한국항행학회 논문지 제13권 제1호 2009년 2월

마커 없는 증강 현실 구현을 위한 물체인식

Object Recognition for Markerless Augmented Reality Embodiment

폴 안잔 쿠마*, 이형진*, 김영범*, 이슬람 모하마드 카이룰*, 백중환*

Anjan Kumar Paul*, Hyung-Jin Lee*, Young-Bum Kim*,
Mohammad Khairul Islam* and Joong-Hwan Baek*

요 약

본 논문에서는 마커 없이 증강 현실을 구현하기 위한 물체 인식 기법을 제안한다. 먼저 SIFT(Scale Invariant Feature Transform)알고리즘을 사용하여 물체 영상으로부터 특징점을 찾는데, 이러한 특징점들은 비율, 회전 또는 이동시에도 그 특징이 변하지 않는 장점이 있다. 또한 조도의 변화에도 일부는 변화지 않는 특성을 갖는다. 추출된 특징점의 독립적인 특성을 이용해 화면내의 다른 이미지의 매칭 포인트를 찾을 수 있는데, 학습된 영상과 매칭이 이루어지면, 매칭된 점을 이용해 화면내의 물체를 찾는다. 본 논문에서는 장면의 첫 프레임에서 발생하는 템플릿 이미지와의 매칭을 통해 현재의 화면에서 물체를 인식하였다. 네 종류의 물체에 대해 인식 실험을 한 결과 제안한 방법이 우수한 성능을 갖는 것을 확인하였다.

Abstract

In this paper, we propose an object recognition technique for implementing marker less augmented reality. Scale Invariant Feature Transform (SIFT) is used for finding the local features from object images. These features are invariant to scale, rotation, translation, and partially invariant to illumination changes. Extracted Features are distinct and have matched with different image features in the scene. If the trained image is properly matched, then it is expected to find object in scene. In this paper, an object is found from a scene by matching the template images that can be generated from the first frame of the scene. Experimental results of object recognition for 4 kinds of objects showed that the proposed technique has a good performance.

Key words : Augmented Reality, Scale Invariant Feature Transform, Difference of Gaussian

I. Introduction

Augmented reality is becoming important in different industrial applications for developing, production and servicing. In Augmented Reality, the user can see the real world around him, with computer graphics superimposed or composited with the real world. The goal of augmented reality is to add information and meaning to a real object or place. Unlike virtual reality, augmented reality does not create a simulation of reality. Instead, it takes a real object or space as the foundation and incorporates technologies that add contextual data to

* 한국항공대학교
· 제1저자 (First Author) : 폴 안잔 쿠마
· 투고일자 : 2009년 2월 9일
· 심사(수정)일자 : 2009년 2월 10일 (수정일자 : 2009년 2월 16일)
· 게재일자 : 2009년 2월 28일

deepen a person's understanding of the subject. Like by superimposing imaging data from an MRI onto a patient's body, augmented reality can help a surgeon pinpoint a tumor that is to be removed. In this case, the technology used might include headgear worn by the surgeon combined with a computer interface that maps data to the person lying on the operating table. In other cases, augmented reality might add audio commentary, location data, historical context, or other forms of content that can make a user's experience of a thing or a place more meaningful. In this way Augmented reality has been put to use in a number of fields, including medical imaging, where doctors can access data about patients aviation, where tools show pilots important data about the landscape they are viewing; training, in which technology provides students or technicians with necessary data about specific objects they are working with; and in museums, where artifacts can be tagged with information such as the artifact's historical context or where it was discovered.

