한국항행학회 논문지 제12권 제6호 2008년 12월

비디오 검색을 위한 얼굴 검출 및 인식

Face Detection and Recognition for Video Retrieval

이슬람 모하마드 카이룰*, 이형진*, 폴 안잔 쿠마*, 백중환*

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

요 약

본 논문에서는 비디오 검색을 위한 새로운 얼굴 검출 및 인식 방법을 제안한다. 인물 정합은 비디오 프레임에서 어떻게 얼굴을 정확하게 찾아내는가에 달려있다. 얼굴 영역은 Adaboost 알고리즘으로 부스트된 viola-jones의 특징을 이용하여 비디오 프레임에서 검출한다. 얼굴 검출 후 조명 보정을 하고 PCA(Principal Component Analysis)로 특징점을 추출하고 SVM(Support Vector Machine)으로 사람의 신원을 분류한다. 실험 결과 제안한 방법이 정합율면에서 우수한 성능을 보였다.

Abstract

We present a novel method for face detection and recognition methods applicable to video retrieval. The person matching efficiency largely depends on how robustly faces are detected in the video frames. Face regions are detected in video frames using viola-jones features boosted with the Adaboost algorithm. After face detection, PCA (Principal Component Analysis) follows illumination compensation to extract features that are classified by SVM (Support Vector Machine) for person identification. Experimental result shows that the matching efficiency of the ensembled architecture is quit satisfactory.

Key words : video retrieval, face detection, illumination, face recognition, PCA, SVM.

I. Introduction

Multimedia can be considered as a huge gathering of information like audio, graphics, image, video and animation. Additionally, there has been a remarkable increase in the availability of this information through cable and the Internet and also in every moment a large amount of information is being added in this cluster. As a result, the amount of multimedia information nowadays is getting enormous size. Video on demand(VOD) systems allow users to select and watch video and clip contents over a network as part of an interactive television system. As the volume increases, the complexity for accessing increases and makes it a more challenging task.

With efficient indexing, the user can extract relevant content and navigate effectively in large amounts of available data. Thus, there are great efforts for developing automated techniques for indexing and organizing visual data, and for developing efficient tools for browsing and retrieving contents of interest. These techniques must be diffused by the production on

* 한국항공대학교 정보통신공학과(Dept. of Information & Telecommunication Engineering, Korea Aerospace University)
· 제1저자 (First Author) : Mohammad Khairul Islam
· 투고일자 : 2008년 10월 2일
· 심사(수정)일자 : 2008년 10월 6일 (수정일자 : 2008년 11월 20일)
· 게재일자 : 2008년 12월 30일

commercial scale and new techniques must be developed for an efficient multimedia indexing and retrieval. The typical research for content-based video indexing includes face detection, speaker identification and character recognition.

Humans detection have been proved to be one of the most challenging tasks because of the wide variability in the appearance due to environmental condition, clothing, illumination and view point variant shape characteristics. So, the detector need to be robust to a large range of lightening variations, noise and partial occlusion.

State-of-the-art human detection algorithms typically provide information regarding the location and scale of the detected person. However these approaches do not provide any information regarding the shape of the detected object. In this paper we applied adaboost algorithm using viola-jones features for the face detection and support vector machines using eigenfaces for face matching.

Ⅱ. Related works

While numerous detection methods have been shown to be effective at detecting people in outdoor scenes [4,3], a template-based method for pedestrian detection was proposed in [5] that provided some shape information by matching the detected object with the most similar template used during training. The implicit shape model proposed in [1] addresses the tasks of detection and segmentation using prototypical image patches and their spatial distribution around the object centroid. However, this class of techniques also requires fully segmented object regions during training. In a related approach [2], discriminative boundary fragments were used to learn the object geometry. Another approach using object boundaries was proposed in [6].

Among the existing face recognition techniques, subspace methods are widely used in order to reduce the high dimensionality of the raw face image [11]. Commonly used subspace methods are Linear Discriminant Analysis (LDA) or Fisherface [13], Bayesian algorithm using probabilistic subspace[14], and Eigenface method [12]. Among the available subspace learning methods, we use PCA [15] for dimensionality reduction. After getting the features in the reduced dimensional subspace we classify them for face recognition. There are several methods for classifying the features to find the test face group where it actually belongs. We apply a most widely used classifier named Support Vector Machine. It has been recently proposed by Vapnik and his co-workers [16] as a very effective method for general purpose pattern recognition. Intuitively, given a set of points belonging to two classes, a SVM finds the hyperplane that separates the largest possible fraction of points of the same class on the same side, while maximizing the distance from either class to the hyperplane.

