논문 2011-48SP-4-2

비디오 감시 응용에서 확장된 기술자를 이용한 물체 검출과 분류

(Object Detection and Classification Using Extended Descriptors for Video Surveillance Applications)

모하마드 카이룰 이슬람*, 파라 자한*, 민 재 홍*, 백 중 환**

(Mohammad Khairul Islam, Farah Jahan, Jae Hong Min, and Joong Hwan Baek)

요 약

본 논문은 비디오 감시 장치에 사용되는 효율적인 물체 검출 및 분류 알고리즘을 제안한다. 이전 연구는 주로 Scale Invariant Feature Transform (SIFT)나 Speeded Up Robust Feature (SURF)와 같은 특정 형태의 특징을 이용해 물체를 검출하거나 분류하였다. 본 논문에서는 물체 검출 및 분류에 상호 작용하는 알고리즘을 제안한다. 이는 로컬 패치들로부터 얻어지는 텍스쳐나 컬러 분포 같은 서로 다른 특성을 갖는 특징값을 이용해 물체의 검출 및 분류율을 높인다. 물체 검출에는 특징점들의 공간적인 클러스터링을, 이미지 표현이나 분류에는 Bag of Words 모델과 Naive Bayes 분류기를 사용한다. 실험을 통해 제안한 기법이 로컬 기술자를 사용한 물체 분류기법보다 우수한 성능을 나타냄을 보인다.

Abstract

In this paper, we propose an efficient object detection and classification algorithm for video surveillance applications. Previous researches mainly concentrated either on object detection or classification using particular type of feature e.g., Scale Invariant Feature Transform (SIFT) or Speeded Up Robust Feature (SURF) etc. In this paper we propose an algorithm that mutually performs object detection and classification. We combinedly use heterogeneous types of features such as texture and color distribution from local patches to increase object detection and classification rates. We perform object detection using spatial clustering on interest points, and use Bag of Words model and Naive Bayes classifier respectively for image representation and classification. Experimental results show that our combined feature is better than the individual local descriptor in object classification rate.

Keywords : Local descriptor, Color Histogram, SIFT, SURF, Bag of Words.

Ⅰ. Introduction

A video surveillance system is the monitoring of the behaviors, activities, or other changing information, usually of people and often in a surreptitious manner. Traditional video surveillance system equip with several closed-circuit televisions covering important areas and human operator(s) or guard(s) for observing these monitors. However, the concurrent observation of several monitors and the long-term exhausting visualization cause the problem of decaying attention. To release a human being from this boring and labor intensive job, automatic video surveillance systems rely on the ability to detect and classify moving object in the video stream. The most

* 학생회원, ** 정회원, 한국항공대학교 정보통신공학과
(Dept. Information & Telecommunication Eng., Korea Aerospace University)

※ 본 논문은 경기도지역협력연구센터 (GRRC) 프로그램에 의해 한국항공대학교 차세대방송미디어기술 연구센터의 지원으로 수행되었으며, 일부 KAGERIIC의 지원에 의함.
접수일자: 2010년12월17일, 수정완료일: 2011년5월9일

common approaches consist of four major steps: interest point detection, interest region description, global representation, and classification. A recent focus has been given on improving region descriptors.

Image features are calculated based on visual cues extracted locally from image regions. Visual cues could be meaningful knowledge gained from the spatial arrangements of the "shape features" such as the edge elements, boundaries, corners, and junctions, or the brightness or color[1]. Significant efforts have been paid for long time to get meaningful knowledge from images and represent them. It is the key issue in computer vision[2]. Due to many challenges such as variations in illumination, viewpoint, scale etc this task is still under magnificent researches.

