE CATALUNYA

Doctorat en AUTOMÀTICA, ROBÒTICA I VISIÓ

Proposta de tesi doctoral

3D Pose Estimation in Complex Environments

Adrián Peñate Sánchez

Advisors

Juan Andrade Cetto, Ph.D.(IRI) and Francesc Moreno Noguer, Ph.D.(IRI)

Juny 13, 2013

Contents

1 Introduction

In Computer Vision, pose estimation consists in estimating the relative position and orientation of an object with respect to the viewing sensor. It is one of the most elemental problems in computer vision with applications in video-games, mobile phones and robotics. Pose estimation has been a research subject for a long time, from opticians in the late 19th century to the actual computer science researchers that apply a variety of approaches.

Although pose estimation is already a reality that can be seen in numerous everyday applications it still suffers from limitations. This limitations usually appear in non-controlled environments. For this reason, we seek to improve the state of the art algorithms for pose estimation to perform better in real life conditions. Some of the issues we want to tackle are: dependence on camera calibration and feature point extraction.

In this work we will focus on the use of perspective cameras. Camera calibration is an essential ingredient in pose estimation. It basically consists on obtaining the geometric parameters that define the perspective camera that we are using. Without knowing this parameters one cannot solve for the rotation and translation of the object that we seek to find. Usually, camera calibration has to be solved with user-assisted methods like [56]. There are pose estimation methods in the literature that solve the camera calibration problem [2, 21, 45, 51] but still camera calibration is rarely done without human intervention, for this reason we seek to further improve the case of uncalibrated pose estimation methods.

The other crucial factor in modern pose estimation algorithms is the extraction and matching of feature points. A feature point is composed of two parts, the point of interest and a feature descriptor. The point of interest provides the coordinates of a very recognizable point in an image and the feature descriptor is the way in which we represent such point. A good point of interest extractor is able of finding the same part of an object (points in our case) in different images of the same object. A good feature descriptor should be discriminant enough for the correct matching of points of interest between two different images of the same object with high repeatability.

Feature points are often subject to viewpoint changes, lighting changes and occlusions. In this work we intend to cope with increasing difficulty in obtaining such feature points. As our focus is to improve the performance in real case scenarios we will center our efforts in rigid objects. The additional variability of non rigid objects or deformable objects is out of the scope of this thesis proposal. We have defined three main objectives to achieve in this thesis, and are as follows.

1 Uncalibrated pose estimation from known correspondences.

2 Uncalibrated pose estimation from unknown correspondences.

3 Pose estimation without points of interest.

We will first attempt to improve uncalibrated solutions to the pose estimation problem using feature points. That is, supposing that we know the correct matching between points in an image and their spatial representation on the surface of an object, in this case a 3D model. Once we achieve results on this matter we will assume that feature descriptors are not reliable, this will leave us with only the coordinates of the points of interest, but without knowing the one-to-one matching between them. And finally, we will attempt to solve the pose estimation problem in situations where feature points cannot be detected, such as in motion blurred images. When solving for the pose without either points of interest or feature descriptors, we need to be very cautious. The quality of the images in which feature points cannot be used could be of such low quality that it might not be possible to distinguish the subtle differences produced by changes in

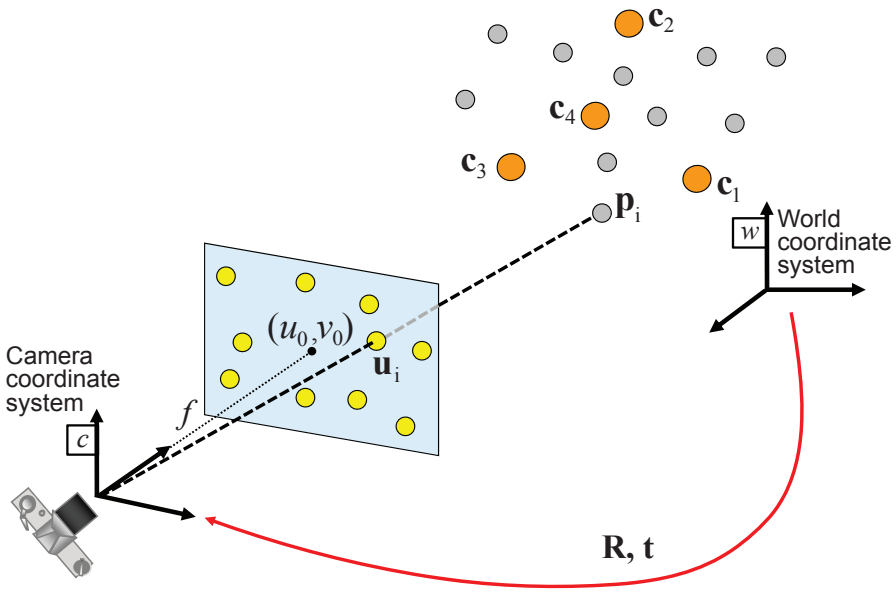

Figure 1: **Problem Formulation:** Given a set of correspondences between 3D points $\mathbf{p}_i$ expressed in a world reference frame, and their 2D projections $\mathbf{u}_i$ onto the image, we seek to retrieve the pose ($\mathbf{R}$ and $\mathbf{t}$) of the camera w.r.t. the world and the focal length f.

the camera geometry. We will try to cope with uncalibrated cameras in this case, but achieving a correct camera pose in such conditions with a calibrated camera is by itself a huge feat.

We find this order of tasks the most appropriate to handle the goals of our work. The progression in difficulty will help the student understand the actual problems of pose estimation in order to identify the requirements for the next step. This approach also helps the assimilation of the state-of-the-art and the acquisition of the know-how necessary to handle the task at hand. We will now explain each of the three tasks separately dedicating a separate section for each one of them. We will define the state of the art for all three tasks, we will define the proposed methodologies to improve the state of the art and we will mark the expected objectives to accomplish.

2 Uncalibrated pose estimation from known point correspondences

2.1 Outline

Estimating the camera pose from n 3D-to-2D point correspondences is a fundamental and well-understood problem in computer vision. Its solution is relevant to almost every application of computer vision in the era of smart phones. The most general version of the problem requires estimating the six degrees of freedom of the pose and five calibration parameters: focal length, principal point, aspect ratio and skew (see Fig. 1). This can be established with a minimum of 6 correspondences, using the well known Direct Linear Transform (DLT) algorithm [18].

