The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Using Feature Scale Change for Robot Localization along a Route

Andrew Vardy
Department of Computer Science
Faculty of Engineering & Applied Science
Memorial University of Newfoundland
St. John's, Canada
av@mun.ca

Abstract— **An appearance-based similarity measure for localizing a robot along a route is presented. This measure assesses the likelihood that the robot lies between a pair of positions where snapshot images were captured during training. The change in the scale parameter of matched SIFT features is used to determine whether the robot lies ahead or behind each snapshot. Experimental results in two different indoor environments suggest that this similarity measure will improve localization accuracy in situations where there is a large distance between snapshot positions.**

I. INTRODUCTION

Autonomous navigation of a robot along a trained route is an ability with a variety of potential applications, such as autonomous driving, security, and environmental monitoring. Two general approaches to the problem have emerged. The quantitative approach seeks to describe the position of the robot and all sensed landmarks within the same global coordinate frame. The qualitative or appearance-based approach describes the robot's location with respect to a set of stored sensory snapshots captured during training. Methods based on the quantitative approach generally adapt techniques for Simultaneous Localization and Mapping (SLAM) to the route following problem. Consequently, these methods also inherit the computational burden of trying to reconstruct the geometry of the route from a set of noisy samples—a process that often requires offline solution (e.g. [1]). However, it has been demonstrated recently that long routes can be followed robustly without requiring a correct global reconstruction, thereby reducing overall computational cost [2].

The qualitative approach requires no reconstruction, but it does require sensory snapshots to be captured along the route with sufficient frequency. Localization involves comparing the current sensory snapshot with some subset of the stored snapshots. This comparison of snapshots may operate directly on images or range data by using some sort of correlation measure [3], [4], [5]. It is perhaps more common to compare visual sensory snapshots by using image features such as vertical edges [6], KLT windows [7], or SIFT keypoints [8], [9]. Route following methods recently proposed by Chen and Birchfield [7] and Zhang and Kleeman [5] adhere to the qualitative framework and are particularly impressive. They have demonstrated robust performance along routes that are hundreds of metres in length. Nevertheless, both report that errors in localization do occur and occasionally cause their methods to fail.

Our approach to route following relies on visual homing to guide the robot from its current position to the next snapshot along the route [10], [11]. The route is represented by a sequence of snapshot positions at which images are captured during training (see figure 1). The method described in this paper estimates the probability that the robot lies between each pair of adjacent snapshot positions along the route.

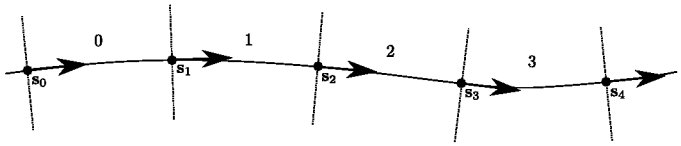


Fig. 1. A route is represented by a sequence of snapshot positions denoted $\mathbf{s}_i$. The arrows indicate the orientation of the robot at each $\mathbf{s}_i$. Directions perpendicular to the route through each $\mathbf{s}_i$ are given by dashed lines. Each segment of the route is given an integer label i corresponding to the pair of snapshot positions $(\mathbf{s}_i, \mathbf{s}_{i+1})$.

If the robot is believed to lie between two snapshot positions, $\mathbf{s}_i$ and $\mathbf{s}_{i+1}$ then to travel forwards along the route it should home to the image captured at position $\mathbf{s}_{i+1}$. If the wrong pair is chosen then the robot may go off the route. Some form of similarity measure is therefore required to compare the current image with all pairs of snapshot images. Inaccuracies in the similarity measure may be compensated by probabilistic techniques based on Bayes filter [12] (such as the Kalman filter used by Zhang and Kleeman [5]). Such techniques work well to filter out distant possibilities by considering the robot's motion constraints. However, if the similarity measure indicates that the robot is between $\mathbf{s}_i$ and $\mathbf{s}_{i+1}$ when it actually lies between $\mathbf{s}_j$ and $\mathbf{s}_{j+1}$ then a localization error can still occur if the previous belief combined with the motion model fails to rule out $(\mathbf{s}_i, \mathbf{s}_{i+1})$.

