University of Glasgow at ImageCLEF 2009 Robot Vision Task

Yue Feng, Martin Halvey and Joemon M. Jose
Multimedia Information Retrieval Group
University of Glasgow, Glasgow, G12 8RZ, United Kingdom
{yuefeng, halvey, jj}@dcs.gla.ac.uk

Abstract

The submission from the Multimedia Information Retrieval Group at the University of Glasgow for the ImageCLEF 2009 Robot Vision Task combines point matching methodologies with rule based decision techniques. Instead of the whole image we use a large set of interesting points extracted from the image contents to represent each image. The points of interest are extracted using an edge corner detector. The RANAC method [3] was then applied to estimate the similarity between the test and training images based on the number of matched pairs of points. The location of robot is then annotated based on the training image that contains the highest number of matched point pairs with the test image. A set of decision rules with the respect to the trajectory behaviour of robot's motion are defined to refine the final results. An illumination filter was also applied for two of the runs in order to reduce the illumination effect. Three runs were submitted using the different of combination of the above approaches.

Categories and Subject Descriptors

I.4 [Image Processing and Computer vision]: I.4.8 Scene Analysis; I.4.9 Applications; I.5 [Pattern Recognition]: I.5.4 Applications

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Robot Vision, Computer Vision, Illumination Filter, Decision Rules.

1. Introduction

In this paper, we describe the system and examine the approaches and results for the three independent runs submitted by the Multimedia Information Retrieval group at the University of Glasgow for the ImageCLEF 209 Robot Vision task.

The goal of the ImageCLEF 2009 Robot Vision Task is to address the problem of topological localisation of a mobile robot using visual information. Specifically, participants were asked to determine the topological location of a robot based on images acquired with a perspective camera mounted on a robot platform [2]. Given a test sequence, a system/algorithm must be able to provide information about the location of the robot, where the relevant location information is available in a training sequence. Our strategy is to analyse the visual content of the test sequence and compare it with the training sequence to decide the location. Given a set of training sequences which are annotated with location information and an unannotated video frame captured by the robot, the best solution for identifying the location of this unannotated frame is to find the closest frame in the training set. Two possible techniques could be applied to solve this problem: classification and image matching. The classification approach [6], which is famous for classifying images into one of a trained set of groups, according to feature distance to each group, requires an initial training step to obtain the class properties. However, its performance for identifying images belonging to an unknown class is not very good due to a lack of training samples, which are not available at the application stage. The image matching [7] approach is another potential solution, which is able to automatically and efficiently detect that two images are similar, or find similar images to a test image within a database of images. This approach estimates the visual distance between the unannotated frame and each frame in the training sequence and returns a ranked list, where the location of the unannotated frame could be annotated as the same as training image with the highest score in the rank list. In order to annotate the unknown frames, which are captured in an unknown room, the computed rank list is not reliable since all the training images on the list are all annotated. Thus, a similarity threshold could be introduced to filter out the false matches and annotate the unknown rooms. Thus, our intuition is that an image matching approach is a better approach than a classification approach for the ImageCLEF 2009 Robot Vision Task.

It is well known that when an object/robot moves through its environment it does not move randomly. Instead, it usually follows specific trajectories or motion patterns corresponding to its intentions [1]. Knowledge about such patterns can be used to assist any system to robustly keep track of robot in its environment and identify the location of the robot.

Motivated by the above work [1, 6, 7], our adopted approach combines image matching with a rule based decision approach. In addition, an illumination filter is integrated into one of the runs in order to minimise lighting effects with the goal of improving the predictive accuracy.

The rest of the paper is organised as follows. The next section introduces our proposed methodology for the robot vision task. In Section 3, we describe our submitted runs and the results of those runs. Finally, in Section 4 we present a conclusion.

2. Methodology

For our approach image matching is first applied to find the most closely matching frames from the training sequences in order to annotate the test frame. A rule based decision maker is then applied in order to refine the results based on the movement behaviour of the robot. In addition, an illumination filter is applied to pre-process the sequences in order to reduce the lighting effectiveness. Each of these steps is described in more detail in the following subsections.

