AUDIO-VISUAL SPEECH RECOGNITION INCORPORATING FACIAL DEPTH INFORMATION CAPTURED BY THE KINECT

Georgios Galatas [1,2], *Gerasimos Potamianos* [3,1], *Fillia Makedon* [2]

[1] Institute of Informatics and Telecommunications, NCSR “Demokritos”, Athens, Greece
[2] Heracleia Lab., Dept. of Computer Science and Engin., University of Texas at Arlington, Texas, USA
[3] Dept. of Computer and Communication Engin., University of Thessaly, Volos, Greece

Emails: ggalatas@mavs.uta.edu , gpotam@ieee.org , makedon@uta.edu

ABSTRACT

We investigate the use of facial depth data of a speaking subject, captured by the Kinect device, as an additional speech-informative modality to incorporate to a traditional audio-visual automatic speech recognizer. We present our feature extraction algorithm for both visual and accompanying depth modalities, based on a discrete cosine transform of the mouth region-of-interest data, further transformed by a two-stage linear discriminant analysis projection to incorporate speech dynamics and improve classification. For automatic speech recognition utilizing the three available data streams (audio, visual, and depth), we consider both the feature and decision fusion paradigms, the latter via a state-synchronous tri-stream hidden Markov model. We report multi-speaker recognition results on a small-vocabulary task employing our recently collected bilingual audio-visual corpus with depth information, demonstrating improved recognition performance by the addition of the proposed depth stream, across a wide range of audio conditions.

Index Terms— Audio-visual automatic speech recognition, depth information, multi-sensory fusion, linear discriminant analysis, Microsoft Kinect.

1. INTRODUCTION

Introducing the visual modality to the task of automatic speech recognition (ASR) has been repeatedly shown in the literature to improve ASR accuracy and robustness to audio noise, aiming towards more natural, speech-based human-machine interaction [1, 2]. Typically the incorporated visual speech information is extracted from traditional planar video data of the speaker’s facial region, captured in the visible spectrum. Few only datasets and experimental results have been published that deviate from this paradigm, by utilizing some sort of 3D information from the speaker’s face. Such are, for example, the Australian English speech data corpus (AVOZES), where stereo cameras are used for video recording [3], the WAPUSK20 database recorded with a Bumblebee stereo camera [4], and the in-car Spanish database AV@CAR, where the subject’s image is captured from six different angles in order to reconstruct a 3D textured mesh of the speaker’s face [5]. In this paper, we also deviate from the traditional visual stream paradigm, by incorporating facial depth data, captured by a novel multimodal recording device, the Microsoft Kinect. In particular, we build on prior work [6], where we described the collection of a small-vocabulary bilingual audio-visual corpus with depth information (BAVCD) employing the Kinect, to investigate the use of such data to the problem of audio-visual automatic speech recognition (AVASR). Our approach leads to a multi-sensory, multimodal ASR system, where speech information extracted from the audio, visual, and depth data streams is fused to yield utterance transcripts. To our knowledge, this constitutes the first such effort in the AVASR literature.

An important aspect of the effort is related to extracting speech informative features from the visual and depth streams. For this purpose, various feature selection and data transformation techniques have been adopted from the literature, where various schemes have been proposed, such as, for example, the use of genetic algorithms for feature selection and principal component analysis for feature transformation [7]. In our approach, we employ appearance based features, obtained from the discrete cosine transform (DCT) of the mouth region-of-interest. A straightforward feature selection method of the resulting DCT coefficients is the use of feature energy as a measure of information content [8]. According to this technique, features with higher energy values over time are more informative, and thus their selection based on energy sorting can be effective. The process is further facilitated by the use of linear discriminant analysis (LDA) that has been observed to benefit automatic speechreading performance [8, 9]. In our work, LDA is applied both within each frame and across temporally adjacent feature frames to capture dynamic speech information discriminatively. Furthermore, this two-stage LDA is also applied on the depth data, after appropriately mapping the tracked mouth region-of-interest from the traditional video to the depth data stream.