There are, though, several simplifications to the problem which turn into an extensive list of different algorithms that improve the accuracy of the DLT. The most common simplification is

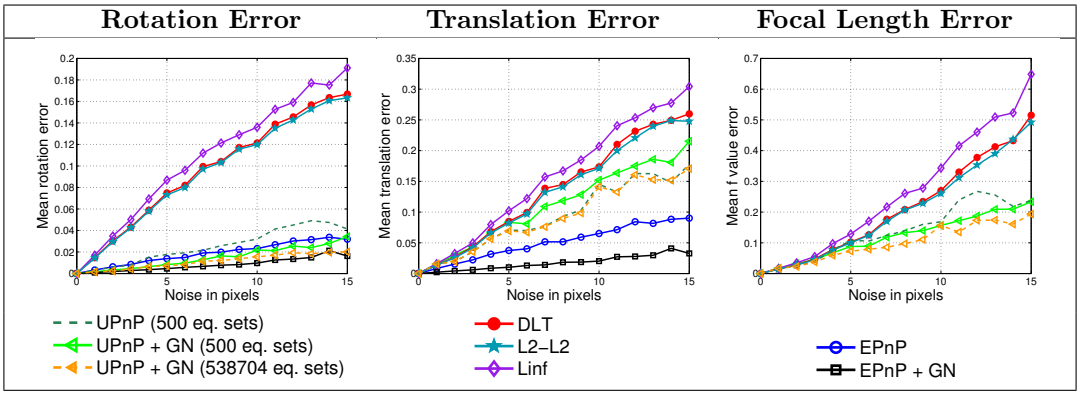

Figure 2: Results on synthetic data for non-planar distributions of points, our approach is dubbed UPnP. Mean rotation, translation and focal length errors for: increasing levels of image noise on 10 2D-3D correspondences, and two different focal lengths. Each tick in the plot represents the average over 100 experiments with random points.

to assume known calibration parameters. This is the so-called Perspective-n-Point problem, for which three point correspondences suffice in its minimal version [15]. There exist also iterative solutions to the over-constrained problem with $n > 3$ point correspondences [10, 19, 29] and non-iterative solutions that vary in computational complexity and accuracy from $O(n^8)$ [1] to $O(n^2)$ [13] down to $O(n)$ [26].

For the uncalibrated case, given that modern digital cameras come with square pixel size and principal point close to the image center [3, 18], the problem simplifies to the estimation of only the focal length. Solutions exist for the minimal problem with unknown focal length [2, 24, 45, 51], and for the case with unknown focal length plus unknown radial distortion [3, 4, 21, 51].

Unfortunately, in the presence of noise and mismatches, these solutions to the minimal problem become unstable and may produce unreliable pose estimates. This is commonly addressed including an extra RANSAC [14] iterative step for outlier removal, either taking minimal or non-minimal subsets [47], but at the expense of high computational load. Recent approaches have reformulated the problem as a quasi-convex optimization problem, allowing for the estimation of global minima [7, 22, 23]. Yet, while this is a very attractive idea, the iterative nature of these approaches makes them unpractical for real-time applications, unless a very small number of correspondences is considered.

We would want to advocate for an efficient solution that can handle an arbitrarily large point sample, thus increasing its robustness to noise. Using a large point set may be especially useful for current applications such as 3D camera tracking [25] or structure-from-motion [55], which require dealing with hundreds of noisy correspondences in real time.

The approach we proposed fulfilled these requirements: it allows estimating pose and focal length in bounded time, and since it is a non-minimal solution, it is robust to situations with large amounts of noise in the input data. Drawing inspiration on the EPnP algorithm [26, 31], showed that the solution of our problem belongs to the kernel of a matrix derived from the 3D-to-2D correspondences, and thus can be expressed as a linear combination of its eigenvectors. The weights of this linear combination become the unknowns of the problem, which can be solved applying additional distance constraints.

However, solving also for the focal length has the effect that the linearization and relinearization techniques used in [26, 31] to estimate these weights no longer were valid. Several factors contribute to this: (1) the new polynomials that needed to be considered are of degree four, in

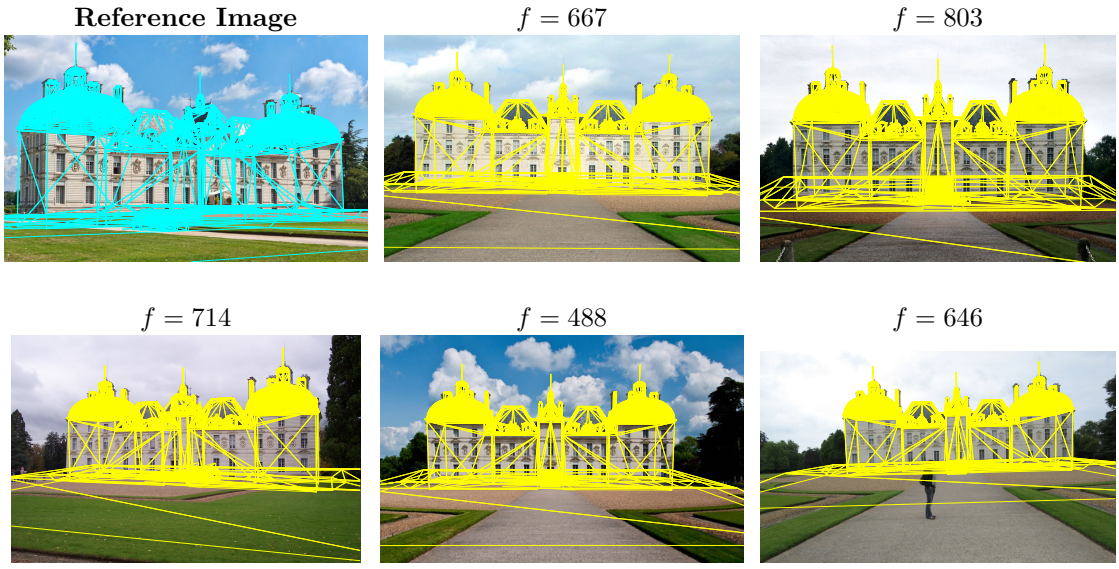

Figure 3: Example of validation on a real set of images obtained from *Flickr* with a 3D model obtained from *GoogleEarth*. 3D model reprojected onto the reference and testing images.

contrast to those in the EPnP that were of degree two; (2) the variables being computed differ in several orders of magnitude and small inaccuracies in the input data may propagate to large errors in the estimation; and (3) the number of possible combinations in the solution subspace explodes combinatorially for large kernel sizes. All these issues make that a naïve selection of equations for back substitution after linearization produces unreliable results. Moreover, a least squares solution of the kernel weights was not viable since it equally ponders constraints that involve variables with different orders of magnitude. We needed to develop new techniques to solve the limitations of linearization and relinearization, and thus, solving the uncalibrated PnP problem robustly.

The proposed approach, compares favorably in terms of accuracy to the DLT algorithm, the only closed-form solution we are aware that is applicable for an arbitrary number of correspondences. This was the case because the least squares solution of the DLT algorithm chooses an optimal solution only in the direction along the vector associated with the smallest singular value of the linear system of equations built from the 3D-to-2D correspondences. In contrast, our proposed approach considered all directions of the kernel of the system, which for the ideal case is of size one [41], but for noisy overconstrained systems grows in size [26]. We compared our approach against global optimal methods like [22] and [23], which are algorithms that guarantee maximum error tolerance, but which are computationally expensive. Also, in real experiments, robustness to outliers is only possible inside an extra RANSAC loop at additional computational cost, we provided experiments that showed the performance of our approach against all compared methods in this environment.

