Local Visual Feature Detection and Description for Non-rigid 3D Objects

KAIMAN ZENG, NANSONG WU, KANG K. YEN
Florida International University
Department of Electrical and Computer Engineering
10555 West Flagler St. EC 3900, Miami
USA
kzeng001@fiu.edu, nwu001@fiu.edu, yenk@fiu.edu

Abstract: Feature extraction is an essential step in various image processing and computer vision tasks, such as object recognition, object tracking, image retrieval, augmented reality, and so on. Design of feature extraction method plays the most significant role in achieving high performance of various tasks. Different applications create different challenges and requirements for the design of visual features. In this paper, we explored and investigated the effectiveness of different combinations of promising local feature detectors and descriptors for non-rigid 3D objects. Different configurations of visual feature detectors and descriptors have been enumerated, and each configuration has been evaluated by image matching accuracy. The results indicated that the scale-invariant feature transform feature detector and descriptor achieved the best overall performance in describing local features of non-rigid 3D object.

Key–Words: Feature Extraction, Local Feature Detector, Local Feature Descriptor, Image Matching

1 Introduction

Feature extraction plays an important role in many computer vision tasks. A good feature should properly represent the image characteristics, be repeatedly detected in images that capture the same objects/scenes while under different imaging condition, and also be distinctive so that it could distinguish it from other similar images. Besides, an ideal feature should be robust to imaging variations, such as rotation, viewpoint changes, illumination changes and occlusions. There is no universal defined feature, since different problems and different types of applications often have different characteristics. When the application domain changes, it usually requires re-designing feature detector and descriptor to capture features and achieve high performance. A feature is referred to as an interesting point/region in an image. Interesting points/regions are visually salient. Design of feature extraction method is probably the single most important factor in achieving high performance of various computer vision tasks [1]. Given the large number of feature extraction methods researched in the literatures, which feature extraction method is the best for a given application? This question leads us to characterize the available feature extraction methods, so that the most promising methods could be sorted out. In this paper, we concentrated on 3D object under different viewpoint. In particular, we are interested in recognizing 3D objects whose shape is neither fixed nor known a priori. Previous work on object recognition has concentrated on rigid objects of known 3D shape to simplify the task [2, 3]. These approaches have difficulty in dealing with unstructured objects, and thus cannot be applied to more generic categories of objects. Non-rigid object is a significant challenge because of its large variation and deformation within the object classes. The non-rigid deformation often observes large variation globally. Their local structures are somewhat more invariant to the changes. On that basis, our focus is on non-rigid 3D object recognition with local features.

Image local feature extraction usually consists of two stages: feature detection and feature description. A local feature commonly refers to a local pattern in an image that changes from its direct neighborhood in property or multiple properties of intensity, color, and texture simultaneously. Feature detection is algorithms that compute abstractions of image information and make local decisions at every image pixels whether there is an image feature of a given property type. The resulting features are subsets of the image domain, often in the form of isolated points, continuous curves or connected regions. Once the feature is detected, the local image patch around the feature is extracted and generated as the feature descriptor.

In this paper, the effectiveness of several promis-

ing local features on 3D non-rigid objects are explored and investigated. We configure different visual feature detectors and descriptors, and evaluate each configuration in detail. To the best of our knowledge, existing research on the comparison of visual feature detectors and descriptors are conducted for other computer vision tasks. In literature [3] the effectiveness of different visual feature detectors and descriptors are compared for mobile visual search of rigid product like books and CDs. The comparison study in literature [4] is focused on the visual object categorization. Neither of these comparisons targeted the effectiveness of 3D object recognition, the focus of this paper. The performance of different combination of visual feature detectors and descriptors on non-rigid 3D object has not been fully understood. The contribution of our work is filling this knowledge gap. Different combinations of detector and descriptor are enumerated and evaluated by the accuracy of image matching. This accuracy indicates how accurately the repeatable salient local features can be detected, described, and matched from one imaging condition to another.

The paper is organized as the follows. In Section 2, we have a literature review of classic and recent feature extraction techniques. Section 3 discusses the details of the researched feature detectors and descriptors. In Section 4, several experiments of different combination of feature detector and descriptor are conducted on the benchmark datasets. And their performances are compared in the forms of accuracy of image matching. Finally, we conclude comparison results with promising feature extraction techniques and discuss future works in Section 5.