Image statistics like color, gradient, and filter responses are used as images features in computer visions for long time[7~9]. Color histogram is a classical image feature and used for object tracking[3], texture representation[4~5], matching[6] and other problems in the field of computer vision. However, these features are not robust in the presence of illumination changes and non-rigid motions. Recently, SIFT, a scale and rotation invariant descriptor, has been successfully applied in various general object recognition tasks[10~11, 18~19]. This approach extracts blob-like local features from an image, and represents each blob structure at an appropriate scale with a mechanism of automatic scale selection[12]. It is computationally expensive. Regarding computational speed, another robust feature named SURF outperforms SIFT on general purpose computers[17]. Recently Bag-of-Words has been used for the recognition of scenes by Sivic et al.[13]. Nister and Stewenius[14] describe a fast and accurate implementation allowing real-time searching of image databases.

Since feature extraction methods are individually limited to specific cues, they are not robust to real world scenarios. For instance, SIFT, and SURF efficiently work on the images with significant amount of textures and poorly work on non-textured but colorful images. In the above circumstances, a feature extraction scheme should carry the color information and texture pattern of an image in such a way that the feature contains the texture information as well as color information.

In this paper we investigate object detection and classification using various types of features extracted from an image separately and propose a combined feature approach. In order to make the article more self-contained, we briefly discuss concepts of various technologies in the following sections. Section Ⅱ of this paper depicts block diagram of our proposed approach. Section Ⅲ describes feature extraction methods, and section Ⅳ explains the Bag of Words model. Section Ⅴ illustrates object detection and classification techniques which is followed by experimental results and conclusion in sections Ⅵ and Ⅶ.

Ⅱ. Proposed Approach

Given a test image, our object detection and classification approach aims to determine the presence of any object of interest in the image, locate its area and classify it into one of a set of predefined object categories. In this approach we extend one of the state-of-the-art feature extraction technologies, adopt object classification model, and propose a new object detection technology. Fig. 1 gives a pictorial description of our proposed method which is broadly subdivided into training and test stages. In this figure. FE stands for Feature Extraction, BoW for Bag of Words, CHN for Cumulative Histogram Normalization, RS for Region Splitting, ROI for Region of Interest, RG for Region Growing and NB for Naive Bayes. Training stage consists of feature extraction, vocabulary generation, and object modeling and test stage has feature extraction similar to training, object detection, and classification.

In train stage, features are extracted from object and background imageries separately aiming to reduce false alarm caused by cluttered background.

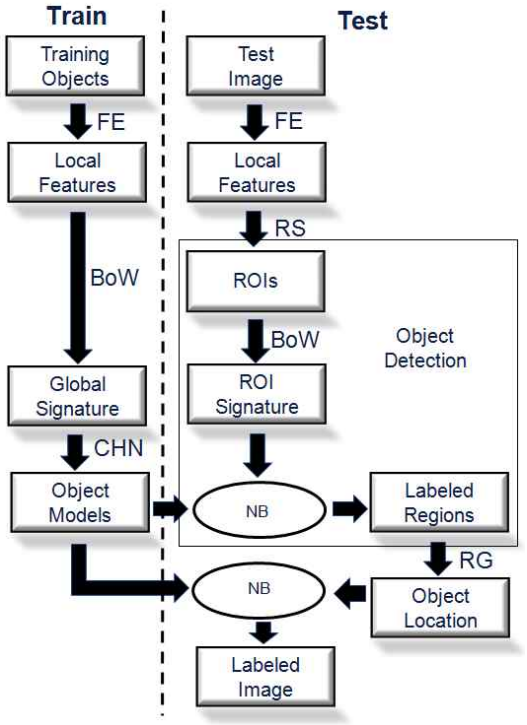


그림 1. 물체 검출과 분류에 대한 블록도
Fig. 1. Block diagram of object detection and classification approach.

We adopt the BoW concept for vocabulary generation which is used for image signature generation both in the training and test stages. Image signatures obtained from each category is summated and normalized to obtain a probabilistic model of that category.

In test stage, features are extracted from a given input image aiming to detect the object area in it and then classify it using the information learned at the training stage. For object detection, all interest points obtained during feature extraction are split into smaller groups based on their spatial Euclidean distance. Each group is then classified using BoW concept, and weighted by the number interest points belonging to the group. Finally, a majority voting and region merging technique results object area in the image. A validation test is performed on the object area to confirm the object class or category. In the following sections we describe the details of our approach.

