비디오 감시 응용을 위한 텍스쳐와 컬러 정보를 이용한 고속 물체 인식

Fast Object Classification Using Texture and Color Information for Video Surveillance Applications

이슬람 모하마드 카이룰*, 자한 파라*, 민재홍*, 백중환*

Mohammad Khairul Islam*, Farah Jahan*, Jae-Hong Min*, and Joong-Hwan Baek*

요 약

본 논문에서는 텍스쳐와 컬러 정보를 기반으로 비디오 감시를 위한 빠른 물체 분류 방법을 제안한다. 영상들로부터 SURF와 색 히스토그램의 국부적 패치들을 추출하여 그들의 장점을 이용한다. SURF는 명암 내용 정보를 제공하고 색 정보는 패치에 대한 특이성을 증강시킨다. SURF의 빠른 계산뿐만 아니라 객체의 색 정보를 활용한다. 국부적 특징을 이용하여 관심 영역 혹은 영상의 전역적 서술자를 생성하기 위해 Bag of Word 모델을 이용하고, 전역적 서술자를 분류하기 위해 Naïve Bayes 모델을 이용한다. 또한 본 논문에서는 판별적인 기술자인 SIFT도 성능 분석한다. 네 종류의 객체에 대한 실험결과 95.75%의 인식률을 보였다.

Abstract

In this paper, we propose a fast object classification method based on texture and color information for video surveillance. We take the advantage of local patches by extracting SURF and color histogram from images. SURF gives intensity content information and color information strengthens distinctiveness by providing links to patch content. We achieve the advantages of fast computation of SURF as well as color cues of objects. We use Bag of Word models to generate global descriptors of a region of interest (ROI) or an image using the local features, and Naïve Bayes model for classifying the global descriptor. In this paper, we also investigate discriminative descriptor named Scale Invariant Feature Transform (SIFT). Our experiment result for 4 classes of the objects shows 95.75% of classification rate.

키워드 : SURF, SIFT, 색 히스토그램, 백 오브 워즈, K-평균, 나이브 베이즈

Keywords: SURF, SIFT, Color Histogram, Bag of Words, K-Means, Naïve Bayes.

I. Introduction

Traditional video surveillance system equip with several closed-circuit televisions in important areas and a human operator for observing these monitors. However, the concurrent observation of several monitors and the long-term exhausting visualization cause problem of decaying human attention. To release a human being

* Dept. of Information & Telecommunication Engineering, Korea Aerospace University, Goyang-city, 412-791, Korea

· 제1저자 (First Author) : 이슬람 모하마드 카이룰(Mohammad Khairul Islam)
· 투고일자 : 2011년 1월 31일
· 심사(수정)일자 : 2011년 1월 31일 (수정일자 : 2011년 2월 23일)
· 게재일자 : 2011년 2월 28일

from this boring but labor intensive job, automatic video surveillance systems rely on the ability to detect and describe moving object in the video stream which is a relevant information extraction step in a wide range of computer vision applications. The major jobs in this domain are generating discriminative signature from an object and then classification. Object signature is generated based on visual cues extracted locally from image pixels. 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 features [1]. It is the key issue in computer vision [2]. The majority of feature extraction approaches focus on detecting local regions such as Difference of Gaussian (DoG) regions [3], saliency regions [4], or other types of local patches. Scale Invariant Feature Transform (SIFT) uses DoG and has been successfully applied in various general object recognition tasks. It is computationally very expensive. Regarding computational speed, another robust feature named Speeded Up Robust Feature (SURF) outperforms SIFT implementations on general purpose computers. High level color description often provides links to image content [5] which can be used for image signature. Descriptors representing only either shape features or color features are not enough to discriminately represent an image. The representation scheme should carry the color information and its pattern of appearance on the object in such a way that the description contains the texture information as well as color information.

Bag-of-Words has been used for the recognition of scenes by Sivic et al. [6]. Nister and Stewenius [7] describe a fast and accurate implementation allowing real-time searching of image databases. However, the extracted features depend largely on local regions, such as corners and textured patches, therefore they are able to recognize objects only from one viewpoint and might not be accurate for recognizing objects when the viewpoint changes [6].

Therefore, we study computational models and techniques to merge color and shape invariant information for object recognition. In this paper we propose an integrated descriptor containing speeded up robust features and color information for object classification. Section II of this paper depicts basic building blocks of our approach. Section III describes feature extraction methods, and section IV describes Bag of Words model. Section V illustrates classification technique and section VI shows experimental result. The conclusion and future plan are briefly mentioned in section VII.

