

Cooperative CONDENSATION-based Recognition

D Santosh Kumar, C V Jawahar
Center for Visual Information Technology
International Institute of Information Technology
Hyderabad 500032, India
{santosh@students.,jawahar@}iiit.ac.in

Abstract. This report presents a new technique for visual matching of images for the purpose of object recognition and shape matching by employing a multi-hypothesis approach. The approach is based on an analogy between multi-robot localization and object recognition. The algorithm is recursive with each individual step verifying fragments of different objects while guiding the focus towards discriminative regions and the sequence of resulting steps producing the overall recognition result. The method provides a novel framework to utilize the complimentary techniques developed in the field of multi-robotics to improve the solutions to conventional recognition problems. Experimental results on standard image databases validate our approach and demonstrate its efficacy.

1 Recognition Vs Multi-Robot Localization

Object recognition is an important task in computer vision. Different schemes have been proposed in the past for performing the recognition task, ranging from nearest neighbor search and Hough-based grouping [10] to the use of vocabulary trees [11]. Although these approaches have helped us gain a state of maturity in this field, they are limited in their ability to discriminate among objects with high appearance similarity. Most approaches use feature descriptors to distinguish between objects, which only describe a small amount of information that could be used for discrimination. For instance, SIFT-based features are not detected at edges and other salient regions, which are crucial in the identification of the objects. Another drawback of these approaches is that the image is processed in a sequence of steps avoiding the use of feedback information. Feedback helps in guiding the focus towards distinguishable regions enabling discrimination even in presence of high appearance similarity. While there has not been much research to tackle these problems in the recognition community, there has been significant amount of work in the recent past in the area of multi-robotics, addressing similar issues in the context of robot localization [6, 14]. In this paper, the complimentary techniques developed in this field are adapted to tackle the above problems.

The motivation behind our approach is based on the realization that recognition and multi-robot localization attempt to solve similar problems. Analogous to the way a swarm of robots explore their surroundings and find a match between their local map (deduced from their sensor data) to the global map provided to

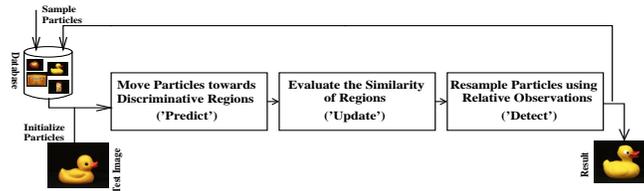


Fig. 1. Schematic description of the approach

them; in object recognition, a swarm of ‘points of focus’ explore an image and find its corresponding match in the image database based on the local features extracted by them. The additional complexity here is introduced by the projective distortions of the input image rather than simple rigid transformations.

Our approach utilizes the recent advances in the field of collaborative multi-robot localization [6] and exploration [3] to perform the recognition task. An initial set of feature correspondences is first generated. The method anchors on them and gradually explores the surrounding areas, trying to construct more and more matching features. As the extent of matched features increases, so does information available to judge their individual correctness. Overtime, the confidence of the correct model image increases. In addition, coordination between different focus points is introduced by way of sharing the knowledge of their relative poses. In similar images, this implies that particles can quickly discover a disambiguating region. This aspect is exploited to direct the focus towards such regions to yield a high recognition accuracy in spite of similarities in appearance. Fig. 1 illustrates our approach.

2 Cooperative CONDENSATION-based Techniques

CONDENSATION-based techniques belong to the general class of Monte Carlo filters [14] that have been recently used in computer vision for target-tracking [8], vision-based localization [5] etc. In the context of localization, CONDENSATION is used to estimate the pose x (position and orientation) of a robot at the current time-step t , given the knowledge about the initial state and all measurements $Z_t = z_t; i = 1 \dots t$ up to the current time. This is done by approximating the density (or *belief*) $p(x_j|Z^t)$ with a set of random particles drawn from it. Initially, the particles are spread uniformly in the environment and as the robot explores its surroundings and acquires new measurements, the particles converge to the actual pose. This process is carried out iteratively in two phases, namely *prediction* and *update*. For a detailed review, the reader may refer to [5, 14].