Ⅲ. Feature Extraction

1. Scale Invariant Feature Transform

It is a robust feature extraction method which is performed in the following ways:

Scale-space extrema detection: At this stage, an image is blurred with Gaussian filters at different scales as in Eq. (1) and the difference of successive Gaussian-blurred images called Difference of Gaussian (DoG), are taken as in Eq. (2). In DoG images, each pixel is compared with its 26 neighbors in 3×3 regions at the current and adjacent scales to find the extremum points.

$$L(u,v,\sigma)=\frac{1}{2\Pi\sigma^2}e^{-\frac{u^2+v^2}{2\sigma^2}}*I(u,v) \tag{1}$$

$$D(u,v,\sigma)=L(u,v,s\sigma)-L(u,v,\sigma) \tag{2}$$

In the above equations (u, v) is pixel location or image point in the image plane I, σ is the standard deviation of the Gaussian distribution and s is scale.

Keypoint localization: Too many extremum points are detected at the previous stage; some of them are not good enough. For example, points with low contrast or localized along edges are sensitive to noise, so they should be eliminated. Aiming to do this a low contrast point is fitted to nearby data using a quadratic approximation function. If it is a local minima, it is removed. Points at edges with low cornerness are also removed.

Orientation assignment: For each image point, magnitude and orientation are computed as mentioned in Eq. (3) and Eq. (4). A histogram weighted by magnitude and gaussian window (σ is the half the window size), of local gradient directions is computed at selected scale and the orientation bin holding the peak of the histogram is assigned as keypoint orientation.

$$m(u,v)=((L(u+1,v)-L(u-1,v))^2+(L(u,v+1)-L(u,v-1))^2)^{\frac{1}{2}} \tag{3}$$

$$\theta = \arctan((L(u,v+1)-L(u,v-1))/(L(u+1,v)-L(u-1,v))) \quad (4)$$

Keypoint descriptor: Typical keypoint descriptor consists of 16 (44 grid) gradient histograms each with 8 bins. Each histogram is created over a window of 4×4 pixels. The resulting SIFT feature vector have 16×8 or 128 elements.

2. Speeded Up Robust Feature

It is also a scale and rotation-invariant interest point detector and descriptor[17]. This algorithm works in the following ways:

Integral image: For fast computation, it uses integral images. Given an input image I with resolution wxh and if a point in the image plane is denoted by (u,v), integral image I_{Σ} is calculated as in Eq. (5).

$$I_{\Sigma} = \sum_{u=1}^{u \le w} \sum_{v=1}^{v \le h} I(u,v) \quad (5)$$

Interest point detection: It is based on hessian matrix. Given a point $X=(u,v)$ in an image I, the Hessian matrix $H(x,\sigma)$ in x at scale σ is defined as Eq. (6).

$$H(X,\sigma) = \begin{bmatrix} L_{uu}(X,\sigma) & L_{uv}(X,\sigma) \\ L_{uv}(X,\sigma) & L_{vv}(X,\sigma) \end{bmatrix} \quad (6)$$

where $L_{uu}(u,\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\delta^2}{\delta u^2} g(\sigma)$ with the image I at point X, and similarly for $L_{uv}(X,\sigma)$ and $L_{vv}(X,\sigma)$. A pixel at which the Hessian's determinant satisfies a certain threshold is considered as an interest point.

Dominant orientation: Circular neighborhood of radius $6s$ around an interest point (s = the scale at which the point was detected) is selected to find dominant orientation. Haar wavelet responses with side length of $4s$ in horizontal and vertical directions are computed. Sum of all responses within a sliding orientation window covering an angle of 60 degree yields a vector. The longest vector is the dominant orientation.