In this paper we describe a novel similarity measure based on the scale change of Lowe's SIFT features [13]. We utilize omnidirectional images as sensory snapshots. These images have forward-facing, and backward-facing halves. Prior to reaching the position of a snapshot along the route, we expect to see the same features that lay in the forward half of the snapshot image, only reduced in scale. Similarly, features in the back half of the snapshot image should appear enlarged in scale. Applying this style of reasoning for a

pair of snapshots allows us to compute a simple similarity measure that incorporates, the appearance, position, and scale of features along the route.

In the next section we describe our similarity measure. The notation used for the remainder of the paper is first presented, followed by a discussion of how feature scale change can be used as a stand-in for distance change measurements. The details of our localization technique are then described. We then present experimental results, discussion, and conclude with some directions for future work.

II. METHOD

A. Notation

Positions in space, such as the current position $\mathbf{c}$ and snapshot positions $\mathbf{s}_i$ and $\mathbf{s}_{i+1}$, will be indicated in bold lower-case. Images captured from these positions will be given in upper-case (e.g. C, S_i, and S_{i+1}). Features extracted from an image will be denoted with the same symbol, with a superscript giving the index of the feature. For example, S_i^j indicates the j^{th} feature extracted from image S_i.

B. Feature scale change

In the description of our method below, we make geometric arguments on the basis of whether a perceived feature has expanded or contracted. That is, whether the object that generated the feature is closer or further from the robot at the current position than at some reference position. As opposed to estimating the distance to the feature, we use the change in the scale parameter of SIFT features to indicate whether the feature has expanded or contracted. Consider C^j the j^{th} feature extracted from the current image:

$$C^j = \{C^{j,x}, C^{j,y}, C^{j,\theta}, C^{j,\sigma}, C^{j,d}\} \quad (1)$$

The feature's image location is $(C^{j,x}, C^{j,y})$, its orientation is $C^{j,\theta}$, its scale is $C^{j,\sigma}$, and its descriptor vector is $C^{j,d}$. As far as we are aware, our previous visual homing algorithm [11] was the first visual navigation method to make explicit use of the scale parameter $C^{j,\sigma}$ (henceforth referred to as σ if the context is clear). Here we use it again, only for the purpose of localization as opposed to homing. Informally, σ is the effective amount of Gaussian blurring required for a feature's distinctive characteristic to emerge (the distinctive characteristic being that the point is a local extrema with respect to both scale and space). Consider a landmark which yields one or more SIFT features. If the landmark is approached, it will take more blurring for the corresponding features to be detected. Thus, σ increases as the distance between the landmark and viewer decreases.

For our purposes we need only determine whether the distance to a landmark has increased or decreased with respect to a reference location. We utilize σ for this purpose. This substitution is valid as long as σ decreases monotonically as distance increases. Figure 2(a) shows a selection of panoramic images captured in the lobby of the S.J. Carew building at Memorial University. A total of 10 images were captured at increasing distances from a plaque on the wall. The top image shows the positions of SIFT features extracted from the vicinity of this plaque. Subsequent images show the matched features for images at distances of 2.4, 4.8, and 7.2m from the top image. Figure 2(b) shows the scale σ of matched features versus distance from the reference location. A clear trend of decreasing scale with increasing distance is observable. Although, there are a few exceptions such as the feature indicated by the heavy trace.