The organization of the remaining paper is as follows. Section 3 describes the method of face detection. Illumination compensation, feature extraction from faces and their classification method are explained in section 4. Experimental results are depicted in section 5. Finally, section 6 gives the concluding remarks.

Ⅲ. Face Detection

3-1 Adaboost Algorithm

Adaboost is a boosting algorithm. It is originated from Probably Approximately Correct (PAC) learning and one of the most popular and efficient learning machines based approaches for detecting faces. The Adaboost algorithm was introduced by Freund and Schapire [7] which was extended by Viola[8] who introduced the cascading of weak classifiers. It is used to solve the following three fundamental problems: (1) choosing magnificent features from a large feature set; (2) constructing weak classifiers, each of which is based

on one of the selected features; and (3) cascading the weak classifiers to construct a strong classifier. Each classifier is trained using Haar-like features. Viola and Jones made effective computation of a large number of such features. A hit from the first classifier leads to the evaluation of a second classifier which is used to yield very high detection rates and so on. A miss at any point leads to the immediate rejection. The cascaded stages are constructed by training classifiers and adjusting the threshold to minimize false positives. The Algorithm can be described as follows [10]:

Input:

Training data $D=(x_1,y_1),...,(x_1,y_1)$ where $x_k=(x_k^1,x_k^2,...,x_k^m)$ is a feature vector, $y_k \epsilon Y=-1,+1$ is a corresponding label (+1 corresponding to face, -1 corresponding to non-face).

Weak classifiers: $h_j=X\rightarrow -1,+1$

Output:

The strong classifier $H(x)$

Algorithm:

Initialize the weight vector

$$w_k^1=\frac{1}{N},\quad where\quad k\in 1,...,N$$

Building a strong classifier using T weak classifiers:

For $t=1,2,...,T$ do

1. Compute the error ϵ_j corresponding to the weak classifier

$$\epsilon_j=\sum_{k=1}^{k=N} w_k^t|h_j(x_k)-y_k|$$

2. Find the weak classifier h_t that minimizes the error ϵ_j
3. If $\epsilon_j \geq \theta$ (an error threshold) then stop
4. Compute the confidence

$$\alpha_t=\frac{1}{2}ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

5. Update the weights

$$w_k^{t+1}=w_k^t\begin{cases}\frac{1}{2(1-\epsilon^t)} & \text{if } h_t(x_k)=y_k\\ \frac{1}{2-\epsilon^t} & \text{if } h_t(x_k)\neq y_k\end{cases}$$

The final strong classifier is

$$H(x)=sign\left(\sum_{t=1}^{t+T}\alpha_t h_t(x)\right)$$

3-2 Feature Extraction

Feature extraction is an important step toward object detection. In previous techniques, the systems measured nodal points on the face, such as the distance between the eyes, the shape of the cheekbones and other distinguishable features. These nodal points are then compared to the nodal points computed from a database of pictures in order to find a match. But, these systems were limited based on the angle of the face captured and the lighting conditions present. Well recognized Haar-like feature is described by the shape, coordinate relative to the search window origin and the scale factor of the feature. Viola and Jones proposed 4 basic templates of scalar features for face detection depicted in Fig. 1.

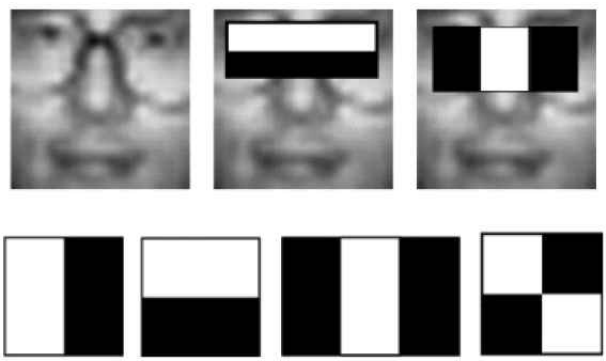
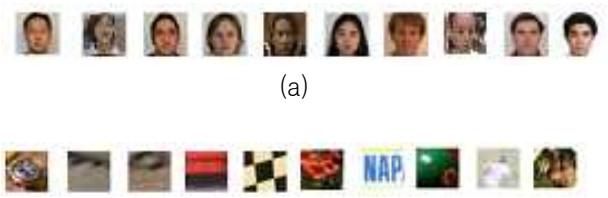