2.2 State of the art

Both algebraic and geometric solutions exist for the minimal solution of the calibrated case. A representative case is Gao's solution [15], in which a triangular decomposition of the P3P equation system is given. Many iterative methods to solve the calibrated case for large values of n exist [10, 19, 29]. Of these, the most representative method is Lu's method [29], which is one

of the fastest and most accurate methods producing slightly better results than the non-iterative EPnP algorithm [26]. Accuracies become similar when Gauss-Newton optimization is applied to the EPnP solution at negligible cost.

Whilst iterative methods might get trapped in local minima, efforts are also aimed at finding globally optimal solutions, for instance, casting it as a positive semidefinite quadratic optimization problem [42]. This method resorts to a parameterization of rotations using quaternions, and it is unclear whether quaternion unit norm constraints might hinder convexity of the solution search space.

For the uncalibrated case, our problem of interest, minimal solutions exist for the computation of the 5- and 6-point relative pose problems with unknown focal length that use Groebner bases to solve large systems of polynomial equations [24, 45]. In both cases, the problems are casted as polynomial eigenvalue computations, but differ from ours in that they compute only the relative pose between two views instead of the absolute camera referential.

Bujnak *et al* [2] give a general solution to the minimal problem that is closest to ours –P4P for a camera with unknown focal length–. Their system reduces to finding the solution to a system of 5 equations with 4 unknowns and 20 monomials, for which they compare two methods, a hidden variable solver, and also the use of Groebner bases to solve large systems of 154 polynomial equations with 180 monomials and 4 unknowns. Our method would need to compare favorably to both with respect to computational load, and robustness to pixel noise. Bujnak solves only the minimal problem ($n = 4$). More recently, other solutions have been proposed for the mininal P4P problem with unknown focal length and radial distortion [3, 4, 21] that also need an extra robust optimization loop.

2.3 Contributions and results

We extended the EPnP algorithm [26] to give an equivalent solution to the uncalibrated case, by defining a system of equations using the intrinsic calibration parameters as further unknowns and using different equation solving techniques. We needed to develop alternative solutions to linearization and relinearization in order to circumvent their limitations by systematically exploring the solution subspace, for this purpose we created two techniques, dubbed *exhaustivelinearization* and *exhaustiverelinearization*. The proposed approach is also a fast solution to the problem of recovering the pose and focal length of a camera, given n 3D-to-2D correspondences. We proved that the uncalibrated PnP can be expressed as the solution of a fixed-size linear set of equations independent of the number of points, similar to the EPnP algorithm for the fully calibrated case.

Validation of the proposed resulting approach was done over synthetic data and real images. For validation on synthetic data we simply created random sets of points and calculated a new pose that then we needed to find with our approach. The advantage of synthetic validation is that ground truth is obtained easily so intense testing can be performed without cost. Results on synthetic data can be observed in Fig. 2, we show that we perform better than state of the art uncalibrated methods and we also obtain comparable results again calibrated methods. An example of validation on real data can be seen in Fig. 3. Here we show how we took a 3D registered image as reference and used our PnP algorithm to find the rotation, translation and focal length of the camera.

Figure 4: **Uncertain 3D model**. **Left:** 3D model acquired with a Kinect camera. Regions in which the 3D data is most uncertain are depth discontinuities. **Center:** We detect the uncertain regions –shown in red– computing depth covariances within local neighborhoods. **Right:** A 3D covariance can be assigned to each 3D model point and propagated to the image plane. This will be used to limit the area where to search for potential match candidates.

3 Uncalibrated pose estimation from unknown point correspondences

3.1 Outline

In the above-mentioned case, we assumed that the image to 3D model point correspondences were given. When these correspondences are not known in advance, we need robust methods to establish them. There are cases however in which the feature descriptors are not reliable enough to establish robustly such pairing, or even cases in which feature detection is possible but appearance information is very poor. In such cases, we need to establish feature correspondences with the aid of geometry as well.

We can resort to choosing small sets of points as putative candidate matches, and fit them robustly to a geometrical model (using fo instance RANSAC) [8, 14].

Matching methods based on RANSAC rely on having a small percentage of outliers, this way the probability of obtaining a correct minimal set of points is high enough to be achieved in fixed time. If the percentage of outliers increases, the number of RANSAC loops needed to obtain a correct set of minimal points grows exponentially. Also, as shown in [38], using the minimal set of points might not yield the best pose if there is noisy data in the system. The percentage of outliers depends on several factors: the precision recall of the point descriptor, changes in appearance, changes in points of view, repeated patterns, self occlusions, etc. We want to develop a matching algorithm that aims to perform robust matching under the presence of a high percentage of outliers. Our approach would be a generalization of the work in [32] to deal with uncalibrated cameras and noisy 3D information. The noisy 3D data will be obtained with a Kinect camera [30]. In Fig. 4 we can see how the Kinect sensor produces noisy data that makes pose estimation hard. In Fig. 5 we show the difference we expect to make between a SIFT [28] robust matching using RANSAC and our proposed approach.

To tackle all these issues we propose to split the initial prior distribution of the combined pose and focal length estimate into an arbitrary large number of Gaussian priors. These priors are spread within very rough bounds of where the pose and focal length are expected to be. At runtime, each of these priors is used to guide the search for the 3D-to-2D correspondences, while progressively pruning the number of potential candidate matches and refining the pose and focal length values. Repeating this process for each of the priors guarantees an exhaustive exploration of the solution space at a limited computational cost. An example of how we plan to reduce the number of candidates in the proposed approach can be seen in Fig. 6.

Experiments in both synthetic and real data must show that the proposed approach is robust

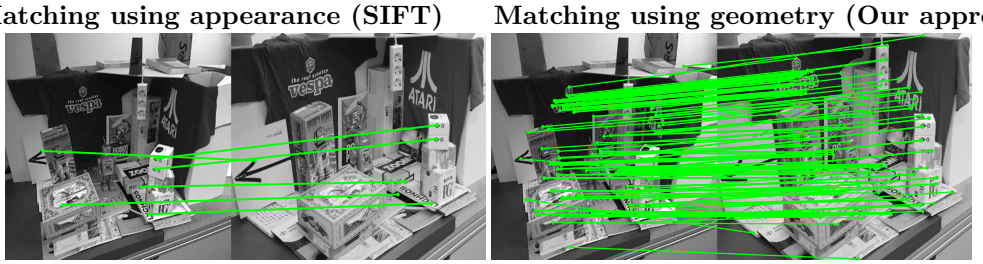

Figure 5: Inlier correspondences of a matching algorithm that uses appearance information (left) and a purely geometric matching approach (right) that uses the 3D and 2D location of the points to search for the correspondences. Matching was made between the intensity component of a Kinect camera and a Canon EOS 60D camera. Together with the viewpoint changes and the self-occlusions, differences in terms of image noise and resolution jeopardize appearance matching producing a very small set of correct matches. In contrast, our method based purely on geometric information would be able to retrieve a much larger number of correspondences, accurately computing the relative pose and focal length under such conditions.

to large levels of 2D and 3D noise and clutter, yielding reasonable results for high outlier rates. Validation will be performed by testing against RANSAC based approaches [6, 8], global methods [23] and other purely geometrical calibrated methods [32]. Comparisson against calibrated methods will give us a baseline to show the real impact of selecting a non calibrated camera.