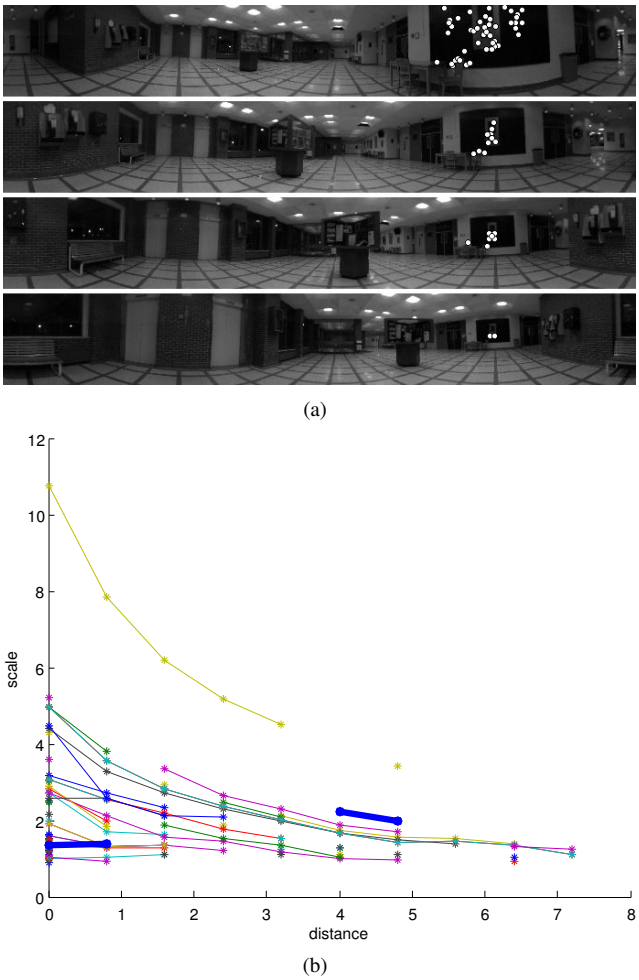


Fig. 2. (a) Images taken from the lobby of the S.J. Carew building of Memorial University. Overlaid are the locations of features extracted from the vicinity of a plaque on the wall. (b) Plot of the relation between spatial distance and feature scale for the features extracted from the top image in (a).

C. Concept for localization

As shown in figure 1, each snapshot position s_i has an associated heading, which is just the orientation of the robot at the time that the image S_i was captured. The heading at these positions is important as it is used to define the front and back of each snapshot image. Dashed lines through each s_i indicate the direction orthogonal to the heading. These lines divide the route up into segments labelled by index i. Each segment i is associated with the pair of snapshots $(\mathbf{s}_i, \mathbf{s}_{i+1})$. The true segment that the robot lies on is given by the discrete state variable x_t. Our task is to estimate x_t given the set of features extracted from the current image.

We define *sift* as a function that extracts a set of SIFT features from an image. Thus, the set of features from the current image is $(C) = \{C^j\}$. For each possible value of i we estimate $p\left(x_t = i|\{C^j\}\right)$ which can be decomposed according to Bayes rule:

$$p\left(x_t|\{C^j\}\right) = \frac{p\left(\{C^j\}|x_t\right)p(x_t)}{p\left(\{C^j\}\right)} \quad (2)$$

$$\propto p\left(\{C^j\}|x_t\right)p(x_t) \quad (3)$$

Since we are interested in the maximum value of $p\left(x_t|\{C^j\}\right)$ the term $p\left(\{C^j\}\right)$ is omitted. If we have no prior notion of the location of the robot then we have a *global localization* problem to solve. In this case $p(x_t)$ will be equal for all i and can be omitted. If we are *tracking* the movement of the robot over time then $p(x_t)$ can be replaced with $\overline{p}(x_t)$ representing the predicted probability of the robot lying at x_t, which is obtained by taking the estimate from the last time step and incorporating the robot's most recent movement [12]. In either case, we focus on $p\left(\{C^j\}|x_t\right)$ which is certainly required to determine $p\left(x_t|\{C^j\}\right)$.

To compute $p\left(\{C^j\}|x_t = i\right)$ we assume that the robot is located on segment i of the route between s_i and s_{i+1}. To be more specific, the robot lies in front of s_i and behind s_{i+1}.

$$p\left(\{C^j\}|x_t = i\right) = p\left(\{C^j\}|x_t > i-1\right) \cdot p\left(\{C^j\}|x_t < i+1\right) \quad (4)$$

The probability of computing the set of features $\{C^j\}$ from such a position depends on the features computed from images S_i and S_{i+1}. We consider first determining the probability of obtaining $\{C^j\}$ given that the robot is in front of $\mathbf{s}_i$. That is, we compute $p\left(\{C^j\}|x_t > i-1\right)$.

We employ omnidirectional images (cf. figure 2(a)) captured from a digital camera mounted upwards on a robot to point at a hyperbolic mirror. Since the forwards direction corresponds to a fixed position in the image, it is always possible to separate the image into its front and back halves. We separate the set of features $\text{sift}(S_i) = \{S_i^j\}$ into those from the front half of the image and those from the back. Let $\{F_i^j\}$ be the set of features from the front half of S_i and $\{B_i^j\}$ be the set from the back half of S_i.

$$\{F_i^j\} = \text{front}(\{S_i^j\}) \quad (5)$$

$$\{B_i^j\} = \text{back}(\{S_i^j\}) \quad (6)$$

Next we determine the set of correspondences between $\{F_i^j\}$ and $\{C^j\}$ and between $\{B_i^j\}$ and $\{C^j\}$. We utilize the standard match criterion described by Lowe [13] which accepts a match only if it is significantly better than the second closest match. This generates a set of correspondences M_{F_i} from $\{F_i^j\}$ to $\{C^j\}$ and M_{B_i} from $\{B_i^j\}$ to $\{C^j\}$. Both M_{F_i} and M_{B_i} are sets of ordered pairs (a, b) where a is the index of the reference feature (F_i^a or B_i^a) and b is the index of the matching feature from the current image C^b.