Fig. 1. Types of features

3-3 Training the classifiers

The number of viola-jones features derived from each prototype is quite large and differs from prototype to prototype. So the significant features are selected from the large set of simple features by learning the training samples. A cascade of classifiers is degenerated decision tree where at each stage a classifier is trained to detect almost all objects of interest while rejecting a certain fraction of the non-object patterns [11].

Figure 2 shows the training images. A cascade of N classifiers depicted in Figure 3 is trained with the samples in Figure 2. A set of weak classifiers use one feature from our feature set in combination with a simple binary thresholding decision. At each round the classifier that best classifies the weighted samples is added. As the number of stages is increased, the number of weak classifiers is increased.



(a)

(b)

Fig. 2. Training images. (a) face images and (b) non-face images.

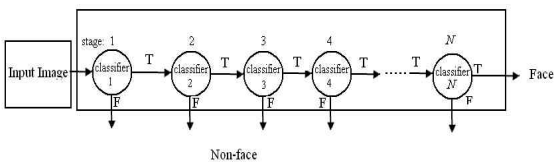


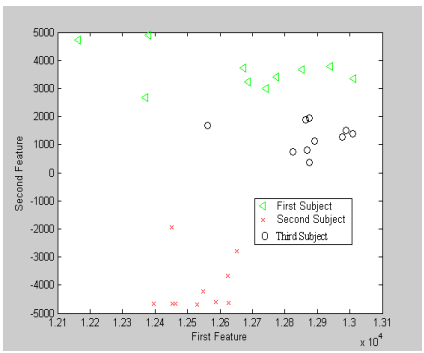
Fig. 3. Cascade of classifiers with N stages. At each stage a classifier is trained to achieve a hit (T) or miss (F).

Ⅳ. Face Recognition

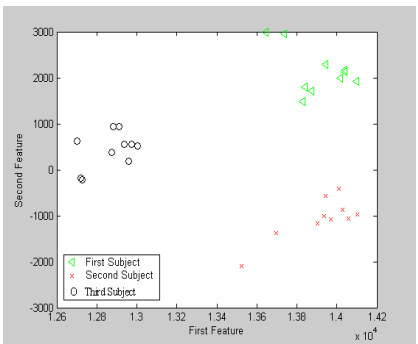
4-1 Illumination Compensation

Face images extracted from video sequences suffer with illumination variations. Thus, it affects the features extracted from faces which in result hampers the face identification performance. So, before extracting features, we apply illumination compensation.

Here, we at first segment the input face into blocks, then we determine minimum the luminance value in each block. These values are then bilinearly interpolated to construct luminance image. The luminance image is then subtracted from the original image. After that, we apply histogram equalization on the residue. Thus we get illumination compensated image. An example of the distribution of features after and before the illumination compensation on face images is illustrated in Fig. 4. In the figure, the distribution of the features in (b) of each class are closer than in (a). So, finding the discriminating boundary is easier in illumination compensated images than images without illumination compensation. That means, the features in (b) is more distinctive than (a) and discrimination boundary in Fig. 4(b) would perform better than in Fig. 4(a).



(a)



(b)

Fig. 4. Extracted features of three subjects of movie frames: (a) before and (b) after illumination compensation

4-2 Eigen Face

From the detected faces we extract important features. Every pixel in a face image has its own feature values. So, if a face image $I(x,y)$ be a two-dimensional N by N array of intensity values, it can be represented by a row or a column vector of dimension N^2. In such case, an image maps to a point in this huge space where each dimension represents a feature. But, dealing with a huge dimensional space is very difficult in respect of computation. In our method we use Principal Component Analysis (PCA) also called the Karhunen-Loève Transform for achieving low dimensional subspace. PCA works in the following way in our face space transformation process:

Step 1. Obtain face images $I_1, I_2, ..., I_M$ (training faces) where all images are of same size.

