Analysing False Positives and 3D Structure to Create Intelligent Thresholding and Weighting Functions for SIFT Features

Michael May, Martin Turner, and Tim Morris

The University of Manchester, UK
michael.may@student.manchester.ac.uk
{martin.turner,tim.morris}@manchester.ac.uk
http://www.cs.manchester.ac.uk
http://www.michael-may.co.uk

Abstract. This paper outlines image processes for object detection and feature match weighting utilising stereoscopic image pairs, the Scale Invariant Feature Transform (SIFT) [13,4] and 3D reconstruction. The process is called FEWER; Feature Extraction and Weighting for Enhanced Recognition. The object detection technique is based on noise subtraction utilising the false positive matches from random features. The feature weighting process utilises a 3D spatial information generated from the stereoscopic pairs and 3D feature clusters. The features are divided into three different types, matched from the target to the scene and weighted based on their 3D data and spatial cluster properties. The weightings are computed by analysing a large number of false positive matches and this gives an estimation of the probability that a feature is matched correctly. The techniques described provide increased accuracy, reduces the occurrence of false positives and can create a reduced set of highly relevant features.

1 Introduction

The scale invariant feature transform (SIFT) [13] is used as a detection algorithm for finding correspondence between features within parts of images thereby allowing image matching to occur. In this paper we consider the specific matching problem of a target stereoscopic image pair of a 3D object within a hand-held stereoscopic video sequence. This paper introduces novel techniques for object detection and feature weighting. The process is called FEWER; Feature Extraction and Weighting for Enhanced Recognition.

For the detection process a set of random features are matched to the scene and the ratio of matches to the number of target features is used as a baseline for noise as these are false positives. Subtracting this noise correspondence ratio from the correspondence ratio calculated from a target image acts as a threshold to indicate if the object is present in a scene.

For the weighting process a 3D point cloud is constructed from target and scene stereo pairs and the features are clustered. For each image the features are divided into three different types, matched from the target to the scene and

Y.-S. Ho (Ed.): PSIVT 2011, Part I, LNCS 7087, pp. 190–201, 2011.

weighted based on the their 3D and spatial cluster properties. This weighting gives an estimation of the probability that a feature is matched correctly. The technique, similar to the previous one, utilises the expected rate of false positives found by studying how randomly selected features match to a scene, creating noise property statistics.

The paper is structured as follows; background work, an explanation of the noise subtraction based object detection, the feature weighting process, with an explanation of the technique by which the weightings are calculated, followed by evaluation and conclusions.

1.1 Background

The SIFT feature detection algorithm developed and pioneered by David Lowe [4,13] is a process that creates unique and highly descriptive features from an image. These features are designed to be invariant to rotation and are robust to changes in scale, illumination, noise and small changes in viewpoint. The features are used to indicate if there is any correspondence between areas within images. This allows object recognition to be implemented by comparing a set of features generated from input images to a set of features generated from images of target objects.

As the target and scene data both consist of stereoscopic pairs a structure from motion (SfM) system (Bundler API utilising a modified version of the sparse bundle adjustment [7] as the optimisation engine) is used to detect different types of matches and produce 3D geometric reconstruction.

Object recognition work using multiple views of a scene has been carried out [5,8,18] using multiple images and rough registration information to determine possible corresponding detections across multiple viewpoints. Work on integrating information across many images has been conducted using Bayesian strategies to combine uncertain information between views [10,19]. Combining data across multiple frames of a video to obtain depth information has also been studied [1,20]. Many other papers show that the use of 3D depth information [3,6,8,11,12,16,17] can be applied successfully to aid object recognition.

Although the processes in this paper use SIFT they could be applied to many other feature detectors such as SURF [2], GLOH [15] or FAST [9].