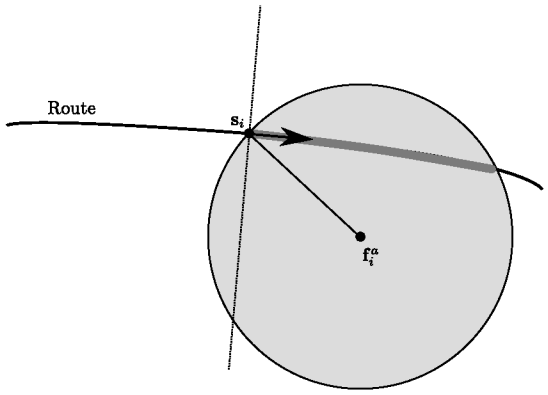
We now identify those features from F_i which have expanded in C and those from B_i that have contracted:

$$Exp(F_i) = \{F_i^a : (a,b) \in M_{F_i} \text{ and } F_i^{a,\sigma} \leq C^{b,\sigma}\} \quad (7)$$

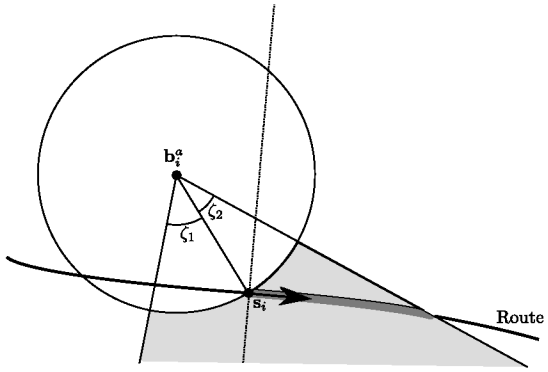
$$Con(B_i) = \{B_i^a : (a,b) \in M_{B_i} \text{ and } B_i^{a,\sigma} \geq C^{b,\sigma}\} \quad (8)$$

Each feature $F_i^a \in Exp(F_i)$ has expanded, meaning that the matching feature in the current image C^b has a larger value of σ. Consequently the distance from the robot to the object that generated the feature is smaller than the distance at $\mathbf{s_i}$. We denote the object that generated F_i^a as $\mathbf{f}_i^a$. The fact that F_i^a has expanded implies that $\mathbf{c}$ lies inside the circle centred at $\mathbf{f}_i^a$. This circle is shown in figure 3(a). The portion of the route within this circle is shaded dark grey. If the robot lies on the route in front of $\mathbf{s}_i$ we assume that features from the front of S_i will have expanded. Therefore the proportion of expanded features should be proportional to the probability that the robot lies along the route ahead of $\mathbf{s}_i$.

$$p\left(\{C^j\}|x_t > i-1\right) \propto \frac{|Exp(F_i)|}{|\{F_i^j\}|} \quad (9)$$



(a) Feature from the front half of S_i has expanded



(b) Feature from the back half of S_i has contracted

Fig. 3. (a) The feature F_i^a has expanded, meaning that $\mathbf{c}$ lies within the circle centred at the position of the feature-generating object $\mathbf{f}_i^a$. (b) The feature B_i^a has contracted, meaning that $\mathbf{c}$ lies outside the circle centred at $\mathbf{b}_i^a$. The feature is assumed to be visible only within the cone defined by ζ_1 and ζ_2. The light grey region indicates the possible locations of $\mathbf{c}$ with the region along the route shaded in dark grey.

