A planar object tracking approach robust to total occlusion

João Marcelo Teixeira, Judith Kelner
Virtual Reality and Multimedia Research Group
Computer Science Center, UFPE
Recife, Brazil
{jmxnt, jk}@cin.ufpe.br

Veronica Teichrieb
VOXAR Labs
Computer Science Center, UFPE
Recife, Brazil
vt@cin.ufpe.br

Abstract—This paper presents a novel algorithm for performing planar object detection in partial/full occlusion situations. Firstly, all scene features are computed and matched against the reference template ones. Secondly, a test identifies if the behavior of the features surrounding the template is similar to the reference template and then the reference feature database is updated. Experiments on both real-world and synthetic data show that our approach proved very successful in dealing with different occlusion scenarios.

Keywords-component; image features, tracking, occlusion

I. INTRODUCTION

There are a vast number of object trackers. The most used ones usually follow the object’s movement (tracking step), requiring an initialization (detection) step responsible for parameters regarding the start position [1]. A less used approach for performing object tracking is to solely rely on information obtained by object detection. This way, only the detection step is used. One advantage of such approach is that the object can suffer large displacements (wide baseline) and still be tracked.

Three main characteristics affect tracking quality: image noise, object occlusion and lighting variation. Many methods can deal with occlusion, from motion estimation to whole environment perception. The interesting thing about lighting variation is that when the processing of luminance change is considerable, it can be dealt as an occlusion problem as well. Considering this reason, this work focuses on improving the tracker robustness under partial and full occlusion scenarios.

During a video sequence, while recording some object of interest, it may not be visible in all captured frames. When partial occlusion happens, one can rely on the remaining visible features of the reference template to still perform the tracking. This approach fails when the object is completely occluded. One manner to tackle this problem is to understand the relationship of the object and its surroundings, in a way that new features can be learned in a seamless way. This work proposes a technique of identifying such spatial behavior between objects and how to take benefit of this information to improve object detection and tracking.

The remainder of this paper is organized as follows. Section 2 details the proposed scheme and how it differs from conventional tracking-by-detection approaches. Section 3 illustrates some of the experiments performed under different environmental conditions. At last, section 4 highlights the benefits of the proposed approach and points some directions for future work.

II. PROPOSED APPROACH

The purpose for using a tracking-by-detection scheme as basis of the proposed approach is due to the necessity of having to calculate the image features periodically (image keypoints and their descriptors are computed every new frame). Additionally, only the feature matching process over the detected image features needs be done in order to find (track) the desired object.

The tracking-by-detection scheme works as follows. Initially, the reference template features are extracted. These features are stored in a database of features, which will be further used during the matching process (find the correspondence between template and image features). The system’s “training” phase is simple because it only requires that the template features be extracted and stored. The online tracking phase consists, for each input frame, of the extraction of all image features and their further matching with the features stored on the database. With the calculated matches on hand, a robust estimator, such as RANSAC [2], can be used in order to find the parameters of the transformation that brings the points from template to image coordinate space, besides removing the detected outlier matches. The online phase is repeated for each new input frame captured. An architecture commonly used by tracking-by-detection approaches is shown in Figure 1.

The proposed tracking scheme is generic enough to enable the use of any feature detector and descriptor. For example, SIFT [3] or others can be used. Initially we intend to validate the idea for tracking planar objects, but eventually a similar approach will be extended for dealing with more complex objects.

The extraction of image features is the key of the proposed approach, because it focuses on the update of the feature database. The goal of the update stage is to add to the template new features that were not originally present on the template (i.e., features located outside the template area), but can be found on the input images. Only image features that have behavior similar to the template features are added to the database. Similar behavior means that both (template and new features) should have the same transform (same homography). The addition of new features on the database will allow the increase of tracking robustness, in a way that it will be possible to keep detecting the template even if its original features are no longer visible on the input images.
Instead of using the RANSAC algorithm we followed the idea of [4], in which they use the Hough transform [5] to reduce the number of outliers found after a feature matching procedure. A keypoint can be basically defined by the tuple $(x, y, \sigma, \theta)$, in which $x$ and $y$ correspond to the position of the keypoint, $\sigma$ to its scale and $\theta$ to its orientation. A simple translation plus scale change transform can be written as

\[
\begin{pmatrix} x' \\ y' \\ \end{pmatrix} = e \begin{pmatrix} x_{\text{ref}} \\ y_{\text{ref}} \\ \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ \end{pmatrix}
\]

