Object Recognition Based on ORB Descriptor for Markerless Augmented Reality

4 authors:

Mahfoud Hamidia
University of Science and Technology Houari Boumediene

Nadia Zenati
Centre Algérien de Développement des Technologies Avancées

Hayet Belghit
Centre Algérien de Développement des Technologies Avancées

Abdelkader Bellarbi
Centre Algérien de Développement des Technologies Avancées

Object Recognition Based on ORB Descriptor for Markerless Augmented Reality

Mahfoud HAMIDIA[1,2], Nadia ZENATI-HENDA[1]
[1]Centre de Développement des Technologies Avancées, CDTA, B.P. 17, 16303, Baba-Hassen, Algiers, Algeria.
[2]Faculty of Electronics and Computer Science, USTHB, B.P. 32, 16111, Bab-Ezzouar, Algiers, Algeria.
[2]mhamidia@usthb.dz

Hayet BELGHIT[1] ,Abdelkader BELLARBI[1]
[1]Centre de Développement des Technologies Avancées, CDTA, B.P. 17, 16303, Baba-Hassen, Algiers, Algeria.
[1]{mhamidia, nzenati, hbelghit, abellarbi}@cdta.dz

***Abstract*—This paper investigate a binary local image descriptor for Augmented Reality (AR) applications. Recently, various fields are benefit from AR. This technique can enhance the real environment by inserting virtual objects generated by computer. Temporal coherence between virtual and real objects must be ensure in AR system realization. In this paper, object recognition based on extracted natural features is presented. Experimental results indicate better performance of object recognition using ORB(Oriented FAST and Rotated BRIEF) descriptor compared to the SURF(Speed Up Robust Features) descriptor in AR applications.**

Keywords—Augmented reality; object recognition; local binary descriptor; feauture matching; ORB; SURF.

I. INTRODUCTION

Augmented reality (AR) takes an interest part of studies in recent years, due to the fast emergence of technologies and a wide use in different applications. AR is defined as a combination between a real scene viewed by the camera and a virtual scene generated by computer as is shown in Figure 1, which presents an example of a desk augmentation.

AR enhances the user's perception and interaction with the physical-real world and simplifying his life by inserting virtualinformation into his viewed scene [1]. Azuma [2] defined AR systems that have three characteristics: they combine real and virtual elements, they are interactive in real time and they are registered in 3D.

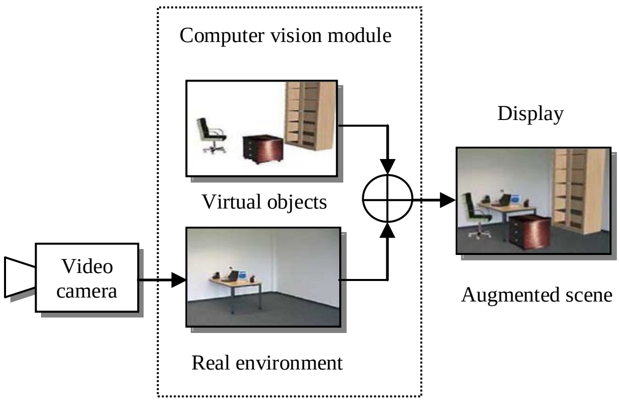

Figure 1. Global structure of AR system.

Different displays devices are used in AR, can be cited for example: the well-known device Head Mounted Displays (HMD), and handheld devices (personal digital assistants, smart phone, tablet-PC, etc.).

Coherence spatio-temporal is one of the most challenging tasks in AR. In order to, the real and computer-generated objects must be accurately positioned relative to each other in the real time.

In this context, two kinds of AR can be cited: marker AR and markerless AR.

Marker augmented reality is based on the use of coded tags, inserting in the real word for fast recognizing a right position of an inserted virtual object. In fact, different fiducial markers have been proposed, ARToolKit [3] is the most marker used in AR applications. This marker consists of a black square border a round of a pattern. Marker recognition based on the matching between the captured markers in pre-recorded models of database. Moreover, to enhance marker recognition in different changes of image, various markers are used such as, Cyber Code [4], QR codes [5], ARtag [6], etc. The main advantage of these marker is fast is fast to recognize in a real scene. In other hand, these marker have some disadvantages such as, sensitive of illumination change, partial occlusion and unrealistic to use these marker for each augmentation in complex environments.

For this raison, markerless augmented reality is investigated. Markerless tracking based on image processing uses natural features that, color, shape, texture and interest point, in images to calculate acamera's pose.

