A HIGHLY EFFICIENT AUDIO-VISUAL RECOGNITION SYSTEM BASED ON MULTIBAND FEATURE

D. Manivannan, N.K. Manikandan and Ravikumar S
Department of Computer Science and Engineering
Veltech Dr.RR & Dr.SR University, Chennai, India

ABSTRACT

This paper presents a extremely high-octane audio-visual detection system in compression domain. For face recognition systems, the multiband feature fusion method selects the wavelet sub bands that are equable to illumination and facial expression variations. These sub bands will be distilled directly from the inverse quantization in the compression system. By taking the inverse quantized wavelet coefficient of the video as the input, the inverse wavelet transform which equivalents to image reconstruction is omitted. As a result, the computational complexity of the conventional video-based face recognition system is attenuated. To present a set of new face localization methods to localize the facial wavelet coefficients from the wavelet sub band image. The additional feature (Audio) will be added into the facial feature to enhance the recognition performance of the system. The dual optimal multiband feature fusion method is then used to fuse the two set of wavelet coefficients and generate the scores. The proposed system achieves high recognition accuracy in audio-visual database and also provides low computational complexity.

Index Terms: Audio-Visual Recognition, Computational Complexity, Face Localization, Face Segmentation, Video-Based Face Recognition, Wavelet Transform.

http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=10

1. INTRODUCTION

The appeal for better security frameworks supports the improvement of biometric acknowledgment frameworks. A biometric acknowledgment framework is a computerized framework that checks or recognizes a man's personality utilizing a man's physiological attributes and additionally behavioral qualities. Single model biometric framework has impediments. For instance, acknowledgment execution of the speaker acknowledgment framework is profoundly reliant on the foundation clamor level. To defeat the constraints, the
multimodal biometric framework is utilized. There are many sorts of multimodal biometric frameworks. For instance, face and unique mark, face and palm print, face and iris, face and discourse, and face, unique finger impression, and hand geometry. Among these frameworks, the varying media acknowledgment framework is well known in light of the fact that the acoustic and visual biometric signs can be acquired utilizing easy to understand methodology and ease sensors.

The varying media acknowledgment framework is shaped by holding the speaker acknowledgment and visual acknowledgment system. The framework utilizes sound (discourse) and visual (facial) highlights to perceive a man's character. The facial highlights utilized as a part of the framework can be refined from a solitary static picture or from a video succession. Face acknowledgment framework which utilizes a solitary picture (still picture) and a video succession for acknowledgment are known as picture based acknowledgment and video based acknowledgment [1],[2], separately. Contrasted with video based acknowledgment, picture based acknowledgment is more inclined to false framework infiltration in light of the fact that pre-caught picture is less demanding to be produced than continuous video succession [4].

Varying media acknowledgment framework that embraces video-based face acknowledgment supports two issues. To begin with, video grouping generally contains various visual varieties. This beguiles an awesome test to the vast majority of the current face acknowledgment strategies [5] which are just strong to one specific visual variety. Analysts mindful of the issues and, in this manner, video-based face acknowledgment frameworks [1],[2] are proposed to address the issues as of late. Second, computational many-sided quality in the video-based face acknowledgment framework is typically high. This is on account of handling numerous edges of video with expansive information measurement takes longer time than preparing a solitary picture. Moreover, in the majority of the applications, video is should have been compacted and transmitted if long separation access to the acknowledgment framework is essential. Be that as it may, the computational multifaceted nature issue is from time to time being tended to. All the more as of late, Tang and Li [4] decreased the computational multifaceted nature of the video-based face acknowledgment framework by grasping sound flag. A video outline transient synchronization conspire is utilized to adjust edges of comparative pictures over the video successions. Consequently, just the chose video outlines are refined for acknowledgment.

In this paper, a very high-octane varying media acknowledgment framework in pressure area is proposed to address the two issues specified previously. For confront acknowledgment, the video based face acknowledgment approach is received. The computational intricacy of the regular video-based face acknowledgment is constricted in pressure space. To do as such, we distil the converse quantized wavelet coefficient sub band of the info video succession for confront acknowledgment. By refining wavelet coefficients specifically after opposite quantization, the backwards wavelet change which counterparts to picture remaking is overlooked. Albeit a few papers have proposed confront location and moving article following techniques in pressure area, none of these strategies address the varying media acknowledgment in pressure space as in this paper.