To track any object correctly, first task is to recognize the object perfectly. In the applications of augmented reality where virtual objects are generated by the computer on the real scene, the position and tracking the real scene has its great significance. In traditional Augmented Reality applications marker based approach are very popular. The marker is attached to different objects and by detecting the marker, computer finds out position of the object and generates the virtual object on it. This approach is suitable for indoor augmented reality and simple scenario. The main drawback of the marker based system is the size and appearance of the marker. Most of the cases the markers need to have certain dimensions and appearance. Sometimes in the indoor environments where there are multiple objects are existed, markers may be detected wrong with other similar kind of objects. On the other side for outdoor augmented reality applications it is not convenient to hang markers to all objects like roads, buildings. So marker based system is not robust approach for outdoor environment. Robust Augmented Reality system must have the capability to recognize and match natural objects both in indoor and outdoor environment. Thus in the application for Augmented Reality for indoor and outdoor environment the important task is to recognize natural objects from the scene. Computer vision algorithms can be applied for recognizing natural objects. Suppose in the application of automatic navigation where we need to display the address or name of any building. To implement this kind of system, first we must need to create a database of the object image, extract the distinct features of these objects and match those features in real scene. This can be implemented for the indoor environment also. Many Researchers worked with natural object recognition without markers. The 2D feature extraction and matched with the priori known 3D models proposed by D.Beier [3] for markerless Augmented Reality Applications .Quan Wang et el. [4] used Multiple View Kernel Projection that combines a multiple view training stage and kernel projection for feature description for Augmented Exhibitions .Schiele et el.[5] used multidimensional receptive field histograms for object recognitions . David Marimon [6] used orientation histogram based matching for region finding. There are also several other object matching algorithms based on area, histogram and so on. Aibing Rao [7] used spatial color histograms; Cordelia Schmid Et el. [8] worked with Harris corner detector. They used rotational invariants features at corner points. These approach works well for rotation invariant but not worked well for scale invariant. It is also sensitive to viewpoint and illumination changes. These algorithms are limited in many aspects, in the context of scaling, rotation, distortion, background, illumination, etc. Many researchers worked with feature extraction. We used SIFT algorithm proposed by David Lowe [1]for extracting the features from images of objects. We first make some template images from different objects and find out the keypoints of those template images. We

tried to match based on still images but varying the scale, and alignment. Then we tried to match the template with the scene capture by the webcam. We get very impressive result, and in the case of scene matching we found the exact object from multiple objects those do not coincide with the other objects those presented in the scene.

II. Scale Invariant Feature Transform

Image matching is fundamental in computer vision for object recognition. For recognition of any object the matching must be done based on some features that can be extracted from the images. These image features may have many properties. For getting robust recognition performance these features must be invariant to image scaling and rotation and partially invariant to illumination changes.

The scale invariant feature transform (SIFT) (Lowe, 1999-2004) aims to resolve the practical problems in low level feature extraction and their use in matching images. SIFT involves two stages, feature extraction and description. The description stage concerns use of the low level features in object matching .Low-level feature extraction within the SIFT approach selects salient features in a manner invariant to image scale (feature size) and rotation and partial invariance to change in illumination. The Algorithm has four stages.

1. Scale-Space Extrema Detection
2. Keypoints Localization
3. Orientation assignment
4. Keypoint descriptor

2-1 Scale Space Extrema Detection

In this first stage the interest points, keypoints are detected within the SIFT framework. The image is convolved with Gaussian filters at different scales. The difference of successive Gaussian-blurred images is computed. Keypoints are then computed as maxima/minima of the Difference of Gaussians (DoG) that occur at multiple scales. A Difference of Gaussian image $D(x,y,\sigma)_i$ is defined as the follow.

$$D(x,y,\sigma) = L(x,y,k_i\sigma) - L(x,y,k_j\sigma) \quad (1)$$

Here $L(x,y,k\sigma)$ is the convolved output of original image $I(x,y)$ with the Gaussian blur $G(x,y,k\sigma)$at scale kσ, i.e

$$L(x,y,k\sigma) = G(x,y,k\sigma)*I(x,y) \quad (2)$$

So, A DoG image between scales kiσ and kjσ is just the output from difference of the Gaussian-blurred images at scales kiσ and kjσ. To detect scale-space extrema in the SIFT algorithm, the image is first convolved with Gaussian-blurs at different scales. The convolved images are grouped by an octave. An octave corresponds to doubling the value of σ. The value of ki is selected so that a fixed number of convolved images per octave can be obtained. Then the Difference-of-Gaussian images are taken from adjacent Gaussian-blurred images per octave.

Once DoG images have been obtained, keypoints are identified as local minima/maxima of the DoG images across scales. This is done by comparing each pixel in the DoG images to its eight neighbors at the same scale and nine corresponding neighboring pixels in each of the neighboring scales. If the pixel value is the maximum or minimum among all compared pixels, it is selected as a candidate keypoints.

