Robust Head Detection in Collaborative Learning Environments Using AM-FM Representations

Wenjing Shi[1], Marios S. Pattichis[1], Sylvia Celedón-Pattichis[2] and Carlos LópezLeiva[2]
{wshi, pattichi, sceledon, callopez}@unm.edu
[1] image and video Processing and Communications Lab (ivpcl.unm.edu)
Dept. of Electrical and Computer Engineering
University of New Mexico, United States.
[2] Dept. of Language, Literacy, and Sociocultural Studies
University of New Mexico, United States.

***Abstract*—The paper introduces the problem of robust head detection in collaborative learning environments. In such environments, the camera remains fixed while the students are allowed to sit at different parts of a table. Example challenges include the fact that students may be facing away from the camera or exposing different parts of their face to the camera. To address these issues, the paper proposes the development of two new methods based on Amplitude Modulation-Frequency Modulation (AM-FM) models. First, a combined approach based on color and FM texture is developed for robust face detection. Secondly, a combined approach based on processing the AM and FM components is developed for robust, back of the head detection. The results of the two approaches are also combined to detect all of the students sitting at each table. The robust face detector achieved 79% accuracy on a set of 1000 face image examples. The back of the head detector achieved 91% accuracy on a set of 363 test image examples.**

***Keywords*-head detection; face detection; AM-FM representations;**

I. Introduction

Head detection in general videos can be very challenging. Many of the detection challenges come from the requirements that the methods should work on images captured from different viewpoints, with varying illuminations, variable distances to the subjects, where the heads themselves may be partially occluded. Here, we restrict our attention to head detection in collaborative learning environments where students are being taught how to program.

We present different head detection examples in Fig. 1. The examples demonstrate occlusion, the need to work at different scales, hair color variation, and slight illumination variations. In our examples, the camera was fixed on a tripod and no attempt was made to track the students. Thus, some students are facing the camera while others are looking away. In terms of structural noise, the videos are also characterized by complex background motions. For example, in Fig. 1(a) and (d), we have people moving in the background.

The majority of prior research has placed substantial constraints on head positions. For example, in [1], the authors constrained viewpoint variation to the range of -90° to +90°. In [2], the authors considered methods for head detection based on depth images. In [3], the authors developed a 3D ellipsoidal model for head detection in a controlled environment. Thus, in prior research, we did not find the complex background motions and strong viewpoint variations that we consider in this paper.

(a) (b) (c) (d)

Fig. 1: Robust head detection in a collaborative learning environment. (a) Occluded face and back of the head example. (b) Tilted head with different hair color and occluded back of the head. (c) Face example. (d) Tilted heads and different distances from the camera.

It is common to use skin color for face detection as discussed in [4]. Texture based face detection is discussed in [5]. In [5], the authors used an FFT based approach and local phase quantization for face representations under varying illumination. A deep learning approach to face detection is discussed in [6]. A front-face hair dection approach is discussed in [7].

In our paper, we will consider the combined use of color models with a texture-based FM model for robust face detection. Here, we note that the FM components tend to be

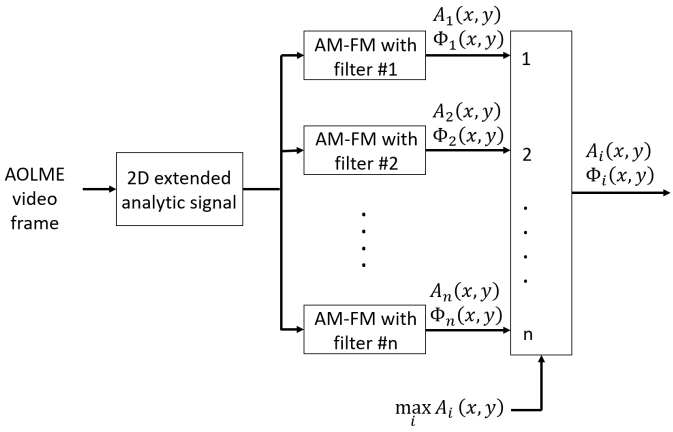


Fig. 2: AM-FM demodulation using a family of Gabor filters. Refer to [8] for the new daisy-petal filter bank design that was used.

illumination invariant, and we shall describe in the text, they can provide effective face descriptors without being affected by skin color. For the back of the head detection in the wild, there has been very little research. Here, we develop a very effective approach based on the use of AM and FM components that provide effective representations for hair strands.

The rest of the paper is divided into five sections. In section II, we summarize the AM-FM demodulation approach that is based on a new daisy-petal filter bank design described in [8].

We describe the proposed head detection algorithms in section III. We provide results in section IV and provide concluding remarks in section V.