2 Noise Subtraction for Object Detection

The initial basis for this work is a novel method to detect the presence of an object using the ratio of matched features to the total number of features in a target image. The target image is that of the object being searched for in a pair of scene images. By dividing the total matched features by the total features in the target image the correspondence ratio can be found. This normalises the number of features matched therefore different target images with varying numbers of features can be compared. For example, an image with five hundred features may have fewer matches to a scene than an image with two thousand features, but may have a higher correspondence ratio. The higher absolute number of

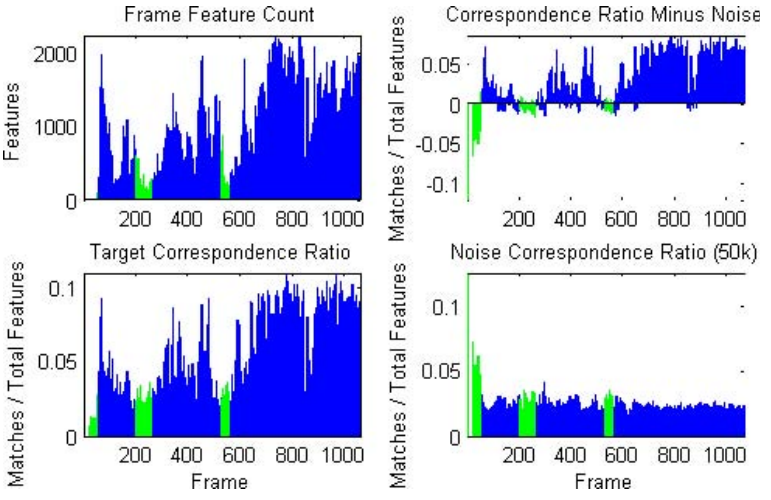

Fig. 1. The results of noise subtraction across 1062 frames of a video sequence. The top left graph shows the feature count for each frame. The top right hand graph shows the result of noise subtraction where the peaks indicate the presence of the object and the bottom left graph shows the target correspondence ratio for the object being identified (524 features) and the bottom right graph shows the false positive noise from a set of fifty thousand randomly collected features. Green (lighter) areas highlight those frames where the object is not present at all and these are shown to be negative on the top right, noise subtraction, graph.

matches in the second images may be noise (false positives) as the larger number of features available means more false positives will occur.

The technique uses the correspondence ratio for a large numbers of randomly collected features as a noise baseline for a particular scene. The features were collected automatically by randomly downloading large numbers of images from Flickr and applying SIFT to them. As these features are known to be random they are unlikely to match. This means that the ratio of matches indicates a level of matches that are statistically insignificant for an object that is being detected. As such, a ratio greater than this baseline of noise plus the average standard deviation can be deemed statistically significant (1σ) for detection. Tests have shown that using SIFT's default parameters has an average false positive rate of 0.024 and an average standard deviation of 0.007 for a random set of one million features. It has also been calculated that as few as ten thousand random features are enough to achieve these noise characteristics. This therefore means that on average a correspondence ratio greater than 0.031 is required for the number of matches to a scene to be deemed statistically significant.

By subtracting the noise correspondence ratio from the actual target correspondence ratio the data is automatically thresholded such that many false positives from the target to the scene will be ignored. Fig. 1 demonstrates this for a target image matched to 1062 frames of a video sequence where the object is present in most but not all of the frames.

3 FEWER: Feature Extraction and Weighting for Enhanced Recognition

Following this initial technique for subtraction of SIFT noise a second process has been developed which utilises the 3D stereoscopic image pairs of the target and scene to specify weighted feature matches to indicate confidence in their accuracy. This is called FEWER; Feature Extraction and Weighting for Enhanced Recognition. A pair of target images of the object that is being detected and a pair (or stream of pairs) of stereo images of a scene are used. Simply put, if a feature doesn't match well to its counterpart in a stereo pair the chances of it being stable are lower. The process has nine stages:

Extract SIFT Features. Extract the features from the target and scene stereo pairs as shown in Fig. 2.