Step 2. Represent every image I_i as a vector Γ_i.

Step 3. Compute the average face vector Ψ which gives mean image.

$$\Psi = \frac{1}{M}\sum_{i}^{M}\Gamma_i$$

Step 4. Subtract the mean face from original face

$$\Phi_i = \Gamma_i - \Psi$$

Step 5. Normalize the image difference.

Step 6. Compute the covariance matrix C :

$$C = \frac{1}{M}\sum_{n=1}^{M}\Phi_n\Phi_n^T = AA^T$$
$$where\ A = [\Phi_1\Phi_2...\Phi_M]$$

Step 7. Find the K best eigenvectors of AA^T: $u_i = Av_i$. u_i is called the eigenface.

Step 8. Normalize u_i such that $||u_i|| = 1$

Each face Φ_i in the training set can be reconstructed as a linear combination of the best K eigenvectors as:

$$\hat{\Phi} = \sum_{j=1}^{K} w_j u_j \quad (where \quad w_j = u_j^T\Phi) \qquad (1)$$

4-3 Classification

After extracting the features we fed the features to the classifier Support Vector Machine (SVM) to recognize the actor. These are a set of related supervised learning methods used for classification and regression [17].

For binary classification, consider the problem of separating the set of training vectors belong to two separate classes,

$$(x_1,y_1),...,(x_k,y_k), where\ x_i \in R^n, y_i \in -1,+1$$

with a $wx + b = 0$. The set of vectors is said to be optimally separated by the hyperplane if it is separated without error and the margin is maximal. According to [17] a canonical hyperplane has the constraint for parameters w and b as in the following equation.

A separating hyperplane in canonical form must satisfy the following constraints,

$$y_i(wx_i + b) \geq 1, i = 1,...,k \qquad (2)$$

Where the vector w defines a direction perpendicular to the hyperplane and b is the distance of the hyperplane from the origin.

The distance of a point x from the hyperplane is,

$$d(w,b,x) = \frac{|wx + b|}{\|w\|} \qquad (3)$$

The margin is $2/\|w\|$, according to its definition.

The OSH is given by:

$$w_{\Phi} = \sum_{i=1}^{k} \alpha_i y_i x_i \qquad (4)$$

Where α_1 is the Lagrange multiplier.

Thus,

$$b_{\Phi} = -\frac{1}{2} w_{\Phi} [x_{sv1} + x_{sv2}] \qquad (5)$$

Where X_{sv1} and X_{sv2} are support vectors, satisfying,

$$\alpha_{sv1}, \alpha_{sv2} > 0, y_{sv1} = 1, y_{sv2} = -1$$

For a new data point $Z_{\neq w}$, the class of the sample y_{new} is then found using:

$$y_{new} = sign(w_* z_{new} + b_*) \qquad (6)$$

V. Experimental Results

Illumination compensation is performed to reduce the effect of lighting on the face images. Fig. 6 shows the recognition performance with respect to the number of features without illumination compensation. The figure illustrates that performance varies irregularly with the number of features. On the contrary, Fig. 7 shows the recognition performance after applying illumination compensation in the faces. In Fig. 8, gray level performs better than RGB.

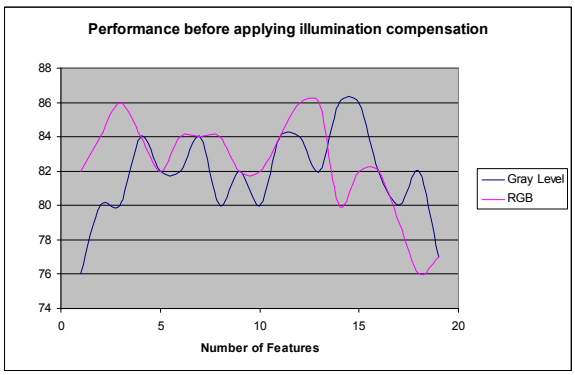


Fig. 6. Number of Features vs. Performance for gray level and RGB images before applying illumination compensation.