© EURASIP, 2012 - ISSN 2076-1465

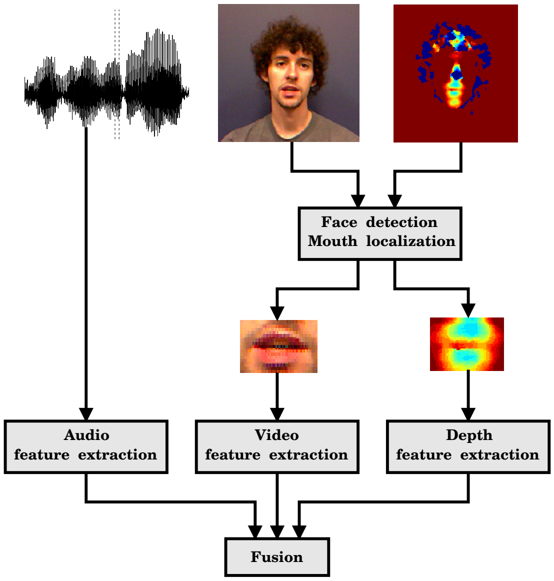

Figure 1: Proposed multi-modal ASR system overview.

The rest of the paper is structured as follows: Details about the data acquisition of the BAVCD database are presented in Section 2. In Section 3, the system architecture is described, as well as the visual front-end, statistical speech modeling, and fusion. Section 4 is devoted to our experiments, and, finally, Section 5 concludes the work.

2. THE BAVCD DATABASE

The database used in our experiments is the "Bilingual Audio Visual Corpus with Depth information" (BAVCD), consisting of connected digit utterances in both English and Greek [6]. More specifically, its English part, used here, contains approximately 4.5k connected digits from 15 speakers, and its Greek part contains approximately 2k connected digits.

The devices employed in capturing the data were a Microsoft Kinect, a Canon Vixia HF100 HD camera, and a Zoom H4 sound recorder. The MS Kinect is a novel device, mainly used to control videogames through gesture recognition. Its design is based on the PrimeSensor device [10], and it can capture both VGA resolution video, as well as depth images at the same resolution. In order to acquire the latter, the Kinect utilizes a laser, an IR camera, and the structured light methodology [11]. The effective range of the depth camera is 0.7m-6m, but its depth resolution decreases with increasing object distance. Therefore, in our experiments the Kinect was placed at approximately 0.9m from the speaker's face. The data streams captured by the Kinect were 640×480 pixel, 24-bit RGB at 20 fps color video, and 640×480 pixel, 11-bit at 20 fps depth information. The audio was captured by the Zoom sound recorder, which incorporates a pseudo X/Y condenser microphone configuration that exhibits nearly uniform directionality and flat frequency response. The device yielded two tracks of 16-bit, 44.1 kHz, PCM encoded audio. Finally, note that the recorded HD camera video stream was not used in our experiments.

All data were collected at the Vision Capture and Human Tracking Laboratory of the Computer Science and Engineering Department at the University of Texas at Arlington. The recording environment offered controlled illumination, clean acoustics, and a solid blue background, simplifying somewhat the face detection and tracking task.

3. SYSTEM ARCHITECTURE

Our system consists of the visual front-end implementing the region-of-interest (ROI) detection, the feature extraction and transformation module, and the statistical ASR module for model training and testing on features fused across all data streams. A system overview is depicted in Fig. 1. The modules are described in more detail in the following.