However in certain cases, the samples cannot converge to the correct density if only a single belief is maintained to sample the PDF. Specifically, in case of localization in a large unknown environment, a single robot takes long time to localize itself accurately as it requires to explore large parts of its surroundings to gain sufficient measurements. Further in presence of symmetries or ambiguities in the environment, localization is not possible without additional *a priori*

information. To circumvent this problem, the idea of multiple robots simultaneously localizing in an environment was proposed in [6]. Using several robots introduces a redundancy in the framework, which not only makes it more fault-tolerant but also helps to accomplish the task faster. In this case, a single belief over all robot locations *i.e.*, $L = L_1 \times L_2 \times \dots \times L_n$ is computationally demanding. To circumvent this problem, a factorial representation is employed for updating the beliefs, which assumes that the overall distribution is the product of its \mathcal{N} marginal distributions *i.e.*, $P(L_1, \dots, L_n|d) = P(L_1|d) \dots P(L_n|d)$ [6]. Apart from a *prediction* and *update* phase, there is now additionally a *detection* phase.

Detection Phase When one robot determines the location of another relative to its own, information from one belief function is transferred to another and both robots refine their internal beliefs based on each other’s estimate using an observation (or detection) model as

$$Bel_n(x_t) = Bel_n(x_t) \int p(\mathcal{X}_n = x_t | \mathcal{X}_m = x'_t, r_m). Bel_m(x'_t) dx'_t. \quad (1)$$

Here $p(\mathcal{X}_n = x_t | \mathcal{X}_m = x'_t, r_m)$ specifies the conditional probability of how the weights of the particles of robot n should be updated when robot m has detected it at a relative pose r_m . In addition to the motion model, a motion strategy also has to be defined that describes how the robots should explore their environment. Rather than independent motion of each robot, coordinated motion would better explore the environment avoiding overlaps between their paths and facilitating larger areas to be explored in a shorter time [2, 3].

Cooperative strategies have gained recent popularity in computer vision [16] and related areas [2]. In [16], cooperation was utilized for ‘mutual calibration’ where panoramic cameras on two cooperative moving platforms are dynamically calibrated by looking at each other. As summarized in [2], cooperation is not just that different techniques are needed to tackle different aspects of a complex task; rather complementary techniques can assist each other, thus extending the total effectiveness.

3 Cooperative CONDENSATION-based Recognition

A single robot cannot localize itself accurately in presence of symmetries in the environment; similarly, a single point of focus cannot discriminate between similar objects in the database (See Fig. 2). Thus multiple ‘cooperating’ points of foci are employed to accomplish the recognition task. We refer to the resulting approach as Cooperative CONDENSATION-based Recognition. In the same way

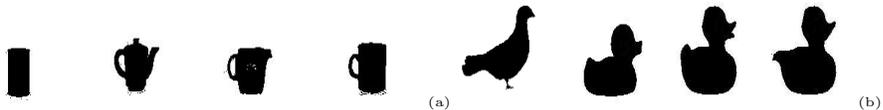


Fig. 2. Each sub-figure shows similar class of objects in our dataset. Discrimination among these objects is difficult without employing cooperative strategies

as multi-robot localization uses a set of particles to approximate a probability distribution for the robots’ positions in an environment, multiple particles are propagated to localize the pose of *points of focus* in an image, which iteratively converge to the correct model image. The different mechanisms involved in performing the recognition task are explained below.

Representation A ‘point of focus’ is described simply using its pose \mathcal{P} in the input image, while its corresponding particles are represented using their pose \mathcal{P}' , the index of the model image i and the affine mapping \mathcal{A} relating the input to the model image *i.e.*, $(\mathcal{P}', \mathcal{A}, i)$. As the input image can be deformed by an affine transformation, the affine mapping is necessary in this context (unlike in the case of localization). Localization now implies identifying the correct model image i and finding the corresponding pose within it for different focus points iteratively (See Fig. 3).

