비디오 인덱싱을 위한 얼굴 검출 및 매칭

모하마드 카이룰 이슬람, 이순탁, 윤재웅, 백중환
한국항공대학교 정보통신공학과

Face Detection and Matching for Video Indexing

Mohammad Khairul Islam, Sun-Tak Lee, Jae-Yoong Yun, Joong-Hwan Baek
Department of Information and Telecommunication Engineering
Hankuk Aviation University

Abstract

This paper presents an approach to visual information based temporal indexing of video sequences. The objective of this work is the integration of an automatic face detection and a matching system for video indexing. The face detection is done using color information. The matching stage is based on the Principal Component Analysis (PCA) followed by the Minimax Probability Machine (MPM). Using PCA one feature vector is calculated for each face which is detected at the previous stage from the video sequence and MPM is applied to these feature vectors for matching with the training faces which are manually indexed after extracting from video sequences. The integration of the two stages gives good results. The rate of 86.3% correctly classified frames shows the efficiency of our system.

Keywords : Video Indexing, Face Detection, PCA, MPM

I. Introduction

Multimedia retrieval and browsing techniques improve a service quality of the multimedia and value of contents that the media provider possesses, and it is a salient research topic in the multimedia service industry. So, research for multimedia indexing and retrieval has to be worked continuously. There is a recent research of semantic based video analysis. In the semantic video indexing, adjacent shots are clustered and clustered shots compose the story units [2]. Video summarization based story units provides a higher-level video context, however, not every shots contains a meaningful thematic topic [1]. For that reason, specific event detection methods are demanded for a higher semantic level so as to better reveal, represent and abstract the video content. The typical research for content-based video indexing includes face detection [3], speaker identification and character recognition [4].

In this paper, we propose an efficient face region detection and matching algorithm using visual information like skin color. The paper is organized as follows. Section 2 demonstrates the framework of the approach. Section 3 face detection. Section 4 and 5 describes PCA (Principal Component Analysis) and MPM (Minimax Probability Machine) respectively. In Section 6, an experimental result shows that the proposed method has a high performance. Finally, the conclusion and discussion of future work are given in Section 7.

II. Framework for our approach

Given an arbitrary video sequence, the goal of this approach is to determine whether or not there is any actor in the sequence, if present, identify the actor. The following figure illustrates the working procedure of our approach.

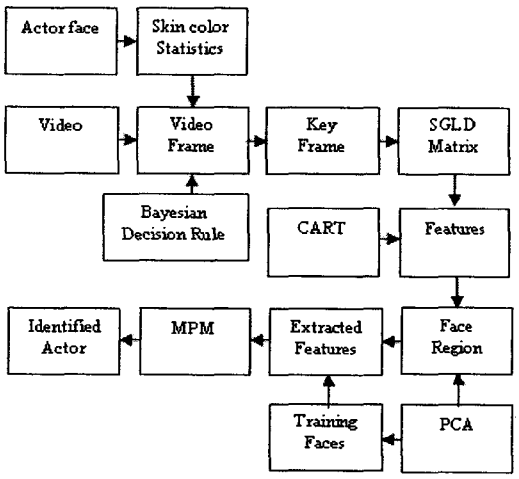

Fig. 1.General schema of our approach to video indexing.

III. Face Region Detection

In this paper, for face candidate, we utilize YCbCr color space that is normalized from -0.5 to 0.5 and analyze statistical attribution for skin color. Skin region is classified using Bayesian decision rule based on the statistical attribution. As result of skin detection, we get binary image. To eliminate noise and isolated lines, cross median filter is applied to the binary image. Then the binary image is segmented by horizontal and vertical projections to remove non-skin blocks precisely. Among the face candidate regions, to extract only the face regions SGLD matrix is used for features extraction and classification tree for minimizing the false detected region.

A. Feature extraction from SGLD

For an image with level [0, L-1], let I(i, j) be the gray level value at pixel (i, j). As shown in equation (1), the number of occurrence in two neighboring pixels displaced by a vector (m, n) for m = 1, 2,···, M and n = 1, 2,···,N can be calculated. It is denoted as P_{ab}(m, n) and called as SGLD matrix. In equation (1), # means the counting number of set {I(i, j) = a, I(i+m, j+n) = b}. W and H denote the width and the height of the image, respectively. SGLD matrix is normalized approximately by equation (2). Textural feature values are derived from normalized matrix N_{ab}(m, n).

$$P_{ab}(m,n) = \#\{(i,j),(i+m,j+n)\} \in (W \times H) \quad I(i,j) = a, I(i+m,j+n) = b\} \tag{1}$$

$$N_{ab}(m,n) = P_{ab}(m,n)/(W \times H) \tag{2}$$

Textural features derived from SGLD matrix include energy, entropy, inertia, inverse difference, and correlation. We utilize inertial, inverse difference, correlation shown in equation (3), (4), and (5), respectively.