Interest points of the environment to be augmented for tracking are mostly use [7]. That represents distinctive visual characteristics. The main purpose is to track a real object by recognizing it in each frame of video stream.Several methods of interest point detection are available in the literature. Harris [8] is the most popular corner detector using in computer vision.

The local interest point description consists of the presentation of the region around of the keypoint (patch) using neighborhood information. SIFT (Scale Invariant Feature Transform) algorithm proposed by Lowe [9], it is efficient to find interest points using the differences of Gaussians of the image, and a very robust descriptor based on histogram of gradient. SURF (Speed Up Robust Features) algorithm is presented in [10], which approximates to SIFT andoutperforms other method. Although these histogram gradient basedalgorithms show competitive performance, but they have high computational complexity [11].

Recently, several binary keypoint descriptors are proposed for efficient local feature matching in real-time applications. These descriptors have some advantages, for example low computational complexity for vector descriptor calculates.

In this paper, ORB (Oriented FAST and Rotated BRIEF) [12] local binary descriptor is used for object recognition based on feature matching, which to apply in AR applications. The main purpose is to recognize the real scene and keeps a right position of virtual object into the real object in the real time. In which realize a realistic augmentation of a real environment.

The remainder of this paper is organized as follows: Object recognition based on interest points is presented in Section II. In Section III, ORB feature descriptor is described, then; the experimental results are discussed in Section IV. Finally, the paper is concluded in Section V.

II. Object Recognition Based on Feature Matching

Figure 2 show the block diagram of object recognition process in the video sequence based on interest pointmatching. The principle of this technique is based on the features extraction from the reference image (object), and features extraction from each frame of a video sequence (real-time image or query image). After that, the matching between these features is determined by the union of corresponding points using a similarity measure. For recognize the object in the real-time image, homography matrix is calculated to find the object corner in this latter.

In markerless AR, the interest points are identified and tracked in real-time and for each frame of the video sequence todetermine the position and orientation (pose) of the observing camera, which can to insert the virtual object in the right position.

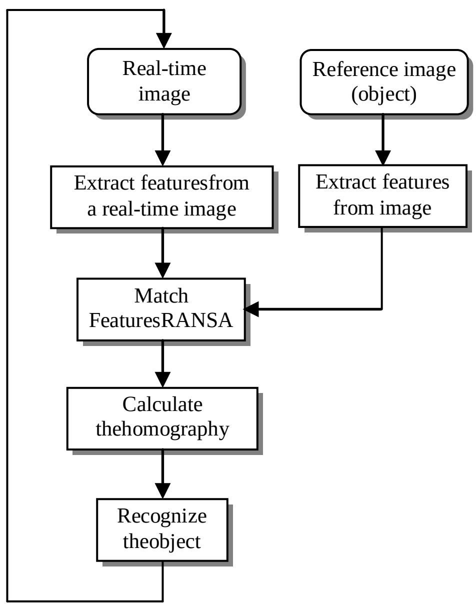

Figure 2. Block diagram of object recognition using feature matching in the real-time.

A. Feauture extraction

Feature extraction consist of two steps: interest points detection from the image, and interest points description for generating vector descriptors. These features must be robust to various variations in viewpoint; pose and lighting condition of image for enhance the performance of object recognition.

1) Interest points detection

Akeypoint (or interest point) is defined by some particular image intensitiesaround it, such as a corner [13]. During the past few years, interest points detectors have been very popular in computer vision, and have been widely applied to object recognition. Different techniques have been proposed to detectinterest points inimages.We can cite the Harris detector [8],FAST (Features from Accelerated Segment Test) [14],and local scale invariant features detectors:SIFT [9], SURF [10]. Object recognition using interest points presents numerous advantages: invariant to the image change (translation, scale and rotation), including robustness to partial occlusion without significant increase in complexity.

2) Interest points description

A wide variety of feature description algorithms have been presented over the years. Local descriptor aim at describing a region around of the keypoint for generates the description vector.An ideal descriptor should achieve two competing goals: high quality description and low computational complexity [15]. We can distinguish to classes of interest point descriptors. These are: histogram gradient based descriptors, such as SIFT, GLOH (Gradient Location and Orientation Histogram) [16], SURF, and local binary descriptors, we can cite, BRIEF (Binary Robust Independent Elementary Features) [17], ORB, BRISK (Binary Robust Invariant Scalable Keypoints) [18], FREAK (Fast REtinAKeypoint) [19].