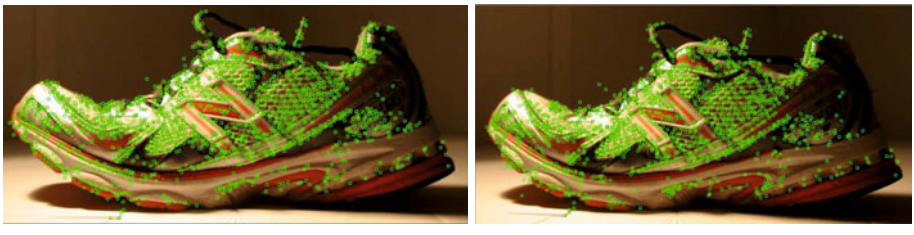

Fig. 2. A stereo pair of target images displaying the SIFT features extracted from them. There are 2176 in the left image and 2087 in the right image.

Calculate 3D Positions. For both the target and scene pairs a 3D point cloud is generated from the features as shown in Fig. 3.

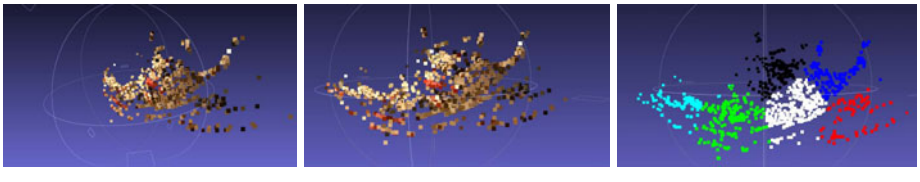

Fig. 3. The set of 3D feature positions generated from the stereo pair in Fig. 2 using the Bundler API [7]. The first two images show two different angles for the same data and the curvature of the shoe is clearly visible. This is a subset of the total features extracted from the original images and consists of 885 features. The right hand image shows the *type*3 features spatially clustered.

Cluster 3D Data. The 3D matched features are then spatially clustered in 3D space (using k-means [14]) to separate and label various 3D aspects of the scene. Clusters help differentiate between foreground and background objects.

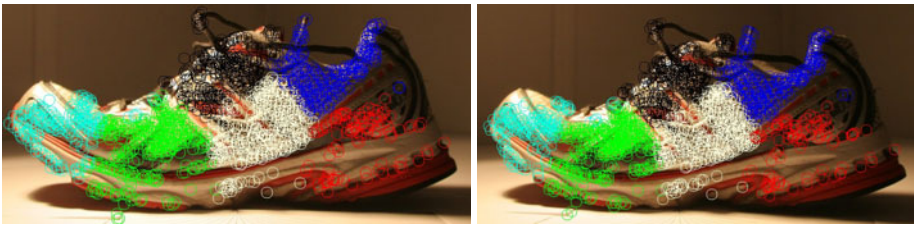

Fig. 4. The final set of clustered *type*2 and *type*3 features for the left and right images in a stereo pair. There are 1802 in the left image and 1782 in the right image.

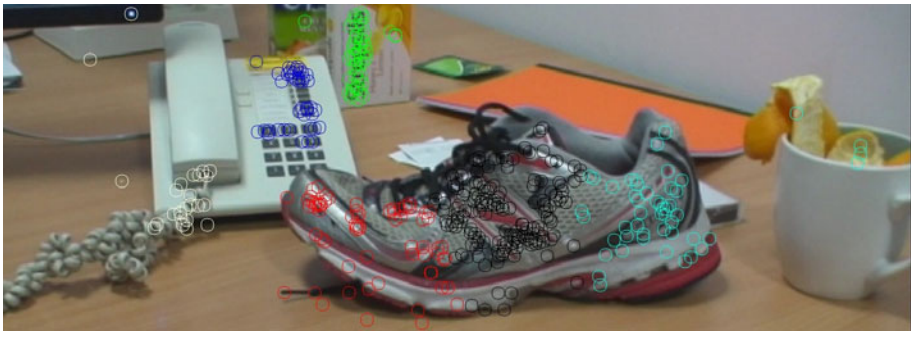

Fig. 5. The set of clustered features in a scene input image. The advantages of spatial clustering are clearer here as various objects have are roughly separated by the different clusters so as to provide more information when matching features.