$$B_I(m,n) = \sum_{a=0}^{L-1}\sum_{b=0}^{L-1}(a-b)^2 N_{ab}(m,n) \tag{3}$$

$$B_D(m,n) = \sum_{a=0}^{L-1}\sum_{b=0}^{L-1}\frac{1}{1+(a-b)^2} N_{ab}(m,n) \tag{4}$$

$$B_C(m,n) = \frac{1}{\sigma^2}\sum_{a=0}^{L-1}\sum_{b=0}^{L-1}(a-\mu)(b-\mu)N_{ab}(m,n) \tag{5}$$

The classification performance varies depending on M, N, and normalized image size. In our experiment, when SGLD matrix with M=N=6 is extracted from 6078 normalized image, the result of classification has the highest performance.

B. Classification trees

Classification tree grows through the splitting and pruning process [5]. In splitting process, each node of the tree chooses a feature decreasing the impurity as much as possible in the descendent node and splits the data into subsets by the selected feature. We measure impurity using Gini impurity function. shown as equation (6) that is generalized as the variance impurity function useful in the two-class case. $P(\omega_i)$ of (6) is the fraction of patterns at node N that are in class ω_i.

$$i(N) = \sum_{i \neq j} P(\omega_i)P(\omega_j) = 1 - \sum_j P^2(\omega_j) \tag{6}$$

The tree growing goes on until all terminal nodes are pure.

IV. PCA (Principal Component Analysis)

Let a face image I(x, y) be a two-dimensional N by N array of intensity values, or a vector of dimension N2. A typical image of size 78 by 60 describes a vector of dimension 4,680, or, equivalently, a point in 4,680-dimensional space. Images of faces will not be randomly distributed in this huge image space, rather can be represented by a relatively low dimensional subspace. The main idea of getting low dimensional subspace is to extract the important features from the face space. This can be done using the using Principal

Components Analysis (PCA) also called the Karhunen-Loève transform for achieving low dimensional subspace. PCA finds the important components of a dataset in their descending order.

Let the training set of face images be the vectors $\Gamma_1,\Gamma_2,\Gamma_3,\cdots,\Gamma_M$. The average face of the set is defined by $\psi=\frac{1}{M}\sum_{n=1}^{M}\Gamma_n$. The covariance matrix is calculated as $C=\frac{1}{M}\sum_{n=1}^{M}\int \Phi\Phi_n^T=AA^T$, where $\Phi_i=\Gamma_i-\Psi$ and $A=[\Phi_1,\Phi_2,\cdots,\Phi_m]$. Let, u_i and λi be the eigenvectors and associated eigen values of the C matrix respectively. But, AAT is impractical. So, if we consider v_i as eigen vectors of the AT A, then we can find $u_i=Av_i$. Thus M largest orthonormal vectors u_i and their associated values λi are taken for processing because they can describe the images significantly. Thus the calculations are greatly reduced from the order of the number of pixels in the images N^2 to the order of the number of images in the training set (M). In face recognition arena, u_i is called eigenfaces. Each face image can be represented by combining these vectors linearly by equation (7) and normalized training face by equation (8).

$$\hat{\Phi}=\sum_{j=1}^{K}w_j u_j, \quad (where \quad w_j=u_j^T\Phi) \tag{7}$$

$$\Phi_i=\begin{bmatrix}w_1^i & w_2^i & \Lambda & w_K^i\end{bmatrix}^T, \quad i=1,2,\Lambda,M \tag{8}$$

These ω_i's are the extracted feature for face recognition.

V. MPM (Minimax Probability Machine)

When constructing a classifier, the probability of correct classification of future data points should be maximized. In face recognition, two kinds of well-known classifiers are Neural Network and Support Vector Machine. The research work in [6] demonstrated the success of eigenface, SVM, NN on face recognition applications. On the other hand, MPM is a very new classification technique. It enjoys competitive classification performance comparing with most of state-of-the-art classification techniques. The most attractive properties of MPM are that it can explicitly provide a worst-case bound on the probability of misclassification of future data when the mean and covariance matrix of the data are known.

The goal of MPMC is to find a decision boundary $H(a,b)=\{z|a^Tz = b\}$ such that the minimum probability H of classifying future data correctly is maximized. If we assume that the two classes are generated from random vectors x and y, we can express this probability bound just in terms of the means and covariance of these random vectors.