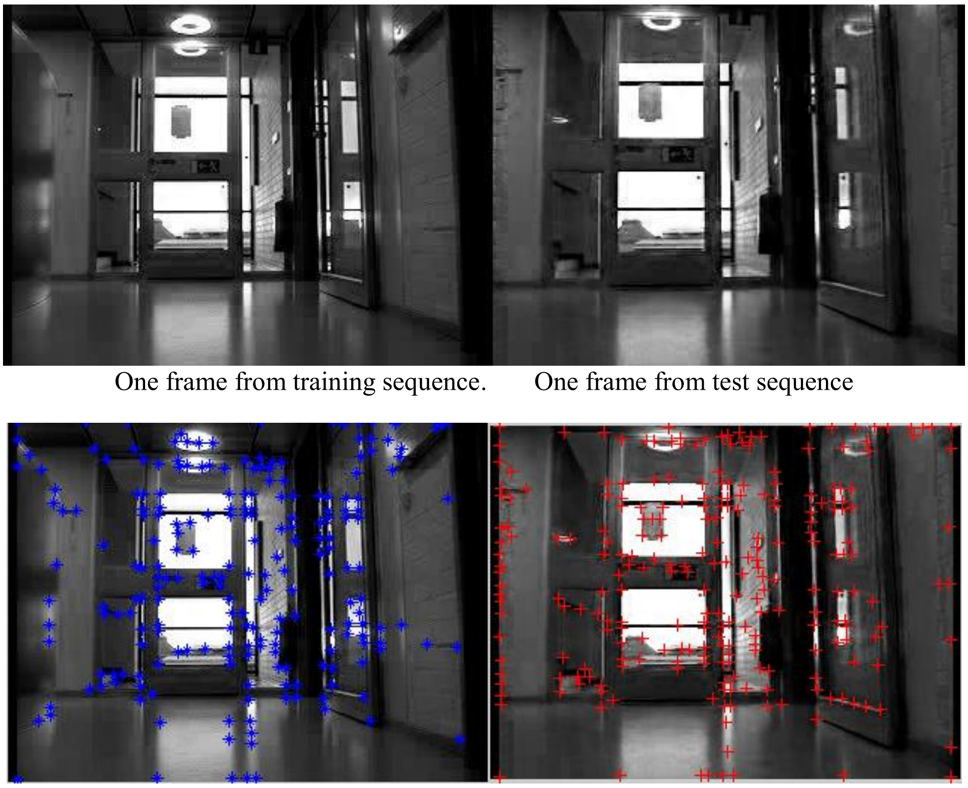
2.1. Points Based Image Matching

Considering both the training and test sequences are captured using the same camera in the same geographic condition, we assume that frames taken in the same location will contain similar content and geometric information.

Motivated by the above assumption, the proposed image matching algorithm is designed in the following successive stages: (1) A corner detection method is first used to create an initial group of points of interest; (2) The Ransac algorithm [3] is then applied to establish point correspondences between two frames and calculate the Fundamental matrix [4]. This matrix encodes an epipolar constraint which is applied to the general motion and rigid structure; this is used to compute the geometric information for refining the matched point pairs. (3) The number of refined matched points will be regarded as the similarity between two frames.

2.1.1. Points of Interest

In order to reduce the computation cost, we use points of interest (POI) rather than all pixels in the frame. Considering the fact that corners are local image features characterised by locations where variations of intensity in both *X* and *Y* directions are high, it is easier to detect and compare the POI within and around the corners, such as edges or textures. To this end, for this work we employ Harris corner detector [5] to initiate the POI, since it has strong invariance to rotation, scale, illumination variation and image noise. The Harris corner detector uses the local autocorrelation function to measure the local changes of the signal to detect the corner positions in each frame. More details can be found in [5]. Results of the corner detection are shown in Figure 1, where the detected corners are represented by '*' and '+'.