Prediction and Update Phase In analogy with the formal filtering problem, the algorithm proceeds in two phases. In the first phase, when a focus point undergoes a movement, the motion model is used to predict the new pose of the particles x_t from the set of particles computed in the previous iteration. The particles’ motion is dependent on the affine transformation relating its pose to the focus point. In the second phase, a measurement z_t is obtained by applying a feature extractor around the particle and the measurement model $p(z_t|x_t)$ is used to obtain the posterior belief. The new set of particles are now obtained by re-sampling this weighted set (See Fig. 3(b) and Fig. 3(c)).

Detection In the task of localization, additional sensors are employed by a robot to detect the presence of other robots in its vicinity [6]. However, in the current context, as the relative pose of the focus points is already maintained in their state information, no separate detection process is required. By directly using the knowledge of the points’ state, the relative pose between them can be deduced. Nevertheless, in case of an unsegmented image, a focus point should be constrained to detect only other points that are exploring the same object in the test image. To ensure this ‘visibility’ constraint, a coarse edge segmentation is performed to extract high-level boundaries of various objects and a focus point is allowed to detect other points that lie within the same contour.

Detection Phase When a focus point n is detected by the m th focus point and the relative pose between them is r , the beliefs can be updated using the incremental update equation as in (1). Here the detection model D denotes the m th particle’s belief about the detected n th particle’s pose. We use a normal distribution centered around r as our detection model. The same process is applied to constrain the m th particle’s pose based on the belief of the n th particle. Notice that (1) requires multiplication of two sample sets. Since samples in $Bel_n(x_t)$ and $Bel_m(x_t)$ are drawn randomly, it is not straightforward to establish correspondence between individual samples in both the distributions. The fusion of such an observation is done by transforming the detection model into a density tree and the density values obtained from this tree are then multiplied into the weights of the samples of the detected point [14]. Fig. 3(c) demonstrates the

effect of a detection phase, where the reduction in uncertainty of the particles' pose can be observed.

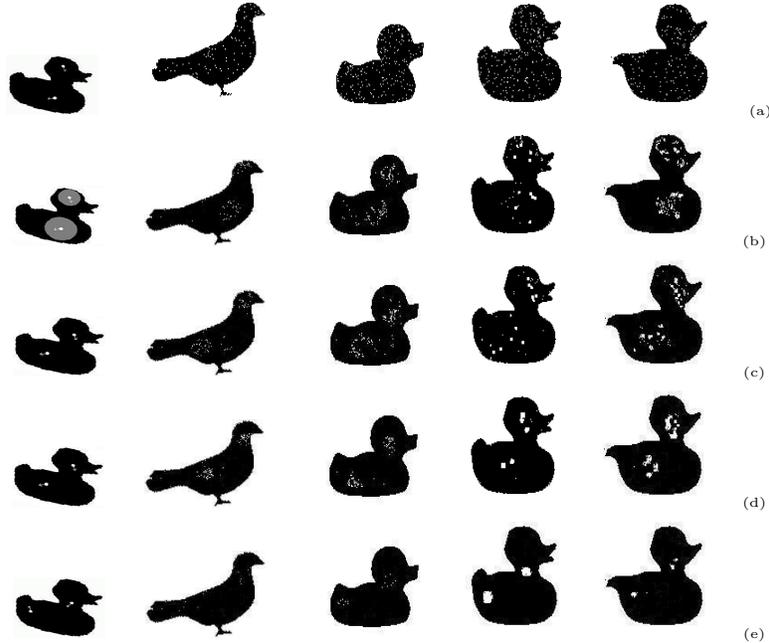


Fig. 3. Graphical illustration of one complete iteration of the algorithm using two focus points. The first figure in each row shows the test image, while the rest four display the candidate images. (a) Sampling of particles (b) Measurement update (c) Motion update (d) Detection update (e) Result after third iteration. In spite of similarities, discrimination is accomplished in very few iterations. Observe in (d) that uncertainty of focus point (in breast region) is reduced by its detection by the localized focus point (in head region).

Re-sampling After each iteration, all particles that are connected to a focus point are re-sampled based on their posterior beliefs, no matter which model image they belong to. Consequently, a wrong model image loses all its particles, which get clustered in the correct candidate image. Further, re-sampling is also performed on the focus points. Notice that as the algorithm progresses, the number of particles attached to a focus point exploring an incorrect image region decrease. The re-sampling process ensures that such points are eliminated and the weights of other focus points are renormalized, thus directing the focus towards the regions of interest. This equips the method to tackle unsegmented images.

