DETECTION OF FACIAL COMPONENTS IN A VIDEO SEQUENCE BY INDEPENDENT COMPONENT ANALYSIS

Kunio Takaya and Kyung-Yung Choi†

Electrical Engineering, University of Saskatchewan
57 Campus Drive, Saskatoon, Sask. S7N 5A9, Canada
Phone: (306) 966-5392 Fax: (306) 966-5407
E-mail: kunio_takaya@engr.usask.ca

ABSTRACT

This paper presents two approaches for detecting facial components in the images contained in a video sequence by Independent Component Analysis (ICA). The ultimate objective is to map detected facial components such as eyes and mouth to a 3D wireframe model to be used in facial animation. One approach is to use face localization technique whereas the other deals with an entire image without cropping face part, but filtering with wavelet subband filters. Face localization is achieved by using the human skin color filtering in YCbCr subspace with other relevant facial information. The wavelet subband filter in the latter approach removes facial parts occupying a large area that contribute in ICA as large basis vectors. The method of Independent Component Analysis is then applied to identify main facial components for subsequent geometrical shape analysis.

1. INTRODUCTION

In MPEG-4, a new feature of the face and body animation was added to meet the demands such as telecasting. A customized 3-D agent model as often referred to as avatar can replace a real newscaster in T.V. program, or can be used as a storyteller. In order to take full advantage of this newly added feature in MPEG-4, it is crucially important to develop a means to automatically produce a 3-D facial model from a video sequence. Such facial models must be described with 3-D geometrical wire frame structures with human like textures associated with the face. From the obvious reason that video images are 2-D whereas the target model is 3-D, the artificial intelligence aided by a data base needs to supply any missing information to construct such a 3-D wire frame facial model.

To construct 3-D facial model, initially, the face needs to be detected and localized in the video sequence. This is also an essential step required for facial feature extraction. Various approaches face detection and localization have been reported, and a comprehensive survey is presented by Yang et al [13]. The human skin color filtering in sub-color spaces has been proved to be effective, and has been adopted in most of the face detection systems at the front end in conjunction with other attributes of the face. Once the face is detected and localized, the stage of facial feature extraction follows to extract specific facial components, which will be subsequently used for synthesizing a face model and for tracking their motion.

Principal Component Analysis (PCA) is an approach that takes dominant information of data and reconstructs it with reduced dimensionality in terms of the eigen-space analysis that minimizes the reconstruction error. Although PCA has been successfully applied to many face recognition systems with the 2nd order statistics concept, a number of papers argued that PCA method might lose important information in higher order statistics of facial images. On the other hand, a recently developed technique, called Independent Component Analysis (ICA) [2-4] is a powerful method to represent signals or images, retaining higher order statistics. Blind Source Separation (BSS) and feature extraction are its main application area. In the field of feature extraction using ICA, many papers showed its superior face recognition capability than that of the eigen-face approach based on PCA.

In this paper, we address the method of presenting the facial images with some localized facial features such as eyes and mouth using an ICA approach, which

The authors would like to thank TRlabs for cntinued support in the form of fellowship†.

extracts the statistically independent basis image sets from multiple image presented. Initially skin color filtering is performed in YCbCr color space to a video sequence under a certain constrained conditions. Some other relevant information such as size, shape and geometrical features are combined to localize a face. After the ICA processing on the localized facial images in a video sequence, we will have some localized facial features among all the independent basis image sets that can be used as the key to identify major components of the facial images in the video sequence. The chosen facial features represented by associated coefficients are the specific parts in a face, eyes and mouth parts.

The face object decoding table and definition set in the MPEG-4 standard can be applied to the wide range of facial expression, to have authentic facial expression of an individual. When facial features extracted from a video sequence are mapped onto the facial animation feature point set defined in the MPEG-4, the extracted features play a key role in controlling facial motion.

2. FACE LOCALIZATION

Although the color discrimination for detecting face has some fundamental problems such as the lack of reliability in varying luminance conditions, background color close to skin color, and different facial colors, it is still a favorable method because of its computational efficiency and robustness against orientations. A number of color spaces have been examined including HSV, normalized RGB, YES, CIE XYZ and other subspaces for constructing the human skin color model. The color descrimination method is usually applied in conjuction with other relevant information describing a face [5-12]. The method used in this paper for the face localization has a constrained condition: There is only one frontal face shown through the video sequence. Though this condition excludes many video clips even with a face or faces, the cases of telecasting or storyteller satisfy the condition. Constrained under this condition, the location of the face within the image was determined, considering such characteristics as skin-like color, an elliptical shape and the ratio between major and minor axis of the ellipse model. The task of skin color filtering was performed on the subtracted image of two different frames in video sequence as shown in Figure 1.