II. AM-FM Demodulation

We begin with current AM-FM demodulation methods and provide an overview of the new filterbank that was designed for the current paper. Our approach provides an extension of the approach described in [10] and a new filter-bank design that is covered in [8] (see Fig. 2).

We begin with an introduction to AM-FM representations. Let $I(x,y)$ be expressed as a sum of AM-FM components given by:

$$I(x,y) = \sum_{n=1}^{N} A_n(x,y)\cos\left[\phi_n(x,y)\right] \tag{1}$$

where $A_n(x,y)$ represents instantaneous amplitude (IA) and $\phi_n(x,y)$ represents instantaneous phase (IP). The extended analytic signal with 2D discrete Hilbert transform is provided to estimate the IA and IP components:

$$I_{AS}(x,y) = I(x,y) + j\mathcal{H}[I(x,y)] \tag{2}$$

where:

$$\mathcal{H}[I(x,y)] = \frac{1}{\pi x} * I(x,y) \tag{3}$$

The resulting signal is processed using a new Gabor filter-bank based on 54 filters arranged in a daisy petal design. The daisy-petal design allows for tighter fitting of the 2D frequency plane by alternating the placement of the Gabor filters between consecutive orientations. The result is a better coverage of the 2D frequency plane. Unfortunately, we do not have sufficient space to fully describe the filterbank. For more detalis on the filter-bank design, refer to the M.Sc. thesis given in [8].

We approximate the output I_{n_AS} of the n-th filter using

$$I_{n\_AS}(x,y) \approx A_n(x,y)\exp\left[j\phi_n(x,y)\right]. \tag{4}$$

For each filter, we estimate the amplitude and phase components using:

$$A_n(x,y) = \left|I_{n\_AS}(x,y)\right| \tag{5}$$

and

$$\phi_n(x,y) = \arctan\left[\frac{\text{imag}\left(I_{n\_AS}(x,y)\right)}{\text{real}\left(I_{n\_AS}(x,y)\right)}\right]. \tag{6}$$

For our purposes, we adopt a dominant component analysis approach where we generate a single AM-FM component for the entire image. The basic approach is demonstrated in Fig. 2. At each pixel, we identify the channel (i) that has the maximum IA ($A_i(x,y)$). Then for this pixel, we select the IP and IA outputs from the n-th channel. The collection of all the outputs from all the pixels forms a single AM-FM component: $a(x,y)\cos\phi(x,y)$. In what follows, we will use the AM ($a(x,y)$) and FM ($\cos\phi(x,y)$) components to build robust detectors.

III. Head Detection

We decompose the problem of head detection as two separate sub-problems. First, we develop a robust face detector based on both color and FM texture information. Second, we develop a robust back of the head detector based on AM-FM texture information. Then, we combine the results from the two detectors to detect all of the heads present in the videos. In what follows, we provide more details for each detector.

A. Face Detection

We summarize the basic system in Fig. 3. Here, we consider the application of two different approaches and combine the results using an `and` operation. Thus, we require agreement from both detectors to detect a face. Following detection, we use cross-correlation to track the faces through the video for the next ten frames. After ten frames, we repeat the face detection process.

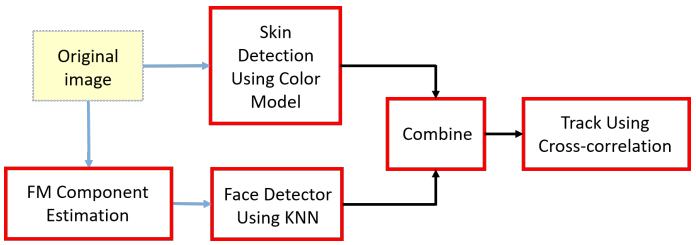


Fig. 3: Robust face detection based on color and FM texture information.

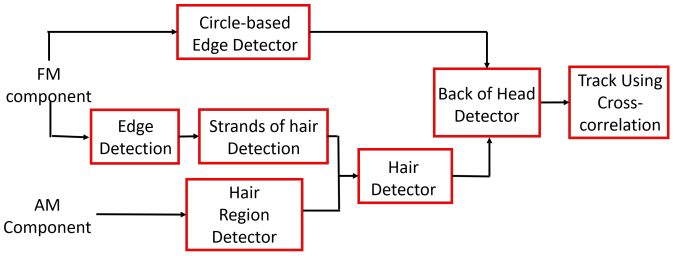


Fig. 4: Robust back of the head detection using an AM-FM model

To recognize the effectiveness of the proposed FM approach, we refer to the FM component in Fig. 5(a). From Fig. 5(a), it is clear that the FM component captured the face characteristics extremely well. The eyes, hair, mouth, and face outline is nicely outlined. Furthermore, note that the FM component does not suffer from any illumination variations, and it is not affected by skin color. Instead, it simply extracts the essential face characteristics. Furthermore, it is quite clear that the FM outline captures sufficient detail to identify the student.