3.2 State of the art

Pose estimation techniques that maximize image similarity, such as [5], are not applicable in our context due to their limited ability to deal with significant differences in appearance and reduced capture range, as that of Fig. 5. We shall therefore consider only techniques that explicitly perform matching using the geometric structure of the 2D and 3D point sets.

The robust estimation of correspondences between two sets of points has been historically solved by hypothesize and test algorithms such as RANSAC [14] and Least Median Squares [40]. They rely on a random sampling of minimal subsets to generate model hypotheses, and favor the one that best explains most of the data points. Unfortunately, in these methods, computational complexity scales exponentially with the number of model parameters and the size of the point set. Among the several variations of the original RANSAC algorithm, Guided-MLESAC [49] and PROSAC [8] avoid sampling unlikely correspondences by using appearance based scores and thus are not applicable to our problem. Similarly, GroupSAC [33] uses image segmentation to sample more efficiently the data. Other techniques of the same family such as Preemptive RANSAC [34] or ARRSAC [39] work within a limited time scenario thus increasing the probability of not reaching the best estimate. Finally, MultiGS [6] accelerates the search strategy by guiding the sampling with information from residual sorting and is able to account for multiple structures appearing in the scene.

In the absence of robust appearance information graph matching can be used, as proposed by [11]. Yet, due to its high computational cost, this methods are only applicable to small graphs. While this approaches allow global optimization, they cannot be used with large intra-image distances due to very different points of view of the scene or when the number of outliers is excessive, which are the cases we consider in this paper.

The L_∞ technique proposed in [23], uses second-order cone programming, and guarantees optimality under the L_∞ norm, for different geometric structure and motion problems, including

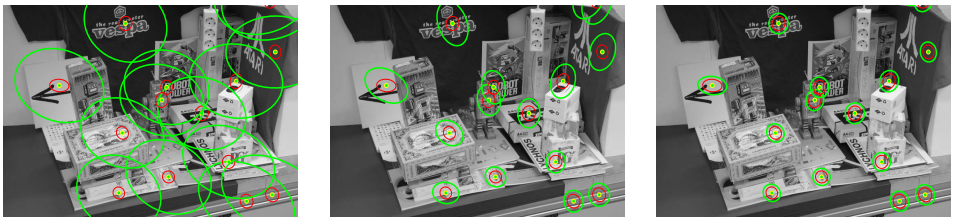

Figure 6: Limiting the number of potential candidates by refinement of the search space, after establishing correspondences.

the camera pose estimation considered in this work. However, this particular metric is highly sensitive to outliers, as pointed out in [17]. Even when it is possible to address in part the outlier removal problem as proposed in [44], the L_∞ solution for the camera pose estimation only performs as well as the standard L_2 norm.

Other approaches simultaneously solve for pose and correspondences purely from geometric point matching. Of these, SoftPOSIT [9] uses an iterative technique to generate correspondence candidates, but the global minimum can not be guaranteed. Our approach is inspired in the Blind PnP algorithm [32], where local optimality is alleviated introducing the scene geometry as pose priors, modeled as a Gaussian mixture model, and progressively refined by hypothesizing correspondences. Incorporating each new candidate in a Kalman filter rapidly reduces the number of potential 2D matches for each 3D point and makes it possible to search the pose space sufficiently fast for the method to be practical. More recent techniques [43] use robust estimation in a final stage to refine the pose. Unfortunately, the approach cannot be applied straightforward to the uncalibrated case due to the ambiguities between focal length and the pose translation vector.

3.3 Expected contributions and partial results

Simultaneously estimating the camera position, orientation, focal length and establishing 3D-to-2D correspondences between model and image points, poses a challenging optimization problem which can hardly be solved without prior information. Most current approaches rely on appearance information to first solve the correspondences and then retrieve the pose and focal length while rejecting missmatches. Yet, there are many situations in which the appearance is either not available or not a reliable cue.

In the absence of appearance, we propose to use only geometric priors, which are just rough approximations of the pose and focal length solution space. By progressively exploring these priors we are able to efficiently prune the potential number of 3D-to-2D matches, while reducing the uncertainty of the pose and focal length estimates. The method is shown to be highly resilient to clutter and noise, on the image features and in the 3D model. The latter is especially suited for dealing with 3D models obtained from noisy range sensors, such as the Kinect or Time of Flight cameras.

We expect to obtain a new method to perform uncalibrated pose estimation that can handle high values of outliers using only geometric information. To validate this approach we will compare against RANSAC based approaches [6, 8], global methods [23] and other purely geometrical calibrated methods [32], in both real case scenarios and synthetic data. For real case scenarios we will use a Kinect sensor to obtain the 3D model, we will also assess that we are able of handling the noise inherent in the system.

The fact that we are using uncalibrated cameras and textureless features, opens new research

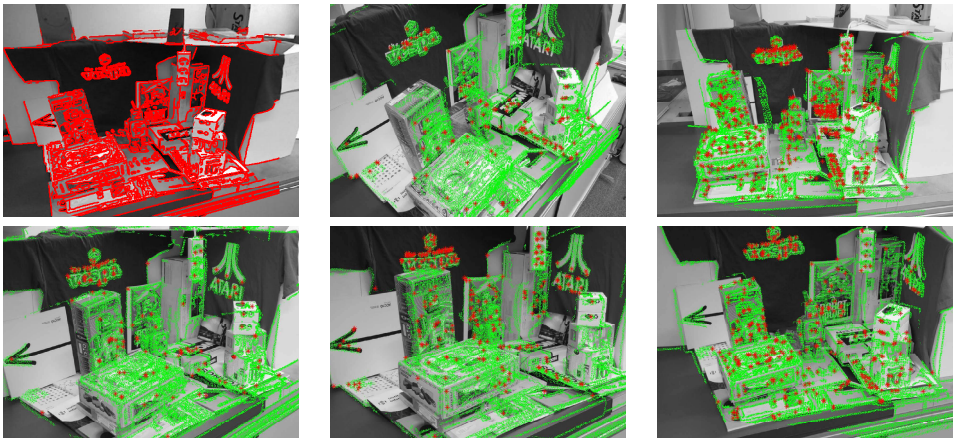

Figure 7: Real experiments. **Top Left:** Reference image registered to the 3D model. **Others:** Reprojection of the 3D model onto the input images after estimating pose, focal length and 3D-to-2D matches with our approach.

areas for the future. This kind of techniques could be used within an active exploration setting, to estimate focal length for cases where the camera zoom is actuated to enforce good feature tracking. It also could be integrated in a setting in which generic appearance models can be aggregated for the same keypoint as observed from multiple vantage points. These generative appearance models could be used in turn as additional priors within our or competing frameworks to speed up match search. In Fig. 5 and in Fig. 7 we show the partial results we have already obtained, as it can be seen our approach shows to perform robustly on noisy sensors.