One frame from training sequence. One frame from test sequence

The detected corners of the above frames

Figure 1: The original frames and results for corner detection

2.1.2. Point Matching

The next step of our approach is to use a point matching technique to establish point correspondences between two frames. The point matching method generates putative matches between two frames by looking for points that are maximally correlated with each other inside a window surrounding each point. Only points that correlate strongly with each other in both directions are returned. Given the initial P points of interests, a parameter X is used for checking whether this point is fitted or not, is first estimated using N points chosen at random from P. It is then found how many points in P fit the model with values of X within a tolerance value T given by the user. If the number is satisfactory, it is regarded as a fit and the operation terminates with success. Such operations are carried on looping through all the POI. In the present work, T is set at 95%. Setting such a high threshold value reduces the number of points of interest, so only the points with 'good match' are kept.. Examples of matched points can be seen in Figure 2a, where a pair of matched points is represented by '*' and '+'.

The initial matching pair may contain many mismatches, thus a post processing step for refining the results is needed. Given the assumption in Section 2.1 that frames taken in the same location will contain similar geometric information, we applied the fundamental matrix [4]. The fundamental matrix was initially designed for finding corresponding points in stereo images, where it helps to find the matching pairs using the computed geometric information.

Given the initial matching points, the fundamental matrix F can be estimated given a minimum of seven point's correspondences. Its seven parameters represent the only geometric information about the cameras that can be obtained through points correspondences alone.

Given the computed *F*, we applied it on all the matching pairs to eliminate the incorrect matching pairs, where the matched point pair should satisfied with the following formula,

$$rTFx = 0$$

$r\,x'$ are the corresponding points in two images, *Fx* describes a line (an epipolar line), where the other corresponding point x' on the other image must lie. The results of the matching point pairs after the refining step are illustrated at Figure 2b.

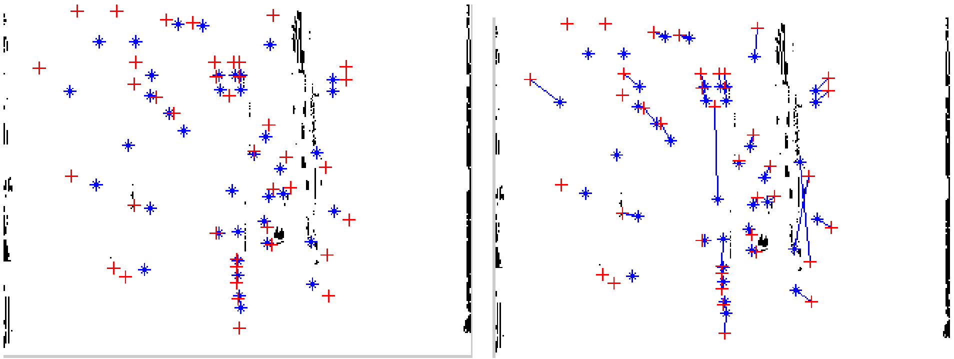


Figure 2: (Left most) the initial matching map and (right most) the matching map after filtering, the points linked with lines are correctly matched.

Thus, after applying the matrix *F* on all the paired points to filter out the mis-matched points, the number of matched point pairs remaining will be regarded as the similarity between two images.

In order to localise the robot, we assume that the position can be retrieved by finding the most similar frame in the training sequence to the test sequence. Thus, the most similar frame should contain the most matched points with the testing frame.

2.2. Decision Rules

Given the results of point matching, each test frame can be annotated as being from one of the possible rooms and the trajectory of the robot could be generated. The trajectory could be represented using the extracted annotation information frame by frame. An example (Example 1) of the extracted trajectory could be represented as:

ppppppppppppppppppppppppp	*period 1*
cccc	*period 2*
pp	*period 3*
cccccccccccccccccccc cccccccccccccccccccc	*period 4*
eeeee eeeee eeeee eeeee eeeee eeeee eeeee eeeee	*period 5*
kkkkkkkkkkkkk kkkkkkkkkkkkk	*period 6*
cccccccccc ccccccccccccc	*period 7*

Where *p*, *c, e* and *k* represent ‘printing area’, ‘corridor’, ‘kitchen’, and ‘two-person office’ respectively. In this example, the robot travels continuously from ‘printing area’ to ‘corridor’, then back to ‘printing area’, ‘corridor’, ‘two-persons office’, ‘kitchen’ and ‘corridor’. In order to study the movement behave of robot, each continuous moving shot is considered as one period, this sequence contains seven periods in different rooms.

