Speeded-Up Object Recognition and Pose-Estimation using SURF

Tayyab Bin Tariq
tayyabt@gmail.com

Naveed Ejaz
naveed.ejaz@nu.edu.pk

Department of Computer Science, FAST NUCES,
Islamabad, Pakistan

Abstract

Object recognition is one of the most important problems in computer vision, with wide ranging applications such as content based search, automated surveillance, action recognition etc. In this paper we present a framework for object recognition and pose estimation using SURF features. In this framework we make four novel contributions. Our feature-reduction process allows a speed-up of matching speed-up of 634.8% by using only the most repeatable features for matching. The noise-reduction process allows a further increase in matching speed-up reducing the false positive rates by 50%. A modified definition of the second-neighbor in the in the nearest neighbor ratio matching strategy allows matching with increased reliability. We also introduce a hierarchal approach for feature database storage that presents an easy way for pose estimation of objects.

1 Introduction

Object recognition is a very important part of many computer vision applications. Content based multimedia search, automated monitoring of surveillance videos, action recognition and video understanding are a few of these applications. Owing to this importance a lot of work has been done in this field, however, it remains a challenge due to the significant variation shown by real world objects and intra-class similarity.

SURF is a very important object recognition technique that belongs to the category of local interest point descriptors. SURF first localizes discriminative and repeatable interest points. These interest points are described by the Haar-wavelet responses using a 64 dimensional vector of floating point values (1). These descriptors are similar in function to those developed by Lowe (2). Lowe reported that an object can be identified using as little as three matched features (interest points). The same holds true for SURF, since SURF features are more discriminative than SIFT features.

This paper presents a framework for recognition and pose-estimation of multiple objects using SURF features in an efficient and reliable manner. The technique achieves a speed-up of up to 7 times as compared to simple SURF, during the matching phase. During the training phase two techniques, feature reduction and feature noise-reduction, are employed to reduce the number of features in the database. This allows for the speed-up during the matching phase. During the matching phase, a modified definition of second nearest neighbor (1) is used. This increases the reliability of matching process as a larger number of features are correctly matched. Finally, pose-estimation is done with the help of a hierarchal arrangement of the object/feature database.

SURF finds a large number of highly discriminative features from an object. However, an object can be recognized using as little as three features. So, in theory, we only need to store the three most repeatable features for every object in the feature database. However, the identification of the three most repeatable features is a daunting task. This work presents a technique for the identification of the most repeatable features under affine transformations. The technique allows approximately 634.8% reduction in search time with less than 2% reduction in accuracy.

Identification of false positives in feature matching is another challenge while using SURF and related techniques. Features are matched by comparing features from the test image with the features in the database using Euclidean distance between them. A matching pair is detected, if its distance is closer than 0.7 times of the distance with the second nearest neighbor (1). This is the nearest neighbor ration matching strategy. This paper presents a definition of the second neighbor based in k-Means clustering of SURF features as opposed to a definition based on object classes. This allows larger number of features to be matched.

This paper also presents a method for arranging the object database in a hierarchal manner. This arrangement of the database helps in the reliable recognition and pose-estimation of objects. It also improves the ability to generalize the matching of objects outside of the training set. The hierarchal arrangement of the database and the noise-reduction process are described in details in the design section.

The rest of this paper is organized as follows. Section 2 presents the detailed design of the technique discussed above. Section 3 presents the experimental setup and results. Section 4 concludes the paper with the future work.

2 Proposed Design

This section presents the design of the framework for reliable and efficient object recognition using SURF. As in all recognition systems there are two phases, Learning Phase and the Classification Phase. Figure 1 shows this framework. The rest of this section describes the two phases and feature database in detail.

2.1 Learning Phase

The first step in the learning phase is feature extraction from tagged training images using SURF (1). The second step is reduction of SURF features to the most repeatable ones under affine transformations. The third step removes noisy features using k-Means clustering. Lastly, these features are stored in the feature database along with object tags. Each of these steps is explained in detail below.