Coordinated Motion Strategy A key question when employing multiple foci is to move them in an optimal manner to explore the entire image and facilitate faster and improved discrimination. The problem here is to not only devise an efficient motion strategy for a single focus point but also coordinate the actions of multiple points. The idea is to ‘actively’ choose different actions for the individual foci so that they simultaneously explore different areas of the input image [3]. To

achieve this, we maintain a global map not only to keep track of explored areas but also to plan and coordinate paths of different foci. Occupancy grids [14] are employed to integrate information from across multiple focus points. In this context, the cells of the grid correspond to the pixels in the image. The explored area by each point is tracked and the possible target locations are identified using a frontier cell approach [14]. Instead of moving all particles to the target points that have the minimum travel cost, the ‘utility’ of unexplored positions is also considered. It provides information about the area that is expected to be ‘visible’ when a focus point reaches a target point. If a focus point has already moved to a particular frontier cell, the utility of that cell (and other cells in its vicinity) can be expected to be lower for other focus points. Based on this information-theoretic measure, different target positions are allotted for the focus points, thus avoiding multiple foci to cluster in a small region.

Algorithm 1 Determining the Optimal Action for each focus point

```

Determine the set of frontier cells
Compute for each focus point  $i$  the cost  $V_t^i$  for reaching each frontier cell
Set the utility  $U_t$  of all frontier cells to 1
while there is one focus point left without a target point do
    Determine a focus point  $i$  and a frontier cell  $t$  which satisfy:  $(i, t) =$ 
         $arg \max_{(i', t')} (U_{t'} - V_{t'}^{i'})$ 
    Reduce the utility of each target point  $t'$  in the visibility area according
        to  $U_{t'} \leftarrow U_{t'} - P(\|t - t'\|)$ 
end while
end

```

The cost of reaching the frontier cells is directly proportional to the distance between the current and the target cell. It is computed using a variant of the value iteration scheme [14]. The utility of a cell depends on the probability that the cell is visible from a target cell that is assigned to another focus point. If a target point t' is selected for a particular focus point, the utility of the adjacent frontier cells at distance d from t' is reduced according to a probability distribution $P(d)$. In our experiments, we set $P(d)$ as $1.0 - \frac{d}{maxRange}$, where $maxRange$ is the maximum possible distance between two foci. Thus every focus point is assigned to a target location that has the best trade-off between the utility of the location and cost of reaching it. Algo. 1 summarizes the devised motion strategy. Observe in Fig. 3(b,c), that the focus point exploring the breast region of the bird could have chosen a target point towards the neck region. This would have made both the focus point cluster together, thus delaying the discrimination of the objects. By introducing the modified strategy, the focus point was directed towards the tail region facilitating faster discrimination. It must be noted that the above devised strategy is inspired from the exploration algorithms in the robotics domain. Simpler strategies may be formulated in the current context as the motion of focus points is less constrained than that of robots.

4 Experimental Results and Analysis

In this section, we present sufficient analysis to demonstrate the advantages of our approach and also compare it to the state of the art methods. The approach was implemented and tested on standard image databases. The basic implementation of the algorithm can be summarized into the following steps.

1. Initialize focus points in the test image and its corresponding particles in the model image
2. For each focus point V , perform the following steps
3. Obtain the optimal action for V using Algo. 1 and move V accordingly
4. For each corresponding particle m associated with V , perform the following steps
 - Predict its new pose using the Motion Model
 - Update its weight using the Measurement Model
5. Perform a detection and update the weights using the Detection Model
6. Re-sample the particles based on their posterior beliefs

We report our results on the Amsterdam Library of Object Images [7] (ALOI) and the Columbia Object Image Library [1] (COIL). A combination of images (for each object) were chosen as the model images, while another set of images were selected as test cases. Before analyzing our results, we describe the affine mapping, the measurement model and the motion model employed in the experiments.