Description: An interest region is split into 4×4 square sub-regions with 5×5 regularly spaced sample points inside. Haar wavelet responses d_u and d_v weighted with a Gaussian kernel centered at the interest point are calculated. Sum of responses for d_u, $|d_u|$, d_v, and $|d_v|$ creates a feature vector of 16×4 or 64 elements and the sum of responses d_u, and $|d_u|$ computed separately for $d_v < 0$ and $d_v > 0$ and similarly for the sum of d_v, and $|d_v|$ creates a feature vector of 128 elements.

3-3 Color Histogram

If the color pattern is unique compared with the rest of the data set, color histogram serves as an effective representation of the color content of an image. The color histogram is easy to compute and effective in characterizing both the global and local distribution of colors in an image. In addition, it is robust to translation and rotation about the view axis and changes only slowly with the scale, occlusion and viewing angle. So color histograms are widely used for the content-based image retrieval. In this paper, we integrate color information with texture information extracted from local patches centered at extremum points. In this case we consider extremum points obtained at scale-space extrema detection stage of SIFT extraction. A region of resolution 16×16 around each extremum point is considered as a patch for color histogram. We explore both RGB and

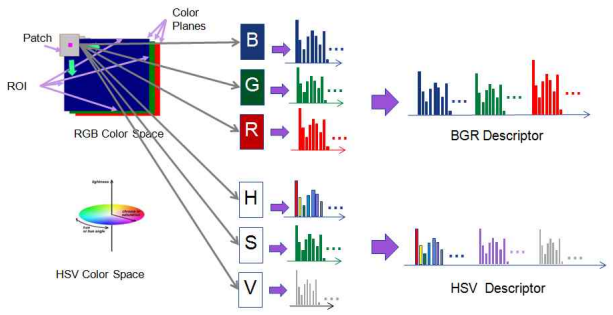


그림 2. 컬러 기술자
Fig. 2. Color descriptor.

HSV color spaces to use their distinctive power in object classification. Each patch is split into R (Red), G (Green), and B (Blue) or H (Hue), S (Saturation), and V (Value) planes. A frequency histogram of every plane is computed. Each histogram has same number of bins. These histograms are concatenated and thus a color descriptor is obtained. Fig. 2 gives a pictorial view of color descriptor construction process.

3-4 Proposed descriptor

Color descriptor is appended at the end of visual descriptor such as SIFT. Up to this stage, we get a descriptor for each interest point. Fig. 3 shows an example of combined descriptor. Where first 128 elements comes from SIFT feature and the other $3n$ elements from color. Here, n is the number of bins from each plane and we get best result when n is set to 16.

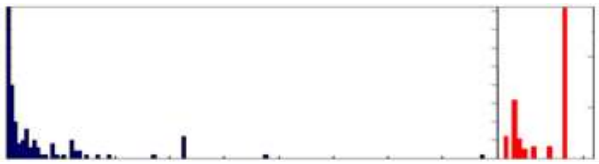


그림 3. 확장된 기술자
Fig. 3. Extended descriptor.

Ⅳ. Bag of Words Model

BoW is a famous document classification method. It is recently being used in image classification. A pictorial representation of BoW is given in Fig. 4 where NN and NB stand for Nearest Neighbor and Naive Bayes respectively. The whole process in BoW can be sub-divided into Codebook Generation, Object Modelling, Image Representation, and Classification. In codebook generation, all local features extracted from all training images are clustered using k -means[10] algorithm. The clustering process gives a feature space holding k cluster centers in it. These centers altogether are called 'codebook' or 'dictionary' and each cluster is considered as a codeword. BoW builds a probabilistic model of each object category

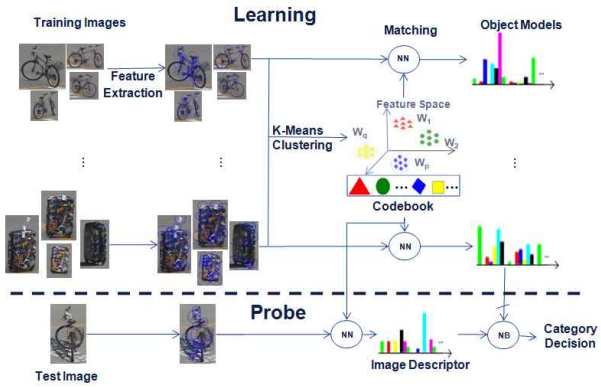


그림 4. Bag of Words 모델
Fig. 4. Bag of Words model.