2 Related Work

Local feature, representing local patches of an image, has shown promise in many tasks of computer vision, such as image match, object recognition, image registration and so on. Feature detection is utilized as the initial step in local feature extraction algorithms. It is a classic research area in image processing and computer vision. And there are a variety of different types of features, e.g. edges, corners/keypoints, regions of interest and ridges. The corner/keypoint is treated as the same concept since a corner can be not only considered as an intersection of two lines, but also a point that has two different edge directions within a local window of the point. Likewise, a keypoint can be defined as a corner, line endings, a point of local intensity maximum or minimum, or a point on a curve where the curvature is local maximum. As a result, the corner/keypoint detection is mainly divided into edge-based method and gray density based method. Current research is focused on gray density based corner/keypoint detection, since a small degree variation of the target object leads to great difference in edge extraction, and the edge extraction is computationally expensive [5, 6]. Gray density based approach detects the corner/keypoint by calculating the curvature and gradient of points. Moravec operator, Forstner operator, Harris operator and SUSAN operator are some of the examples. Harris operator [7] is the most classic detector among them. Mikolajczyk takes the scale space theory into consideration and proposes Harris-Laplace detector, which applies Laplace-of-Gaussian (LoG) for automatic scale selection [8]. It obtains scale and shape information and can represent local structure of an image. Lowe applies Difference-of-Gaussian (DoG) filter, an approximate to LoG, in the SIFT algorithm to reduce computational complexity [9]. Also, in order to increase the algorithm efficiency, Hessian Affine, FAST, Hessian-blobs, and MSER are further proposed. In [10], Mikolajczyk et al. extract 10 different keypoint detectors within a common framework and compare them for various types of transformations. Van de Sande extracts 15 types of local color features, and examines their performance on transformation invariance for image classification. Many detection methods are studied seeking a balance between keypoint repeatability and computational complexity [11].

After the keypoint detection, we compute a descriptor on the local patch. Feature descriptors can be divided into gradient-based descriptors, spatial frequency based descriptors, differential invariants, moment invariants, and so on. Among them, the histogram of gradient-based method has been wildly used. The gradient histogram is used to represent different local texture and shape features. The Scale Invariant Feature Transform (SIFT) descriptor proposed by Lowe is a landmark in research of local feature descriptor. It is highly discriminative and robust to scaling, rotation, light condition change, view position change, as well as noise distortion [9]. Since then, it has drawn considerable interests and a larger number descriptors based on the idea of SIFT emerges. SURF [12] uses the Haar wavelet to approximate the gradient SIFT operation, and uses image integral for fast computation. DAISY [13, 14] applies the SIFT idea for dense feature extraction. The difference is that DAISY use Gaussian convolution to generate the gradient histogram. Affine SIFT [15] simulates different perspectives for feature matching, and obtains good performance on viewpoint changes, especially large viewpoint changes. Since SIFT works on the gray-scale model, many color-based SIFT descriptors are proposed to solve the color variations, such as CSIFT, RGB-SIFT, HSV-SIFT, rgSIFT, Hue-SIFT, Opponent

SIFT, and Transformed-color SIFT [11, 16, 17]. Most of them are obtained by computing SIFT descriptors over channels of different color space independently; therefore they usually have higher dimension (e.g. 3×128 dimension for RGB-SIFT) descriptors than SIFT. The color boosted SIFT introduced in [18] involves the amended color histogram factor based on RGB color space model into the SIFT. It retains sufficient color information and is robust to photometrical variations. Song et al. proposed compact local descriptors using an approximate affine transform between image space and color space [19]. Burghouts et al. performed an evaluation of local color invariants [20].

3 Local Feature Extraction for Non-rigid Object

In this section, we discuss the visual features considered in our work. The feature detectors include Harris, FAST, SIFT, SURF, and BRISK detectors. For the descriptions, the BRISK, SIFT, and SURF feature descriptors are considered. We choose these feature detectors and descriptors for the following reasons. First, Harris detector is the best-known operator around. The SIFT is the most widely used and successful detector developed in recent decade for different computer vision. The FAST, SURF, and BRISK detectors achieve a good balance between the detection performance and computation complexity. Second, the selected feature descriptors have the potential to handle the task of object recognition based on previous studies of other researchers. For instance, Chandrasekhar et al. [3], compared several feature descriptors for visual search application, and reported the SIFT feature descriptor as one of the promising one. The SIFT and SURF are concluded in Lankinen?s work [4] as the top two reliable descriptors for visual object classification. The BRISK descriptor is considered in our work because of its big advantage in computation speed.