Ⅱ. Proposed Approach

Our proposed approach consists of training and test stages. Training is done in offline and test in online. In training stage, (i) we manually crop object areas from training images, (ii) extract features from objects, (iii) using Bag of Words (BoW) model a signature of an object is generated from the extracted features and (iv) finally, the signatures from all object imageries are used to construct object models which are normalized frequency histograms of local features mapped to a visual dictionary which is previously built by K-Means clustering in bag of words model. The major steps in test stage are (i) features are extracted images, (ii) a region of interest (ROI) is set centering at each interest point with an average size of training objects, (iii) features

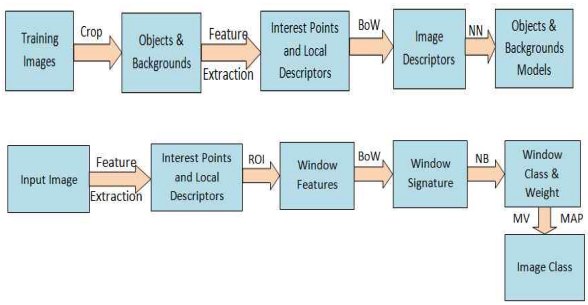

그림 1. 제안된 시스템의 블록도
Figure 1. Block diagram of the proposed system.

extracted from the region generate a signature using BoW model, (iv) the signature is then classified by Naïve Bayes classifier (NBC), (v) an object category achieving maximum number of windows is labeled as image description. Figure 1 shows block diagram of training and testing procedures of our approach.

Ⅲ. Feature Extraction

Speeded Up Robust Feature (SURF): It is a scale and rotation-invariant interest point detector and descriptor. This section presents a brief summary of SURF extraction process. (i) Interest Point Localization: Interest point detection is based on hessian matrix. Given a point $X=(x,y)$ in an integral image I, the Hessian matrix $H(X,\sigma)$ in X at scale σ is defined by Eq. (1).

$$H(X,\sigma)=\begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \quad (1)$$

Here $L_{xx}(x,\sigma)$ is the convolution of the Gaussian second order derivative with the image at point x, and similarly for $L_{xy}(x,\sigma)$ and $L_{yy}(x,\sigma)$. (ii) Dominant Orientation: A circular neighborhood of radius, 6 times the scale, around the interest point is selected for orientation calculation. x- and y-response are computed with side length of 4. Sum of all responses within a sliding orientation window covering an angle of 60 degree is calculated. The longest vector is the dominant orientation. (iii) Interest point Description: SURF descriptor is extracted from an interest region aligned to the orientation. A descriptor of length 64 or 128 is calculated based on Haar wavelet responses dx, |dx|, dy, and |dy| over sub-regions.

Scale Invariant Feature Transform (SIFT): SIFT extraction method is performed in the following steps: (i) Scale-space extrema detection: difference of Gaussian-blurred images in successive scales is taken as in Eq.(2) where (i,j) presents pixel location. (ii) Keypoint localization: it discards the keypoints with low contrast. (iii) Orientation assignment: for each image point, magnitude and orientation are computed to find dominant orientation. (iv) Keypoint descriptor: a region around a keypoint is split in a 4×4 grid and each grid generates a vector of length 8, resulting SIFT feature of length 128 in total.

$$D(i,j,\sigma)=L(i,j,k\sigma)-L(i,j,\sigma) \quad (2)$$

Color Histogram: High level color description often provides links to image content. Color histogram serves as an effective representation of the color content of an image if the color pattern is unique compared with the rest of the data set. It 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. In this paper, we build color histograms of small patches centered on interest points. We crop an image patch of 16x16 around an interest point, and then split the colors into H, S, and V color planes. We calculate histogram of individual color and concatenate them to build a color descriptor. Fig. 2 shows color feature extraction method.

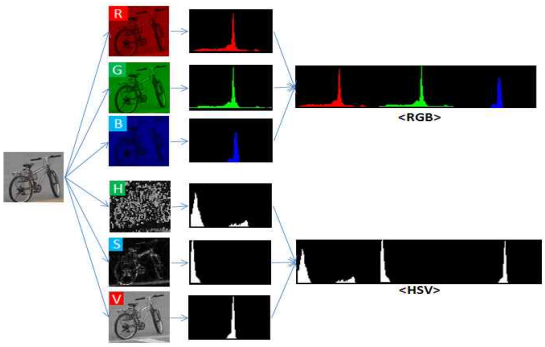

그림 2. 컬러 히스토그램 생성
Figure 2. Color histogram construction.