Feature Labelling. Three different feature types are defined depending on their 3D and cluster properties. *Type*3 features are labelled by mapping the 3D features back to their 2D image locations for each image. *Type*3 features are those which have 3D information associated with them and therefore match to the other stereo image. To define *type*2 features a distance threshold is used to find other features near each of the *type*3 features and they are added to the clusters. These features are likely to be part of the same object as they are nearby but as they do not match to the other stereo image they can be considered less reliable. These are therefore labelled as *type*2 and a secondary cluster index is generated for them. The remaining features are then labelled as *type*1 and they do not have any cluster information relating to them.

Target to Scene Matching. Feature matching is performed for each target to scene combination; left target to left scene, left target to right scene, right target to left scene and right target to right scene. This is done using the nearest neighbour technique described by Lowe [13].

Initial Weighting. Each target image has its own set of weighting for matches to both of the scene images. Thus four sets of weightings are calculated. The initial weightings for each feature are given by which type they are and which

type they match too. A *type*3 target to *type*3 scene match will have a larger initial weighting than a *type*3 target to *type*1 scene match. There are therefore nine possible combinations of matches each with their own weighting.

Type 3 Mismatches. The weightings are then adjusted by checking if matching pairs of *type*3 features from each target image match to similar positions in the scene images. Fig. 6 illustrates these cases. If the same *type*3 feature in both of the target images matches to different points in the scene the weighting is reduced as the likelihood of one or either being correct is reduced. The weighting is effected differently if the single scene feature is *type*3 or not *type*3.

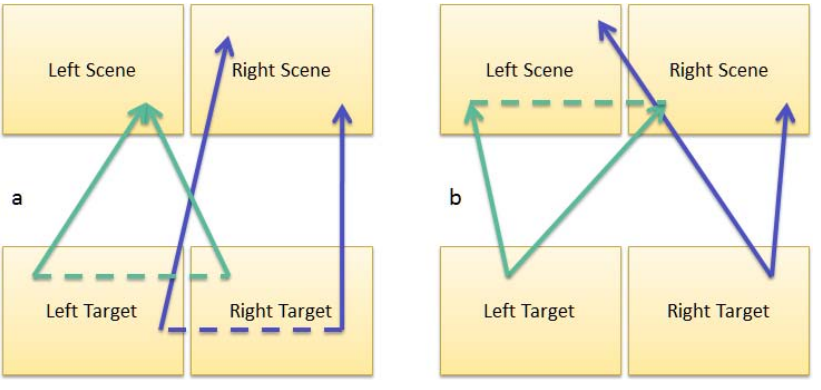

Fig. 6. This shows the two cases for *type*3 mismatches. Case *a* shows the correct (lighter) and incorrect (darker) matches from *type*3 features in the target images to any type of scene feature. Case *b* shows the correct and incorrect matches from the target image to the *type*3 scene features.

A secondary check is carried out for each target feature which matches to a *type*3 scene feature. If the feature matches to both corresponding *type*3 scene features then the weighting is increased. If a target feature matches a *type*3 scene feature and also matches a different feature in the other scene image then the weighting is reduced. There is no effect if the target feature matches one scene but not the other. Again the weighting is affected differently if the single target feature is *type*3 or not *type*3.

Cluster Weightings. The next stage is to adjust the weightings based on the 3D spatial cluster that a feature is in and how groups of features in the same cluster match. The basic hypothesis is that as more features in a target cluster match to a specific scene cluster the more likely it is that there is correspondence between these areas of the scenes. The confidence weighting is calculated as follows:

$$\text{confidence} = \frac{\text{signal}}{\text{noise}} \times \sqrt{\text{sample size}} \tag{1}$$

where *signal* is the correspondence ratio from a target cluster to a scene cluster, *noise* is the correspondence ratio from the target cluster to every other scene

cluster and *sample size* is the total correspondence ratio from the target cluster to all of the scene clusters. A confidence value is calculated for each target cluster to every scene cluster. This equation means that the sample size and the signal both have to be significantly large to generate a high confidence thus a low numbers of matches will not be statistically significant when calculating a feature's weighting. This confidence value is thresholded so that a high confidence cluster pair will result in a higher weighting for features which match between them. The boundaries and distribution of the clusters can affect the performance of this technique and as such there is no negative weighting for low confidence.