Figure 1. The image subtraction.

The image subtraction simplified the localization algorithm for as false detection was avoided at the skin color-like background. The search space where skin color filtering is applied, is also reduced from a whole image to only a moved image block. Besides that, potential ability of modeling hair is gained as well. The moved block to which skin color filtering is applied is roughly identified by checking the values in a subtracted image. After applying skin color filtering in the normalized YCbCr color space: $\bar{Cb} = Cb/(Y+Cb+Cr)$ and $\bar{Cr} = Cr/(Y+Cb+Cr)$, a skin color region in the image is obtained by simply applying threshold values to $\bar{Cb}$ and $\bar{Cr}$ components: Figure 2.

Figure 2. Skin color filtered image.

The YCbCr color space has been chosen because it is not only closely related to MPEG coding also has a compact skin color distribution in chrominance plane. The luminance component is not useful to find skin color region itself, however, it can extract head-only region in some case. For example, the shade right below the chin in a certain lightening condition and the area below the chin can be removed with luminance information combined with chrominance component. This combination of luminance and chrominance finds the exact head position to cut out a face image by the size filter. Then, an ellipse, initially set as a circle, is palced to bring the upper end to the upper vertex of skin-like region. Keeping this upper vertex of the ellipse model fixed, this ellipse is enlarged to come to the bottom of skin-color region or until a pre-defined ratio between the major and minor axis of this ellipse model is reached. During this enlarging process, the ratio of the two areas, skin color region and an ellipse, is calculated. Since this ratio generally exhibits a cubic function, the most negative derivative gives the best match of the ellipse to the skin color region. The facial images are thus localized by skin color, then normalized to a pre-assigned size for the subsequent step of ICA analysis. Figure 4 shows the results of face localization.

3. FEATURE EXTRACTION USING ICA

The ICA method used in this paper for finding facial features, or equivalently finding the independent basis image sets in the video sequence is conceptually based on the method developed by Bartlett et al [1]. Consider the following ICA model.

$$S = B^T X \tag{1}$$

where $X = [x_1^T, x_2^T, \cdots, x_n^T]^T$ and $S = [s_1^T, s_2^T, \cdots, s_n^T]^T$. The row vector X, $x_i \in R^N$ is a zero-mean vector and is the concatenation of the column vectors representing the localized facial image matrixes in a video sequence. The row vector S, $s_i \in R^N$ is the ICA basis vector and B is a weight matrix to be determined. The fundamental concept in [1] is to represent image data with independent component bases and their associated coefficients. ICA is performed on $m(\leq n)$ principal eigenvectors of the covariance matrix of input images. However, in this paper, the dimension of independent basis sets is reduced to an appropriate number, m, by using a whitening matrix, $M \in R^{m\times n}$, if necessary. Note that PCA approach is applied to find the whitening matrix.

$$M = D^{-1/2} V^T \tag{2}$$

Where, $D = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_m)$ has the eigen values of covariance matrix of X in descending order and $V = [v_1, v_2, \cdots, v_m]$ is the eigen vector matrix corresponding to the eigen value. The whitened input given by

$$\bar{X} = MX. \tag{3}$$

is the input to be supplied to the ICA algorithm. Computationally efficient Hyvärinen's algorithm [2] described by

$$W(k+1) = E\{\bar{X}(W(k)^T \bar{X})^3\} - 3W(k) \tag{4}$$

was adopted in this paper. The ICA basis are obtained by

$$S = W^T \bar{X}. \tag{5}$$

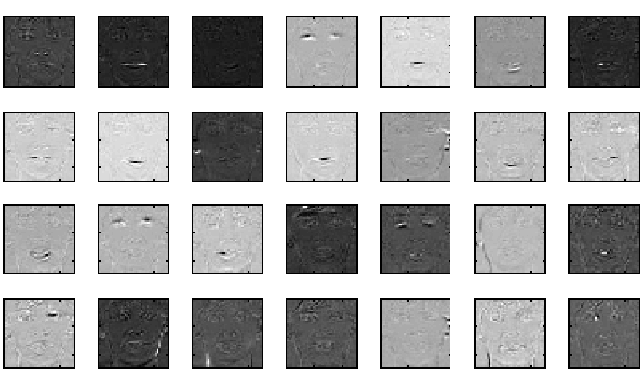

Figure 3. The ICA basis.

