# Multimodal Age-Group Recognition for Opinion Video Logs using Ensemble of Neural Networks

Sadam Al-Azani[1], El-Sayed M. El-Alfy[2]
College of Computer Sciences and Engineering,
King Fahd University of Petroleum and Minerals
Dhahran 31261, Kingdom of Saudi Arabia

***Abstract*—With the wide spread usage of smartphones and social media platforms, video logging is gaining an increasing popularity, especially after the advent of YouTube in 2005 with hundred millions of views per day. It has attracted interest of many people with immense emerging applications, e.g. filmmakers, journalists, product advertisers, entrepreneurs, educators and many others. Nowadays, people express and share their opinions online on various daily issues using different forms of content including texts, audios, images and videos. This study presents a multimodal approach for recognizing the speaker's age group from social media videos. Several structures of Artificial Neural Networks (ANNs) are presented and evaluated using standalone modalities. Moreover, a two-stage ensemble network is proposed to combine multiple modalities. In addition, a corpus of videos has been collected and prepared for multimodal age-group recognition with focus on Arabic language speakers. The experimental results demonstrated that combining different modalities can mitigate the limitations of unimodal recognition systems and lead to significant improvements in the results.**

***Keywords*—Multimodal recognition; opinion mining; age groups; word embedding; acoustic features; visual features; information fusion; ensemble learning; Arabic speakers**

## I. Introduction

Due to the increasing adoption of mobile and web technologies, people tend to share their opinions and interact online on various aspects of their lives through a variety of social media platforms and websites, e.g. reviewing products, rating movies, or evaluating services [1]–[3]. Examples of major social media platforms and blogging websites include Twitter, Facebook, Google+, Instagram, Pinterest and LinkedIn. Over the past years, there has been a growing interest in social media analysis ranging from simple stats dashboards, to more advanced sentiment analysis and topic trending, to more incredible recommendation systems. The aim is transformation of available raw data into insightful information of relations and content to support decision making and guide strategic planning. This plays important role in business intelligence to set up plans and strategies to leverage marketing campaigns and enhance customer satisfaction.

Most of the state-of-the-art techniques for sentiment analysis have focused on textual data analysis of people's comments or feedback. Sentiment analysis is concerned with analyzing, evaluating and understanding the opinions, attitudes, appraisals towards different entities, aspects or features [4]. These techniques are based on using natural language processing, text mining, and computational linguistics to identify subjective information, determine opinions polarities (e.g. positive or negative) or affective states (e.g. happiness, sadness, fear, anger, surprise or disgust) for a given text, recognize sentiments on different aspects of a product, etc. Lots of work have been carried out in this regard. For a rigorous survey on sentiment analysis, we refer interested reader to [5], which reviews over one hundred articles published from 2002 to 2015 and organizes them based on tasks, approaches and applications. As Twitter has been one of the most prevalent microblogging platforms, several studies have focused on sentiment analysis of tweets. In [6], the authors presented a recent survey of algorithms and sentiment related tasks for Twitter such as tracking sentiments over time, irony detection, emotion detection, and tweet sentiment quantification.

Due to the limitations and challenges facing textual based sentiment analysis, researchers have been more recently attracted to consider other sources of information that are becoming more popular in social media such as voice, images and videos. For instance, there has been a great interest in information fusion for affective computing utilizing more than one modality or information channel [7]. Several factors can negatively affect the unimodal analysis and recognition systems, including noisy sensor data, non-universality, and lack of individuality. Each modality has its own challenges. For example, recognition systems based on voice might be affected by different attributes such as low voice quality, background noise and disposition of voice-recording devices. Text-based recognition systems also suffer from several issues related to morphological analysis, multi-dialects, ambiguity, temporal dependency, domain dependency, etc. This is also true regrading recognition systems based on visual modality, which can also suffer from illumination conditions, posture, cosmetics, resolution, etc. In consequence, this leads to inaccurate and insufficient representation of patterns. So, incorporating different modalities for an entity can overcome such issues because each source of information can replenish each other. This might result in developing more accurate and robust recognition systems. It provides several evidences for the same identity which can lead to significantly improving the performance as compared to unimodal systems.

Several approaches, resources and techniques have been provided, designed and conducted to address sentiment analysis. Current sentiment and opinion mining based approaches evaluate peoples opinions in different analysis levels including document, sentence and aspect/ feature using different approaches: lexicon-based, machine learning based and hybrid based approaches. However, they don't take into account the impact of users' age on the analyzed opinions. Multimodal age recognition systems can be beneficial in such applications to tune the analysis and decisions towards particular age groups in order to meet their needs. For example, some products are specific for young people and reviews on these products by elder people are biased and may provide incorrect indicators for decision makers. Governments can also benefit from these systems to explore political decisions or services related to their citizens according to their age groups. Adaptive educational systems will be smarter when they consider age of the learner alongside the emotion.