3.1 Harris Detector

Harris detector, proposed by Harris and Stephens [7], is developed from the auto-correlation matrix, also called the second moment matrix. Given an image I, an approximation to the local auto-correlation matrix of I is computed at every pixel (x, y):

$$M(x,y) = \begin{bmatrix} \sum w_{u,v} I_x^2(x_r,x_y) & \sum w_{u,v} I_x(x_r,x_y) I_y(x_r,x_y) \\ \sum w_{u,v} I_x(x_r,x_y) I_y(x_r,x_y) & \sum w_{u,v} I_y^2(x_r,x_y) \end{bmatrix} \quad (1)$$

where I_x and I_y are the partial derivative of image $I(x, y)$ with respect to x and y. $(x_r, y_r) = (x + u, y + v)$ and $w(u, v)$ is the weighting function. $w(u, v)$ can be a constant or a Gaussian function $\exp(\frac{-(u-x)^2-(v-y)^2}{2\sigma^2})$.

M presents the gradient distribution in a local neighborhood of an image pixel (x, y). The image pixel can be classified into three regions according to the eigenvalues λ_1 and λ_2 of M. If both λ_1 and λ_2 are small, the image pixel belongs to flat region. If λ_1 is far larger than λ_2 or vise versa, the image pixel is located in edge region. If both λ_1 and λ_2 are large and $\lambda_1 \approx \lambda_2$, the pixel is the corner in the image. In order to reduce the computation cost, Harris proposed a cornerness measure that derived from two eigenvalues:

$$\begin{aligned} c(x,y) &= \lambda_1\lambda_2 - \alpha(\lambda_1+\lambda_2)^2 \\ &= \det(M(x,y)) - \alpha \cdot [\mathrm{trace}(M(x,y))]^2 \end{aligned} \quad (2)$$

where $c(x, y)$ denotes the cornerness measure, $\det(M(x, y))$ is the determinant of $M(x, y)$, and $\mathrm{trace}(M(x, y))$ is the trace of $M(x, y)$. α is the experience constant, typically ranging from 0.04 to 0.06.

Then, non-maximum suppression is performed in a 3×3 or 5×5 neighborhood, and the local maxima of the corneress function forms the corner features of the image.

3.2 Features from Accelerated Segment Test

FAST is a high-speed corner detector developed by Rosten and Drummond [21]. The detection is performed on a discrete Bresenham circle around a candidate image pixel p. If there is a set of contiguous pixels at least nine on the circle around p, and they are all brighter or darker than the intensity of p by a predefined threshold t, then p is considered as a corner candidate. Besides, the algorithm is accelerated with a decision tree to reduce the number of pixels that need to be processed. Subsequently, the following score is computed at each corner candidate to remove the false candidates:

$$s(p \max(\sum_{q \in S_+} |I_q - I_p| - t, \sum_{q \in S_-} |I_q - I_p| - t) \quad (3)$$

where S_+ is the subset of contiguous pixels that are brighter than p by t on the circle. S_- is the subset of contiguous pixels that are darker than p by t on the circle. The corner candidates, who have an adjacent corner with a higher score, will be removed. Then, non-maximum suppression is applied to locate corner features.

3.3 Binary Robust Invariant Scalable Key Points

BRISK, proposed by Leutenegger et al. [22], is a binary local feature detection and description method

with very high computational efficiency. The first step is to create a scale space pyramid, generally consisting of 4-layer octave images and 4-layer intra-octave images. Each octave is half-sampled from previous octave, and each intra-octave is down-sampled so that it is located between two octaves. Next, the FAST detector score s is computed at each octave and intra-octave to generate the keypoint candidates. Non-maximum suppression is then performed at each octave and intra-octave so that score s is the maximum within a $3times3$ neighborhood; and score s is the largest among the scales above and below. These maxima are then interpolated using a 1D quadratic function across scale spaces and the local maximum is chosen as the scale for the feature found.

Given a set of the detected keypoints, the BRISK descriptor is constructed as a binary descriptor by simple brightness comparison tests. The brightness comparison test is performed on the samples in a pattern. This pattern is defined as N equally spaced locations on circles concentric with the keypoint.

3.4 Scale-Invariant Feature Transform

SIFT, introduced by Lowe [9], is a scale invariant feature detector with highly distinctive feature descriptor. In order to achieve scale invariance, a scale space pyramid of images is first built through convolutions of image I with differences of Gaussians (DoG) at different scales σ:

$$DoG_{k,\sigma}(x,y) = G(x,y,k\sigma) - G(x,y,\sigma) \qquad (4)$$

Then, each sample is compared with its 3×3 neighbors at current layer I_n, as well as the 3×3 neighbors from layers above and below (I_{n-1} and I_{n+1}) at the same octave. These local extrema are considered as keypoints. Further, the keypoint location is refined by interpolating the sample points and its direct neighbors. Keypoints with low contrast and small ratio of principal curvatures are removed. Subsequently, the gradient magnitudes and orientations of the remaining keypoints are computed. The orientations are then weighted by a Gaussian window and the gradient magnitude, and the dominant orientations are sorted out from the histogram of the weighted orientations. If multiple dominant orientations exist at a keypoint, for every dominant orientation an additional keypoint are generated.