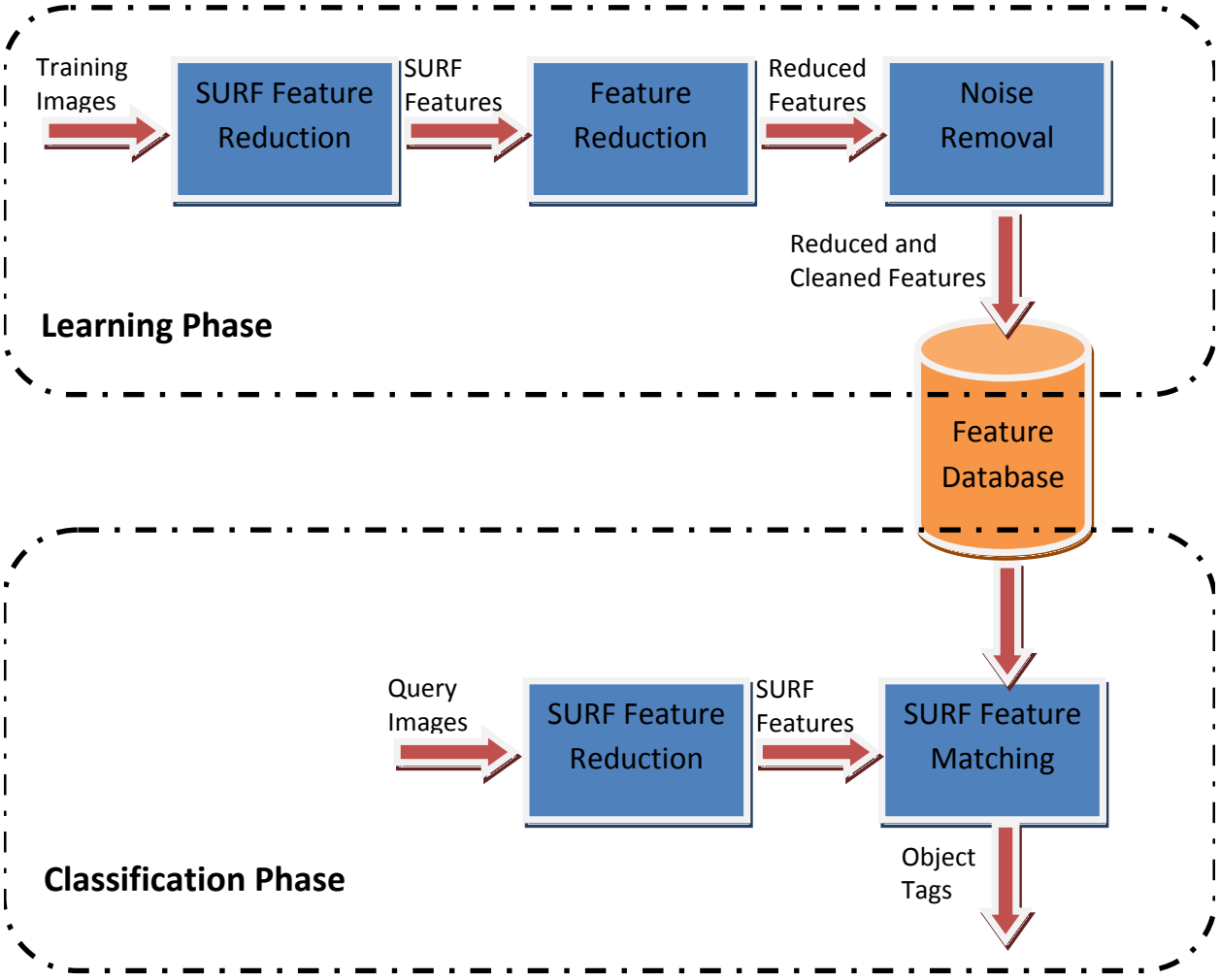

Figure 1: Framework for Object Recognition using SURF

2.1.1 SURF Feature Extraction

The first step of the learning phase is feature extraction from training images. First the interest points in an image are localized (1). These interest points are then described using SURF feature descriptors.

2.1.2 Feature Reduction

The feature reduction phase finds the features most repeatable under affine transformations. This is the most important step in the training phase. The most repeatable features are found by comparing the features extracted from the original training image with the feature extracted from the affine variants of the images. These affine transformations include 2x scale-up, 2x 25° clockwise rotation and XY-skew. On an average the number of features is reduced by about 7 times during this step.