Detecting users' age-groups through emotional modalities makes the problem more interesting and significant, especially with the revolution in social media platforms nowadays. Several social media platforms are being used to support opinion videos such as YouTube, Vimeo, Twitter, Facebook, Instagram, Flickr, etc. Thus, it is highly important to exploit such data for mining significant information and insights. Profiling user identification such as recognizing age group from emotional modalities is a challenging task because it relies on several attributes that are hard to model such as feeling, thought, behavior, mood, temperament [8]. The research of user profiling identification and detection for Arabic language is even more scarce [9], [10]. This is another motivation for this study to build a dataset of multimodal age-group identification for Arabic opinion videos and present a multimodal age-group identification system. To our knowledge this is the first study to present a multimodal age-group identification approach specific from opinion videos. Additionally, it is the first study to present a multimodal for Arabic videos in this concern, in general. It evaluates the capability of audio, textual and visual features individually to detect age-group for the same entity. Then, it presents an ensemble neural network method to fuse different modalities in order to improve the performance of the individual modalities. Several experiments are conducted to evaluate the proposed approach.

The rest of the paper is organized as follows. The most related work is briefly reviewed in Section II. Section III describes the methodology. Section IV presents the experimental work and results. Finally, Section V concludes the paper.

## II. Related Work

Age identification is considered as a task of user profiling detection and has received a growing attention in social media and human-computer-interaction systems with rising need for personalized, reliable, and secure systems. In the literature, this problem is addressed in a variety of ways. Some research work considered it as a classification problem to predict age group of a given user, e.g. [11], whereas others addressed it as a regression problem to predict the age in years, e.g. [12], [13]. Most of existing methods have mainly focused on single modalities or single mediums including texts [14], [15], images [16]–[19], voice/speech [11], [20], and meta-data of users on Twitter [21].

Safavi et al. [11] presented a method to detect age-group from children's speech using the OGI Kids dataset. They applied Gaussian Mixture Model-Universal Background Model (GMM-UBM), and Gaussian Mixture-Support Vector Machine (GMM-SVM) with i-vector systems. Regions of the spectrum containing important age information for children are identified by conducting Age-ID experiments over 21 frequency subbands. The main findings were the GMM-UBM and i-vector system significantly performed better than the GMM-SVM system with an accuracy of 85.77%. An approach for age estimation from telephone speech patterns based on i-vectors was also presented in [13]. Each utterance was represented by i-vector and Support Vector Regression (SVR) is applied to estimate the age of speakers. Bocklet et al. [20] present a method to detect a person's age and gender from his\her voice. As acoustic features, they applied Mel Frequency Cepstrum Coefficients (MFCCs), Perceptual Linear Prediction (PLPs) and Temporal Patterns (TRAPS)-based features. Different models were generated and combined at feature level and score level fusion and evaluated using GMM. They reported that combining different acoustic models led to improve the results with minor differences between feature level and score level fusions.

Different age categories were considered in the literature to address the age recognition task. For example, the authors in [14] ran their experiments on three categories: 10s (13-17), 20s (23-27) and 30s (33-42) years old. In Safavi et al. [11], three age groups are considered: (5-9 years old), (9-13 years old), (13-16 years old).

Multimodal recognition systems are still in their early stage and started applying for different tasks as gender detection [22], sentiment analysis [7], [23]. Our work differs from the literature in several aspects. First, it recognizes age-groups from three modalities for the same user and compares the effectiveness of these modalities with each others. It explores different features for representing modalities such as word embedding based features for textual modality, dense optical flows for visual modality and a combination of several types of acoustic features. It builds a corpus for opinions videos of Arabic speakers. Moreover, it explores a novel ensemble of a neural network approach to combine different modalities.

## III. Methodology

In this study, the age-group recognition task is addressed as a classification problem. This can be useful in applications such as targeted marketing which is directed to certain age groups rather than specific ages. For example, companies can tune their products to meet the needs of a specific age-group of people. Fig. 1 depicts the general framework of the proposed multimodal age-group recognition system. Some preprocessing operations are conducted to come up with three modalities for each video: audio, text, and visual. Each audio input is in WAV format with 256 bits, 48000Hz sampling frequency, and a mono channel. This is followed by the transcription task to generate texts corresponding to each video. Each video input is then resized into 240 × 320 after detecting faces. A feature extractor is constructed for each input source. The acoustic feature extractor constructs feature vectors of 68 features for each input. Moreover, a textual feature extractor is implemented to extract textual features with a dimensionality of 300 features for each instance. The visual feature extractor generates 800 features for each input.

A fusion method based on ensemble neural network is proposed to combine the different modalities. It is based on two levels; the first level is trained using the training dataset and gives a score for each age group from each modality (visual, text and audio). The resulting scores from the first