Other than computational intricacy issue, to likewise address the visual varieties issues in this paper. In the proposed framework, the double ideal multiband include technique [6] is grasped to unravel the brightening and outward appearance varieties issues. This strategy securities the wavelet sub groups which are invariant to enlightenments and outward appearance varieties and appoints the weights to the sub band's scores adaptively. The multiband include combination technique [7] is held to choose the two arrangements of wavelet sub groups and these wavelet sub groups are refined after the opposite quantization in the pressure framework.
A Highly Efficient Audio-Visual Recognition System Based On Multiband Feature

for confront acknowledgment. Other than facial district, a large portion of the information video successions brag complex foundation which sophisticates the acknowledgment procedure. Since the vast majority of the face restriction calculations are incited in light of reproduced picture, they can't be held straightforwardly in the pressure area. In this manner, we introduce an arrangement of face restriction and division strategies that works simply on the pressure space to limit the facial district in wavelet sub band. After the face district is restricted, the eyes and mouth areas are found by absorbing the power and movement data in the wavelet sub band picture. In light of the area of the eyes and mouth, the wavelet sub groups which brag just facial highlights will then be standardized and divided to a standard size for arrangement.

2. MOTIVATIONS

For confront acknowledgment framework, the confined wavelet change form the video outlines into wavelet sub band coefficients. The multiband include combination strategy [7] chooses the wavelet sub groups, OMF–I which is invariant to enlightenment varieties and OMF–E which is equable to outward appearance variety. OMF–I is shaped by linking HALL and AALH sub groups while OMF–E is framed by the ALL sub band ALL alludes to the estimation sub band of level-2 wavelet deterioration. Lobby alludes to the flat sub band of further trim of ALL sub band while AALH alludes to the estimation sub band of further embellishment of ALH sub band. Parallel outspread premise work (RBF) neural systems are utilized to arrange the OMF–I and OMF–E. The choice scores initiated by the RBF neural systems are straightly reinforced through an arrangement of combination weight by the double ideal multiband include strategy. In this technique, the weights are settled by the brightening variety estimator where the light variety factor will be allocated in view of the enlightenment variety level of the info picture. For instance, the weight allotted to the score of the OMF–I is higher than OMF–E if the information picture is in high enlightenment variety.

For speaker recognition system, the audio signal is feature distilled by the mel-frequency cepstrum coefficient (MFCC) and linear discriminant analysis (LDA) is used to further degraded the data dimension. The RBF neural network is used for audio classification. The scores induced by the face recognition system and speaker recognition system will be bonded by the sum rule decision fusion method.

![Figure 1](image-url) Block diagram of the audio-visual recognition system of our research work.

For video-based face recognition, video sequence provides more information than a single image, the tradeoff of this advantage is the increase of computational complexity in the recognition system. Since video needs to be compressed to enable efficiency in storage and transmission, we present a video-based face recognition system that makes use of the readily available technology of compression system to degrade the expensive computation involved in the system. This system will then be instrumented in the proposed audio-visual recognition system to increase the system efficiency.
Figure 2: Discrete wavelet sub band locations of (a) OMF−E, ALL which is invariant to facial expression variation at level-2 decomposition and (b) OMF−I, HALL and AALH which are invariant to illumination variation at level-3 decomposition.

3. PROPOSED AUDIO-VISUAL RECOGNITION SYSTEM IN COMPRESSION DOMAIN

Fig. 3 demonstrates the proposed varying media acknowledgment framework in the pressure area. The Dirac and MPEG sound layer 2 pressure frameworks are utilized as a part of the face acknowledgment and speaker acknowledgment framework, separately. The Dirac pressure framework is a long-gathering of picture video codec that utilizations wavelet changes and movement pay together with entropy coding. It is helpful in our proposed framework since it utilizes wavelet change which is now demonstrated its reasonability in confront acknowledgment framework as the change coding [6], [7].

Figure 3: Block diagram of the proposed audio-visual recognition system in compression domain.