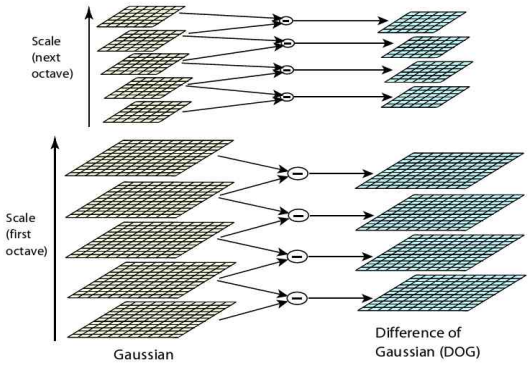

Fig1. 키포인트를 찾는 쉬프트 검출기를 이용한 디오지 영상 피라미드

Fig 1. The DoG image pyramid used by the SIFT detector to locate keypoints.

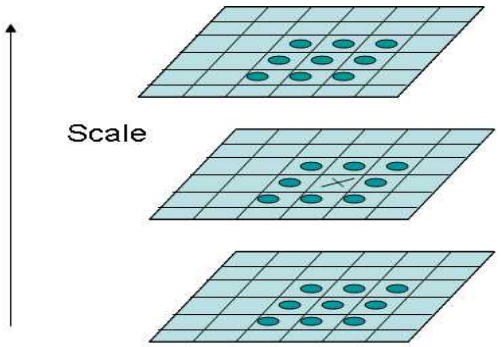

그림 2.

키포인트의 최대값과 최소값은 26개의 이웃 픽셀값으로 설정

Fig. 2. A keypoint must be a local minimum or maximum of its 26 neighbors.

2-2 Key Point Localization

Scale-space extrema detection produces too many keypoint candidates, some of those are unstable. It needs to perform a detailed fit to the nearby data for accurate location, scale, and ratio of principal curvatures. Keypoint localization gives the information of those points to be rejected having low contrast, sensitive to noise and poorly localized along an edge.

For each candidate keypoint, interpolation of nearby data is used to accurately determine its position. The interpolated location of the maximum is calculated, which substantially improves matching and stability. The interpolation is done using the quadratic Taylor expansion of the Difference-of-Gaussian scale-space function, D(x,y,σ) with the candidate keypoint as the origin. This Taylor expansion is given by

$$D(x) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X \quad (3)$$

$$\hat{X} = \frac{\partial^2 D^{-1}}{\partial X^2} \frac{\partial D}{\partial X} \quad (4)$$

where D and its derivatives are evaluated at the candidate keypoint and $X = (x, y, \sigma)$ the offset from this point. The location of the extremum $\hat{X}$ is determined by taking the derivative of this function with respect to X and setting it to zero. If the offset X is larger than 0.5 in any dimension, then that's an indication that the extremum lies closer to another candidate keypoint In this case, the candidate keypoint is changed and the interpolation performed instead about that point. Otherwise the offset is added to its candidate keypoint to get the interpolated estimate for the location of the extremum. To discard the keypoints with low contrast, the value of the second-order Taylor expansion is computed at the offset . If this value is less than 0.03, the candidate keypoint is discarded. Otherwise it is kept, with final location $y + \hat{x}$ and scale σ, where y is the original location of the keypoint at scale σ.

2-3 Eliminating edge responses

The DoG function has strong responses along edges, even if the candidate keypoint is unstable to small amounts of noise. Therefore, in order to increase stability, it needs to eliminate the keypoints that have poorly determined locations but have high edge responses. For poorly defined peaks in the DoG function, the principal curvature across the edge would be much larger than the principal curvature along it. To find out principal curvatures amounts we need to solve for the eigenvalues of the second-order Hessian Matrix H.