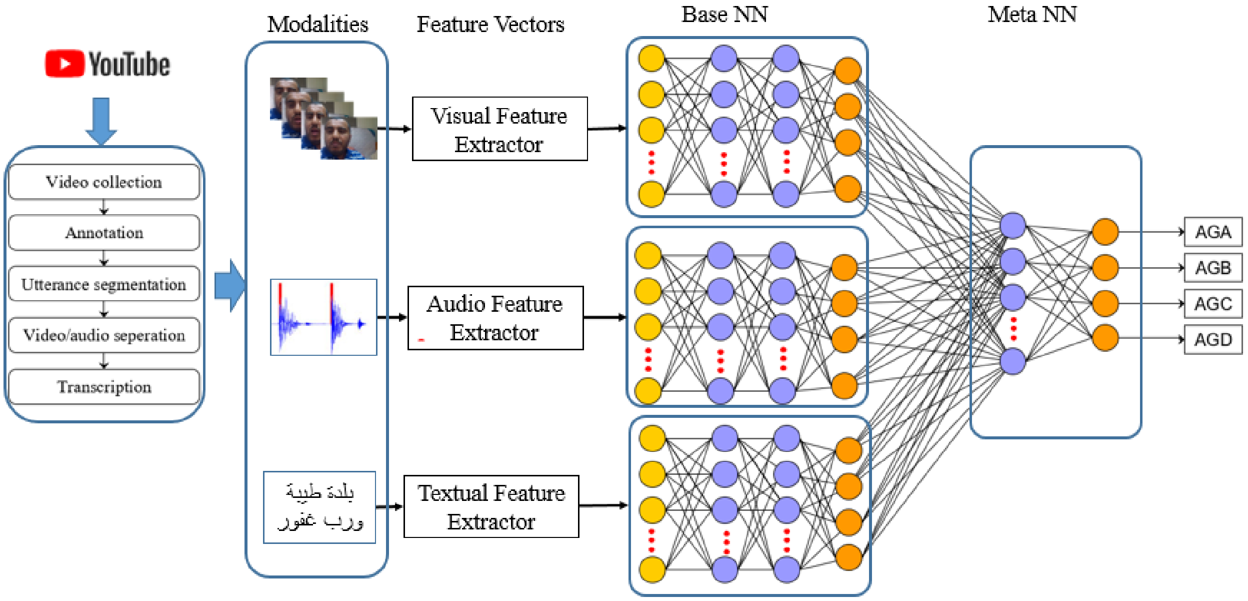

Fig. 1. Multimodal age identification system from opinion videos

level are combined using a meta-learner in the second level to produce the final scores. The predicted age group is determined corresponding to the maximum final score.

### A. Multimodal Age-groups Recognition Dataset

A video corpus is collected from YouTube. It is composed of 63 opinion videos expressed from both females and males in different domains including reviews of products, movies, cultural views, etc. Using various settings, the collected videos were recorded by users in real environments including houses, studios, offices, cars or outdoors. Users express their opinions in different periods. The videos are segmented into 524 utterances. The age-groups are specified as four classes as described in Table I. The instances are manually labeled into the considered age-groups carefully and systematically. First, for the well-known speakers, we looked for their ages in their profiles and assigned their age by subtracting date of recording videos from their birthdays. For the remaining speakers who we couldn't find their birthdays, three human annotators were involved to assign their age-group labels, using majority votes to break ties.

### B. Feature Extraction

*1) Acoustic features extraction:* The input audio is split into frames with size of 50 millisecond with a frame step of 20 millisecond. For each generated frame, a set of 34 features are computed: (1) ZCR (Zero Crossing Rate), (2)Energy, (3) Entropy of Energy, (4)Spectral Centroid, (5)Spectral Spread, (6) Spectral Entropy, (7) Spectral Flux, (8) Spectral Rolloff, (9-21) MFCCs (22-33) Chroma Vector, and (34) Chroma Deviation. Then statistics are computed from each audio's segment to represent the whole audio using one descriptor; in our study we used the mean and standard deviation. Thus, each input audio is represented by $34 \times 2 = 68$ features. This process of audio feature extraction is illustrated in Fig. 2.

*2) Visual feature extraction:* Two main steps are involved: face detection and visual feature extraction. The general frontal face and eye detectors [24] are utilized to detect the face of the speaker and segment faces from the rest of given frame based on HAAR features [25], which increasingly combines more complex classifiers in a cascade to detect the face. In addition, an eye detector detects eye positions which provide significant and useful values to crop and scale the frontal face to a size of $240 \times 320$ pixels in our case.

Then, optical flow is considered to extract the visual features from the videos processed in the previous step. Optical flows are, first, computed for each frame in a video and then used to compute histograms. They measure the motion relative to an observer between two frames at each point of them. At each point in the scene, the magnitude and the direction values are obtained which describe the vector representing the motion between the two frames. This leads to $NoF \times W \times H \times 2$ dimensions to describe each video, where $NoF$ represents the number of frames in a video and the $W \times H$ represents the resolution of the frame. In our case frames are scaled into the resolution of $240 \times 320$. To describe each video as a single feature vector (descriptor), a histogram of the optical flows per video is calculated. The scene is split into a grid of $10 \times 10$ with considering eight directions: $\{0-45, 46-90, 91-135, 136-180, 181-225, 226-270, 271-315, 316-360\}$. Consequently, each scene is represented by 800 features and to represent the whole input video the average of the histograms is calculated. The face detection and visual feature extraction is illustrated in Fig. 3.