Now, the located keypoints have been assigned with orientations and scales. A local coordinate system can be defined to compute the SIFT descriptor. A new orientation histogram is computed within a 16×16 local window and then 4×4 sub windows. For each sub window, the orientation histogram is calculated with 8 bins and weighted again by a Gaussian window and corresponding gradient magnitude. This yields the SIFT descriptor of length 128 ($4 \times 4 \times 8$).

3.5 Speeded-Up Robust Features

SURF, designed by Bay et al. [12], is similar to SIFT with faster feature detection and description. SURF detector is developed from the determinant of the Hessian matrix. It then employs box filters to approximate the second order Gaussian partial derivative for scale space analysis. The score in SURF is defined as:

$$\begin{aligned} s(x,y,\sigma) &= D_{xx}(\sigma) \cdot D_{yy}(\sigma) - [0.9D_{xy}(\sigma)]^2 \\ &\approx \det(H(x,y,\sigma)) \end{aligned} \qquad (5)$$

where D_{xx}, D_{yy} and D_{xy} are the convolution of the image using box filters. Constant factor 0.9 is chosen to make the approximate solution closer to $\det(H(x,y,\sigma))$. Then, a non-maximum suppression is performed in a $3 \times 3 \times 3$ neighborhood, and the resulted maxima are interpolated across scale spaces to localize the keypoints. Once the SURF features are localized, the SURF description is addressed in two steps: first, extracting an orientation according to the information from a circular region around the keypoints; second, defining a square region oriented along the formed orientation, and computing the SURF descriptor from the square region. Specifically, the circular region in the first step is convoluted with Haar wavelet along x and y axes. The radius of the circular neighborhood is decided by the scale, at which the keypoint is detected. So do the sampling step and wavelet response. The wavelet response is then weighted with a Gaussian, and represented as a vector with response strength along x and y axis. The dominant orientation is determined by the sum of all responses within a rotating square window. Next, this orientation window is further split up to 4×4 sub square windows, and the descriptor vector is defined as:

$$v = \left[\sum d_x \quad \sum d_y \quad \sum |d_x| \quad \sum |d_y| \right] \qquad (6)$$

d_x and d_y denote the Haar wavelet responses in x and y directions for each sub square region. The generated descriptor vector has a length of 64 ($4 \times 4 \times 4$).

4 Experiments and Analysis

4.1 Data Set

In order to evaluate the performance of different feature detectors and descriptors, we conducted several experiments of image matching on the benchmark dataset of Oxford Dataset [23]. We also perform experiments on the benchmark dataset of Columbia

Object Image Library - COIL 100 [24]. Figure 1 show typical images selected from these datasets. The Oxford dataset has been widely used for evaluating performance of local image descriptors. It contains image pairs under various image transformations, including scale, rotation, image blur, illumination, JPEG compression and viewpoint changes. The dataset also contains ground truth homographies corresponding to the image pairs. Figure 1 (a) shows some image pairs under different image transformations in this dataset. COIL 100 is a database of color images of objects. The objects are placed on a motorized turntable against a black background. The turntable is rotated through 360 degrees to vary object pose with respect to a fixed color camera. Images of the objects are taken at pose intervals of 5 degrees. This corresponds to 72 poses per object and the images are size normalized.

(a)

(b)

Figure 1: Typical images selected from the datasets

4.2 Experimental Evaluation and Analysis

In this experiment we implement 5 feature detectors (Harris, BRISK, FAST, SIFT and SURF) and 3 descriptors (BRISK, SIFT, and SURF) in MATLAB. All combinations are evaluated except for the SIFT-BRISK, since the SIFT detector is not compatible with BRISK descriptor.

Table 1: Average accuracy for different combinations of feature detectors and descriptors

	Descriptor		
Detector	BRISK	SIFT	SURF
Harris	0.3351	0.3264	0.3018
BRISK	0.4288	0.4113	0.3907
FAST	0.4637	0.5021	0.4579
SIFT	N/A	0.5137	0.3725
SURF	0.4110	0.4556	0.4232

The average accuracy of image matching for every combination of feature detectors and descriptors are recorded in Table 1. The results show that the combination of SIFT-SIFT provides the most accurate matching features at matching rate of 0.5173. Following it, the combination of FAST-SIFT achieved comparable performance of matching rate 0.5021. With the same detector, SIFT descriptor and BRISK descriptor performs better than SURF descriptor generally, except for the case of SURF detector.