B. Features matching

Correspondences between interest points can be found by matching of their features (vector descriptors) from a reference and real-time images. This matching between two features is computed by involving similarity measure between vector descriptors. For computing similarity measure, numerous measures are employed such as distance (Minkowsy, Euclidean, chi-squared or Hamming in Local binary descriptor matching), normalized cross correlation. Different features matching approaches can be cited.These are: nearest-neighbor, ratio-based nearest-neighbor, and threshold-based matching.To improve the matching even more, the random sample consensus (RANSAC) algorithm [20] can be used to perform outlier filtration.

C. Homography matrix estimation

Homography is a linear invertible projective transformation that maps points from one plane into another plane. So, for recognizing the object must be finding the homography transformation between feature points on the reference image and feature points on each frame of a video sequence. To estimate this matrix, four point matches, are needed.

SingularValue Decomposition (SVD) method is generally used to calculate it.

III. ORB DESCRIPTOR

Local binary features have several advantages over histogram of gradient based features as they can be faster to compute, more compact to store, and more efficient to compare.Although, it is fast to compute the Hamming distance between pairs of binary features.Binary descriptor bit string $BD = (bit1, \ldots bitM)$ is obtained by a multitude of image intensity comparison between pair pixels (p_i, q_i)around of the keypoint position after smoothing the image I. Let $bit_i = 1$ If $I(p) - I(q) > 0$ and 0 otherwise.

BREIF descriptor used a random order for pairs of pixel locations. Thus, scale or rotation invariance was notintended by BRIEF descriptor.

ORB is a modified version of FAST detector for computing orientation during detection step, with an efficient extension of the feature descriptor BRIEF, This approach tries to merge the rotation and scale invarianceof SIFT and the computational efficiency of FAST detector.

ORB improves rotation invariance by computingan orientation vector based on intensity centroid method defined in [21]. Thismethod proposes form a vector from the center of the patch O to the centroid point C computed by using central moments [22], defined by the following equation:

$$m_{n_1 n_2} = \sum_{xy} x^{n_1} y^{n_2} I(x, y) \tag{1}$$

where$I(x, y)$ represents intensity of pixels at position (x, y) of image I and $n_1 = n_2 = 1,2,3$.

Thecentroid C is determined by:

$$C = \left(\frac{m_{10}}{m_{00}}, \quad \frac{m_{01}}{m_{00}}\right) \tag{2}$$

Join points O and C we have the vector $\overrightarrow{OC}$, representing dominant orientation, and angleθ can be computed as:

$$\theta = \text{atan } 2\ (m_{01}, m_{10}) \tag{3}$$

where atan 2 (m_{01}, m_{10}) is the angle between the *difference vector*(m_{01}, m_{10}) and the x-axis.

Moment calculationsare performed within a circular region of radius r centered on point (x, y). Finally,ORB performs a Gaussian smoothing of image patch before BRIEF descriptor computation, in order to increase robustness against digital noise.

IV. RESULTS AND DISCUSSION

This section presents experimental results and discussion of object recognition based on feature matching using ORB descriptor, then an example of scene augmentation.

These results are performed using visual studio 2010 C# environment with the popular open source library for computer vision OpenCV [23]. The computer processor is writing Figure Intel ® core TM, i3, M380, 2.53 GHz. The Logitech Quick Cam Pro 9000 webcam is used with the resolution 320x240.

In augmentation task, AR systems need to know the relationship between the real scenes (3D world) and the image corresponding (2D) of this one [24]. So, transformation matrix must be calculated using the least squares method to obtain the matrix which relates 3D points with theirprojections. To insert the virtual object in the right into a real scene, we use the obtained reference point from the object to be recognized.

The experimental results are shown in the following figures, where the graffiti image presents the object to be recognized and tracked using matching feature.

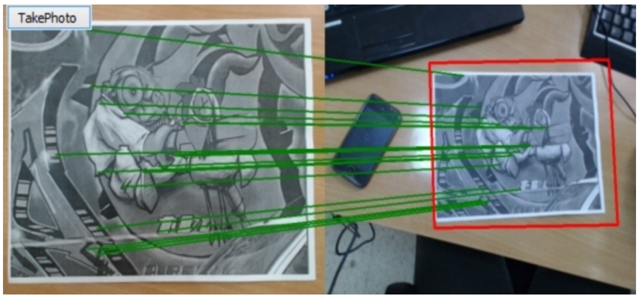

Figure 3. Object rocognition for one frame: (right) reference image, (left) real-time image, (red) object to be recognized, (green) the matching.