*3) Textual features:* The word embedding technique skip-grams word2vec [26], [27] is employed to extract textual features. Embedding techniques are recognized as an efficient method for learning high-quality vector representations of words/terms/phrases from large amounts of unstructured text data. They refer to the process of mapping words, terms or phrases from the vocabulary to real-valued vectors such that

TABLE I. DESCRIPTION OF DATASET

| Class | Description | #Samples |
|---|---|---|
| AGA | Age-group A: 15-29 years old | 128 |
| AGB | Age-group B: 30-39 years old | 159 |
| AGC | Age-group C: 40-49 years old | 142 |
| AGD | Age-group D: greater than 49 years old | 95 |
| Total | | 524 |

TABLE II. THE PERFORMANCE OF STANDALONE MODALITIES USING DIFFERENT NETWORKS STRUCTURES; HIGHEST RESULTS ARE MARKED IN BOLD.

| | Structure# NN1 | | | Structure#NN2 | | | Structure# NN3 | | | Structure# NN4 | | | Structure# NN5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Prc$ | $Rec$ | $F_1$ | $Prc$ | $Rec$ | $F_1$ | $Prc$ | $Rec$ | $F_1$ | $Prc$ | $Rec$ | $F_1$ | $Prc$ | $Rec$ | $F_1$ |
| Audio (A) | | | | | | | | | | | | | | | |
| AGA | 0.8235 | 0.8750 | 0.8485 | 0.8077 | 0.8203 | 0.8140 | 0.8400 | 0.8203 | 0.8300 | 0.7197 | 0.7422 | 0.7308 | 0.8739 | 0.8125 | 0.8421 |
| AGB | 0.8373 | 0.8742 | 0.8554 | 0.8137 | 0.8239 | 0.8188 | 0.8012 | 0.8616 | 0.8303 | 0.7529 | 0.8050 | 0.7781 | 0.7989 | 0.8994 | 0.8462 |
| AGC | 0.8593 | 0.8169 | 0.8375 | 0.8345 | 0.8169 | 0.8256 | 0.8085 | 0.8028 | 0.8057 | 0.7879 | 0.7324 | 0.7591 | 0.8667 | 0.8239 | 0.8448 |
| AGD | 0.9080 | 0.8316 | 0.8681 | 0.8404 | 0.8316 | 0.8360 | 0.8851 | 0.8105 | 0.8462 | 0.7889 | 0.7474 | 0.7676 | 0.8791 | 0.8421 | 0.8602 |
| Weighted-Avg | 0.8527 | 0.8511 | **0.8512** | 0.8227 | 0.8225 | 0.8226 | 0.8279 | 0.8263 | 0.8264 | 0.7608 | 0.7595 | 0.7595 | 0.8501 | 0.8473 | 0.8473 |
| Macro-Avg | 0.8570 | 0.8494 | **0.8524** | 0.8241 | 0.8232 | 0.8236 | 0.8337 | 0.8238 | 0.8280 | 0.7624 | 0.7567 | 0.7589 | 0.8547 | 0.8445 | 0.8483 |
| Textual (T) | | | | | | | | | | | | | | | |
| AGA | 0.6695 | 0.6172 | 0.6423 | 0.5549 | 0.7500 | 0.6379 | 0.5923 | 0.6016 | 0.5969 | 0.6316 | 0.5625 | 0.5950 | 0.5299 | 0.5547 | 0.5420 |
| AGB | 0.5054 | 0.5912 | 0.5449 | 0.5105 | 0.4591 | 0.4834 | 0.5290 | 0.5157 | 0.5223 | 0.4945 | 0.5660 | 0.5279 | 0.4535 | 0.4906 | 0.4713 |
| AGC | 0.5217 | 0.5070 | 0.5143 | 0.4880 | 0.4296 | 0.4569 | 0.4753 | 0.5423 | 0.5066 | 0.5093 | 0.5775 | 0.5413 | 0.4631 | 0.4859 | 0.4742 |
| AGD | 0.6463 | 0.5579 | 0.5989 | 0.6265 | 0.5474 | 0.5843 | 0.6234 | 0.5053 | 0.5581 | 0.6866 | 0.4842 | 0.5679 | 0.6087 | 0.4421 | 0.5122 |
| Weighted-Avg | 0.5755 | 0.5687 | **0.5702** | 0.5363 | 0.5382 | 0.5323 | 0.5470 | 0.5420 | 0.5428 | 0.5668 | 0.5534 | 0.5552 | 0.5029 | 0.4962 | 0.4968 |
| Macro-Avg | 0.5857 | 0.5683 | **0.5751** | 0.5450 | 0.5465 | 0.5406 | 0.5550 | 0.5412 | 0.5460 | 0.5805 | 0.5476 | 0.5580 | 0.5138 | 0.4933 | 0.4999 |
| Visual (V) | | | | | | | | | | | | | | | |
| AGA | 0.5827 | 0.6328 | 0.6067 | 0.6419 | 0.7422 | 0.6884 | 0.5556 | 0.7422 | 0.6355 | 0.5567 | 0.8438 | 0.6708 | 0.6389 | 0.5391 | 0.5847 |
| AGB | 0.6562 | 0.5283 | 0.5854 | 0.6074 | 0.6226 | 0.6149 | 0.5917 | 0.4465 | 0.5090 | 0.6794 | 0.5597 | 0.6138 | 0.6200 | 0.5849 | 0.6019 |
| AGC | 0.5279 | 0.7324 | 0.6136 | 0.6452 | 0.5634 | 0.6015 | 0.6230 | 0.5352 | 0.5758 | 0.7080 | 0.5634 | 0.6275 | 0.5492 | 0.7465 | 0.6328 |
| AGD | 0.8500 | 0.5368 | 0.6581 | 0.6404 | 0.6000 | 0.6196 | 0.5946 | 0.6947 | 0.6408 | 0.6860 | 0.6211 | 0.6519 | 0.7397 | 0.5684 | 0.6429 |
| Weighted-Avg | 0.6386 | 0.6107 | 0.6114 | 0.6320 | 0.6317 | 0.6301 | 0.5919 | 0.5878 | 0.5819 | 0.6584 | 0.6412 | **0.6383** | 0.6271 | 0.6145 | 0.6135 |
| Macro-Avg | 0.6542 | 0.6076 | 0.6159 | 0.6337 | 0.6321 | 0.6311 | 0.5912 | 0.6047 | 0.5902 | 0.6575 | 0.6470 | **0.6410** | 0.6370 | 0.6097 | 0.6156 |