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \quad (5)$$

The eigenvalues of H are proportional to the principal curvatures of D. The ratio of the two eigenvalues, $\gamma = \alpha/\beta$ where α is the larger one and β the smaller one, is sufficient for SIFT's purposes. From the Hessian matrix H, $D_{xx} + D_{yy}$ is the sum of the two eigenvalues and it is trace of H; $D_{xx}D_{yy} - D_{xy}^2$ is the Determinant. The ratio $R = Tr(H)^2/Det(H)^2$ can be shown to be equal to, $(r+1)^2/r$ which depends only on the ratio of the eigenvalues rather than their individual values. The ratio R reached to minimum when both eigenvalues are equal. Therefore the higher the absolute difference between the two eigenvalues, which is equivalent to a higher absolute difference between the two principal curvatures of D, the higher the value of R. It follows that, for some threshold eigenvalue ratio r_{th}, if R for a candidate keypoint is larger than $(r_{th}+1)^2/r_{th}$ that keypoint is poorly localized and hence rejected.

2.3.1 Orientation Assignment

Each keypoint is assigned one or more orientations based on their local image gradient directions. This has done for achieving rotation invariant $L(x,y,\sigma)$. The Gaussian-smoothed image at the keypoint's scale σ is taken so that all computations are performed in a scale-invariant manner. The gradient magnitude $M(x,y)$ and orientation $\theta(x,y)$ are pre computed using pixel differences for an image sample $I(x,y)$ at scale σ. Let consider, $A = (L(x+1,x) - L(x-1,y))$ and $B = (L(x,y+1) - L(x,y-1))$, then magnitude and orientation can be represented by

$$M(x,y) = \sqrt{A^2 + B^2} \quad (6)$$

$$\theta(x,y) = \arctan\frac{B}{A} \quad (7)$$

The magnitude and direction calculations for the gradient are done for every pixel in a neighboring region around the keypoints in the Gaussian-blurred image L. An orientation histogram is formed and the total number of bins is 36, every bin covering 10 degrees. Each samples in the neighboring window added to a histogram bin. Every sample is weighted by its gradient magnitude and by a Gaussian-weighted circular window with an σ that is 1.5 times that of the scale of the keypoint. The peaks in this histogram correspond to dominant orientations. The orientations corresponding to the highest peak and those local peaks within 80% of the highest peaks are assigned to the keypoint, once the histogram filled. For multiple orientations an additional keypoint is created having the same location and scale as the original keypoint for each additional orientation.

2-4 Keypoint descriptor

The computation of descriptor vectors for these keypoints is important because the descriptors are highly distinctive and partially invariant to the remaining variations, like illumination, 3D viewpoint, etc. This step is pretty similar to the Orientation Assignment step. The feature descriptor is computed as a set of orientation histogram (4 x 4) pixel neighborhoods. The orientation histograms are relative to the keypoint's orientation. The orientation data comes from the Gaussian image closest in scale to the keypoint's scale. Each pixel is weighted by the gradient magnitude, and also by a Gaussian with σ 1.5 times the scale of the keypoint. Each Histogram contains 8 bins, and each descriptor contains a 4x4 array of 16 histograms around the keypoint. This leads to a SIFT feature vector with (4 x 4 x 8) =128 elements. This vector is normalized to enhance invariance to changes in illumination.

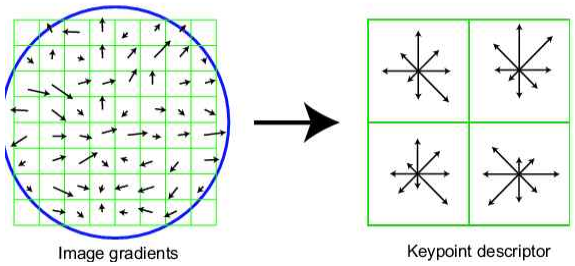

그림 3. 이미지 증감도와 키포인트 기술어
Fig. 3. Image gradients and KeyPoint Descriptor.

III. Matching of Keypoints

The dimension of the descriptor is 128 that are high. Descriptors with lower dimension than this don't perform as well across the range of matching task. The computational cost remains low due to the approximate Best Bin First (BBF)method used for finding the nearest-neighbor. Best Bin First is an approximate algorithm which returns the nearest-neighbor for a large fraction of queries and a very close neighbor. To test the distinctiveness of the SIFT descriptors, matching accuracy is also measured against varying number of keypoints in the testing database, and it is shown that matching accuracy decreases only very slightly for very large database sizes.