by using the local features of all images of the category. Each local feature is mapped to its closest cluster in the feature space and counted in the corresponding bin in a k-bin frequency histogram. Finally, we obtain the category model after normalizing the histogram. Similarly, an image is globally represented by a k-bin frequency histogram without normalization. Given an input image for classification, a global signature of the image is calculated. Naive Bayes classifier is then used to determine the image category. It is a simple and popular classifier in the field of computer vision. The following content gives a brief idea of NB method.

NB is a special form of Bayesian network that is widely used for classification[15] and clustering[16]. It represents a distribution as a mixture of components where there is a 'naive' assumption that within each component all variables are 'independent' of each other. It is so named because of its "naïve" assumption of independence. Given a component $X=\{X_i|\, i=1,2,\ldots,n\}$ where all variables X_i are mutually independent, NB finds Posterior probability of each category for X. Maximum A Posterior gives the category of the component. Eq. (7) shows the classification process of the given component X where C denotes a set of categories.

$$Classify(X)=\underset{c}{\arg max}\, p(c)\prod_{i=1}^{n} p(X_i|c) \tag{7}$$

V. Object Detection and Classification

In object detection, we segment object area from the background. An efficient object detection helps to reduce false alarms in classification. Prior to object segmentation objects as well as background models are constructed using BoW. We adapt the concept of divide and conquer algorithm in object detection. In this method, we iteratively split up an image into small regions until any of them become small enough to be an object or object-part. The object detection consists of the following steps: feature extraction, region splitting, classification, majority voting, and merging. Interest points obtained in feature extraction are grouped into regions using their spatial information. We calculate Euclidean distance for spatial grouping. For example, if two points have a distance less then a *threshold* they belong to same group. A minimum rectangle holding all points in a group is considered as a candidate or ROI for object category; so it is classified by BoW. Since, features are once extracted from the candidate regions before splitting, we do not need to extract features from the region again in BoW. For each candidate region BoW gives object category as well as posterior probability obtained by the category. Number of feature points belonging to the candidate region is counted for the obtained category. All regions in the image are classified in the same way. If there is no candidate of any object category, distance threshold for region splitting is reduced to obtain smaller candidate regions. It is done until any object is found or to some iteration which makes sure that there is no object in the image.

If object is found in an iteration, majority voting is applied to decide image category. In majority voting, we find the category obtaining maximum number of feature points. If multiple categories obtain maximum vote, then the image goes to the category obtaining maximum posterior probability among them.

The final object area is obtained by merging small regions classified as final image category. A minimum rectangle holding the regions(s) in the image is denoted as object area.

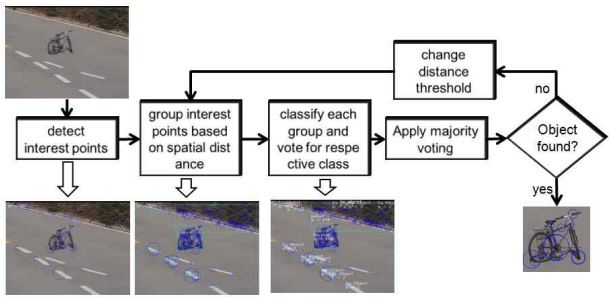


그림 5. 물체 검출 블록도
Fig. 5. Block diagram of object detection.

The image category is confirmed after applying a simple validation process. In this stage, the final object area is classified by BoW and if previous image category and the category obtained at this stage are same then the image is labeled by the obtained category otherwise it is labeled as background. Fig. 5 depicts a block diagram of object detection method.