TABLE III. BIMODAL AND MULTIMODAL RESULTS USING ENSEMBLE CLASSIFIER

| Modalities | A-T | | | T-V | | | A-V | | | A-T-V | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Measures | $Prc$ | $Rec$ | $F_1$ | $Prc$ | $Rec$ | $F_1$ | $Prc$ | $Rec$ | $F_1$ | $Prc$ | $Rec$ | $F_1$ |
| AGA | 0.8682 | 0.8750 | 0.8716 | 0.7090 | 0.7422 | 0.7252 | 0.8923 | 0.9062 | 0.8992 | 0.9120 | 0.8906 | 0.9012 |
| AGB | 0.8580 | 0.8742 | 0.8660 | 0.6772 | 0.6730 | 0.6751 | 0.8805 | 0.8805 | 0.8805 | 0.8712 | 0.8931 | 0.8820 |
| AGC | 0.8493 | 0.8732 | 0.8611 | 0.6739 | 0.6549 | 0.6643 | 0.8750 | 0.8873 | 0.8811 | 0.8611 | 0.8732 | 0.8671 |
| AGD | 0.9080 | 0.8316 | 0.8681 | 0.7234 | 0.7158 | 0.7196 | 0.9011 | 0.8632 | 0.8817 | 0.8804 | 0.8526 | 0.8663 |
| Weighted-Avg | 0.8672 | 0.8664 | 0.8664 | 0.6924 | 0.6927 | 0.6925 | 0.8856 | 0.8855 | 0.8855 | 0.8801 | 0.8798 | 0.8798 |
| Macro-Avg | 0.8709 | 0.8635 | 0.8667 | 0.6959 | 0.6965 | 0.6960 | 0.8872 | 0.8843 | **0.8856** | 0.8812 | 0.8774 | 0.8792 |

elements with similar meaning to have a similar representation.

Word vectors are positioned in the vector space such that words sharing common contexts and having similar semantic are mapped nearby each other. Skip-grams (SG) is a neural network structure trained to predict a context given a word. Word embedding-based features have been adopted for different natural language processing tasks and achieved high results comparing to other traditional features [28]. In our study, a skip-gram model trained from opinions expressed in Twitter with a dimensionality of 300 [29] is used to derive textual features. A feature vector is generated for each sample by averaging the embeddings of that sample [30]. The main steps of textual feature extraction are shown in Fig. 4.

### C. Classification Approach

This study deals with a multimodal identification system for three modalities. Therefore, seven different main models can be generated as follows. Three models are generated for audio, textual and visual modalities. Three other models are generated for the bimodal approaches of audio-textual, textual-visual, and audio-visual modalities. The seventh model is for the trimodal of audio, textual and visual modalities.

Due to the theoretical foundation underlying neural network research and recently-achieved strong practical results on challenging problems, neural networks have recently been rediscovered as a significant alternative to several standard classification techniques [31]. However, the models need to generated well. Different systematic structures of neural networks are investigated to detect the age group from the considered standalone modality. Three models of feed-forward networks structures, Multilayer Perceptron (MLP) models, are applied. The first model is for visual modality, the second is for the audio modality while the third model is for textual modality. Several factors and decisions should be considered when configuring and setting up the neural network structures including: number of hidden layers to use in the neural network, number of neurons in each hidden layer, etc. Another issue for the multimodal approaches is: should the models be homogeneous or heterogeneous?; the former means using the same structure for each modality while the latter means using different structures. In the case of the heterogeneous models,

TABLE IV. P-VALUES FOR PAIR WISE T-TESTS ON ACCURACY

| Methods | P-Value | Conclusion |
|---|---|---|
| A vs. AT | 0.0303 | Reject H0. The accuracy rate of audio-textual modality is significantly higher than the audio modality |
| A vs. AV | 0.00005 | Reject H0. The accuracy rate of audio-visual modality is significantly higher than the audio modality |
| A vs. ATV | 0.00001 | Reject H0. The accuracy rate of audio-textual-visual modality is significantly higher than the audio modality |
| T vs. AT | 0 | Reject H0. The accuracy rate of audio-textual modality is significantly higher than the textual modality |
| T vs. TV | 0 | Reject H0. The accuracy rate of textual-visual modality is significantly higher than the textual modality |
| T vs. ATV | 0 | Reject H0. The accuracy rate of audio-textual-visual modality is significantly higher than the textual modality |
| V vs. AV | 0 | Reject H0. The accuracy rate of audio-visual modality is significantly higher than the visual modality |
| V vs. TV | 0.279427 | Accept H0. combining visual modality with textual modality has no effect comparing to visual modality |
| V vs. ATV | 0 | Reject H0. The accuracy rate of audio-textual-visual modality is significantly higher than the visual modality |
| AV vs. ATV | 0.010979 | Reject H0. The accuracy rate of audio-textual-visual modality is significantly higher than the audio-visual modality |