IV. Experimental Results

We have implemented using VC++ and OpenCV and tested our system on Intel core (TM2) CPU 6320 of 1.86GHZ with 2GB memory. We used the template image of object from the first frame of the scene. The Figure 4 shows the outputs of the SIFT keypoint detector on two images of CAN from two different viewpoints. Altogether, 497 and 297 keypoints were detected Two keypoints were considered to be a match if their distance in the 128-dimensional space is less than a given threshold value. We use the threshold value 0.49. The pairwise distances of the keypoints descriptors (128-vectors) from the two set of keypoints were computed. Then we take image of different oriented view of the same CAN. Then found out the keypoints and then we match between the two views. The 51 matching keypoints found are shown in Figure 5. Due to the similarities in appearances of portions of the image some incorrect matches are evident. Otherwise, the majorities of the matches are correct, revealing the rotation and translation of the camera between the two views. In Figure 6 Gaussian Blurred images produced by varying the σ value has shown.

We gave some single object testing results of CAN and other objects in Figure 7 and Figure 8 using web camera. We test the recognition system from some complex scene. There are multiple objects are presented. We found the exact object from the scene that we want to match. The system recognizes that object and results shown in Figure 9.

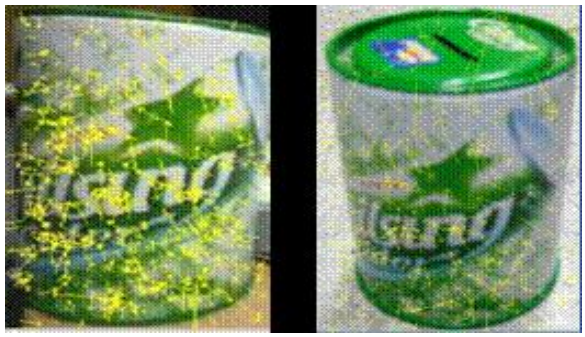

그림 4. 서로 다른 2개 시점 영상의 키포인트
Fig. 4. Showing the keypoints of two Different view image.

그림 5. 템플릿 이미지를 갖는 두 시점의 매칭
Fig. 5. Matching between two views with the template image.

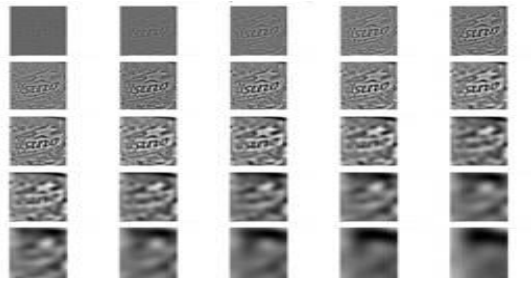

그림 6. 시그마 값의 변화에 따른 가우시안 블러드 이미지

Fig. 6. Gaussian Blurred images produced by varying the σ value.

그림 7. 다른 위치 물체의 매칭

Fig. 7. Single object matching of different orientation.

그림 8. 2가지 예의 물체 매칭

Fig. 8. Two sample object matching.

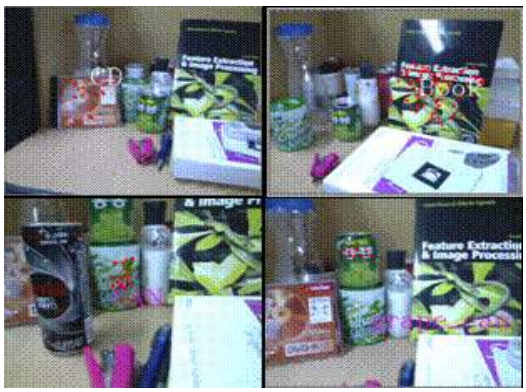

그림 9. 씨디, 책, 캔, 포도음료수 캔 등 서로 다른 물체의 인식

Fig. 9. Different object finding out from the scene. CD, Book, CAN, Grape CAN.