VI. Experimental Results

We capture images from 6 categories (such as Bicycle, Chair, Parcel, Ladder, Luggage, and Pallet) using 2 PTZ cameras. For each category, we capture images in 8 different views, and 3 different zoom factors. Thus we have a total of 2×3×6×8 or 288 images each with resolution of 640×480. Our application is developed using Microsoft Visual C++ 2005. We use a desktop PC containing Intel®CoreTM

그림 6. 표본 영상들(자전거, 의자. 사다리, 팔레트, 수화물, 소포)
Fig. 6. Sample Images (Bicycle, Chair, Ladder, Pallet, Luggage, and Parcel).

2CPU 1.87GHz, 2 GB of RAM. For each image, it takes about 250ms in an average in object detection and classification. Fig. 6 shows a few examples of images used in our experiment.

Fig. 7 depicts object classification result obtained by using SURF, SIFT, and our proposed extended feature concept. For our collected data set we obtain object classification rate of 73.2%, 90.7%, 92.7, and 98.7% using SURF, SIFT, SIFT and RGB, SIFT and HSV histogram respectively. Extended feature increases the overall classification rate by about 8%. In this case, we manually crop object areas from the images and apply 2-fold cross validation. It is apparently seen that our extended feature which integrates SIFT and HSV histogram outperforms other features. The superiority of the combined approach is also validated by testing 4 object classes such as Car-Side, Laptop, Motorbike, and Sunflower from Caltech data set which is depicted in Fig. 8. In the figure object categories are placed along the horizontal axis. From these figures it is seen that SIFT feature is second candidate for object

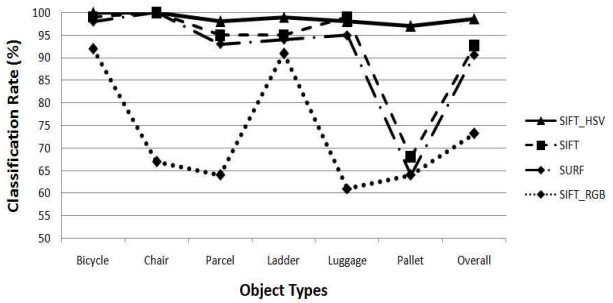


그림 7. 6 클래스의 표본영상에 대한 인식률
Fig. 7. Classification rate for 6 classes sample images.

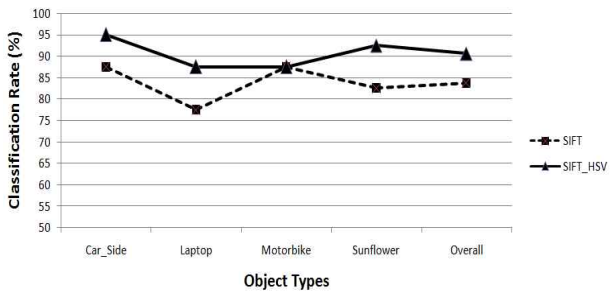
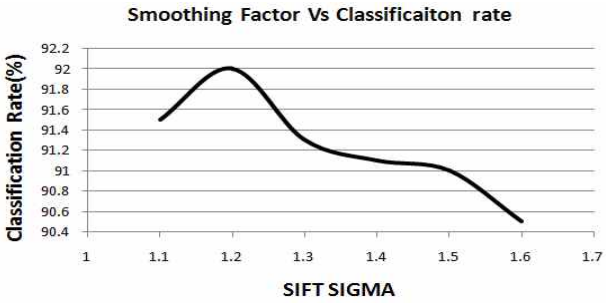
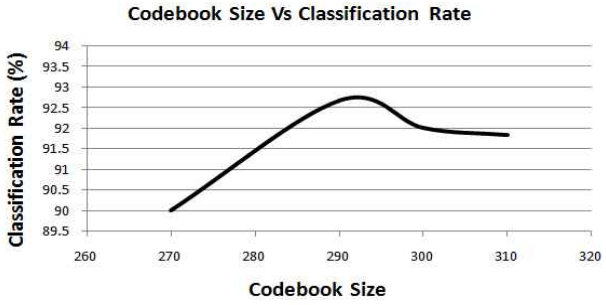