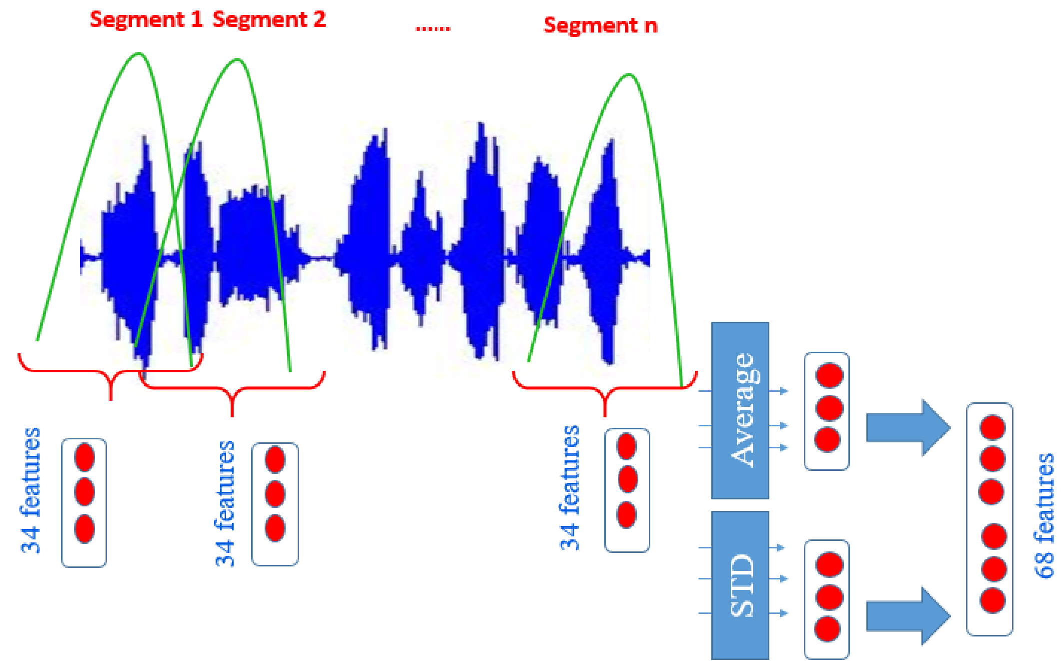

Fig. 2. Audio feature extraction process.

what are the considered attributes. In this study, two hidden layers are considered for each model while several criteria are considered to determine the number of neurons in each layer:

- Same number of neurons in each hidden layer with the same structure.
- Number of neurons is assigned according to the size of inputs for each structure. Several cases are considered including:
  - the number of neurons in a hidden layer is calculated using:

$$N_{h1} = i + o \tag{1}$$

    where $i$ is the size of input, $o$ is the number of classes
  - the number of neurons in a hidden layer is calculated using:

$$N_{h2} = \frac{i+o}{2} \tag{2}$$

  - the number of neurons in a hidden layer is calculated using:

$$N_{h3} = i \times 2 \tag{3}$$

Other parameters are selected and remain the same for all structures to be: activation function = “relu”, alpha = 0.0001, batch_size= “auto”, learning-rate = 0.001, tol = 0.0001, momentum = 0.9, epsilon=$10^{-8}$).

Consequently, five different structures are defined from the aforementioned criteria. The first structure is denoted as NN1 and uses a constant number of neurons in both hidden layers for all modalities. Since the three modalities are trained and evaluated using the same structure, this type is homogeneous. The second structure is denoted as NN2 and uses a number of neurons equals to $N_{h1}$ in the first hidden layer and equals to $N_{h2}$ for the second hidden layer. The third structure is denoted as NN3 and uses a number of neurons equals to $N_{h1}$ in both hidden layers. The fourth structure is denoted as NN4 and uses a number of hidden layers equals to $N_{h2}$ for both hidden layers. The fifth structure is denoted as NN5 and uses a number of hidden layers equals to $N_{h3}$ for both hidden layers. So, all structures except the first one (NN1) rely on the size of the input and are heterogeneous for all modalities. Those structures are considered as baseline models/classifiers.

Another MLP model is constructed as a meta-classifier to ensemble all modalities base models. In this study, a simple structure of one hidden layer is adopted in the second stage. It

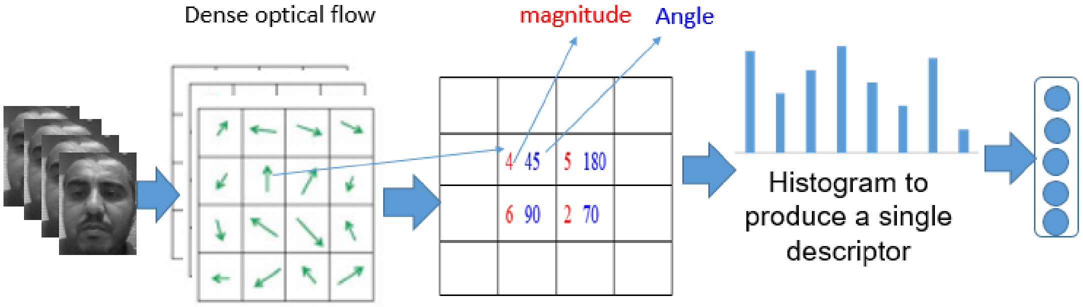

Fig. 3. Face detection and dense optical flow features extraction process.