Affine Mapping For every particle, the affine mapping relating the particle’s pose in the model image to its focus point in the input image is estimated. This mapping represents the hypotheses that the pixels in the model image (around the particle) have to be transformed in order to make the input and model image match each other. It is computed by describing two equilateral triangles, one in the input image and the other in the model image, where the triangle is defined such that the centroid coincides with the location of the particle and one of its vertices coincides with its orientation (See Fig. 4(a)). The transformation is computed such that the vertices of the triangle in the test image map to those in the model image. Depending on the pose of the focus point and the particles, different hypotheses arise. The reason behind computing an affine rather than a projective mapping is that the local image deformations can be reasonably well approximated by affine transformations.

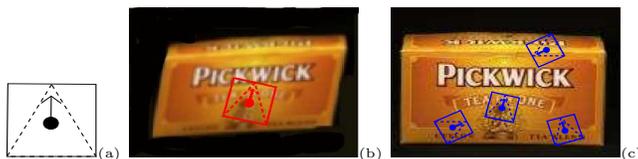


Fig. 4. Measurement Model: (a) Rectangle shows the region in which the histogram of oriented gradients is computed while the triangle displays the affine mapping. The focus point (red) in the input image (Fig.b) has the particles (blue) associated with it in the model image (Fig.c)

Measurement Model The weight of a particle is updated by evaluating the similarity of the measurement acquired by it with that of its corresponding focus point. The measurement is obtained by computing the histogram of oriented gradients (a robust region descriptor invariant to pose and illumination variations) [4] in a rectangular region around the particles. The weight of the particle is set based on the difference in the histograms, subject to the affine transformation (See Fig. 4(b,c)).

Motion Model When a focus point performs a movement in the input image, all its particles will also perform a corresponding movement in the model image, subject to the affine transformation. More specifically, given the translation and rotation of the focus point, the transformation is modified by the particle’s current estimate of the affine transformation and then applied to it.

To handle the large size of the object database, we retrieved the top few candidate images to the given input image from the database using the approach described in [11] and applied our algorithm on the retrieved images. This step limited the number of images to be considered by the algorithm. A set of 500 particles were used for each new point of focus and the number of foci were varied to analyze the performance of the algorithm. The particles were sampled at the initial interest point matches computed using PCA-SIFT [9] as these locations have a high probability of being correct. More precisely, a focus point is placed at each key-point in the input image and is connected to the particles sampled at possible matches in the model images. On the ALOI, a recognition accuracy of 94% was achieved, while an accuracy of 98% was achieved on the COIL. Fig. 3 shows one particular instance of our approach on binary images as an application towards shape matching and recognition. Some sample images of this dataset were shown in Fig. 2. In this case, the measurement model was implemented by letting the particles perform a range scan, measuring the distance to the first edge for each scan line. Notice that although the test image was affine distorted, the algorithm could still converge to the correct model image.



Fig. 5. Each sub-figure shows objects that are largely similar in appearance yet with some distinguishable regions in our dataset. Cooperative techniques are required to achieve accurate recognition result within few iterations.

For comparison, we considered the recognition algorithm based on SIFT-based features (as described in [10]). These algorithms extract local invariant features from both the input and the model images and then match them using

nearest-neighbor techniques. In this case, the considered dataset was a subset of objects selected from the ALOI and COIL database that have high similarities in their appearance and thus provide a challenging test case (See Fig. 5). Using this method, a recognition rate of 65% was obtained. The weak performance was due to the large viewpoint and scale changes apart from the deformations and highly visually-similar elements in the images of the selected imageset. This had resulted in the images having very few keypoint-matches. In addition, the SIFT-based method uses only three keypoint-matches to form the minimal valid cluster making it more likely of finding a matching group in a wrong model image. By applying our algorithm, we achieved a recognition accuracy of 90%. Higher discriminative power was achieved not only due to the richer source of information used (keypoint descriptors + original image data) but also because the decision about the object identity was based on the information densely distributed over the entire visible portion of the object in the test image. This suggests that our approach broadens the range of solvable recognition cases and is worth exploring. It must be emphasized that without cooperation, the particles could not localize themselves accurately and the correct match to the input image could not be found. By exploiting the information of the relative pose, the ambiguities could be easily resolved resulting in high recognition accuracy.