By studying the training sequence released as part of the ImageCLEF 2009 Robot Vision training and test sets, we find (i) the robot does not move 'randomly', (ii) the period of time that the robot stays in the one room is always more than 0.5 seconds, which corresponds to more than 12 continuously frames, (iii) the robot always enters one room from the outside of the room, e.g. *'pa'* or *'corridor'*, instead of from another room, then it exits this room to the place where it came from instead of to a different place.

Based on the above observations, we defined a set of rules to help determine the location of the robot at any given time,

Rule 1: Time length. The robot will not stay in one place for a period less than 20 frames. For instance, in the 2nd and the 3rd period of the above example, the extracted sequence shown that the robot only stayed continuously in these two locations for just four and two frames, respectively, which is satisfied with this rule. Thus, there must be a false detection in the 2nd and the 3rd period.

Rule 2: Jumping room. If the location of the robot changes from room A to room B without passing through the 'corridor' and 'printing area', there must be a false detection. For instance, in the 5th and 6th period of the above example, the extracted sequence shown that the robot moved from 'two-person office' to 'kitchen', which is a typical 'jumping room' case that robot did not enter certain place during changing rooms. Thus, there must be a false detection between the 5th and the 6th period.

Rule 3: Unknown room. Since test sequence contains additional rooms that were not images previously, no correspondence frames in the training set could be used to annotate these rooms. In addition, considering the nature of image matching algorithm that one test frame will be annotated with the most similar frame in the training set even there is only one matched point pair found, a false annotation is not avoidable. In the experiment, we found that the number of detected matched point pair between the unknown frame and the training frames is very limited. Thus, we define one rule that any frame detected less than 15 matched point pairs with the training frames is annotated as an unknown room.

To refine the false detection, if the initial detection does not obey the rule 1, we do as follows:

- If the location before the false detection period is the same as the location behind it, the location of the false period will be revised and annotated the same as the pervious period. For example, the 2nd period will be corrected to the 1st period.

If the initial detection does not obey the rule 2, we do as follows:

- A window with a size of *N* frames is applied on the location boundary to recalculate the similarity.
- The similarity between the test image and the top 10 matched training images is summed as the recalculated similarity. The frames with the highest score will be used to annotate the current frame with the location.

2.3 Illumination Filter

The sequences used in the task are taken in different rooms under different illumination settings, e.g. some of the images are from bright offices where as some are taken in dark corridor. Considering the performance of the image matching method that we have described above is highly dominated by two factors, the number of points of interests and point matching techniques, the more points of interest that are found, the easier the point matching step becomes. Thus, the poor visual quality of the sequence could affect the performance of the edge detector and thus could reduce the number of points of interest for point matching. With the goal of improving the results, a method to reduce the effect of illumination effects was used for one of the runs which were submitted. In order to reduce the illumination effect an illumination filter called Retinex [8], was applied. Retinex can improve visual rendering of a frame in which lighting conditions are not good. Once the illumination condition is improved, the number of points of interest detected by corner detector can be increased, as the increased image quality should result the more corners being detected.

3. Submitted Runs

We submitted three runs for ImageCLEF 2009 Robot Vision task using the different combination of the image matching approach, decision rules approach and illusion filter approach. The detailed descriptions of each run are outlined below:

(i) The first run is regarded as the baseline and it uses every first frame out of every five continuous frames of both the training and testing sequences for image matching, this is followed by the application of the rule based model to refine the results.
(ii) The second run uses all of the frames in training and testing set for image matching, followed by the application of the rule based model. The illumination filter is applied for pre-processing the frames.
(iii) The third run uses every first frame out of every five continuous frames for both training and test sequence for image matching, followed by the application of the rule based model to refine the results. The illumination filter is applied for pre-processing the frames.