A similar situation holds for contracted features in the back of the image. Each feature $B_i^a \in Con(B_i)$ has contracted, meaning that the distance to the feature-generating object $\mathbf{b}_i^a$ has increased. Each such feature was visible from $\mathbf{s}_i$ but we assume that its visibility is limited to a certain angular region. In other words, we assume that features are not radially symmetric. The angles of visibility ζ_1 and ζ_2 are shown in figure 3(b). A contracted feature from B_i indicates that the

robot lies outside of the circle centred at $\mathbf{b}_i^a$ which intersects $\mathbf{s}_i$. Such a feature must also lie within the cone of visibility defined by ζ_1 and ζ_2. The segment of the route within this region is shaded dark grey in figure 3(b). If the robot lies on the route in front of $\mathbf{s}_i$ we assume that features from the back of S_i will have contracted. Therefore,

$$p\left(\{C^j\}|x_t > i-1\right) \propto \frac{|Con(B_i)|}{|\{B_i^j\}|} \tag{10}$$

We combine equations 9 and 10 to obtain the following:

$$p\left(\{C^j\}|x_t > i-1\right) = \frac{|Exp(F_i)|}{|\{F_i^j\}|}\frac{|Con(B_i)|}{|\{B_i^j\}|} \tag{11}$$

It remains to compute $p\left(\{C^j\}|x_t < i+1\right)$, the probability of sensing the current set of features if the robot lies behind s_{i+1}. The logic for this case is symmetric to that described above and leads to the following expression:

$$p\left(\{C^j\}|x_t < i+1\right) = \frac{|Exp(B_{i+1})|}{|\{B_{i+1}^j\}|}\frac{|Con(F_{i+1})|}{|\{F_{i+1}^j\}|} \tag{12}$$

Combining equations 11 and 12 according to equation 4 yields the following:

$$p\left(\{C^j\}|x_t = i\right) = \frac{|Exp(F_i)|}{|\{F_i^j\}|}\frac{|Con(B_i)|}{|\{B_i^j\}|}\cdot \frac{|Exp(B_{i+1})|}{|\{B_{i+1}^j\}|}\frac{|Con(F_{i+1})|}{|\{F_{i+1}^j\}|} \tag{13}$$

We will refer to this measure of $p\left(\{C^j\}|x_t = i\right)$ as *scaleDiff* .

As a benchmark for comparison we use the average percentage of matched features from S_i to C and S_{i+1} to C.

$$\textit{percentMatched} = \frac{1}{2}\left(\frac{|M_{S_i}|}{|\{S_i^j\}|} + \frac{|M_{S_{i+1}}|}{|\{S_{i+1}^j\}|}\right) \tag{14}$$

In our experiments the cost of computing either *scaleDiff* or *percentMatched* is negligible in comparison to the cost of either extracting SIFT features or computing the matches between them.

III. EXPERIMENTAL RESULTS

A. Image sequences

The images used below were collected in the lobbies of the Inco Innovation Centre and the S.J. Carew building, both located on the campus of Memorial University. They were captured by a manually driven robot with an upward-facing camera directed at a hyperbolic mirror. The height of the mirror above the floor is approximately 45 cm. Images are sampled from the raw camera image to yield a rectangular image, with each row corresponding to a constant angular latitude above or below the horizon. An example image from the Inco centre is provided in figure 4. Figure 2(a) provides examples of the robot's view in the lobby of the Carew building. The Inco centre route was 25 m long with images captured every 50 cm. The Carew building route was 39 m long with a capture resolution of 1 m.

Fig. 4. Image 25 from image sequence *inco4* captured in the lobby of the Inco Innovation Centre of Memorial University.

IV. EXPERIMENTS

Figure 5 shows the performance of both *percentMatched* (on the left) and *scaleDiff* (on the right) on the Inco route when every second image is taken as a snapshot. The other images are used to define the test route. The asterisk indicates the actual robot position while the perpendicular diameter of each ellipse is proportional to the value of *percentMatched* or *scaleDifff*. Only some of the data is shown, but in all cases the true position corresponds to the position of maximum *percentMatched* or *scaleDifff*. From this perspective both similarity measures appear equivalent. However, it is clear that the values for *scaleDiff* are more tightly focused, indicating a much greater degree of confidence. The uniformity of the distribution of similarity values can be measured in terms of entropy. Let sim represent a similarity measure (either *percentMatched* or *scaleDiff*). The entropy over the set of n test images $\{T_i\}$ can be expressed as follows:

$$\text{entropy}(\{T_i\}) = -\sum_{i=1}^{n} \text{sim}(T_i)\log_2(\text{sim}(T_i)) \tag{15}$$

The average entropy values are given in the figures corresponding to each experiment (figures 5, 6, and 7). For all experiments described in this paper the average entropy of *scaleDiff* is much lower than for *percentMatched*.