Figure 4: Block diagram of the wavelet coefficient extraction and face localization process in the proposed audio-visual recognition system in compression domain.
A Highly Efficient Audio-Visual Recognition System Based On Multiband Feature

We show an arrangement of face limitation and division strategies and acknowledgment techniques that work absolutely on pressure area. Fig. 4 demonstrates the piece chart of the wavelet coefficient extraction and face confinement process (as alluded in Fig. 3) in the proposed framework. The proposed calculation utilizes space information to help the basic leadership. The ovals speak to the space information. The calculation has four phases, arrange one for wavelet sub band extraction, organize two for facial shading confinement, organize three for confront area division, and stage four for eyes and mouth restriction and sub band standardization.

A. Stage One: Wavelet Sub band Extraction

In the primary phase of the proposed framework, the wavelet sub groups of ALL and ALH of luminance (Y) and ALL of chrominance (Cr) of reference video outline (I) are refined after the backwards quantization. The ALL and ALH of luminance are refined for confront acknowledgment. The ALL of Cr is separated for confront restriction. Since the sub groups (ALL, ALH) required for confront acknowledgment are level-2 disintegration sub groups, the level of decay in the pressure framework is set as 2. Keeping in mind the end goal to distil the ALL and ALH sub groups effectively from the video outlines, the wavelet disintegration coding utilized as a part of a part of the pressure framework is predominant. The lifting wavelet change is utilized as a part of this pressure framework and it is made out of three stages. The initial step is to part the information grouping into even and odd arrangement. The second step is to anticipate the odd grouping utilizing the even arrangement. The distinction between the anticipated odd grouping with the odd arrangement is the high-recurrence coefficient and it is meant as H. The third step is to revamp the even arrangement utilizing the odd arrangement. The distinction between the anticipated odd grouping with the odd arrangement is the high-recurrence coefficient and it is indicated as L. For 2-D picture, the lifting wavelet decay is done first to lines at that point to segments. To begin with, the lines of the picture are separated into first half and second half.

The principal half alludes to the even succession and the second half alludes to the odd sequence. At that point the odd succession will be anticipated and the even grouping will be restored to incite low-recurrence sub groups L and high recurrence sub groups H. This procedure will then rehash itself on the sections of the picture to actuate four wavelet sub groups. In this framework, the Haar channel with no move is utilized. The lifting wavelet coefficients are produced as takes after:

\[
W_{h}^{j} = W_{h}^{j-1}(odd) - W_{l}^{j-1}(even) \\
W_{l}^{j} = W_{l}^{j-1}(even) + \left( (W_{h}^{j-1}(odd) + 1)/2 \right)
\]  

Where j alludes to the level of decay, and are the high and low recurrence coefficients. Consideran input video edge of M× N pixels, the span of the refined ALL and ALH sub groups of luminance segment is M/4 × N/4 the extent of the removed ALL sub band of Cr segment is M/6 × N/6. Fig. 5 shows the yield of this stage.

B. Stage Two: Face Localization

In the second stage, we will talk about in detail how the shading is utilized to spot whether the wavelet coefficients of the photo are a face hopeful or not. Hues are typically repelled into luminance and chrominance parts in video. In our paper, we initially distil the ALL sub band of the Cr segment that conveys the skin shading data. Second, we incite an edge an incentive for the Cr part in wavelet space and utilize it to sift through the non-skin territory. Third, to maintain a strategic distance from wrong face district discernment, basic morphological
operations are utilized to fill in any little gap in the facial range and to evacuate any little protest out of sight region.

To actuate a limit an incentive for the Cr segment for skin shading in wavelet space, we have analyzed the wavelet coefficients esteem in the Cr of the ALL wavelet sub band and we found that there is an edge esteem that can differentiate the face locale and the foundation. Fig. 5(a) demonstrates the information video edge of a subject shows up in non-perplexing and complex foundation. Fig. 5 shows the yield of this stage.

**Figure 5** (a) Input video frames: non-complex background with normal lighting (up) and complex background with illumination (down). (b) ALL wavelet sub band of Cr. (c) ALL wavelet sub band of luminance. (d) ALH wavelet sub band of luminance.

To display the impact of thresholding, we perform color segmentation on the info video outline appeared in Fig. 5(a), and the parallel guide created before thresholding as indicated in Fig. 6(a) and in the wake of thresholding as appeared in Fig. 6(b).