We thus use a simple K-Nearest Neighbors (KNN) template for robust face detection. Initially, the input image is partitioned into 60×60 blocks with 50% overlap. Then each block is classified as a face or non-face block using a KNN ($K = 3$) classifier. The KNN classifier was trained using 2441 face-present blocks and 2410 non-face blocks. The faces were extracted from training videos that were not used during testing videos. Here, for robust detection, we required that the testing videos were selected from different days that were not shared with the training days.

For the color model, we adopted the approach described in [4]. In [4], the authors claimed that their model can address daylight illumination variations. The basic idea is to use thresholds and constrain the component relationships in the HSV, RGB, and YCrCb spaces and then combine all of them using an `and` operation.

Overall, we expect the combination of texture (FM) and color information to lead to more robust face detections. As described above, the FM approach is strongly invariant to illumination variations.

B. Back of the Head Detection

Robust detection of the back of the head requires a more sophisticated approach. We focus on development of a robust method for hair detection that is based on the AM and FM components.

To introduce the approach, we note that the FM components capture the directional information in the hair and head shape information as it is clearly visible in Figs. 5(a) and 6(b). On the other hand, the AM component also captures the hair region variations as shown in Fig. 6(a).

We present the overall method in Fig. 4. To explain the steps, we also present outputs from several steps in Figs. 5 and 6. We begin with FM-based hair detection that is shown

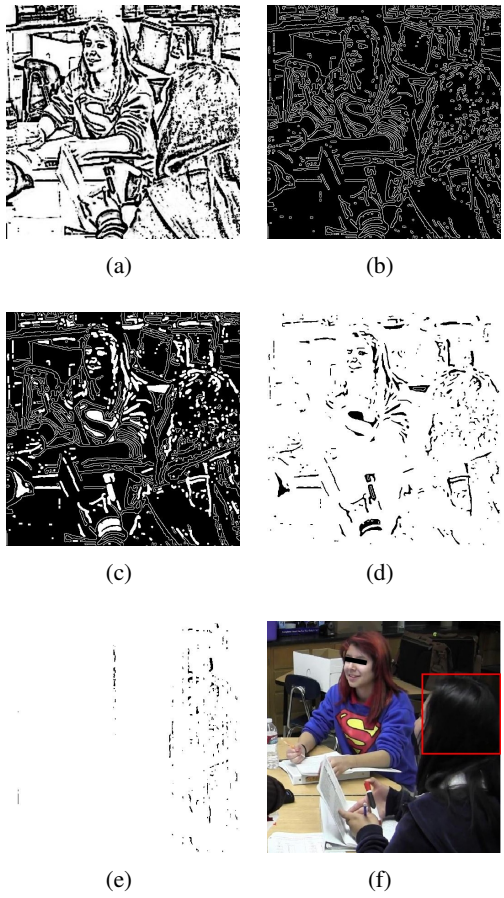
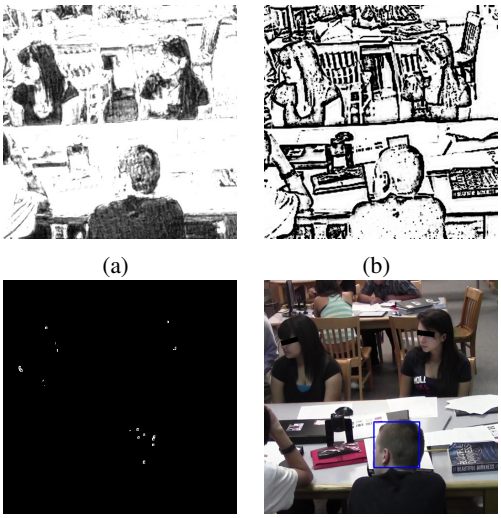


Fig. 5: Back of the head detection by using filled FM components. This method is applicable to long hair detection. (a) FM component image. (b) Edge detection on binarized FM generates closed components. (c) Strands of hair detection by filling detected components. (d) Strands of hair detection made easier to visualize but showing filled components shown as black regions. (e) Strands of hair detection by selecting the columns that correspond to the presence of filled regions. (f) Final detection result.

in Fig. 5. Here, the basic idea is to identify the long vertical components that correspond to the long hair strands. The FM image is thresholded to provide a clear image that is fed to a Canny edge detection to produce closed components (see Fig. 5(b)). The components are filled as shown in Figs. 5(c) and (d). To detect the long hair, we then sum-up the columns (project along the columns) and look for the columns with the maximum number of detected strands of hair pixels as shown in Fig. 5(e). Similarly, the AM component is used to detect strands of hair as illustrated in Fig. 6(c). The hair detector block combines the results from the AM and FM components using an `and` operation.