V. Conclusion

This paper presents a method that recognizes different objects from indoor scene. Traditional marker based system can be replaced by this procedure. We can use this for indoor Augmented Reality Applications like recognition of different books and display their contents, or different objects in the museum. Due to distinctness of recognition capability and freeness from scale change, orientation change this method, will be very good for outside Augmented Reality applications like Road, Building recognitions etc. In future we want to use this method for outside applications and try to combine with GIS and GPS data.

Acknowledgements

This research was supported by the Advanced Broadcasting Media Technology Research Center (ABRC) in Korea Aerospace University, Korea, under the Gyeonggi Regional Research Center (GRRC) support program supervised by Gyeonggi Province.

References

[1] David Ge, "Distinctive Image Features from Scale-Invariant Keypoints", *Int. Journal of Computer Vision,* Vol.60, No.2, pp.91-110, 2004.

[2] David. G. Lowe, "Object recognition from local scale invariant features", *Proc. of the Intnl. Conf. on Computer Vision*, pp. 1150-1157, Corfu, Greece, 1999.

[3] D. Beier, R. Billert, B. Br¨uderlin, D.Stichling, B. Kleinjohann, "Marker-less Vision Based Tracking for Mobile Augmented Reality" *Proc. of IEEE Int. Conf.on Mixed and Augmented Reality,* pp.258 – 259, Oct.2003.

[4] Q. Wang and S. You ,"Real-Time image matching based on multiple view kernel projections" *IEEE Conference on Computer Vision and Pattern*

Recognition, 2007.

[5] Schiele, Bernt, and James L.Crowley," Object Recognition using multidimensional receptive field histograms", *Fourth Europeon Conference on Computer Vision,cambridge,UK(1996)*, pp.67-97.

[6] Aibing Rao, and Rohini K. Srihari, "Spatial Color Histogramsfor Content-Based Image Retrieval", *Proc. of IEEE Int. Conf. onTools with Artificial Intelligence,* pp.183-186, Nov. 1999.

[7] D. Marimon and T. Ebrahimi", Orientation histogram-based matching for region tracking" *IEEE International workshop on Image Analysis and Multimedia interactive services*, pp. 6-8 ,June 2007.

[8] Schmid ,C. and R.Mohr,"Local grayvalue invariants for image retrieval", *IEEE PAMI* ,19,5,1997, pp530 -534.

Anjan Kumar Paul

2000년 1월 : Khulna University, Bangladesh (BE)
2006년 7월 : Indian Institute of Science, India (M.Tech)
2006년 9월 ~ 현재 : 한국항공대학교 정보통신공학과 박사과정
관심분야 : Augmented Reality, 멀티미디어, 컴퓨터비젼

이 형 진 (李炯陳)

2003년 3월 : 천안대학교 정보통신학부(공학사)
2005년 6월 : 천안대학교 정보기술대학원 컴퓨터학과(공학석사)
2005년 6월 ~ 현재 : 한국항공대학교 정보통신공학과 박사과정
관심분야 : 객체 기반 영상처리, 컴퓨터 비젼 및 컴퓨터 그래픽스 응용,

Mohammad Khairul Islam

2000년 7월 : Shahjalal University of Science & Technology, Bangladesh (BS)
2007년 8월 : 한국항공대학교 정보통신공학과 (공학석사)
2007년 9월 ~ 현재 : 한국항공대학교 정보통신공학과 박사과정
관심분야 : 멀티미디어, 영상처리, 컴퓨터비젼

김 영 범 (金榮凡)

1997년 3월 : 한국 항공대학교 통신정보공학과(공학사)
2001년 8월 : 한국 항공대학교 정보통신 공학과(석사)

2008년 3월 ~ 현재 : 한국항공대학교 정보통신공학과 박사과정
관심분야 : 객체 기반 영상처리, 컴퓨터 비젼, AR, 멀티미디어

백 중 환 (白重煥)

1981년 2월 : 한국항공대학교 항공통신공학과(공학사)
1987년 7월 : (미)오클라호마주립대학교 전기 및 컴퓨터공학과(공학석사)
1991년 7월 : (미)오클라호마주립대학교 전기 및 컴퓨터공학과(공학박사)
1992년 3월 ~ 현재 : 한국항공대학교 항공전자 및 정보통신공학부 교수
관심분야 : 영상처리, 패턴인식, 멀티미디어