5 Conclusion

In this paper, we evaluated the effectiveness of different combinations of local feature detectors and descriptors for non-rigid 3D objects. We selected several classic and widely used visual feature detectors (Harris, BRISK, FAST, SIFT, and SURF) and descriptors (BRISK, SIFT, and SURF). The primary difference between this work and the comparison studies of other researchers is that they are targeted in different applications, so that face in different visual characteristics and raise new challenges. It was unclear which feature detection and description methods are best suitable for non-rigid 3D objects. Our evaluation results indicated that the SIFT achieved the best overall performance in describing image local features. This finding could benefit reshaping existing or ongoing other research work based on visual feature, such as non-rigid object visual search. We will use these findings in the future to tune and design new visual features to improve object recognition accuracy and adapt to different applications.

References:

[1] Trier, Ø.D., A.K. Jain, and T. Taxt, Feature extraction methods for character recognition-a survey. Pattern recognition, 1996. 29(4): p. 641-662

[2] Knopp, J., et al. Hough transform and 3D SURF for robust three dimensional classification. in Computer Vision?ECCV 2010. 2010. Springer.

[3] Chandrasekhar, V., et al. Comparison of local feature descriptors for mobile visual search. in Image Processing (ICIP), 2010 17th IEEE International Conference on. 2010. IEEE.

[4] Lankinen, J., V. Kangas, and J.-K. Kamarainen. A comparison of local feature detectors and descriptors for visual object categorization by intra-class repeatability and matching. in Pattern Recognition (ICPR), 2012 21st International Conference on. 2012. IEEE.

[5] He, X.-C. and N.H.C. Yung. Curvature scale space corner detector with adaptive threshold and dynamic region of support. 2004. IEEE.

[6] Mokhtarian, F. and R. Suomela, Robust image corner detection through curvature scale space. Ieee Transactions on Pattern Analysis and Machine Intelligence, 1998. 20(12): p. 1376-1381.

[7] Harris, C. and M. Stephens. A combined corner and edge detector. 1988. Citeseer.

[8] Mikolajczyk, K. and C. Schmid. Indexing based on scale invariant interest points. in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on. 2001. IEEE.

[9] Lowe, D.G., Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004. 60(2): p. 91-110.

[10] Mikolajczyk, K. and C. Schmid, A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005. 27(10): p. 1615-1630.

[11] Van De Sande, K.E.A., T. Gevers, and C.G.M. Snoek, Evaluating color descriptors for object and scene recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010. 32(9): p. 1582-1596.

[12] Bay, H., T. Tuytelaars, and L. Van Gool, Surf: Speeded up robust features, in Computer vision?ECCV 2006. 2006, Springer. p. 404-417

[13] Tola, E., V. Lepetit, and P. Fua, Daisy: An efficient dense descriptor applied to wide-baseline stereo. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010. 32(5): p. 815-830

[14] Winder, S., G. Hua, and M. Brown. Picking the best daisy. 2009. IEEE.

[15] Morel, J.-M. and G. Yu, ASIFT: A new framework for fully affine invariant image comparison. SIAM Journal on Imaging Sciences, 2009. 2(2): p. 438-469.

[16] Behmo, R., et al., Towards optimal naive bayes nearest neighbor, in Computer Vision?ECCV 2010. 2010, Springer. p. 171-184.

[17] Van De Weijer, J. and C. Schmid. Applying color names to image description. 2007. IEEE.

[18] Zeng, K., N. Wu, and K.K. Yen, A Color Boosted Local Feature Extraction Method for Mobile Product Search. Int. J. on Recent Trends in Engineering and Technology, 2014. 10(2): p. 78-84.

[19] Song, X., D. Muselet, and A. Trmeau, Affine transforms between image space and color space for invariant local descriptors. Pattern Recognition, 2013. 46(8): p. 2376-2389.

[20] Burghouts, G.J. and J.-M. Geusebroek, Performance evaluation of local colour invariants. Computer Vision and Image Understanding, 2009. 113(1): p. 48-62.

[21] Rosten, E. and T. Drummond, Machine learning for high-speed corner detection, in Computer Vision?ECCV 2006. 2006, Springer. p. 430-443.

[22] Leutenegger, S., M. Chli, and R.Y. Siegwart. BRISK: Binary robust invariant scalable keypoints. in Computer Vision (ICCV), 2011 IEEE International Conference on. 2011. IEEE.

[23] Oxford Affine Covariant Features dataset. http://www.robots.ox.ac.uk/ vgg/research/affine/.

[24] Columbia University Image Library. http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php.