The well-defined head outlines of the FM components (e.g., see Figs. 5(a) and 6(b)) allow us to use a simple circle-based edge detector to detect the head shape. Our circle detector was implemented using a Hough transform operating on the



(a) (b) (c) (d)

Fig. 6: Back of the head detection using AM and FM components. This method is more applicable to short hair detection. (a) AM component. (b) FM component. (c) Strands of hair region detector. (d) Final detection result.

gradient of the FM component (see Fig. 4). For the last step, the back of the head of the head detector, we combine the results from the circle detector and the hair detector using an `and` operation and then use a sliding window to identify the region with the larger number of detected hair pixels (see Figs. 5(f) and 6(d)). Lastly, as for the face detector, we use a cross-correlation to track the detected heads through the video.

IV. Results

To test the method, we use manual annotation to label difficult examples from ten new videos, taken on different dates (512×512 at $25 fps$). A summary of our test set and the achieved accuracy is given in Table I. Overall, we tested 1363 examples and achieved 82% accuracy. Here, for accurate detection, we require that the detected region has a minimum of 60% overlap with the manual annotation.

We note that the use of the AM-FM models performed extremely well in detecting the back of the heads. We report our results separately for short hair versus long hair. Here, we note that heads with short hair was best detected using the circle-shape detector operating on the FM component. Also, for heads with longer hair, the AM-FM detector is at the bottom part of the flow-chart in Fig. 4. For short-hair, our head detection achieved 100% accuracy. Thus, we would like to dramatically extend our short-hair dataset to further challenge our approach. Similarly, for long-hair, head detection achieved 91% detection that is also good. Overall, please recall that there is very little research on back of the head detection datasets for us to compare these results. Despite the challenges, the robust face detector achieved 79% accuracy. On an i5-7200 CPU @2.50 GHz, the code requires $2.5s$ per frame for computing the AM-FM decomposition (in C++) and an additional $2s$ for the rest of the analysis (done in Python). Refer to [9] for our current efforts to speed up 2D convolutions using FPGAs.

TABLE I: Head Detection Results on Test Images Extracted from Different Videos

	Test Examples	**Accuracy**
Face	$1000 = 500 \times 2$ faces each	79%
Back of Head (long hair)	263	86%
Back of Head (short hair)	100	100%
Back of Head (all)	363	91%
Head (face + back of head)	1363	82%

V. Conclusion and Future Work

The paper proposes the use of new methods for head detection based on the combined use of AM-FM texture models and standard color models for skin detection. The proposed methods gave excellent results in a very challenging dataset composed of students learning in a collaborative environment. We are currently investigating the performance of the methodology in much larger datasets with more complex examples.

VI. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1613637 and Grant No. CNS-1422031.

References

[1] Y. Ishii, H. Hongo, K. Yamamoto, and Y. Niwa, "Real-time face and head detection using four directional features," in *Proc. of Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004.* IEEE, 2004, pp. 403–408.

[2] D. Ballotta, G. Borghi, R. Vezzani, and R. Cucchiara, "Head detection with depth images in the wild," *arXiv preprint arXiv:1707.06786*, 2017.

[3] S. S. Ghidary, Y. Nakata, T. Takamori, and M. Hattori, "Head and face detection at indoor environment by home robot," in *Proceedings of ICEE200*, 2000.

[4] N. A. bin Abdul Rahman, K. C. Wei, and J. See, "Rgb-h-cbcr skin colour model for human face detection," *Faculty of Information Technology, Multimedia University*, vol. 4, 2007.

[5] M. Dahmane and L. Gagnon, "Local phase-context for face recognition under varying conditions," *Procedia Computer Science*, vol. 39, pp. 12–19, 2014.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[7] P. Julian, C. Dehais, F. Lauze, V. Charvillat, A. Bartoli, and A. Choukroun, "Automatic hair detection in the wild," in *20th International Conference on Pattern Recognition (ICPR), 2010.* IEEE, 2010, pp. 4617–4620.

[8] W. Shi, "Human attention detection using am-fm representations," 2016.

[9] C. Carranza, D. Llamocca, and M. Pattichis, "Fast 2d convolutions and cross-correlations using scalable architectures," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2230–2245, 2017.

[10] V. Murray, P. Rodríguez, and M. S. Pattichis, "Multiscale am-fm demodulation and image reconstruction methods with improved accuracy," *IEEE transactions on image processing*, vol. 19, no. 5, pp. 1138–1152, 2010.