4 Pose estimation without points of interest

4.1 Outline of the proposed work

The problem of pose estimation has been addressed from either purely geometric or machine learning perspectives. Geometric methods initially use training data to build a 3D model, and then search for the 2D-to-3D correspondences that best align interest points in the test image with the 3D model [31]. Machine learning approaches on the other hand, annotate training imagery with discrete locations in the pose manifold, and then search globally for this pose-annotated matching of appearance, without resorting to full 3D reconstruction of the object [16, 36, 54]. The advantage of the global methods is that they are less sensitive to precise localization of individual features, which makes them more robust to image degradations than local geometric methods. But in contrast, global methods are not generally robust to occlusions. In addition, they often require splitting the pose space into several classes, and training specific classifiers for each of them, limiting the precision of the estimated pose to that of the granularity between classes and losing the correlation between neighboring poses.

We intend on building a solution that trains in high quality images and tests on any kind of image despite its low quality. We want to combine the strengths of both the geometric and machine learning methods for estimating object pose. We propose using high-definition training images to create a 3D model of the object, and from it, devise a pose-indexed feature extraction scheme that binds image quantities to the object pose. A *single* classifier, common to all the poses, will be trained from these pose-indexed feature vectors.

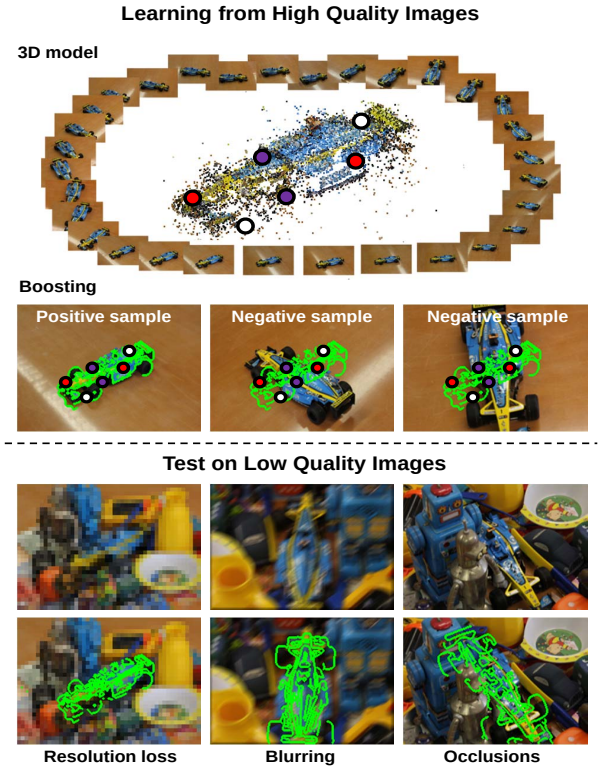

Figure 8: Our approach intends on using high-quality images to train a classifier that is tested on low quality data. The top box shows the built 3D model, and one positive and two negative training image-pose pairs. The fact that features are indexed with pose is indicated by the same location on the three training images of the projection of point pairs, and by the object contour projected in green. Observe that detecting the F1 car in the test images at the bottom box is even difficult for the human eye due to different artifacts. The green contour indicates the correct pose of the F1 car in that image.

We propose to use a new novel procedure, dubbed AbstainBoost, able to cope with incomplete feature vectors, a situation that occurs during self occlusion. During test, given a low quality input image and a hypothetical pose to assess, the pose-indexed feature vector is computed similarly, without the need for detecting and matching points of interest, and fed to the classifier. The object's pose is estimated by measuring the classifier response for all the poses seen in the training images, and then refined by resampling around those poses with maximal classifier response. We insist on the fact that since this optimization is done by visiting multiple poses systematically, we do not need to match points of interest, or perform any type of fragile matching prior to using our predictor.

As shown in Fig. 8, our method intends to estimate the pose even in the presence of severe image artifacts such as motion blur and occlusion. We will need to demonstrate that this approach compares favorably against geometric approaches based on SIFT and DBRIEF features, and also against global approaches based on Bag of Features descriptors [53], GIST [35] and PCA cross-correlation [50]. As the task at hand is highly complex and has not been attempted as so, we will approach it in three stages.

1 Discrete calibrated pose estimation: First we need to prove that we can retrieve the closest pose of a test image from the training set. This is quite important because it will prove that AbstainBoost converges correctly. We will test to see if using bad quality images we can find the correct training image, the precision of this approach will depend on the granularity of the training set. We will also explore the conditions needed to perform the best training possible, and how big is the testing space in which we obtain correct results.

2 Global calibrated pose estimation: Once we prove that our single classifier can obtain correct results in the proximity of the training images we will extend our approach to perform a global search. This will give more precision than only using the training poses as candidates. We expect to do a refinement of the pose by performing gradient descent on the response of the classifier.

3 Extension to uncalibrated pose estimation: If global pose can be achieved with high precision we will evaluate the classifier searching in another additional dimension. Calibrated pose estimation gradient descent would be performed on a 6 dimensional space, adding the focal length would involve adding an additional dimension to the search. The key factor will be if the classifier is able to detect the subtle differences introduced by different geometries of the camera, and even if those changes remain when an image is degraded.

This part of the work is being performed under collaboration with François Fleuret at IDIAP research institute in Martigny, Switzerland.

4.2 State of the art

3D pose estimation methods may be roughly split in those techniques relying on local image features that purely use geometric relations to compute the pose; and methods that compute global descriptors of the image and resort to machine learning tools to estimate the pose.

Local approaches use feature point descriptors to estimate 2D-to-3D correspondences between one input image and one or several reference images registered to a 3D model of the object. PnP algorithms such as the EPnP [31] are then used to enforce geometric constraints and explicitly solve for the pose parameters. On top of that, robust RANSAC-based strategies [8, 32] can be used both to speed up the matching process and to filter outlier correspondences. Yet, while these methods provide very accurate results, they require both the reference and input images to be of high quality, such that local features can be reliably and repetitively extracted. As we will show in the results section, these methods are not applicable for the level of image artifacts we consider in this paper.

By contrast, approaches relying on global descriptions of the object are less sensitive to a precise localization of individual features. These methods typically use a set of training images acquired from different viewpoints to statistically model the spatial relationship of the local features, either using one single detector for all poses [20, 27, 46] or a combination of various pose-specific detectors [36, 37, 48, 54]. Another alternative is to bind image features with poses during training and have them vote in the pose space [16]. These approaches, though, focus on recognizing instances of generic classes and are not designed to deal with image content different from that in the training set.