Threshold Matches. The weighting is normalised transforming its value into the range of 0 to 1. A threshold can now be applied to select a subset of the weighted feature matches.

Table 1. The stages used for extracting and weighting features with FEWER

Stage	**Output**
Extract SIFT Features	Sets of SIFT image features.
Calculate 3D Positions	Relative 3D positions of matched features.
Cluster 3D Data	Index of features indicating the cluster they are contained in.
Feature Labelling	Features labelled by type.
Target to Scene Matching	Indies indicating where features match to the scenes.
Initial Weighting	Weightings for each feature match.
Type 3 Mismatches	Updated weightings based on a disparity in matches.
Cluster Weightings	Updated weightings based on matches between clusters.
Threshold Matches	Set of matched features with weightings above a threshold.

4 Calculating Weightings from Noise

Values for the FEWER weighting adjustment stages described above have to be calculated to weight various characteristics of a matched feature. This is done by studying the noise properties for each stage using a set of stereo features know not to match correctly. By looking at the level of false positives for various feature match types ratios can be calculated which indicate how much more reliable one type of match is than another. The data describes how each type of match is affected by false positives. For the initial weighting stage the correspondence ratio for false positives for each match combination is calculated using large sets of random features. They are matched to videos which are known to contain no correspondence to the scene image. By obtaining the average correspondence ratio across the frames and adding the standard deviation it can be seen for the test data that *type*3 to *type*3 feature matches have a correspondence ratio 16 times less (0.64 / 0.04 from the full set of data listed in Tab. 2) than *type*1 to *type*1 thus the weighting reflects this directly. The weighting (w) is calculated as follows:

$$w = k\frac{1}{\bar{x} + \sigma} \times \frac{\text{relevant matched features}}{\text{total matched features}} \tag{2}$$

where $\bar{x}$ is the mean noise value across a sample, σ is the mean standard deviation of the noise and k is a scaling factor. The *relevant matched features* are the subset of the *total matched features* actually involved in the particular weighting process so that the weightings are scaled accordingly.

Table 2. Weighting values calculated from experimental data for different aspects of the weighting process. The left table shows the initial match weighting values and the right show the *type*3 mismatch weightings. T and S refer to Target and Scene.

	*type*1 S	*type*2 S	*type*3 S
*type*1 T	0.04	0.11	0.26
*type*2 T	0.06	0.41	0.39
*type*3 T	0.11	0.75	0.64

	*type*3 S	*type*3 T
correct *type*3	0.12	0.06
correct not *type*3	0.07	0.01
incorrect	-0.48	-9.3

The same process is used to calculate the weightings for the *type*3 mismatches where the number of false positives matches are used but as only the *type*3 features are involved the *relevant matched features* value reflects this. This incorporates a negative weighting for mismatches which have a relatively high cost as seen in Tab. 2.

For the cluster weightings, analysis has provided data on how well false positive matches cluster and what is the minimum level of cluster matching confidence required to occur beyond random chance. This allowed a cluster confidence threshold to be calculated using the same equation and a weighting for values greater than the threshold to be defined. This only relates to *type*2 and *type*3 features as *type*1 features are not clustered. The threshold was calculated to be 0.00015 and the weighting value added to matches greater than this threshold is 0.4 when using six clusters.

After these three stages the maximum possible weighting that can be achieved using the experimental data weightings is 1.36 and this value is used for normalisation.

5 Results

Following the weighting calculations based on over 2000 frames of video and over 20000 stereo target features, the system has been tested on different target and scene input data within a similar environment. The test involved a 2500 frame stereo video with a target object located within the sequence. Stereo images of the target objects are matched to each frame using the techniques described previously. The system outputs the four match images for each combination of target to scene matches with the matched features drawn using a heat map style colour coding. The colour changes linearly through RGB space from blue to green to red as the weighting increases.