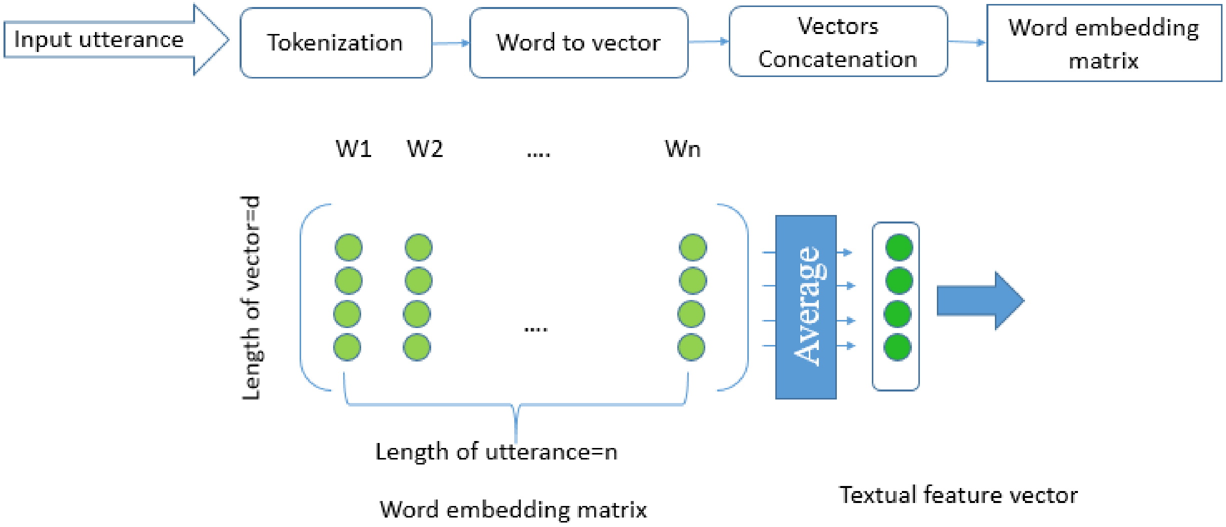

Fig. 4. Textual feature extraction process.

takes as inputs the probabilities of classes produced by the base models. Mathematically, let $Y_V = \{v_1, v_2, v_3, v_4\}$ represents the scores for age groups generated by visual-based model, $Y_A = \{a_1, a_2, a_3, a_4\}$ represents the scores for age groups generated by audio-based model, and $Y_T = \{t_1, t_2, t_3, t_4\}$ represents the scores for age groups generated by textual-based model. These scores are fed into the meta-MLP as input to identify the global age-group of users using the three modalities. So, it takes inputs with size of $M \times N$, where $M$ is the number of modalities and $N$ is the number of classes. As mentioned above, four models can be generated when combining the three different modalities: audio-textual, textual-visual, audio-visual, and audio-textual-visual modalities. In case of bimodal approaches, the size of meta-classifier input is eight while its size is 12 in case of multimodal approach.

## IV. Experiments and Results

The proposed models are evaluated using 10-fold cross validation mode. A prototype is implemented and evaluated for each standalone modality and for the ensemble model in Python using the scikit-learn machine learning package [32]. Several well-known measures are reported to evaluate and compare the performance of various models: Precision ($Prc$), Recall ($Rec$), and $F_1$, which is a weighted average for precision and recall, and is a preferred performance measure for imbalanced class distributions. These measures are computed as follows for each class $c_i$:

$$Prc_i = \frac{\text{\# instances correctly classified as class } c_i}{\text{\# instances classified as class } c_i} \quad (4)$$

$$Rec_i = \frac{\text{\# instances correctly classified as class } c_i}{\text{\# instances actually in class } c_i} \quad (5)$$

$$F_{1i} = 2 \times \frac{Prc_i \times Rec_i}{Prc_i + Rec_i} \quad (6)$$

Besides the per-class performance, we reported the weighted and macro-averages of all classes in each case. The macro-average is unweighted of each class metric without taking class imbalance into account. This measure may overemphasize the low performance of infrequent classes. Hence, we also report weighted average where each class metric is weighted by the support (i.e. the number of true instances for each class).

### A. Unimodal Age-group Recognition Results (Baseline)

Table II shows the results for each modality with different base neural network structures for unimodal age-group identification approaches. The results are presented in terms of precision, recall and $F_1$ for each class as well as weighted and macro averages for all classes. Audio modality achieves the highest results comparing to text and visual modalities in all cases. For audio modality, NN1 achieves the highest results with a weighted $F_1$ average of 85.12%, followed by NN5 which reports a weighted $F_1$ average of 84.73%. However, NN4 achieves the lowest results for audio modality. However, the overall lowest results are obtained using the textual modality. The best performance in the case of textual modality is obtained using NN1 with a weighted average of 57.02%. Regarding the visual modality, the highest results are achieved using NN4 with a weighted average of 63.83%.