As should be obvious from Fig. 6(b), the parallel guide of the input video outline brags different items that are not the face region including foundation questions and shirt. Additionally, the color segmentation may order the eyes as non-skin shading and resulting openings in the face area. Chai and Ngan proposed the morphological strides to expel the clamor in the density map. In our paper, we do the comparative morphological operations, for example, disintegration to expel the little protests and dilation to top off any little openings in the facial area. The result of this stage is shown in Fig. 6(c).

**C. Stage Three: Face Segmentation**
This stage section the face district of from the foundation utilizing the learning based strategy. This strategy is determined utilizing some basic imperatives of the area of the face district in the video outlines. We utilize rectangles with certain angle proportion and we found that the viewpoint proportion of the limited rectangles of 1:2 most appropriate for video outlines tried.
Since in the vast majority of the face acknowledgment frameworks, the face generally situated at the focal point of the handled video outline. Subsequently, in the proposed calculation, we look for confront area which is situated at the center of the video outline. The face confinement limitations that we use to restrict the face district are: 1) the viewpoint proportions of the jumping rectangles ought to be in a scope of 1:2, and 2) the area of the side-limits of the rectangles ought to be constrained from N/4 to 3N/4 and the upper bound and the lower bound ought to be restricted from the 1 to M.

Fig. 6(c) shows that O contains other background region. This non-face region will be included in the rectangles in the face. Therefore, we apply a checking to the sides bound of the rectangles to make sure that the rectangles only include face regions which the horizontal size of a single region is more than 5 pixels. The rectangles are then rescaled to the size of the ALL which is M/4 × N/4 and mapped to the ALL and ALH sub-bands.

D. Stage Four: Eyes and Mouth Localization
The output facial region of stage three comprises the facial region of the subject. However, the face is not in a standard size. In this stage, the eyes and mouth of the output of stage three will be localized and the face will be normalized to a standard size based on the eyes and mouth locations. In the proposed system to incorporate the intensity information to locate the eyes and mouth. The grayscale erosion operation in first applied on the image then the mean gray level value of every row of the image is computed. The two minimum points of the y-axis of the image are positioned as the y-position of eyes and mouth. We denote the y-position of eyes and mouth found by the intensity method as \( Y_{Eye} \) and \( Y_{Mouth} \) respectively. The x-positions of left and right eyes are positioned by averaging every column of the image until \( X_{EyeRight} \). The two minimum positions equivalent to the left and right eyes at x-positions are denoted \( X_{EyeLeft} \) and \( X_{EyeRight} \).

In this observation, the segmented image from stage three comprises not only face but also neck and collar. The nose and the chin may sometime be perceived as the \( Y_{Mouth} \). Since in audio-visual recognition system, the subject is required to utter speeches for audio recognition and the mouth is predicted to move throughout the video. Hence, we propose to incorporate motion information with the guideline from \( Y_{Mouth} \) to find a more accurate mouth position.

4. RESULTS AND DISCUSSIONS
In this section, to evaluate the performance of the proposed system with different audio-visual databases. The experiments comprised in this section are as follows.

- Evaluating and comparing the face recognition performance and matching time of the proposed system with the conventional face recognition systems.
- Evaluating the face segmentation performance of the proposed system.
- Evaluating the recognition accuracy of the proposed audio-visual recognition system.

A. Face Recognition Performance of the Proposed System
The face recognition accuracy of the proposed system is tested on the face datasets of the UNMC-VIER audiovisual database that boast facial expression variation (Set 1) and illumination variations (Set 2) and the face dataset of the CUAVE database. The total number of subjects is 123 and 36 in UNMC-VIER and CUAVE databases, respectively. Fig. 7 reveals the sample snapshots images taken from the videos in CUAVE for (a) training samples and (b) testing samples. The original video size of UNMC-VIER and CUAVE database is 704×576 and 720×480, respectively. The videos are compressed and encoded to the Dirac elementary stream.
and the wavelet coefficients equivalent to the ALL and ALH sub bands of luminance component and ALL of Cr component will be distilled after the inverse quantization of the bit stream. Since the common intermediate format (CIF) video format (352×288) is used and after two-level of wavelet decomposition, the size of each of the distilled ALL and ALH wavelet sub band of luminance component is 88 × 72.