Figs 7 and 8 shows examples of the coloured weightings as feature matches are deemed to be of higher or lower reliability. The images are consistent with the other frames in the sequence and show that incorrect matches are weighted lower.

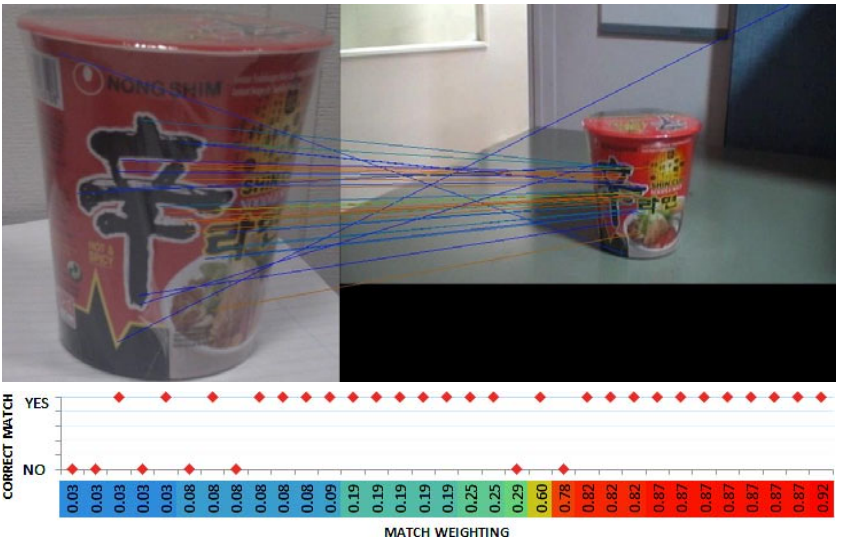

Fig. 7. A typical example of weighted feature matching displaying matches from the left hand target image to the left scene image. Some of the correct matches are green and red indicating higher weightings. The mismatched features in this scene have received low weightings and are coloured blue. The feature matches with low weightings can be removed by adjusting the weighting threshold which is set at 0 in these cases. The graph below shows the weightings for each of the 33 matched features and whether they match correctly.

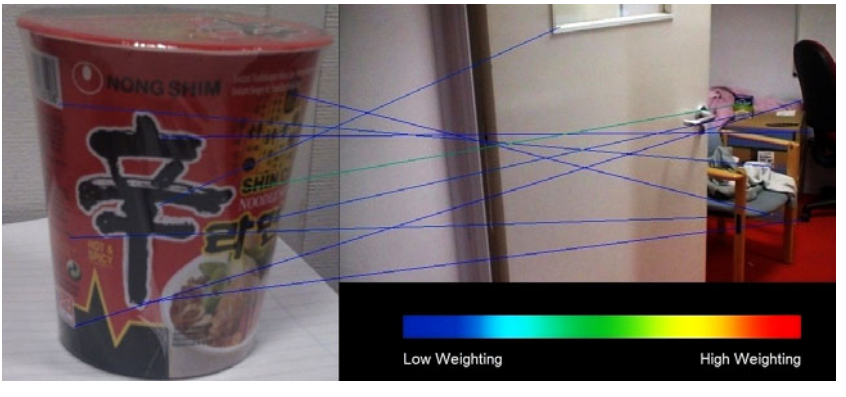

Fig. 8. A typical example of weighted feature matching displaying matches from the left hand target image to the left scene image. This shows false positives matches sucessfully being weighted with lower values.