### B. Bimodal and Multimodal Age-group Recognition Results

The best structures evaluated in the first level classification are then used to represent each modality and fed into the second level classifier. For audio modality and textual modality, NN1 is used while for visual modality NN4 is used. NN1 for audio modality, NN1 for text modality and NN4 for visual modality are fused using the meta-structure. Table III shows the results obtained using bimodal and multimodal approaches, The highest results are achieved using audio-visual (A-V) approach with a weighted average of 88.55% and then audio-text-visual (A-T-V) approach with a weighted average of 87.98%. It can be seen that significant improvements are reported over the baseline performance in Table II. For example, in the worst cases, the highest weighted $F_1$ obtained for baseline textual modality is 57.02% and for visual modality is 63.83% whereas after combining them the results are improved to be 69.25%. In the best cases, the highest weighted $F_1$ obtained for audio modality is 85.12% and for visual modality is 63.83% while combining them leads to improving the results to be 88.55%

It is important to perform statistical test to provide evidence that the improvement of combining different modalities is significant and not by chance. To do so, we re-run the 10-fold cross-validation 10 times for each model. We then used the pairwise t-test to determine how significant is the improvements. Table IV shows the results for the performed t-tests using 95% confidence interval. The reported p-values are less than 0.05 for all the cases except one. Thus, the null hypothesis is rejected and significant improvement is obtained except when textual modality is combined with visual modality (no statistically significant improvement is observed).

## V. Conclusion

We have presented a novel multimodal ensemble neureal network model for detecting users' age-group from opinion videos. For evaluation purpose, three modalities are extracted, namely: audio, text and visual from videos expressed in Arabic language with different dialects. Various ways are adopted to construct different neural network structures for the unimodal recognition as baseline. Then, all modalities are combined using the proposed ensemble neural network approach. For standalone modalities, the audio-based model has achieved the highest performance with the smallest number of features. However, text modality reported the lowest results. Combining different modalities has led to significant improvements in the results in nearly all cases. The highest results have been achieved using the bimodal audio-visual and trimodal audio-textual-visual approaches. As future work, the authors are exploring the impact of age knowledge in opinion mining and sentiment analysis.

## Acknowledgment

The authors would like to thank King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, for support during this work. The first author also acknowledges the scholarship provided by Thamar University, Yemen, for his higher studies.

## References

[1] Z. Xiang and U. Gretzel, "Role of social media in online travel information search," *Tourism management*, vol. 31, no. 2, pp. 179–188, 2010.

[2] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464–472, 2013.

[3] I. Lee, "Social media analytics for enterprises: Typology, methods, and processes," *Business Horizons*, vol. 61, no. 2, pp. 199–210, 2018.

[4] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[5] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[6] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, p. 28, 2016.

[7] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[8] K. Zvarevashe and O. O. Olugbara, "Gender voice recognition using random forest recursive feature elimination with gradient boosting machines," in *IEEE International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, 2018, pp. 1–6.

[9] P. Rosso, F. Rangel, I. H. Farías, L. Cagnina, W. Zaghouani, and A. Charfi, "A survey on author profiling, deception, and irony detection for the arabic language," *Language and Linguistics Compass*, vol. 12, no. 4, p. e12275, 2018.

[10] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, and B. Stein, "Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter," *Working Notes Papers of the CLEF*, 2018.

[11] S. Safavi, M. Russell, and P. Jančovič, "Identification of age-group from children's speech by computers and humans," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] T. Bocklet, A. Maier, and E. Nöth, "Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines/regression," in *Proceedings of International Conference on Text, Speech and Dialogue*, 2008, pp. 253–260.

[13] M. H. Bahari, M. McLaren, D. Van Leeuwen *et al.*, "Age estimation from telephone speech using i-vectors," 2012.

[14] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of age and gender on blogging," in *AAAI spring symposium: Computational approaches to analyzing weblogs*, vol. 6, 2006, pp. 199–205.

[15] M. Potthast, F. Rangel, M. Tschuggnall, E. Stamatatos, P. Rosso, and B. Stein, "Overview of pan'17," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2017, pp. 275–290.

[16] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 578–584, 2008.

[17] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 10–15.

[18] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2015, pp. 534–541.

[19] M. T. B. Iqbal, M. Shoyaib, B. Ryu, M. Abdullah-Al-Wadud, and O. Chae, "Directional age-primitive pattern (dapp) for human age group recognition and age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2505–2517, 2017.

[20] T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth, "Age and gender recognition based on multiple systems-early vs. late fusion," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[21] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," *PloS one*, vol. 10, no. 3, p. e0115545, 2015.

[22] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Multimodal gender detection," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 302–311.

[23] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, 2011, pp. 169–176.

[24] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[25] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001).*, vol. 1, pp. I–I.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at International Conference on Learning Representations*, 2013.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[28] S. Al-Azani and E.-S. M. El-Alfy, "Combining emojis with arabic textual features for sentiment classification," in *9th IEEE International Conference on Information and Communication Systems (ICICS)*, 2018, pp. 139–144.

[29] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "Aravec: A set of arabic word embedding models for use in arabic nlp," in *Proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing 2017)*, vol. 117. Elsevier, 2017, pp. 256–265.

[30] S. Al-Azani and E.-S. M. El-Alfy, "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text," *Procedia Computer Science*, vol. 109, pp. 359–366, 2017.

[31] C. C. Aggarwal, *Data classification: algorithms and applications*. CRC press, 2014.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.