We then considered the accuracy of localization when the distance between snapshots is increased. Tests were done on both routes with the number of images between snapshots increased to four or eight. For the Inco route this corresponds to a distance between snapshots of 2 m and 4 m, respectively, while on the Carew route it becomes 4 m and 8 m. In all cases we select all non-snapshot images to be used as test images.

Figures 6 and 7 show a selection of these results with 8 images between snapshots. The ideal behaviour is for the snapshot pair with maximum similarity measure (indicated by '+') to enclose the true robot position (indicated by 'x'). Otherwise, the estimated and true positions differ and we have a *fault*. No faults occur for either *percentMatched* or *scaleDiff* on the Inco route with 2 m between snapshots. For the Carew route with 4 m between snapshots *percentMatched* experiences 3 / 40 faults whereas *scaleDiff* experiences none.

The faults for the largest tested distance between snapshots (4m for Inco, 8m for Carew) are indicated with stars in figures 6 and 7. On the Inco route, *percentMatched* experiences 10 / 50 faults while *scaleDiff* experiences only one. One of the faults due to *percentMatched* is more serious in that the estimated position and the true position differ by two segments. This fault is indicated with a double star in figure 6. The sole fault for *scaleDiff* occurs when the angle of

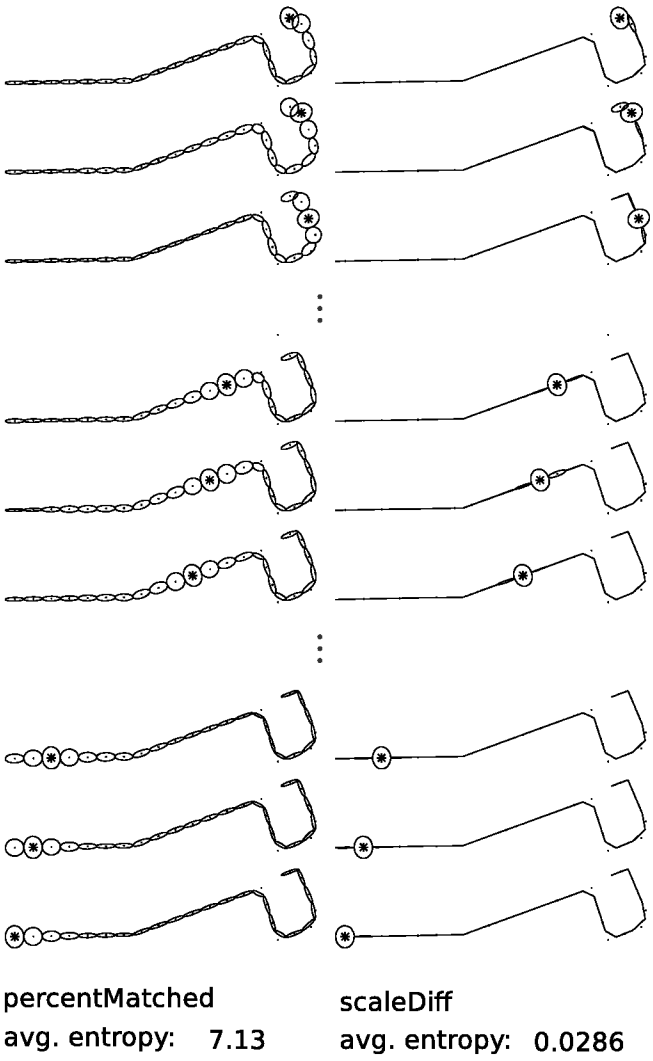


Fig. 5. Comparison of *percentMatched* and *scaleDiff* on the Inco route. Small dotes indicate positions along the test route. Ellipses extend between adjacent snapshot positions with their perpendicular diameters set to be proportional to the normalized value of each similarity measure.

the training route (indicated by the direction of the ellipses) differs strongly from the test route.

On the Carew route, *percentMatched* experiences 13 / 40 faults while *scaleDiff* experiences only two (both of which are shared with *percentMatched*).

V. DISCUSSION AND CONCLUSIONS

A. Discussion

The results indicate comparable performance of these two similarity measures when the distance between snapshots is small. However, as the distance between snapshots grows the number of faults generated by *percentMatched* appears to be much greater than those generated by *scaleDiff*. The only faults experienced by *scaleDiff* occur in the large bend visible on the right side of both routes. This situation can be avoided by increasing the capture frequency in regions with high curvature. Such a step is also necessary to improve the robot's adherence to the route in such regions.