An approximation of the original input image X is reconstructed with the ICA bases and their corresponding coefficient.

$$X = M^{-1} W S \tag{6}$$

The reconstruction itself, however, is not of primary interest. The objective is to find some key components common in facial images such as eyes and mouth. To achieve this goal, only several ICA bases need to be selected among all the ICA bases. Since some ICA bases have highly concentrated values representing specific parts of interest, major components are carefully studied. For instance, the 4th ICA basis in Figure 4 is considered as a eyes feature. Though all ICA bases and associated coefficient vectors are required to reconstruct the original images, only a few selected ICA bases are sufficient to reconstruct localized images of facial parts such as around eyes or mouth area. Because the linear combination of highly localized ICA bases and their associated coefficients reflects only on the highly localized area in the images. To find those ICA basis vectors considered as major components of face such as eyes-like or mouth-like ICA basis vectors, a heuristic approach is used: Eye-like features are found in the upper half and either in the left or right half in a ICA basis matrix. Another similar heuristic is applied to mouth.

Figure 4. The localized facial images.

4. SIMULATION RESULTS

The first 60 frames of 144 × 176 QCIF formatted video sequence of Miss America are chosen. The method mentioned in part 2 was applied to localize the face. A subtracted image was obtained at the interval of 10 frames. The threshold for the normalized chrominance values to extract skin color region is set to be 0.15 $< \bar{C}b <$ 0.3 and 0.3 $< \bar{C}r <$ 0.42. The pre-defined ratio between major and minor axis of the ellipse model is set as 2.5. The dimension of facial feature, m, is chosen to be the same as the number of input image,

n. The localized facial images are normalized to the matrix size of 50×50. The ICA method is applied to the luminance components of these localized facial images to find the ICA basis vectors. Eye- and mouth-like facial component images resulting from projecting the facial images to each of the respective basis vectors are shown in Figure 5 and 6.

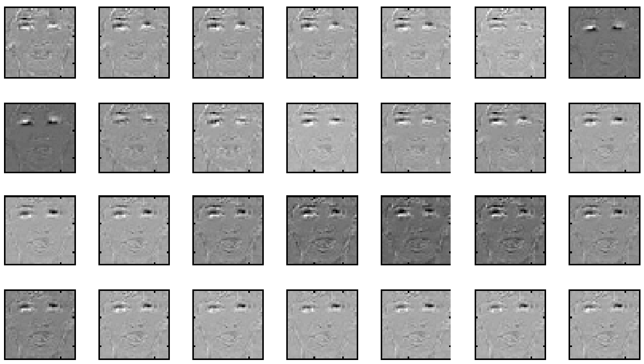

Figure 5. Extracted eyes part in the video sequence.

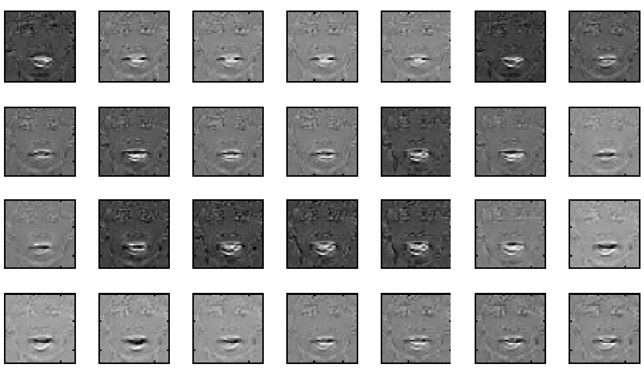

Figure 6. Extracted mouth part in the video sequence.

5. WAVELET TRANSFORM TO ELIMINATE FACE LOCALIZATION

As demonstrated by the simulation results, ICA works well to extract facial components such as eyes and mouth. However, due to the fact that the areas of eyes and mouth are small relative to other areas occupied by such components as hair, face background, neck and collars of the dress as well as the background behind a person, the ICA method requires face localization prior to applying ICA. Head motions associated with a video sequence that contains a single face is another factor requiring face localization. When a face is displaced within a sequence of video, detection of facial components tends to be more difficult unless the face portion is cut out and normalized to a face alone image frame.

Observed is that the magnitudes of the ICA vectors that consists of a facial component is much smaller than other dominating ICA vectors representing the general shape of a face and backgrounds. In another word, signal to noise ratio is too low for successfully obtaining a facial component from the entire image not localized for any specific face region. For example, a large hair area overshadows eyes and mouth. In order to reduce the magnitudes of the ICA basis vectors representing objects occupying a large area, the wavelet subband filtering [15] (Multi Resolution Analysis) is an effective means to filter out anything except the details representing eyes and mouth in particular.