in which $x$ and $y$ are the current feature position on the input image, $e$ represents a scale factor between the images and $t_x$ and $t_y$ map to the relative translation between them. The scale factor and translation parameters can be calculated as

\[
e = \frac{\sigma}{\sigma_{\text{ref}}}
\]

\[
t_x = x - e \times x_{\text{ref}}
\]

\[
t_y = y - e \times y_{\text{ref}}
\]

in which $\sigma$ refers to the scale of the feature found on the image and $\sigma_{\text{ref}}$ to the scale of the feature on the reference database. After finding the parameters for every match, the Hough space is constructed. The matches that map to the same space in Hough coordinates are the ones considered inliers, while the other matches are discarded.

The proposed approach can be divided in the following steps:

1. The features present on the input image are extracted and matched to the original reference features;
2. The same input image features are matched to the modified feature database;
3. For both cases, outliers are discarded using the Hough transform;
4. If an homography cannot be found by only using the reference matches, the second matches are taken into consideration;
5. The Hough transform is again used to compare the behavior between last and current frame;
6. Only features that showed a behavior similar to the ones already in the database are considered;
7. In case a feature maps to a coordinate in reference space that is already occupied, the feature is replaced.

The amount of features added to the feature database can be limited by specifying a fixed area surrounding the reference template. Moreover, the collision between features already present in the database considers a proximity radius (in our tests, good results were obtained using a 5 pixel value). Since the reference template is passes through a “training” phase, we have decided that its features are enough representative of its content. By this reason, the new features added to the feature database must be located outside the template area. Original features are never updated, which means they remain untouchable during the entire application execution.

III. Experiments

Experiments using both real-world and synthetic data were performed in order to illustrate the results that can be obtained by the proposed approach. The SIFT feature detector/descriptor was used in all tests done. Figures 3 to 5 illustrate the three different scenarios analyzed, showing on the first row the use of SIFT alone and on the second row the proposed approach. The first one, referred as “chicken tracking”, comprises a wall image that is artificially translated from right to left. In the middle of the image, a fixed black area was added in a way that different portions of the image are occluded for distinct frames. The template being tracked for this scenario is shown in the left of Figure 2 and the results obtained are illustrated in Figure 3. It is important to notice that in this example, SIFT continues to work when partial occlusions happen, but fails on total occlusion ones.

The second scenario comprises the tracking of a face on a game box captured by performing random movements. This way, it was possible to test the approach using different projective transformations. The template being tracked is shown in the center of Figure 2. As happened with the previous case, SIFT is still able to work relatively well under a partial occlusion scenario, but fails on total occlusion ones.
as shown in Figure 4. It is also important to notice that the detection result by the proposed approach was more coherent than SIFT’s alone. This can be justified by the fact that using only SIFT there are few matches and the homography is not precisely estimated, while with the proposed approach, more features contribute to a better estimation.

The third scenario comprises an outdoor video sequence of a bus on the street. In this case, both camera and object are moving. The template being tracked in this example is the advertisement shown in the right of Figure 2. The results illustrated in Figure 5 showed that in this case, even with partial occlusions, SIFT by itself was not able to recover an adequate homography. The proposed approach continues detecting the object coherently, even when just a few features are available from the bus lateral.

![Figure 2. Reference templates used on experiments. From left to right: chicken drawing, William Adama’s face and bus advertisement.](image)

IV. CONCLUSION

This work showed that it is possible to increase the robustness of object detection/tracking by learning additional features from the environment. The learning process was based on the spatial relationship between planar features. The obtained results were satisfactory, since the proposed approach was able to keep tracking the object in different scenario configurations.

As future work, we intend to research other spatial relationships between objects, in order to further extend the approach to complex (3D) objects. We also intend to evaluate how different detectors/descriptors perform using the proposed scheme. We think it is possible to obtain similar results using different algorithms, with the advantage of reduced computational power.

V. ACKNOWLEDGEMENTS

The authors want to thank FINEPE for financially supporting this research (process 01.10.0584.00) on the scope of the RPCMOD.

REFERENCES


![Figure 3. Chicken drawing experiment. This test involved pure translation movement from right to left. Object detection fails using only SIFT on the second image of the sequence.](image)
Figure 4. Battlestar Galactica game box experiment. Object detection using only SIFT fails on both second and third images of the sequence.

Figure 5. Bus advertisement experiment. Object continues to be tracked even under total occlusion (center image). Object detection using only SIFT fails on all images.