Fig. 9 shows the correspondence ratio across the 2500 frames and the large peak indicates the location of the target. By adjusting the weighting threshold it is shown that the false positive count is reduced leaving many of the most reliable

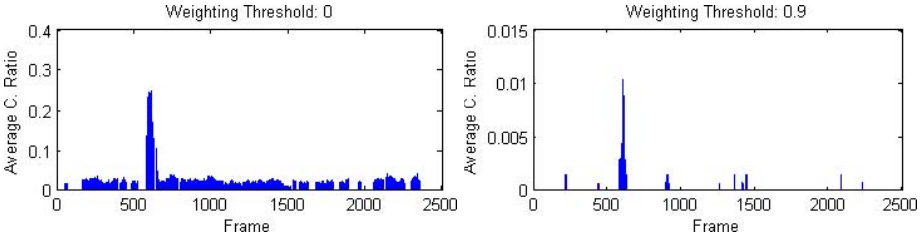

Fig. 9. This shows the correspondence ratio before and after applying a threshold on the feature weightings. The graphs are the mean of the four possible match scenarios (each target to each scene). The peak indicates the location of the object. The left graph shows the correspondence ratio when no threshold has been applied and the right graph shows what happens when a threshold of 0.9 is applied. This reduces the remaining correspondence ratio substantially but the features remaining are of a higher quality and fewer false positives are present across the video sequence.

features. The weighting threshold could be computed adaptively by analysing a set of known false positive feature matches in a similar manner to Section 2 and adjusting the weighting to minimise them.

6 Evaluation

FEWER has been shown to weight the features successfully. It relies on the probability of a feature type being a mismatch therefore, in some cases, incorrect matches can be weighted highly and vica-versa. Investigating how often this occurs will be future work. The weighting threshold provides a sliding scale between a small number of highly reliable matches and a large number of features including more unreliable matches.

The reason FEWER works is that *type*3 features are likely to be more stable than the other features as they correspond between the stereo images and are therefore known to match to a different view of the object. The SfM process [7] could be removed and normal SIFT matching used instead to generate *type*3 features. The SfM process has its advantages for clustering and background separation and is more discriminative when matching than just using SIFT as the matched features have to fit correctly to a 3D model not just match. The *type*2 features are more stable than *type*1 as the features are likely to exist on the objects that have been matched between the stereo objects due to their proximity to the *type*3 features and less likely to be background features. *Type*1 features are the least stable and have no extra properties associated with them. The difference between them is highlighted in Fig. 10.

FEWER allows the system to select a subset of features which are higher in confidence rather than just thresholding using the noise properties in Section 2 which has no indication of which features are likely to be correct. A combination

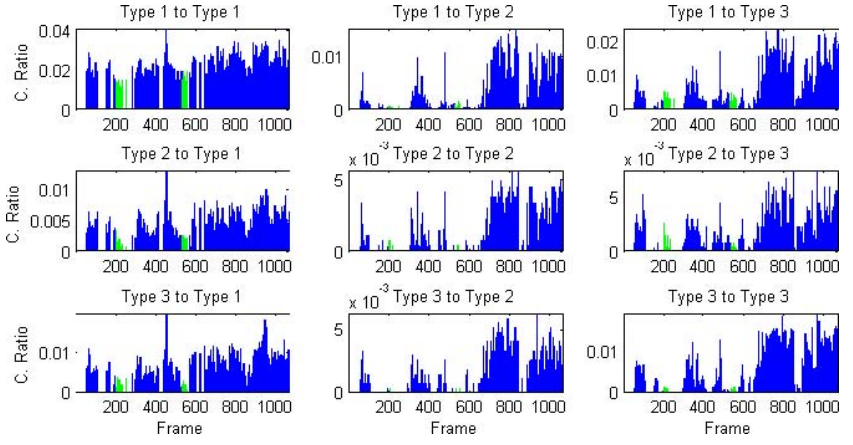

Fig. 10. These are the mean correspondence ratio graphs for the three feature types for matches from both target to both scene images. It can be seen that the *type*1 feature matches have fewer peaks and troughs and the green (lighter) areas, where the object is not present are harder to distinguish than for the *type*3 feature matches. They therefore resulted in lower weighting (see Tab. 2). For the random data used for calculating weightings in Section 4 these graphs are flatter with lower correspondence ratios. They display the random noisy correspondence ratio and give a minimum baseline for noise for each feature type.

of the noise thresholding for detection and FEWER could be used so that the computationally expensive weighting process is only applied to frames which are likely to contain the object to select the best matches.