We will exploit the strengths of both the local and global methods for estimating the 3D pose of a single object. We will use high quality training images to build a 3D model of the object and precisely locate the most discriminant local features. These features will be then combined into strong priors for each training pose. Note that since we focus on one single object, the priors we build can accurately capture the variability of the appearance and generalize to test images with

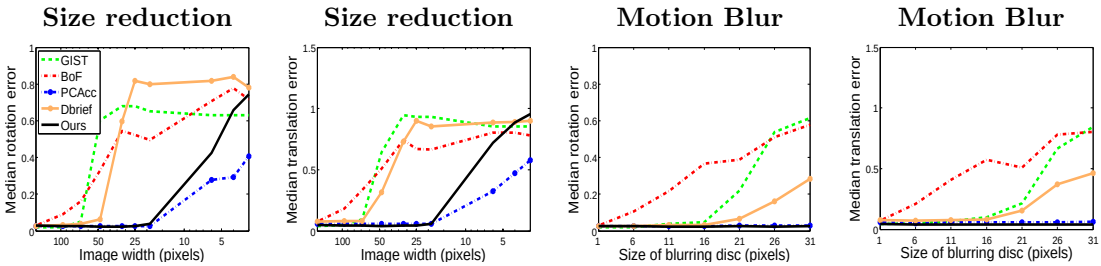

Figure 9: Pose estimation error of our approach and other approaches in experiments with severe degradations of image size and motion blur. We show results over the Sagrada Familia dataset

severe artifacts. In fact, this philosophy is very similar to that of [12, 25]. These approaches, though, again rely in the fact that similar local features appear in both training and test images. We intend of getting rid of this requirement by building specific priors for each training pose, in which we exactly know where the features should appear in the test image. We will therefore evaluate the classifier (which is unique and common to all poses) even when some features have been wiped out or corrupted by image artifacts such as loss of resolution, motion blur or partial occlusions.

Among the methods that compute a global descriptor of the image, we find some holistic representations that do not require extracting points of interest. For instance, the GIST [35] descriptor encodes sustained overall orientation of straight edges on images, rather than localized features. This descriptor is conceived more as a class descriptor than as a unique sample identifier, and is not generally robust for discriminating between poses, mainly because it is built using only 2D intensity data, disregarding visibility information of the 3D model. The same applies to the PCA cross-correlation, used in [50] as a similarity measure between tiny images. In our approach, considering visibility constraints in our pose-indexed feature vectors should bring a remarkable advantage of our approach against such global descriptors, especially under occlusions. This is because, to account for occlusions, we will devote a special treatment to missing data in our feature vector, and design a boosting mechanism able to cope with abstaining weak learners (AbstainBoost).

4.3 Expected contributions and partial results

We will propose a new machine learning paradigm: Learning with high-quality data to be able to test with low quality data. The rationale behind this idea is that inference is possible only from clean data, or using a strong model, and that the latter can be inferred from the former. This by itself will be a big contribution because it will provide the first machine learning pose estimator that uses only one classifier; the learning scheme can also be used for any context in which weak learners can respond neutrally.

We would validate our approach against different state of the art methods, both global and local. Such methods will be: Bag of Features descriptors [53], GIST [35] and PCA cross-correlation [50] as global methods; and DBRIEF [52] as local methods. We expect to compare favorably against them, specially when motion blur, occlusions and low resolution images are tested.

From this general principle, and extending the concept of pose-indexed features to be able to learn them, we expect to derive a novel and very efficient algorithm for the specific problem of pose estimation. With sufficiently good training data, we expect to obtain a good estimate of the object pose, in very low resolution images, and with high levels of noise and occlusion. This

procedure should be ideal as a near-perfect solution to be used in controlled environments such as a factory. Partial results can be observed in Fig. 9 on the Sagrada Familia dataset, we can see that our approach performs really well under both small resolutions and motion blur.

5 Resources and work plan

The proposed research is partially funded by the Spanish Ministry of Economy and Competitiveness under projects PAU+ DPI2011-27510 and MIPRCV Consolider Ingenio 2010 CSD2007-00018, and by the EU projects GARNICS FP7-247947 and ARCAS FP7-287617. A. Penate-Sanchez is the recipient of a JAE-Predoc scholarship funded by the European Social Fund. The work will be developed at the Institut de Robòtica i Informàtica Industrial, UPC-CSIC, in Barcelona.

The algorithms to be developed in this research, and the intense testing derived from them, are computationally intensive and require the availability of a computer able to perform computations in parallel. The facilities of the Institut de Robòtica i Informàtica Industrial include a grid computer that can be used to this end, with eight PC units of two Intel Quadcore Xeon E5310 processors and 4 Gb of RAM each one.

The work plan for the proposed research is divided into 5 main tasks, three of which are subdivided into several subtasks, as described below. The schedule of this plan spans over four years and is presented in Fig. 10 as a Gantt chart. In this chart, **Q1**, ..., **Q4** stand for the four quarters of a year. The work already completed is shown in orange.

Task 1: Initial literature review and initial training

This task initially entails acquiring a general overview on the state of the art in Computer Vision and the tools applied to the field. The study of selected publications, as well as undertaking the required courses as mandated by the doctoral program direction, will establish a solid ground from which to start the proposed research.

In a second stage, the literature review will have to concentrate on methods for Pose Estimation, focusing on previous solutions to the PnP problem, point matching and machine learning, and on the underlying mathematical tools of Projective Geometry and Linear Algebra.

Task 2: Uncalibrated pose estimation from point correspondences

Task 2.1: PnP literature review

We will identify the most representative works on the perspective PnP problem and those which are more widely used. From these works we will identify which are the main weaknesses on which we can make a strong contribution.

Task 2.2: Development of the pose algorithm

This task entails the development of a general algorithm for the PnP problem computation. It will take into consideration the identified weaknesses of previous work and will aim at giving a solution to these issues.

Task 2.3: Assessment of the contribution of the algorithm

We will compare against state of the art pose estimation algorithms and identify the size of the projected contributions.

Task 2.4: Publication of the results

We also plan the estimated time that the preparation of the publication will take. This aspect is commonly overlooked when in fact it takes a great part of the researcher's time.

Task 3: Uncalibrated pose estimation with only points of interest

Task 3.1: Matching literature review

Once the pose estimation literature has been reviewed, we will need to focus on feature points and feature matching literature at this stage.

Task 3.2: Development of the simultaneous pose and matching algorithm

This task entails the development of the proposed approach to solve the simultaneous matching and pose estimation problem. It will take into consideration the identified weaknesses of the previous work and will aim at giving a solution to these issues.

Task 3.3: Assessment of the contribution of the algorithm

We will compare against the state of the art on robust feature point matching algorithms and identify the size of the projected contributions.

Task 3.4: Publication of the results

We also plan the estimated time that the preparation of the publication will take. This aspect is commonly overlooked when in fact it takes a great part of the researcher's time.

Task 4: Pose estimation without points of interest

Task 4.1: Applied machine learning literature review

This literature review will take us away from geometrical pose estimation, we will need to identify the most relevant work on machine learning and see how it is put to use to solve the pose estimation problem. As this are not two fields that have been put together too much in the past it will prove difficult to identify the baseline to which to compare our proposed approach.