B. Conclusions

We have presented a novel similarity measure which enhances the localization performance for a robot travelling along a route. The next step is to test this similarity measure in the context of a temporal filtering technique based on Bayes filter. We are also developing a strategy to interpolate this measure using stored feature scale information for intermediate nodes along the route for which no images are stored. It would also be interesting to compare the performance of our similarity measure with other recently proposed methods based on the visual bag-of-words framework [14], [15], [16].

VI. ACKNOWLEDGEMENTS

The author wishes to acknowledge the great help provided by several anonymous reviewers of the initial draft of this paper.

REFERENCES

[1] E. Royer, J. Bom, M. Dhome, B. Thuliot, M. Lhuillier, and F. Marmoiton, "Outdoor autonomous navigation using monocular vision," in *IEEE/RSJ IROS*, 2005.

[2] P. Furgale and T. Barfoot, "Visual teach and repeat for long-range rover autonomy," *Journal of Field Robotics*, 2010.

[3] Y. Matsumoto, M. Inaba, and H. Inoue, "Visual navigation using view-sequenced route representation," in *IEEE ICRA*, pp. 83–88, 1996.

[4] U. Nehmzow and C. Owen, "Robot navigation in the real world: Experiments with Manchester's fortytwo in unmodified, large environments," *Robotics and Autonomous Systems*, vol. 33, pp. 233–242, 2000.

[5] A. Zhang and L. Kleeman, "Robust appearance based visual route following for navigation in large-scale outdoor environments," *The International Journal of Robotics Research*, vol. 28, no. 3, pp. 331–356, 2009.

[6] L. Tang and S. Yuta, "Indoor navigation for mobile robots using memorized omni-directional images and robot's motion," in *IEEE/RSJ IROS*, 2002.

[7] Z. Chen and S. Birchfield, "Qualitative vision-based path following," *IEEE Transactions on Robotics*, vol. 25, no. 3, pp. 749–754, 2009.

[8] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. Van Gool, "Omnidirectional vision based topological navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 219–236, 2007.

[9] O. Booij, B. Terwijn, Z. Zivkovic, and B. Kröse, "Navigation using an appearance based topological map," in *IEEE ICRA*, 2007.

[10] A. Vardy, "Long-range visual homing," in *Proceedings of the IEEE International Conference on Robotics and Biomimetics*, IEEE Xplore, 2006.

[11] D. Churchill and A. Vardy, "Homing in scale space," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1307–1312, 2008.

[12] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2005.

[13] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.

[15] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Incremental vision-based topological slam," in *IEEE/RSJ IROS*, 2008.