7 Conclusion

The results of this work are promising and provide a technique for identifying and selecting the best feature matches. The results have shown examples of features being weighted to indicate which matches are correct and which are incorrect. The advantages of FEWER are clear as the detection process provides a higher confidence in the matches than standard SIFT matching alone. The system could result in lower data transmission rates as fewer matched features are selected.

Further development of the algorithm will involve data fusion to combine the four output images (left target to left scene etc.) into a single location mapped to a 3D model and superimposed on the 3D scene model. This will provide the user with a consolidated view of the output data to visualise where features match. Also, since the epipolar geometry is available, the weighting could possibly be improved at the matching stage by limiting the search region to a band around the epipolar lines. Comparison will be made to other methods for reducing the number of incorrect mathces using outlier detection methods such as RANSAC alone or the Hough binning used by Lowe [13].

References

1. Andriluka, M., Roth, S., Schiele, B.: Monocular 3D pose estimation and tracking by detection. In: IEEE CVPR, vol. (2), pp. 623–630 (2010)
2. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
3. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and Recognition Using Structure From Motion Point Clouds. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 44–57. Springer, Heidelberg (2008)
4. Brown, M., Lowe, D.: Unsupervised 3D object recognition and reconstruction in unordered datasets. In: 3DIM, pp. 56–63. IEEE Computer Society (2005)
5. Coates, A., Ng, A.: Multi-camera object detection for robotics. In: ICRA, pp. 412–419 (2010)
6. Gould, S., Baumstarck, P., Quigley, M., Ng, A., Koller, D.: Integrating visual and range data for robotic object detection. In: ECCV M2SFA2 (2008)
7. Helmer, S., Meger, D., Muja, M., Little, J.J., Lowe, D.G.: Multiple Viewpoint Recognition and Localization. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part I. LNCS, vol. 6492, pp. 464–477. Springer, Heidelberg (2011)
8. Helmer, S.: Using stereo for object recognition. In: IEEE ICRA, pp. 3121–3127 (2010)
9. Hoiem, D., Efros, A., Hebert, M.: Putting objects in perspective. In: IEEE CVPR, vol. 2, pp. 2137–2144 (2006)
10. Laporte, C., Arbel, T.: Efficient discriminant viewpoint selection for active bayesian recognition. IJCV 68(3), 267–287 (2006)
11. Leibe, B., Cornelis, N., Cornelis, K., Van Gool, L.: Dynamic 3d scene analysis from a moving vehicle. In: IEEE CVPR, vol. 80, pp. 1–8 (2007)
12. Leibe, B., Cornelis, N., Cornelis, K., Van Gool, L.: Dynamic 3D Scene Analysis from a Moving Vehicle. In: IEEE CVPR 2007, pp. 1–8 (2007)
13. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
14. MacQueen, J.: Others: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, p. 14 (1967)
15. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 10(27), 1615–1630 (2005)
16. Quigley, M., Batra, S., Gould, S., Klingbeil, E., Le, Q., Wellman, A., Ng, A.: High-accuracy 3d sensing for mobile manipulation: Improving object detection and door opening. In: IEEE ICRA, pp. 2816–2822 (2008)
17. Sudderth, E., Torralba, A., Freeman, W., Willsky, A.: Depth from familiar objects: A hierarchical model for 3D scenes. In: IEEE CVPR, vol. 2, pp. 2410–2417 (2006)
18. Trajkovi, M., Hedley, M.: Fast corner detection. IVC 16(2), 75–87 (1998)
19. Whaite, P., Ferrie, F.: Autonomous exploration: Driven by uncertainty. IEEE TPAMI 19(3), 193–205 (1997)
20. Wojek, C., Roth, S., Schindler, K., Schiele, B.: Monocular 3D Scene Modeling and Inference: Understanding Multi-Object Traffic Scenes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 467–481. Springer, Heidelberg (2010)