![Figure 7 Sample snapshots of CUAVE database in (a) training dataset and (b) testing dataset.](image)

For the face recognition of our proposed system, the dual optimal multiband feature method is used as the face recognition algorithm and the RBF neural network is used as the classifier. The width of the neural network is chosen to be twice the mean distance between the training samples. The number of hidden neurons and regularization parameter are preferred based on the best recognition accuracy in the cross-validation experiments. The width and regularization parameter of the neural networks for UNMC-VIER and CUAVE database is 15 and 0.01. The number of hidden neurons is 140 for UNMCVIER and 36 for CUAVE database.

The face recognition accuracy of the proposed system in compression domain is examined with the conventional face recognition systems in uncompressed domain. We randomly distil the wavelet coefficients from three reference frames of the training videos as training samples. For face recognition of the proposed system, the segmentation of the facial image is done by the proposed automatic face segmentation method.

### B. Audio-Visual Recognition Performance of the Proposed System

The recognition performance of the proposed audio-visual recognition system is evaluated in the UNMC-VIER, CUAVE and XM2VTS databases in this experiment. The settings of the UNMC-VIER and CUAVE are same with Section IV-A. While for the XM2VTS database, 180 subjects from session 1 and session 2 are used for training and testing, respectively. The recognition accuracy of the speaker recognition system using audio signal is tested in this experiment. The audio streams are decoded and the audio features are extracted by the MFCC and LDA and classified by the RBF neural network. The regularization parameters of the neural networks in the three databases tested in the speaker recognition are 0.01. The width of the hidden neurons is 5 for UNMC-VIER, 1 for CUAVE, and 15 for XM2VTS database. The number of hidden neurons is 123 for UNMC-VIER, 36 for CUAVE, and 190 for XM2VTS database. The white Gaussian noise of signal-to-noise ratio of 10 is added to the testing signal and the recognition rate of the speaker recognition is 90.2% in UNMC-VIER database, 80.6% in CUAVE database, and 72.8% in XM2VTS database.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Recognition Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNMC-VIER(Set 1)</td>
<td>98.4</td>
</tr>
<tr>
<td>UNMC-VIER(Set 2)</td>
<td>92.7</td>
</tr>
<tr>
<td>CUAVE</td>
<td>97.2</td>
</tr>
<tr>
<td>XM2VTS</td>
<td>83.3</td>
</tr>
</tbody>
</table>

Table I Recognition Rate (%) of the Proposed Audio-Visual Recognition System
The face recognition system and the speaker recognition system are then bonded to form the proposed audio-visual recognition system. The scores induced from the face recognition and speaker recognition systems are bonded by the sum-rule decision fusion with equal weights. In this experiment, the recognition rate is acquired by using five video frames since we see improvements in terms of recognition accuracy using five video frames. Table V shows that the proposed system achieves high recognition rates in the tested databases. The proposed system achieves recognition rates of 98.4% and 92.7% in the facial expression and illumination variation dataset of UNMC-VIER database, respectively. In CUAVE and XM2VTS database, the proposed system achieves a recognition rate of 97.2% and 83.3%, respectively. From the results, it can be adhered that with audio signal, the recognition accuracies of the proposed system in all the datasets are enhanced, particularly in dataset which contains illumination variations.

5. CONCLUSION
In this paper, a highly high-octane audio-visual recognition system in compression domain was tendered to address the computational complexity and visual variations problems in the system. For face recognition, the wavelet sub bands which were equable to illumination and facial expression variations problems were directly distilled after the inverse quantization. As a result, the inverse wavelet transform was omitted and the computational complexity of the system was reduced. We then presented a set of new face localization methods that localized wavelet coefficients equivalent to the facial region in the compression domain. After the face region is localized, the eyes and mouth locations were found by assimilating the intensity and motion information in the wavelet sub band. Based on the location of the eyes and mouth, the wavelet sub bands which comprise only facial features will then be normalized and segmented. Then the wavelet sub bands were further perished and fed to the dual optimal multiband feature method and RBF neural network for classification. Experimental results revealed that with lower computational complexity, the proposed system accomplished high recognition accuracy in UNMC-VIER, CUAVE, and XM2VTS audiovisual databases.

REFERENCES