Task 4.2: Development of the algorithms

Task 4.2.1: Development of discrete calibrated pose estimation algorithm

We will develop the previously proposed pose classifier to first solve the discrete calibrated pose estimation. Results will be assessed against state of the art algorithms.

Task 4.2.2: Development of global calibrated pose estimation algorithm

We will develop the previously proposed pose classifier to solve the global calibrated pose estimation. Results will be assessed against the same algorithms as in the discrete approach, we will also compare against the discrete approach to see the incremental in the results.

Task 4.2.3: Feasibility test of the global uncalibrated pose estimation algorithm

In this task we will asses the feasibility of extending the approach into an uncalibrated setting. This will strongly depend on previous results.

Task 4.2.3: Development of the global uncalibrated pose estimation algorithm

If in the previous stage we obtain positive results, we will develop the approach to solve for uncalibrated pose. If results cannot be achieved we would finish our research at this point, we think the amount of research presented is more than sufficient to establish a solid contribution to the field.

Task 4.3: Assessment of the contribution of the algorithm and publication of results

We will compare against state of the art classification algorithms and identify the size of the projected contributions. We think that the size of the contributions which can be made in this area are potentially big, for this reason we expect at least two publications to be produced in this task.

Task 5: Elaboration of the dissertation

The last task of the research is dedicated to the elaboration of the dissertation and the preparation of its public defense. By taking into account the average time taken by colleagues and the time necessary for bureaucratic administration we consider the time needed of that of a at least two to three quarters.

6 Publications

The following is a list of accepted, submitted, or in-preparation publications resulting from the proposed research. The results of Task 2 are collected in [J1]. Partial results obtained so far in Task 3 are detailed in [C1], and those of Tasks 4.2.1 are reported in [C2]. We expect at least another publication to come from Task 3.2.2., we also expect a Journal version of paper [C1].

Conferences

C1. Penate-Sanchez A., Andrade-Cetto J., Moreno-Noguer F. Simultaneous Pose, Focal Length and 2D-to-3D Correspondences from Noisy Observations. (Submitted) In *2013 British Machine Vision Conference.*

C2. Penate-Sanchez A., Moreno-Noguer F., Andrade-Cetto J., Fleuret F. LETHA: Learning from High Quality Images for Pose Estimation in Low Quality Images. (Submitted) In *2013 International Conference on Computer Vision.*

Journals

J1. Penate-Sanchez A., Andrade-Cetto J., Moreno-Noguer F. Exhaustive Linearization for Robust Camera Pose and Focal Length Estimation. In *2013 Pattern Analysis and Machine Intelligence, IEEE Transactions on.*

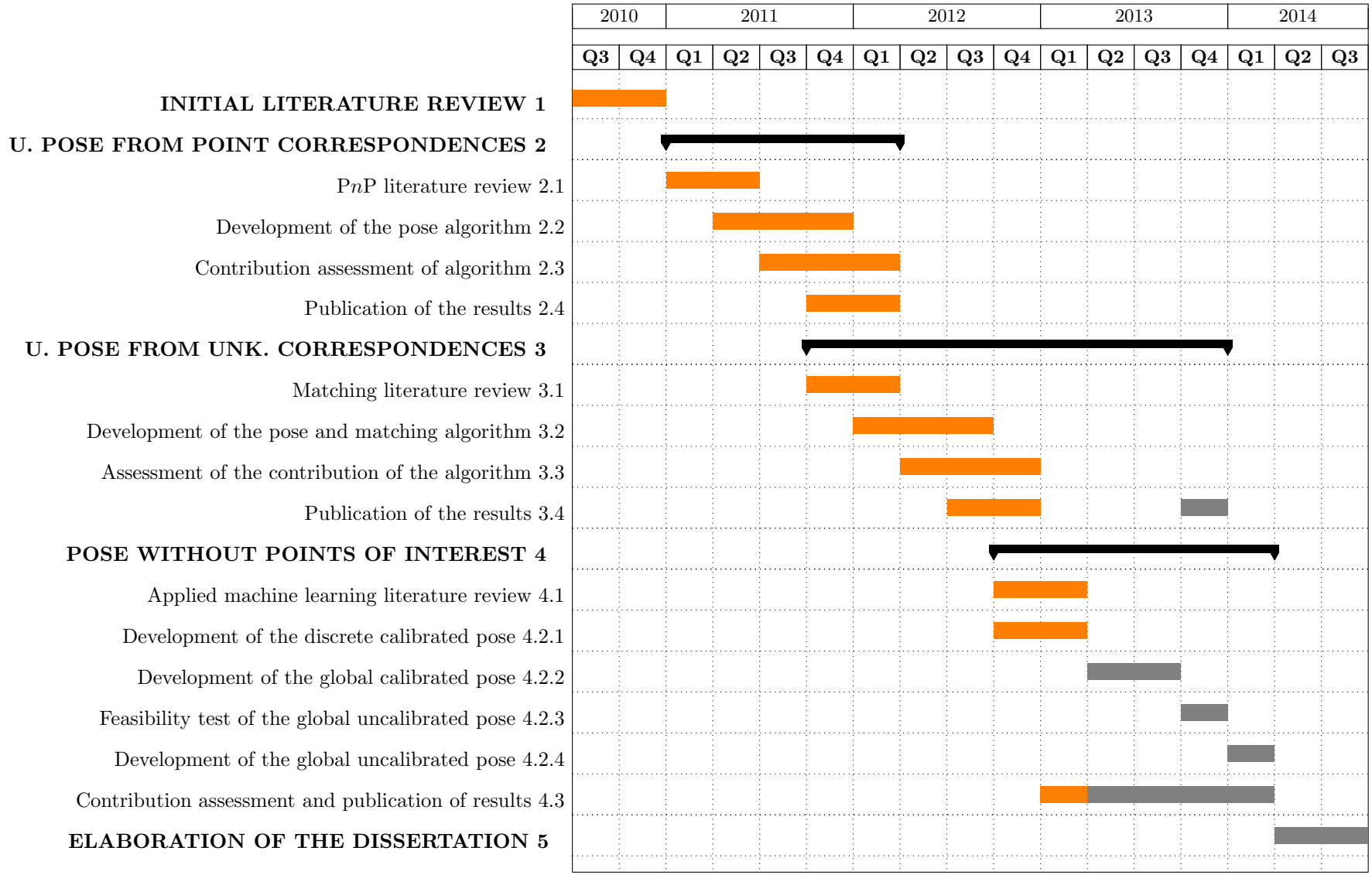

Figure 10: Work plan of the proposed work

References

[1] A. Ansar and K. Daniilidis. Linear pose estimation from points or lines. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 25(5):578–589, 2003.

[2] M. Bujnak, Z. Kukelova, and T. Pajdla. A general solution to the P4P problem for camera with unknown focal length. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[3] M. Bujnak, Z. Kukelova, and T. Pajdla. New efficient solution to the absolute pose problem for camera with unknown focal length and radial distortion. In *Asian Conference on Computer Vision. Vol. 6492 of Lecture Notes in Computer Science*, pages 11–24, 2010.