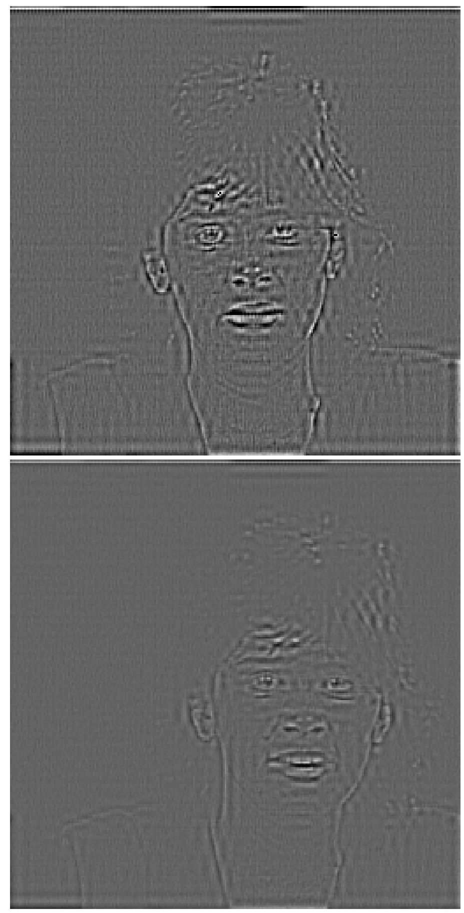

Figure 7. Filtered facial images showing only the details produced by the wavelet transform.

When a scale function $\phi_{j,k}$ is of two-scale, i.e.

$$\phi_{j,k} = \phi(2^j t - k)$$

the wavelet function $\psi(2^j t - k)$ defined by the mother function ψ is orthogonal to $\phi_{j,k}$. When a scaled function of $f(t)$ at the scale j is expanded with a set of coefficients c_k^j,

$$f_j(t) = \sum_k c_k^j \phi(2^j t - k)$$

its dual function $g(t)$ at the same scale j can be expanded with the wavelet functions as

$$g_j(t) = \sum_k d_k^j \psi(2^j t - k).$$

The relationship between adjacent scales,

$$f_{j+1}(t) = f_j(t) + g_j(t)$$

leads to a progressive relationship relating the function $f(t)$ at two different scales at N and M,

$$f_N(t) = g_{N-1}(t) + g_{N-2}(t) + \cdots g_{N-M}(t) + f_M(t).$$

Thus, the wavelet transform, or decomposition is usually calculated as schematically illustrated by

$$\begin{array}{ccccccccc} & & d^{N-1} & & d^{N-2} & & \cdots & & d^{N-M} \\ & \nearrow & & \nearrow & & \nearrow & & \nearrow & \\ c^N & \rightarrow & c^{N-1} & \rightarrow & c^{N-2} & \rightarrow & \cdots & \rightarrow & c^{N-M} \end{array}$$

The inverse wavelet transform, or reconstruction is described schematically as

$$\begin{array}{ccccccccc} d^{N-M} & & d^{N-M+1} & & \cdots & & d^{N-1} & & \\ & \searrow & & \searrow & & \searrow & & \searrow & \\ c^{N-M} & \rightarrow & c^{N-M+1} & \rightarrow & c^{N-1} & \rightarrow & \cdots & \rightarrow & c^N \end{array}$$

The Daubechies wavelet function with 10 for the number of vanishing moments was used to filter out low frequency image components. Since the image size is $256 \times 256 = 2^8$, the image reconstructed from c^7 and d^7, or c^7 with $d^7 = 0$ was used as the image holding only the details instead of the original image c^8. Two filtered images out of 30 video frames used to detect eyes and mouth without face localization are shown in Fig. 7. Note that the position of the face is slightly displaced in these two images. After applying the same ICA procedure targeted to detect mouth, the images shown in Fig. 8 were obtained as results. In spite that no face localization was applied, the mouth image appeared clearly.

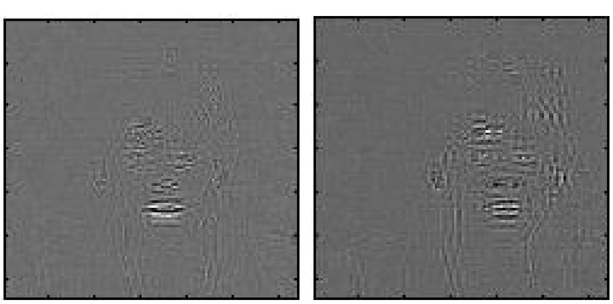

Figure 8. Extracted mouth part obtained from the filtered facial images by the wavelet transform.