Figure 3 shows object recognition and tracking using interest points matching between the reference image (object) and a real-time image. These features are extracted from each frame of the video sequence and are corresponded with the feature of the reference image. The object can be located in the real-time image using homography matrix estimation.

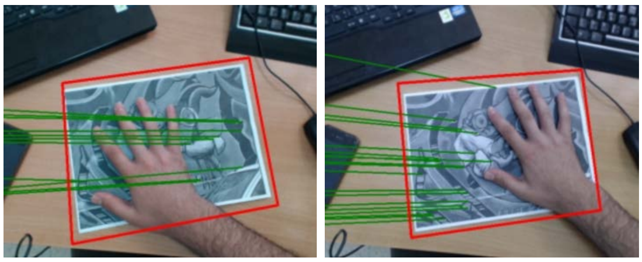

Figure 4. Robustness of the object recognition against partial occlusion.

Performance of object recognition and tracking in term of robustness against partial occlusion is shown in Figure 4. That can recognize object using the appearanceparts of object.

To evaluate the low computational complexity of ORB detector we compare this latter with SURF detector. We use an image of 1024×768 pixels size and we calculate the detection time in function of interest point number. This number of points detected can be controlled by the detection threshold and the image size.

Curves of Figure 5 present the evaluation of execution time in function of number of points, this result indicates that ORB detector if fast then SURF detector. In this latter the complexity increase in function of interest point detected.

In augmented reality task, this result can improve the augmentation and take it realistic.

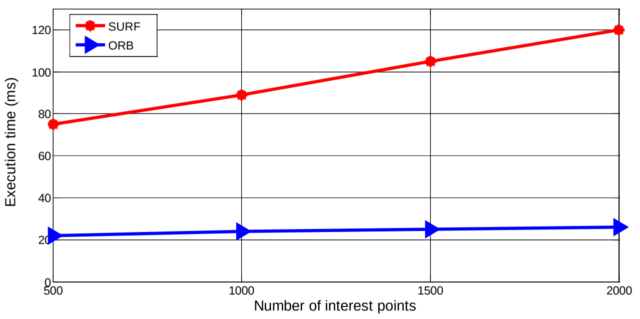

Figure 5. Robustness of the object recognition against partial occlusion.

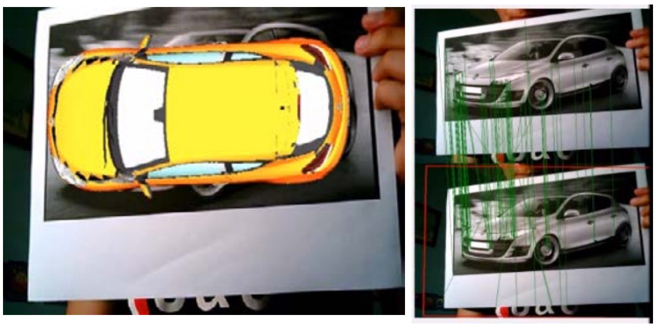

Figure 6. Augmented scene: (right) augmented car, (left) the mathing.

Figure 6 shows an example of augmented car (orange color), a virtual 3D car is inserted into the real-world scene for augmentation purposes, can be benefit in marketing field.

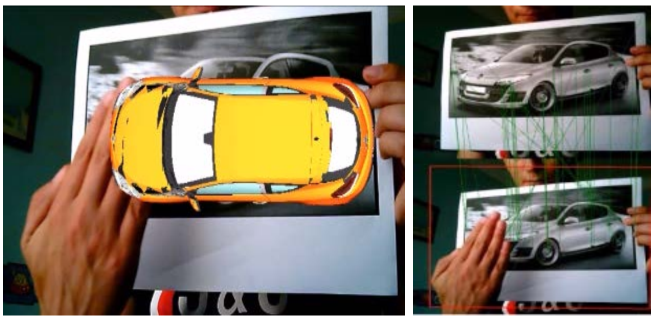

Figure 7. Robustness of Markerless AR against partial occlusion

Robustness against partial occlusion is demonstrated in Figure 7, partial occlusion of real scene without the hidden of virtual object. Thus, markerless AR based on feature matching has better performance compared to marker AR.

V. Conclusion

In this paper, we have investigated the local binary descriptor ORB in Markerless AR. Furthermore, we have focused in spatio-temporal coherence between real and virtual objects using visual feature.The point matching based methodsprovide good recognition and tracking performances, because they are robust to partial occlusion.

The obtained results demonstrate a good performance of object tracking in term of fast recognition and robustness against partial occlusion. Moreover, the advantages of binary features descriptor (ORB) there are: rapid extraction and storage efficiency compared to histogram of gradient based descriptor. ORB descriptor suffers some disadvantages, such as precision and invariance to affine projection.