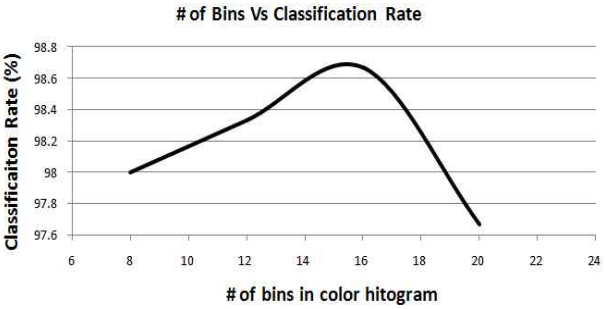
그림 8. Caltech 표본영상 인식률
Fig. 8. Classification rate for 4 categories of objects from Caltech dataset.



(a)



(b)



(c)

그림 9. SIFT 파라미터 변화에 따른 인식률 (a) 가우시안에서 표준편차 (b) codebook 크기 (c) 컬러 히스토그램에서 bin 개수
Fig. 9. Classification rate according to various SIFT parameters: (a) standard deviation in gaussian, (b) codebook size. (c) number of bins in color histogram.

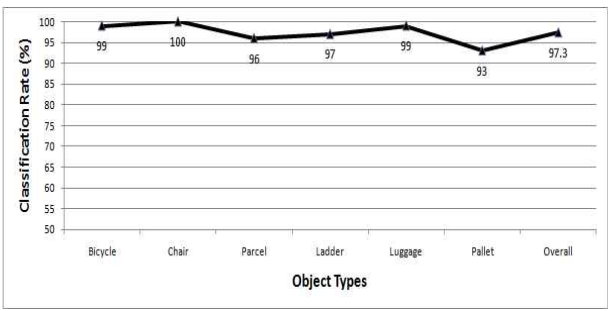


그림 10. 제안된 물체 검출과 분류 알고리즘의 인식률
Fig. 10. Classification rates obtained by applying object detection and classification.

classification. In our experiment, the overall performance is influenced by several parameters such as blurring parameter in SIFT calculation, codebook size k in k-means clustering, number of bins n in color histogram. Fig. 9 shows their influences in classification rates. Fig. 10 finally depicts the overall classification rate after object detection.

Ⅶ. Conclusions

In this paper, we propose an extended descriptor approach for object detection and classification. Experimental results prove the superiority of our proposed descriptor in object classification in real world challenging conditions. Since in real life the images contain object as well as background, and during online testing manual cropping is impossible, we proposed a noble approach for object detection prior to object classification. Our approach for object detection and classification achieves a classification rate of 97.3% during online testing of input images.

For an image of resolution 640×480 our approach takes about 250ms in an average for object detection and classification where almost 230ms is spent in feature extraction. In future, we aim to device some feature extraction technology which will be computationally cheap but holds distinctive power.

References

[1] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Context," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 509-522, April 2002.

[2] S. Ullman, "High-level vision: Object recognition and visual recognition", MIT Press, 1996.

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift", IEEE Conference on Computer Vision and Pattern Recognition, pp. 142 - 149, Hilton Head, SC, 2000.

[4] T. Leung, and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons", International Journal of Computer Vision, vol. 43, pp. 29 - 44, 2001.

[5] M. Varma, and A. Zisserman, "Statistical approaches to material classification", in Proc. of European Conf. on Computer Vision, pp. 167-172, Copenhagen, Denmark, 2002.

[6] B. Georgescu, and P. Meer, "Point matching under large image deformations and illumination changes", IEEE Transaction on Pattern Analyses and Machine Intelligence, vol. 26, no. 6, pp. 674 - 688, 2004.

[7] A. Rosenfeld, and G. Vanderburg, "Coarse-fine template matching", IEEE Transaction on Systems, Man and Cybernetics, vol. 7, pp. 104 - 107, 1977.