In order to determine how signal to noise ratio will improve with respect to mouth image, the power ratio defined by

$$\gamma = \frac{\displaystyle\sum_{\text{basis vectors of mouth}} ||s_i||^2}{\displaystyle\sum_{\text{all basis vectors}} ||s_i||^2}$$

was calculated for three classes of video sequences, (1) face localized and not filtered, (2) face not localized and filtered, and (3) face localized and filtered. Where, s_i is the magnitude of the ICA basis vectors.

type of image	basis	power ratio
face localized, original	9	0.0193
whole image, filtered	6	0.1004
face localized, filtered	6	0.2077

The table indicates that the component basis vectors representing objects occupying a large area in the original images were removed by the wavelet filter, resulting in a greater value of the power ratio γ. It also shows that localizing the filtered image further increase the power ratio.

6. CONCLUSION

To realize authentic facial expression from an individual contained in a video scene for MPEG-4 face animation, it is essential to extract facial features from the video sequence. In this paper, we have examined a face localization technique using skin color filtering in YCbCr color space with some other information relevant to the face, and verified the concept of facial feature extraction using an ICA approach. Results showed the potential capability of this approach for extracting facial features, making facial component images available with representative ICA basis vectors. The performed ICA algorithm seems to regard the movement of facial part in the video sequence as independent signals. The number of ICA basis to extract major components of face should be chosen properly depending on the property of video sequence. This method appears to be promising to capture minute variations involved in eye and mouth motions so that the face animation system can take advantage of the ICA method.

Face localization, however, means an additional step to extract facial component images. Though face localization technique using skin color filtering in YCbCr color space works reasonably well for light skin color without shading, alternative methods that could skip the face localization entirely, are more desireable. The wavelet filtered detail images resulting from Multi Resolution Analysis appear to be such effective alternative as power associated with large objects can be suppressed significantly, increasing the detectability of facial components directly from the original image without localizing a face and cropping it.

7. REFERENCES

[1] M. Bartlett, H. Laddes, T. Sejowski, Independent component representations for face recognition, In Proc. SPIE Conf. On Human Vision and Electronic Imaging III, 3299: 528-539, 1998.

[2] A. Hyvärinen and E. Oja, A fast fixed point algorithm for independent component analysis, Neural Computation, 9: 1483-1492, 1997.

[3] A. Hyvärinen and E. Oja, Independent component analysis: Algorithms and applications, Neural Networks, 13(4): 411-430, 2000.

[4] J. Hurri, A. Hyvärinen, J. Kahunen and E. Oja, Image feature extraction using independent component analysis, Proceedings of IEEE Nordic Conference on Signal Processing (NORSIG'96), 1996.

[5] E. Saber and A. Tekalp, Face detection and facial feature extraction using color, shape and symmetry based cost function, Proceedings of ICPR'96, pp.654-658, 1996.

[6] G. Wei and I. Sethi, Face detection for image annotation, Pattern Recognition Letters, 20: 1313-1321, 1999.

[7] H. Wu, Q. Chen and M. Yachida, Face detection from color images using a fuzzy pattern matching method, IEEE Trans. On Pattern Analysis and Machine Intelligence, 21(6): 557-563, 1999.

[8] C. Garcia, G. Tziritas, Face detection using quantized skin color regions merging and wavelet packet analysis, IEEE Trans. On Multimedia, 1(3): 264-277, 1999.

[9] J. Terrillon, and S. Akamatsu, Comparative performance of different chrominance spaces for color segmemtation and detection of human faces in complex scene images, Proc. 12th Conf. on Vision Interface (VI '99), 2: 180-187, 1999.

[10] Z. Liu, Z. Zhang, C. Jacobs and M. Cohen, Rapid modeling of animated faces from video, Technical Report MSR-TR-2000-11, Microsoft Research, 2000.

[11] S. Tsekeridou and I. Pitas, Facial feature extraction in frontal views using biometric analogies, IX European Signal Processing Conference (EUSIPCO'98), Rhodes, Greece, 1998.

[12] J. Ahlberg, A system for face localization and facial feature extraction, Report no.LiTH-ISY-R-2172, Linkoping University.

[13] M. Yang, N. Ahuja and D. Kriegman, Detecting faces in images: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001.

[14] K. Takaya, N. Aoki, Low bit-rate video by creation of avatars for a certain class of video images - a VRML/JAVA approach, International conference on communication, control and signal processing, pp.239-243, 2000.

[15] M. Vetterli and C. Herley, Wavelets and filter banks: theory and design. IEEE Trans. Acoust. Speech Signal Processing., 40(9) 2207-2232, 1992