Considering that every second of video contains 25 frames and the visual difference between two continuous frames is very limited, a large amount of redundancy is containing in the sequence and increases the computing cost. In order to improve the efficiency, we split one sequence into shots, where every shot contains five continuous frames and the first frame of the five is regarded as the key-frame to represent the shot. Instead of training and annotate using every frame, we now only take the key-frames into account. Once a key-frame has been annotated, all the rest frames in this shot are annotated as the same. Since the computation cost of our approach is linear, the key-frame representation can reduce the processing time by 80%. However, this may result in false detection while the robot changes location. For example, in the example one, the frames from 21st to 25th are annotated as 'printing area'. However, the robot moves out from the printing area to corridor at 23rd frame, since the 21st frame is annotated as 'p' the other four frames are annotated similarly.

4. Results and Evaluation

We have generated multiple runs of annotation results based on the approaches presented before. All three runs are submitted to ImageCLEF for official evaluation. The 'score' is used as the measure of the overall performance of the systems. The following rules are used when calculating the score for a single test sequence:

- +1.0 points for each correctly annotated frame.
- Correct detection of an unknown room is treated the same way as correct annotation.
- -0.5 points for each misannotated frame.
- 0.0 points for each image that was not annotated (the algorithm refrained from the decision).

In addition, we also use 'accuracy' as another way of measure of the performance, which shows the percentage of frames that are annotated correctly.

Table 1 shows the results of the three runs. It can be seen clearly that the second run achieved the highest accuracy and score overall. Compared with the first run, the second run improves the accuracy from 59% to 68.5%, which shows that using more frames for training and testing on all test frames could increase the image matching results. In addition, it also shows that the illumination filter does not improve the system performance.

Table 1 lists the performance of the submitted runs. The baseline run offers a reasonable starting performance for the combination of image matching and rule based decision approaches. After using more frames for training and testing and an illusion filter, the second run obtains a 9.5% and 240 point performance gain in terms of accuracy and score, respectively. In the last run, an illusion filter is used for pre-processing and applied to the same number of frames as the baseline for training and testing the accuracy and score is largely reduced by 33.1% and 838.5, respectively.

	Accuracy	**Score**
Run 1 (baseline)	**59%**	**650.5**
Run 2	***68.5%***	***890.5***
Run 3	**25.9%**	**-188**

Table 1: Results of submitted 3 runs, total frame 1689.

4. Conclusions

The multimedia information retrieval group at the University of Glasgow participated in the ImageCLEF Robot Vision task. In this paper we have presented the methodology, experiments and the preliminary results for the task. The model of combining image matching, decision rules and illumination filter contributes the high effectiveness. The results reflect the magnitude of the difficulty of the problem at robot vision, while we believe we have gained much insight to the practical problems, and future evaluations have the potential to produce much better results.

References

1. M. Bennewitz, W. Burgard, G. Cielniak and S. Thrun; Learning motion patterns of people for compliant robot motion; The International Journal of Robotics Research, Vol. 24, No. 1, 31-48 (2005)
2. B. Caputo, A. Pronobis, P. Jensfelt, Overview of the CLEF 2009 robot vision track, CLEF working notes 2009, Corfu, Greece, 2009.
3. L. Lu, X. Dai and G. Hager, 'Efficient particle filtering using RANSAC with application to 3D face tracking', Image and Vision Computing, Volume 24, Issue 6, 1 June 2006, Pp. 581-592
4. H. X. Zhong, Y.J. Pang and Y.P. Feng, 'A new approach to estimating fundamental matrix', Image and Vision Computing, Volume 24, Issue 1, 1 January 2006, Pp. 56-60
5. C. Harris and M. Stephens, 'A combined Corner and Edge Detector,' Proc. Alvey Conf., pp189-192, 1987
6. Y. Feng, J. Jose, 'A hybrid approach for classification based multimedia retrieval', SAMT'08
7. Y. Feng, J. Jiang, S. S. Ipson: A shape-match based algorithm for pseudo-3D conversion of 2D videos. ICIP (3) 2005: 808-811
8. Z. Rahman, D. J. Jobson, "Retinex processing for automatic image enhancement", Journal of Electronic Imaging, 2004, Vol. 13, No. 1, pp. 100-110