In future works, we plan to proposed new descriptors using hybrid approaches to improve the object recognition task in AR application. So, AR can enhance the human perception, and apply in different applications.

References

[1] B. Furht, “Handbook of augmented reality,” New York: Springer, Vol. 71, 2011.

[2] R. T. Azuma, “Survey of augmented reality,” Presence, Vol. 6, No. 4, pp. 355–385, 1997.

[3] H. Kato, M. Billinghurst, “Marker tracking and hmd calibration for a video-based augmented reality conferencing system,” In proc of the 2nd International Workshop on Augmented Reality, San Francisco, USA, pp. 85–94, 1999.

[4] J. Rekimoto, Y. Ayatsuka, “CyberCode: designing augmented reality environments with visual tags,” In proc of DARE 2000 on Designing augmented reality environments. ACM, pp. 1–10, 2000.

[5] D. Wave, “QRcode.com,” 2003.
Available at www.denso-wave.com/qrcode/indexe. html

[6] M. Fiala, “ARTag, a fiducial marker system using digital techniques,” In Computer Vision and Pattern Recognition (CVPR), Vol. 2, pp. 590-596, 2005

[7] M. Hamidia, N. Zenati-Henda, H. Belghit, M. Belhocine, “Markerless tracking using interest window for augmented reality applications,” In proc IEEE, International Conference onMultimedia Computing and Systems (ICMCS), pp. 20-25, 2014.

[8] C. Harris, M. Stephens, “A Combined Corner and Edge Detector,” In proc of The Fourth Alvey Vision Conference, pp. 147-151, 1988.

[9] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, Vol. 60, No. 2, pp. 91–110, 2004.

[10] H. Bay, T. Tuytelaars, L. V. Gool, “SURF: Speeded up robust features,” In Computer Vision–ECCV, Springer Berlin Heidelberg, pp. 404–417, 2006.

[11] A. Bellarbi, S. Otmane, N. Zenati, S. Benbelkacem, “MOBIL: A Moments based Local Binary Descriptor,” In proc of IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR) pp. 251-252, 2014.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: an efficient alternative to SIFT or SURF,” In proc of the IEEE International Conference on Computer Vision (ICCV), pp. 2564–2571, 2011.

[13] R. Klette, “Concise Computer Vision,” Springer Ed, 2014.

[14] E. Rosten, T. Drummond, “Fusing points and lines for high performance tracking,” In proc of the Tenth IEEE International Conference on Computer Vision (ICCV) , Vol. 2, pp. 1508-1515, 2005.

[15] X. Yang and K. T. Cheng, “LDB: An ultra-fast feature for scalable augmented reality on mobile devices,” In proc of the International Symposium on Mixed and Augmented Reality (ISMAR), pp. 49–57, 2012.

[16] K. Mikolajczyk, C. Schmid, “A performance evaluation of local descriptors,”IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) Vol. *27*,No. 10, pp. 1615–1630, 2005.

[17] M. Calonder, V. Lepetit, C. Strecha, P. Fua, “Brief: Binary robust independent elementary features,” In Computer Vision–ECCV, Springer Berlin Heidelberg, Vol. 6314, pp. 778-792, 2010.

[18] S. Leutenegger, M. Chli, and R. Y. Siegwart, “ Brisk: binary robust invariant scalable keypoints,” In proc of the IEEE International Conference on Computer Vision (ICCV), pp. 2548–2555, 2011.

[19] A. Alahi, R. Ortiz, P. Vandergheynst, “Freak: Fast retina keypoint,” In proc of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 510-517, 2012.

[20] M. A. Fischler, R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” Communications of the ACM, Vol. 24, N°. 6, pp. 381–395, 1981.

[21] P. L. Rosin, “Measuring corner properties,” Computer Visionand Image Understanding, Vol. 73, No. 2, pp. 291 – 307, 1999.

[22] M.-K. Hu, “Visual pattern recognition by moment invariants,” IRE Transactions on InformationTheory, Vol. 8, No. 2, pp. 179–187, 1962.

[23] Open source computer vision library. Avalaibal in: http://www.intel.com/research/mrl/research/opencv

[24] H. Belghit, N. Zenati-Henda, A. Bellabi, S. Benbelkacem, M. Belhocine, “Tracking color marker using projective transformation for augmented reality application,” In proc of IEEE, International Conference on Multimedia Computing and Systems (ICMCS), pp. 372–377, 2012.