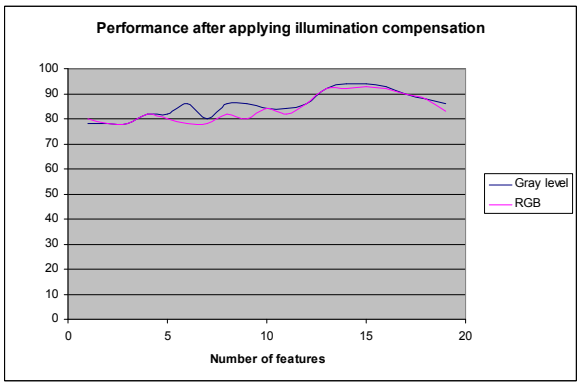


Fig 7. Number of Features vs. Performance for gray level and RGB images after applying illumination compensation.

VI. Conclusions

We propose a person identification method that can reliably index video sequences based on the actors' identity. The system is based on face detection and face recognition scheme which is intended for video retrieval applications and is able to tell whether or not a specific person is present in a video sequence. The concept of illumination compensated face is introduced in order to realize robust face recognition that is insensitive to lightening conditions. Research on video retrieval based on actor-face confronts the full range of challenges found in general purpose, object class recognition. We experiment on various numbers of training classes, features without and with illumination compensation for gray level and RGB face images. The overall result of our experiment is quite satisfactory for video indexing.

Acknowledgements

This research was supported by the Internet information Retrieval Research Center (IRC) in Korea Aerospace University. IRC is a Regional Research Center of Gyeonggi Province, designated by ITEP and Ministry of Knowledge Economy.

References

[1] B. Leibe, A. Leonardis and B. Schiele, "Combined object categorization and segmentation with an implicit shape model", *In Proc. European Conf. Comp*. Vis. Workshop, 2004.

[2] Opelt, A. Pinz, and A. Zisserman, “A boundary-fragment model for object detection”, *In Proc. Eur. Conf.* Comp. Vis., 2006.

[3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Washington, DC, USA, vol. 1, pp. 886–893, 2005.

[4] M. Oren et al., “Pedestrian detection using wavelet templates”, *Proceedings of Computer Vision and Pattern Recognition*, pp. 193–199, 1997.

[5] D. Gavrila, “Pedestrian detection from a moving vehicle”, *Proceedings of European Conference on Computer Vision*, pp. 37–49, 2000.

[6] J. Shotton, A. Blake, and R. Cipolla, “Contour-based learning for object detection”, *Proceedings of International Conference on Computer Vision*, 2005.

[7] R. E. Schapire and Y. Singer, “Improved boosting using confidence rated predictions,” *Machine Learning,* vol. 37, no. 3, pp. 297–336, 1999.

[8] P. Viola and M. Jones, “Robust real-time object detection,” *International Journal of Computer Vision*, 2002.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 1, pp. 511-518, 2001.

[10] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm”, *Proceedings of the Thirteenth International Conference on Machine Learning,* Morgan Kauman, San Francisco, pp. 148-156, 1996.

[11] Z. Li, and X. Tang, "Using Support Vector Machines to Enhance the Performance of Bayesian Face Recognition", *IEEE Transactions on Information Forensics and Security,* vol. 2, no. 2, pp. 174-180, June 2007.

[12] M. Turk and A. Pentland, “Face recognition using eigenfaces,” *IEEE Internation conference on Computer Vision and Pattern Recognition,* pp. 586–591, 1991.

[13] P. Belhumeur, J. Hespanha, and D. Kriegman, “Eigenface vs. Fisherfaces: Recognition using class specific linear projection,” *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp.711–720, July 1997.

[14] B. Moghaddam, T. Jebara, and A. Pentland, “Bayesian modeling of facial similarity,” *NIPS,* pp. 910–916, 1998.

[15] T. Shakunaga and K. Shigenari, “Decomposed eigenface for face recognition under various lighting conditions,” *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

[16] V. N. Vapnik, "Statistical learning theory", John Wiley & Sons, New York, 1998.

[17] Guodong Guo, S.Z. Li, and Kapluk Chan, "Face recognition by support vector machines", *International Conference on Automatic Face and Gesture Recognition,* pp. 196 - 201, 2000.

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

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, 멀티미디어

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

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