Ⅳ. Bag of Words Model

Image representation from the local feature is performed using bag of words model. It generates a codebook or dictionary using K-means clustering [8] over all the local descriptors. Then, all local descriptors from each image in the training set are mapped to their closest codewords in the codebook. The frequency histogram of the features is a vector which represents an image globally and thus called image descriptor or signature. Figure 3 depicts the flow diagram of Bag of Words model.

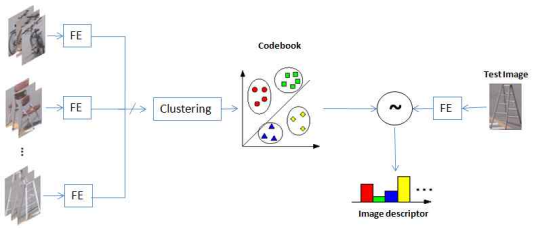

그림 3. Bag of Words 모델
Figure 3. Bag of Words model.

V. Classification

A signature obtained from a ROI or image is classified using Naïve Bayes classifier which is widely used for classification [9] and clustering [10]. Naïve Bayes models are so named for their "naïve" assumption that all variables are mutually independent. Given a set of hypothesis $\{h_i\}$ where $i = 1, 2, \ldots, m$ and data $D < d_1, d_2, \ldots, d_n >$, posterior probability $P(h_i|D)$ of the hypothesis h_i is calculated using Eq. (3) and maximum a posteriori (MAP) h_{MAP} is calculated using Eq. (4).

$$P(h_i|D) = \frac{P(D|h_i)P(h_{i)}}{P(D)} \quad (3)$$

$$h_{MAP} = \underset{h_i \epsilon H}{\arg max}\, P(d_1, d_2, \ldots, d_n|h_j)P(h_j) \quad (4)$$

Given a test image, we locate interest points using SURF/SIFT method, and consider a window or ROI, centered at each interest point with an average size of training objects. The window is classified by Naïve Bayes classifier (NBC) and a weight which is equal to 1 and posterior probability obtained by NBC are assigned to the object class. An object category achieving aximum

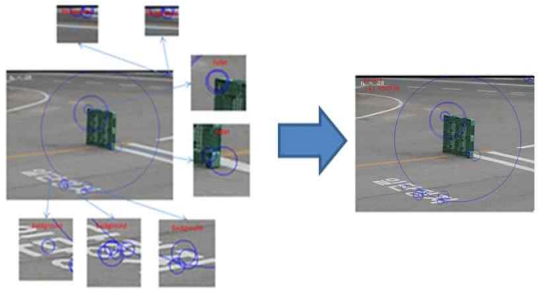

그림 4. Majority voting에 의한 물체 인식
Figure 4. Object classification by majority voting.

weight is labeled as image description. If multiple categories have same weight, maximum a posterior (MAP) is used to decide image label. Figure 4 shows an example of object classification.

VI. Experimental Results

We capture images of 4 categories (such as Bicycle Chair, Ladder, and Luggage) using 2 PTZ cameras. For each category, we capture images of resolution 640x480 in 8~16 orientations, and 3 zoom-in factors. Our application is developed using Microsoft Visual C++ 2005, and OpenCV library. We use a desktop PC containing Intel®Core™2CPU 1.87GHz, 2 GB of RAM. Figure 5. shows some example images used in our experiment.

그림 5. 실험에 사용된 샘플 영상
Figure 5. Sample Images for experiment.

The overall performance of classification is influenced by several parameters. We optimize them and set hessian threshold to 2550 in SURF, sigma to 1.5 in SIFT, descriptor length 128 for both, 3 colors in HSV color space, 16 bins and 300 clusters in K-Means. Fig. 6 depicts object classification result obtained at test stage using SURF, SIFT, and SURF-CH (SURF & Color Histogram), and SIFT-CH (SIFT & Color Histogram).

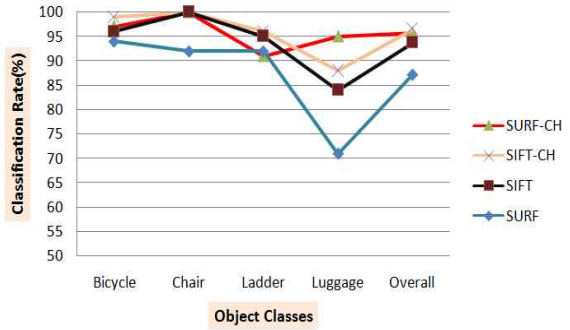

그림 6. 다양한 특징에 대한 인식률
Fig. 6 Classification rates by various feature types.