2.1.3 Noise Removal

The noise removal step aims to remove the features that are likely to have been generated by background or noise in the object images. Once reduced features from all the training images have been stored clustering is performed using the k-Means algorithm. The value of k is determined by the following formula;

$$k = \alpha * |O| * |\bar{I}|$$

In the above equation, $|O|$ is the number of objects in the database and $|\bar{I}|$ is the average number of instances per object in the database. α is a factor used to control the size of the clusters. In most cases 3 is found to be the ideal value for α. Once the features have been clustered, a dominant object is selected for each feature cluster using a majority vote. All the features not belonging to the dominant object in the clustered are removed.

2.2 Classification Phase

The first step in the learning phase is feature extraction from the query images using SURF (1). The second step is matching of these extracted features with the feature in the database. Feature extraction is performed as described in section 2.1.1. The matching of features is described in detail in the rest of this sub-section.

2.2.1 SURF Feature Matching

Bay et al. propose the nearest neighbor ratio matching strategy (1) as used in (2). They compare an interest point in the test image with features in the database by calculating their Euclidean distance. If this distance is less than 0.7 times the distance with the second nearest neighbor, a matching pair is detected.

Lowe et al. define the second closest neighbor as being the closest neighbor known to being come from a different object than the first. The rationale behind such an approach is that the correct matches need to have the closest neighbor significantly closer than the closest incorrect match to achieve reliable matching. For false matches, there will likely be a number of other false matches within similar distances due to the high dimensionality of the feature space (2). The second nearest neighbor can be thought to be providing a measure of the density of the false matches. Using this principle we modify the definition of second neighbor as being the closest neighbor known to have come from a different 'cluster of features' than the first. These clusters are created using the k-Means algorithm as described in section 2.1.3. The rationale behind this modification is that a number of features are not matched using the definition provided by Lowe et al. because other object may have similar features. However, k-Means clustering visually similar features in the same cluster and thus the features are matched. It is observed that the number of features matched are increased substantially using this definition and the number of false positives is also reduced.

An object is said to be found in an image if three or more features from that object are matched. The pose of the object is estimated using the pose information stored in the features database. This is described in detail in the next sub-section.

2.3 Features Database

The features database stores the features extracted from the training images. The features database is organized in as a hierarchical tree-like structure. At the top of the hierarchy is a class of objects such as cars, buildings, road signs etc. Within each category separate instances of objects are arranged. For instance, Ferrari California and Lamborghini Diablo are organized under the cars category. The object instances are further divided into different poses, such as front, rear, side and top. The features extracted from images with object and pose information are stored at this level in the hierarchy.

3 Experiments and Results

This section presents the results of the experimental setup used and the results of the proposed framework. The UK Benchmark Object Recognition Dataset was used for testing the framework. The dataset contains 10200 images of about 2500 objects. There are 4 images of each object. We used one image per object for training and three images per object for testing.

With feature reduction a speed-up of 634.8% percent was achieved with less than 2% reduction in matching accuracy. After applying noise reduction on the reduced features the false positive rate was also

pushed down by about 50%. After noise reduction an overall speed-up of 939.6% is achieved. Table 1 summarizes the results.

	Without Feature Reduction	With Feature Reduction	After Noise Removal
Speed-up (%) compared to SURF		634.8%	939.6%
Matching Accuracy	81.33%	80.00%	79.67%
False Positives	1.33%	6.67%	3.67%%
Average Matching Time	4914 ms	774 ms	523 ms
Average Features Per Object	379.3	50.5	46

Table 1: Comparison of matching accuracy and speed.

4 Conclusion and Future Work

In this paper we present a framework for efficient and reliable object recognition and pose estimation using SURF features. The process of feature reduction is presented that filters out the most repeatable features from an image, allowing a matching speed-up of up to 634.8%. The noise removal process presented allows the false positive rates to be reduced by about 50% and further increases the speed-up to 939.6%. The tree-like hierarchal database organization method presented in this paper allows for pose estimation of objects. Work is underway on two fronts 1) to use multi resolution analysis techniques during the learning phase to improve recognition of objects in low resolution images, 2) to incorporate Fe Fe Li's *hierarchal generative process* (3) in recognition of objects and actions.

5 References

1. *SURF: Speeded Up Robust Features.* **Herbert, Bay, Tinne, Tuytelaars and Luc, Van Gool.** s.l. : European Conference of Computer Vision, 2006.

2. *Distinctive Image Features from Scale Invariant Key Points.* **David G., Lowe.** s.l. : Internation Journal of Computer Vision, 2004.

3. *Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework.* **Li-Jia, Li, Richard, Socher and Li, Fei-Fei.** s.l. : Internation Conference on Computer Vision.