[8] R. Brunelli, and T. Poggio, "Face recognition: Features versus templates", IEEE Transaction on Pattern Analyses and Machine Intelligence, pp. vol. 15, no. 10, pp. 1042 - 1052, October 1993.

[9] R. Mar´ee, P. Geurts, J. Piater, and L. Wehenkel, "Random subwindows for robust image classifi cation", IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA. vol. 1, pp. 34 - 40, June 2005.

[10] M. Brown and D.G. Lowe, "Invariant features from interest point groups", British Machine Vision Conference, pp. 656 - 665, 2002.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, pp. 91 - 110, 2004.

[12] T. Lindeberg, "Feature detection with automatic scale selection," International Journal of Computer Vision, vol. 30, no. 2, pp. 79 - 116, 1998.

[13] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," IEEE International Conference on Computer Vision, pp. 1470-1477, Oct. 2003.

[14] D. Nistier and H. Stewenius, "Scalable recognition with a vocabulary tree," IEEE Computer Vision and Pattern Recognition, pp. 2161-2168, June 2006.

[15] P. Domingos, and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss", Journal of Machine Learning, vol. 29, pp. 103 - 130, 1997.

[16] P. Cheeseman, and J. Stutz, "Bayesian classification (AutoClass): Theory and results", International conf. on knowledge discovery and data mining, pp. 153 - 180, Portland, Oregon, Canada, August 1996.

[17] M. K. Islam, F. Jahan, J. H. Min, and J. H. Baek, “Fast Object Classification Using Texture and Color Information for Video Surveillance Applications”, Journal of Korea Navigation Institute, South Korea, vol. 15, no. 1, pp. 140-146, February 2011.
[18] J. H. Min, M. K. Islam, A. K. Paul, and J. H. Baek, “Realtime Markerless 3D Object Tracking for Augmented Reality”, Journal of the Institute of Signal Processing and Systems, South Korea, vol. 14, no. 2, pp. 272-277, April 2010.
[19] A. K. Paul, M. K. Islam, J. H. Min, Y. B. Kim, and J. H. Baek, “Natural Object Recognition for Augmented Reality Applications”, Journal of the Institute of Signal Processing and Systems, South Korea, vol. 11, no. 2, pp. 143-150, April 2010.

저 자 소 개

Mohammad Khairul Islam(학생회원)
December, 1998 BSc(Engg.) in Electronics and Computer Science, Shahjalal University of Science and Technology, Bangladesh.
August, 2007 MSc in Information and Telecommunication Engineering, Korea Aerospace University, South Korea.
Sept., 2007~Now PhD student in Information and Telecommunication Engineering, Korea Aerospace University, South Korea.
<Research Interest : Multimedia, Image Processing, Computer Vision>

Farah Jahan(학생회원)
December 2005 BSc.(Honors) in Computer Science and Engineering, University of Chittagong, Bangladesh.
September 2009~Now MSc student in Information and Telecommunication Engineering, Korea Aerospace University, South Korea.
<Research Interest : Multimedia, Image Processing, Computer Vision>

민 재 홍(학생회원)
1997년 2월 한국 항공대학교 통신정보공학과 (공학사)
2001년 8월 한국 항공대학교 정보통신공학과(석사)
2008년 3월~현재 한국항공대학교 정보통신공학과 박사과정
<주관심분야 : 객체 기반 영상처리, Augmented Reality, 멀티 미디어, 컴퓨터 비젼>

백 중 환(정회원)
1981년 2월 한국항공대학교 항공통신공학과(공학사)
1987년 7월 (미)오클라호마 주립대학교 전기 및 컴퓨터공학과(공학석사)
1991년 7월 (미)오클라호마 주립대학교 전기 및 컴퓨터공학과(공학박사)
1992년 3월~현재 한국항공대학교 항공전자 및 정보통신공학부 교수
<주관심분야 : 영상처리, 패턴인식, 멀티미디어>