It is seen in Table 1 that classification time is much higher than feature extraction. For example, for SURF feature extraction time is 89.39 ms and classification time is 521.1 ms i.e. calculating global descriptors of all windows using Bag of Words and their classification take (521.1-89.39) ms or 431.61ms.

표 1. 특징 추출과 인식을 위한 처리 시간
Table 1. Processing times for feature extraction and classification.

Feature Type	Feature Extraction Time (ms)	Image Classification Time (ms)	Classification Rates
SURF	89.39	521.1	87.25 %
SIFT	180.4	612.9	93.75 %
SURF-CH	98.2	673.7	95.75 %
SIFT-CH	188.6	759.9	95.75 %

Thus, the total time for image classification depends on the number windows tested in a whole image. But if we skip some window it makes the classification fast while slightly reducing classification rate. Table 2 depicts the computation time and image classification rate with respect to number of windows selected for classification using SURF-CH. We can see that classification rate doesn't remarkably reduce if we reduce number of windows to below 10 at which point computing time reduces from 673.7 ms to 175.0 ms for a single image.

표 2. SURF-CH 사용 시 윈도우 수에 따른 영상 인식 시간과 인식률
Table 2. Image classification time and classification rate for different number of windows using SURF-CH

# of windows	Elapsed Time (ms)	Classification Rate
All	673.7	95.50 %
Average	617.7	95.50 %
20	313.9	95.50 %
10	241.1	95.25 %
5	193.9	92.25 %

VII. Conclusions

In this paper we use SURF, SIFT, SURF -CH, and SIFT-CH to describe local patches of images, apply bag of words model for global signature, Naïve Bayes for classification. For our collected data set we obtain object classification rate of 87.25%, 93.75%, 95.75%, and 95.75% using SURF, SIFT, SURF-CH, SIFT-CH respectively. It is clear that SIFT is much better than SURF in object classification. But when we combine color descriptor with them, classification rate is improved for both of them which are 95.75% and a tie between them. The most interesting fact is, SURF gets strong backup from color information and competes with SIFT. Considering computing time we use SURF in our proposed model. We reduce computing time by skipping window calculation and it takes around 241 ms per image without largely affecting classification rate. For real time application, the computation speed should be reduced. That is why, we need to device some other descriptor

which will be computationally cheap but holds distinctive power.

Acknowledgment

This research work was supported by KAGERIIC program and the Gyeonggi Regional Research Center (GRRC) support program supervised by Gyeonggi Province.

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.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[4] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," *Proc. of European Conference on Computer Vision, Prague, Czech Republic,* pp. 228-241, 2004.

[5] Zhi-Gang Fan, Jilin Li, Bo Wu, and Yadong Wu, "Local patterns constrained image histograms for image retrieval", *IEEE International Conference on Image Processing (ICIP), San Diego, CA,* pp. 941 - 944, December 12, 2008.

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

[7] D. Nistier and H. Stewenius, "Scalable recognition with a vocabulary tree," *IEEE Computer Vision and Pattern Recognition (CVPR),* pp. 2161-2168, Jun. 2006.

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

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

[10] P. Cheeseman, and J. Stutz, "Bayesian classification (AutoClass): Theory and results", *Advances in knowledge discovery and data mining, AAAI Press, Menlo Park, CA,* pp. 153-180, 1996.

Mohammad Khairul Islam

Dec., 1998: BSc(Engg.) in Electronics & Computer Science, Shahjalal University of Science & Technology, Bangladesh.
Aug., 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

Dec., 2005: BSc.(Honors) in Computer Science & Engineering, University of Chittagong, Bangladesh
Sept., 2009 ~ till date : MSc student in Information and Telecommunication Engineering, Korea Aerospace University, South Korea.
Research Interest: Multimedia, Image Processing, Computer Vision.

민 재 홍 (閔載泓)

1997년 2월 : 한국 항공대학교 통신정보공학과 (공학사)
2001년 8월 : 한국 항공대학교 정보통신공학과(석사)
2008년 3월 ~ 현재 : 한국항공대학교 정보통신공학과 박사과정
관심분야 : 객체 기반 영상처리, Augmented Reality, 멀티미디어, 컴퓨터 비젼

백 중 환 (白重煥)

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