The utility of the coordinated exploration strategy was also evaluated. With a simple motion model (e.g., moving forward and reflecting at the image boundaries) and uncoordinated movements, the focus points got clustered in a small region and explored the same image parts. This resulted in a longer time for localization of the particles, delaying the recognition task. By the use of the frontier-based exploration strategy along with coordinated movements, the focus points could optimally explore the entire input image in few iterations. This facilitated faster discrimination accelerating the recognition task (See Fig. 6).

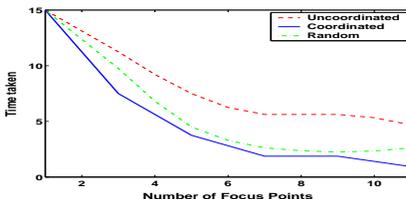


Fig. 6. Average time needed by a team of foci to explore the test image and identify the correct match. Observe that the introduction of coordinated movements accelerates the recognition process

5 Conclusions

In this paper, a novel framework for visual matching of images based on an analogy between the tasks of recognition and multi-robot localization has been developed. As an application of this framework, an object recognition algorithm that was able to discriminate among objects in spite of high appearance similarities was presented. The proposed approach falls under the general class of methods

employing a multi-hypothesis framework to perform the recognition task [12, 13, 15]. A related work to the current method was discussed recently in [15], where a Monte Carlo technique to perform the recognition task was proposed. Though the algorithm was demonstrated to be comparable to the state-of-the-art methods in its accuracy, it was not efficient in its performance due to its high time- and resource-consumption. Further discrimination in presence of appearance similarities could not be achieved. Our approach demonstrates higher (and faster) discriminative power due to the richer source of information used (key-point descriptors + original image data) and because the decision about the object identity is based on the information densely distributed over the entire visible portion of the object in the image. More importantly, the approach provides a means to apply the established techniques developed in the field of probabilistic robotics to improve the solutions to the recognition problems. We believe this framework would open new avenues in the area of recognition and retrieval.

References

1. Columbia object image library. <http://www.cs.columbia.edu/CAVE>.
2. A. Bundy. Cooperating reasoning processes: More than just the sum of their parts. *IJCAI*, pages 2–11, 2007.
3. W. Burgard, M. Moors, C. Stachniss, and F. Schneidery. Coordinated multi-robot exploration. *IEEE Transactions on Robotics*, 21(3):376–386, June 2005.
4. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE CVPR*, 1:886–893, 2005.
5. F. Dellaert, W. Burgard, D. Fox, S. Thrun. Using the CONDENSATION algorithm for robust, vision-based mobile robot localization. *IEEE CVPR*, 2:2588–2594, 1999.
6. D. Fox, W. Burgard, H. Kruppa, and S. Thrun. A probabilistic approach to collaborative multi-robot localization. *Autonomous Robots*, 8(3):325–344, June 2000.
7. J. M. Geusebroek, G. J. Burghouts, and A. M. Smeulders. The amsterdam library of object images. *IJCV*, 61(1):103–112, January 2005.
8. M. Isard and A. Blake. CONDENSATION - conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, August 1998.
9. Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *IEEE CVPR*, 2:506–513, 2004.
10. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004.
11. D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *IEEE CVPR*, 2:2161–2168, 2006.
12. Y. Owechko and S. Medasani. A swarm-based volition/attention framework for object recognition. *IEEE CVPRW*, 3:91-91, 2005.
13. B. Schiele and A. Pentland. Probabilistic object recognition and localization. *ICCV*, 1:177–182, 1999.
14. S. Thrun, W. Burgard, D. Fox. *Probabilistic Robotics*. MIT Press, Cambridge 2005.
15. F. v. Hundelshausen, H. J. Wunsche, M. Block, R. Kompass, and R. Rojas. Mesh-based active monte carlo recognition. *IJCAI*, pages 2231–2236, 2007.
16. Z. Zhu, D. R. Karuppiah, E. M. Riseman, and A. R. Hanson. Dynamic mutual calibration and view planning for cooperative mobile robots with panoramic virtual stereo vision. *Computer Vision and Image Understanding*, 95(3):261–286, 2004.