[4] M. Byrod, Z. Kukelova, K. Josephson, T. Pajdla, and K. Astrom. Fast and robust numerical solutions to minimal problems for cameras with radial distortion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[5] Michael Calonder, Vincent Lepetit, Mustafa Özuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua. Brief: Computing a local binary descriptor very fast. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 34:1281–1298, 2012.

[6] Tat-Jun Chin, Jin Yu, and David Suter. Accelerated Hypothesis Generation for Multi-Structure Robust Fitting. In *European Conference on Computer Vision*, 2010.

[7] K. Choi, S. Lee, and Y. Seo. A branch-and-bound algorithm for globally optimal camera pose and focal length. *Image and Vision Computing*, 28(9):1369–1376, 2010.

[8] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 220–226, 2005.

[9] Philip David, Daniel DeMenthon, Ramani Duraiswami, and Hanan Samet. SoftPOSIT: Simultaneous Pose and Correspondence Determination. *International Journal of Computer Vision*, 59(3):259–284, 2004.

[10] D. DeMenthon and L.S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15(1-2):123–141, 1995.

[11] O. Enqvist, K. Josephson, and F. Kahl. Optimal Correspondences from Pairwise Constraints. In *International Conference on Computer Vision*, 2009.

[12] G. Fanelli, J. Gall, and L. van Gool. Real time head pose estimation with random regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 617–624.

[13] P.D. Fiore. Efficient linear solution of exterior orientation. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 23(2):140–148, 2001.

[14] M.A. Fischler and R.C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[15] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng. Complete solution classification for the perspective-three-point problem. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 25(8):930–943, 2003.

[16] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In *International Conference on Computer Vision*, 2011. 1275–1282.

[17] R. Hartley and F. Schaffalitzky. L_∞ Minimization in Geometric Reconstruction Problems. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[18] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.

[19] R. Horaud, F. Dornaika, B.t Lamiroy, and S. Christy. Object pose: The link between weak perspective, paraperspective, and full perspective. *International Journal of Computer Vision*, 22(2):173–189, 1997.

[20] W. Hu and S.-C. Zhu. Learning a probabilistic model mixing 3D and 2D primitives for view invariant object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 2273–2280.

[21] K. Josephson and M. Byrod. Pose estimation with radial distortion and unknown focal length. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2419–2426, 2009.

[22] F. Kahl, S. Agarwal, M. Chandraker, D. Kriegman, and S. Belongie. Practical global optimization for multiview geometry. *International Journal of Computer Vision*, 79(3):271–284, 2008.

[23] F. Kahl and R. Hartley. Multiple-view geometry under the L_∞-norm. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(9):1603–1617, 2008.

[24] Z. Kukelova, M. Bujnak, and T. Pajdla. Polynomial eigenvalue solutions to the 5-pt and 6-pt relative pose problems. In *British Machine Vision Conference*, pages 56.1–56.10, 2008.

[25] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 28(9):1465–1479, 2006.

[26] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An accurate $O(n)$ solution to the PnP problem. *International Journal of Computer Vision*, 81(2):151–166, 2008.

[27] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1688–1695.

[28] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[29] C.P. Lu, G.D. Hager, and E. Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 22(6):610–622, 2000.

[30] Microsoft Corp. Redmond WA. Kinect for Xbox 360. 2010.

[31] F. Moreno-Noguer, V. Lepetit, and P. Fua. Accurate non-iterative $O(n)$ solution to the PnP problem. In *International Conference on Computer Vision*, pages 1–8, 2007.

[32] F. Moreno-Noguer, V. Lepetit, and P. Fua. Pose priors for simultaneously solving alignment and correspondence. In *European Conference on Computer Vision*, volume 2, pages 405–418, 2008.

[33] Kai Ni, Hailin Jin, and Frank Dellaert. GroupSAC: Efficient Consensus in the Presence of Groupings. In *International Conference on Computer Vision*, pages 2193–2200, 2009.

[34] David Nistér. Preemptive RANSAC for Live Structure and Motion Estimation. In *International Conference on Computer Vision*, pages 199–206, 2003.

[35] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[36] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 778–785.

[37] N. Payet and S. Todorovic. From contours to 3D object detection and pose estimation. In *International Conference on Computer Vision*, 2011. 983–990.

[38] Adrian Penate-Sanchez, Juan Andrade-Cetto, and Francesc Moreno-Noguer. Exhaustive linearization for robust camera pose and focal length estimation. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 99, 2013.

[39] Rahul Raguram, Jan-Michael Frahm, and Marc Pollefeys. A Comparative Analysis of RANSAC Techniques Leading to Adaptive Real-Time Random Sample Consensus. In *European Conference on Computer Vision*, pages 500–513, 2008.

[40] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

[41] M. Salzmann, V. Lepetit, and P. Fua. Deformable surface tracking ambiguities. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[42] G. Schweighofer and A. Pinz. Globally optimal O(n) solution to the PnP problem for general camera models. In *British Machine Vision Conference*, 2008.

[43] E. Serradell, M. Özuysal, V. Lepetit, P. Fua, and F. Moreno-Noguer. Combining Geometric and Appearance Priors for Robust Homography Estimation. In *European Conference on Computer Vision*, pages 58–72, September 2010.

[44] K. Sim and R. Hartley. Removing Outliers Using the L_∞ Norm. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 485–494, 2006.

[45] H. Stewenius, D. Nister, F. Kahl, and F. Schaffalitzky. A minimal solution for relative pose with unknown focal length. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 789–794, 2005.

[46] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *International Conference on Computer Vision*, 2009. 213–220.

[47] T. Thang-Pham, T.J. Chin, J. Yu, and D. Sutter. The random cluster model for robust geometric fitting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 710–717, 2012.

[48] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiel, and L. Van Gool. Towards multi-view object class detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 1589–1596.

[49] Ben Tordoff and David W. Murray. Guided-MLESAC: Faster Image Transform Estimation by Using Matching Priors. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 27(10):1523–1535, 2005.

[50] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[51] B. Triggs. Camera pose and calibration from 4 or 5 known 3d points. In *International Conference on Computer Vision*, pages 278–284, 1999.

[52] T. Trzcinski and V. Lepetit. Efficient Discriminative Projections for Compact Binary Descriptors. In *European Conference on Computer Vision*, 2012.

[53] A. Vedaldi and B. Fulkerson. Vlfeat – an open and portable library of computer vision algorithms. 2010. 1469–1472.

[54] Michael Villamizar, Helmut Grabner, Francesc Moreno-Noguer, Juan Andrade-Cetto, Luc Van Gool, and Alberto Sanfeliu. Efficient 3D object detection using multiple pose-specific classifiers. In *British Machine Vision Conference*, pages 20.1–20.10, 2011.

[55] B. Williams, G. Klein, and I. Reid. Real-time SLAM relocalisation. In *International Conference on Computer Vision*, pages 1–8, 2007.

[56] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 1998.