The following result due to Marshall and Olkin[7] and extended by Bertimas and Popescu[8] provides the theoretical basis for assigning probability bounds to hyperplane classifiers:

$$\Omega_H=\inf_{x\sim(\bar{x},\Sigma_x),y\sim(\bar{y},\Sigma_y)}\Pr\{\mathbf{a}^T\mathbf{x}\geq b\wedge\mathbf{a}^T\mathbf{y}\leq b\} \tag{9}$$

Note that we do not make any distributional assumptions other than that $\bar{\mathbf{x}}$,X, $\bar{\mathbf{y}}$, y are bounded. Exploiting a theorem from Marshall and Olkin[7], it is possible to rewrite(7) as a closed expression: $\Omega_H=\frac{1}{1+m^2}$.

where,

$$m=\min_{\mathbf{a}}\sqrt{\mathbf{a}^T\Sigma_x\mathbf{a}}+\sqrt{\mathbf{a}^T\Sigma_y\mathbf{a}} \quad s.t. \quad \mathbf{a}^T(\bar{\mathbf{x}}-\bar{\mathbf{y}})=1 \tag{10}$$

The optimal hyper plane parameter a_* is the vector that minimizes (10). The hyper plane parameter b* can then be computed as equation (11).

$$\mathbf{b}_*=\mathbf{a}_*^T\bar{\mathbf{x}}-\frac{\sqrt{\mathbf{a}_*^T\Sigma_x\mathbf{a}_*}}{m} \tag{11}$$

A new data point Z_{new} is classified according to sign($a_*^TZ_{new}$-b*); if this yields +1, Z_{new} is classified as belonging to class x, otherwise it is classified as belonging to class y.

VI. Experiments

In our experiment, during detection, we use AT&T face images for face class and Corel images for non-face class. Our system shows precision ratio of 0.85% in an average for face region detection. Figure 2 shows the extracted face regions from several video frames.

We, in our experiment, build a database of 210 face images, including 21 distinct persons from three videos, each with 10 faces that vary in position, rotation, scale, and different expressions. Fig. 3 shows

snapshots of two persons.

Fig. 2. Detected face regions.

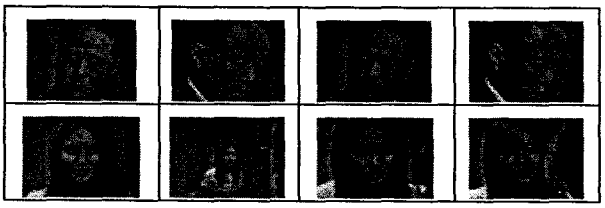

Fig. 3. Snapshots of an actor and an actress from

We construct multi-class classification technique by employing binary MPM classifier. We use the 10-folder cross validation. Table 1 shows correct face recognition rate in percentage, obtained from ten trials, for both AT&T face database and video sequences.

Table 3. Experimental results on the AT&T face database and video sequence.

Subset	AT&T Database	Video sequence
1	96.90%	86.59%
2	96.84%	87.95%
3	95.41%	81.37%
4	95.89%	89.32%
5	96.13%	83.64%
6	96.96%	87.05%
7	96.78%	89.55%
8	96.50%	86.14%
9	99.96%	84.32%
10	97.14%	87.50%
Average	96.50%	86.34%

Ⅶ. Conclusion and Future work

We propose a face matching method after detection of face region for video indexing. In face detection, image segmentation using projection can save a large amount of computer time and space compared to working on the original image because an nxn image contains only about n data elements. PCA and MPM can reliably evaluate future data class; although they have higher computational cost. Experimental result shows that classification rates are 96.50% and 86.3% for AT&T face database and video frames, respectively. The correct classification rate of video frames is less than AT&T, because video frames suffer from variations in lighting, pose, scale etc.

In future work, we will consider event based video summarization and indexing by speaker identification using audio-visual information.

본 논문은 산업자원부 한국산업기술평가원 지정 한국항공대학교 부설 인터넷정보검색 연구센터의 지원에 의함.

References

[1] Azriel Rosenfeld *et al.*, Video Mining, Kluwer Academic Publishkers, 2003.

[2] Minerva Yeung *et al.*, "Segmentation of Video by Clustering and Graph Analysis", Computer Vision and Image Understanding, vol. 71, no. 1, pp. 94~109, 1998.

[3] Hualu Wang *et al.* "A Highly Efficient system for Automatic Face Region Detection in MPEG Video", *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 7, no 4, pp. 615-628, 1997.

[4] Ying Li *et al.*, Video Content Analysis using Multimodal Information, Kluwer Academic Publishers, 2003.

[5] Breiman and Friedman *et al.*, Classification And Regression Trees, CRC Press, 1998.

[6] Phillips, "Support vector machines applied to face recognition," in *Proc. Advances in Neural Information Processing Systems*, 1998.

[7] W. Marshall and I. Olkin. Multivariate chyshev inequalities. Annuals of Mathematcal statistics, 31(4)L1001-1014, 1960.

[8] Popescu and D. Bertsimas. Optimal inequalities in probability theory: A convex optimization approach. Technical Report TM62, INSEAD, Dept. math. O.R., Cambridge, Mass, 2001.