[16] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

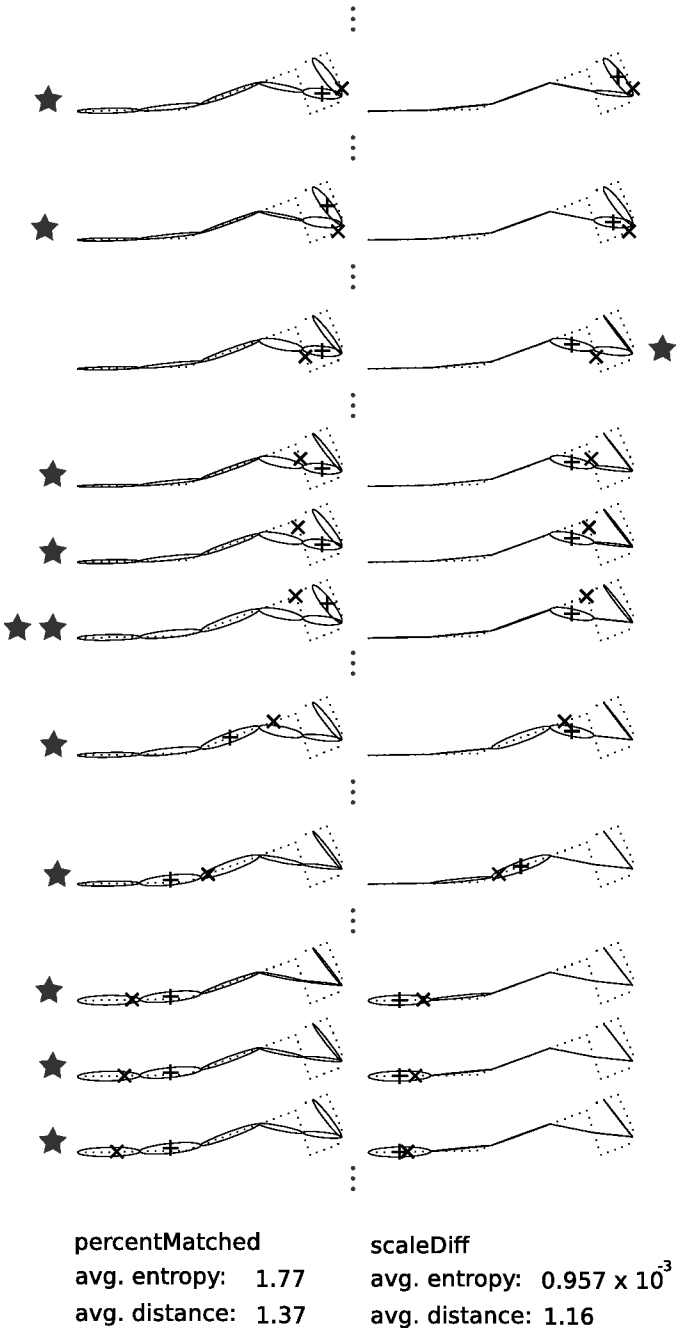


Fig. 6. Comparison of *percentMatched* and *scaleDiff* on the Inco route. Every eighth image from the manually trained route is selected as a snapshot image. The 'x' indicates the robot's current position. The '+' indicates the position of maximum *percentMatched* or *scaleDiff*. Stars indicate faults, as described in the text.

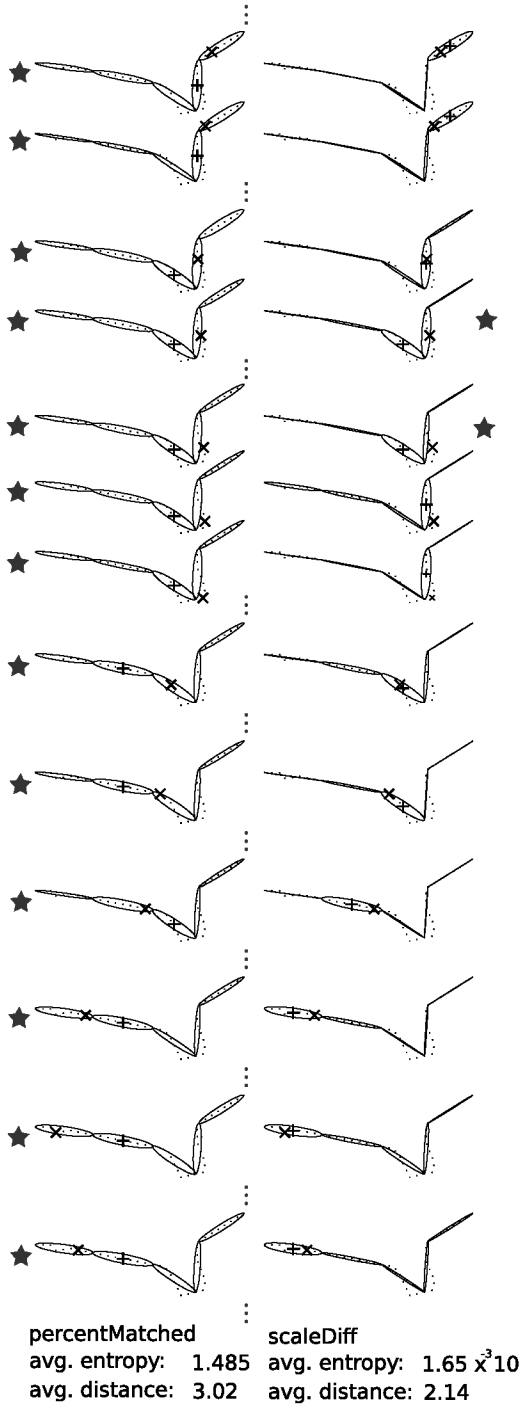


Fig. 7. Comparison of *percentMatched* and *scaleDiff* on the Carew route. Every eighth image from the manually trained route is selected as a snapshot image. See captions of figures 5 and 6 for notation.