3.1. Visual Front-End

Our visual front-end module is based on the Viola-Jones algorithm for ROI detection. The approach employs AdaBoost, using cascades of weak classifiers to achieve high detection performance [12]. In particular, we applied the method on the planar visual data stream in a nested fashion, first employing one detector to localize the face in each video frame, and subsequently another detector to localize the mouth region within the face. In addition, we used a smoothing scheme by median filtering the mouth bounding box coordinates over ten neighboring frames. This way, the influence of false detections was minimized, yielding more robust tracking. The resulting coordinates were also used for locating the mouth region in the depth images, by adjusting the coordinates according to the disparity of the two sensors. The final ROI for both video and depth images was obtained by resizing the respective mouth bounding boxes to 64×64 pixels.

3.2. Feature Extraction

Following ROI extraction, the next step is to obtain a small number of meaningful features, adequately capturing the speech information present in the lip movements. To do so, we employed an appearance-based approach, applying the discrete cosine transform (DCT) on the mouth ROI image extracted from each video and depth frame [8]. We then considered the DCT coefficients in the upper-left corner of the transformed image, having higher energy values and thus capturing more lip movement information. The number of coefficients we extracted with this method was 45 for every frame. We then interpolated the features from 20 Hz to 100 Hz, in order to match the audio feature extraction rate (see below).

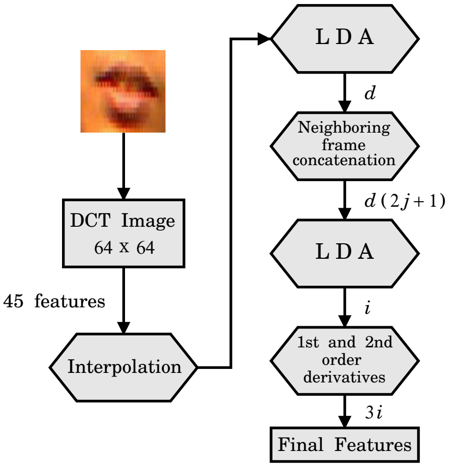

Figure 2: Feature extraction and selection pipeline used in our system for the visual (shown here) and depth data streams.

In order to improve feature selection, we implemented the two-stage LDA based approach depicted in Fig. 2, similarly to [9]. Specifically, at the first stage, we applied LDA on the 45 features of each frame ("intra-frame") selecting $d < 45$ features with the highest eigenvalues. Subsequently, at the second stage, we concatenated j neighboring feature vectors at each side to the vector of the current frame, in order to capture dynamic visual speech information. We then applied LDA ("inter-frame") to the concatenated vector of dimension $(2\,j + 1)\,d$, selecting of course a smaller number of features i with the highest eigenvalues. Finally, we calculated their first and second order temporal derivatives, appending them to the feature vector, thus yielding features of dimensionality $3\,i$. For the planar visual data stream, values $d = 10$, $j = 3$, and $i = 10$ were chosen, whereas for the depth data stream values $d = 15$, $j = 6$, and $i = 10$ were preferred. In both cases, the final features were of dimension $3\,i = 30$.

Concerning the audio features, we calculated the Mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives on windowed speech segments of 25 ms duration and 10 ms overlap. The length of the audio feature vector was 39.

3.3. Statistical ASR

Hidden Markov models (HMMs) are broadly used in ASR applications for modeling speech. The Baum-Welch algorithm is used for training the models and the Viterbi algorithm for recognition [13]. In our experiments, we compared the performance of two types of models, baseline single-stream HMMs (i.e., employing feature fusion) and state-synchronous multi-stream HMMs (two- and tri-stream HMMs, i.e., a decision fusion approach). The latter constitutes an early integration form of decision fusion, where the fused data observation likelihood is expressed as the product of the observation likelihoods of each stream, raised to exponents that express the reliability of each particular stream. In total, thirty context-dependent phonetic models were employed (triphones), each having three states in a left-to-right topology and four Gaussian mixtures per observation stream and state. The hidden Markov model toolkit (HTK) [13] patched with the HTS software [14] was used for training and testing. A free grammar was used at decoding (i.e., no constraints to the length of the recognized digit sequence were imposed).

Visual front-end	Video	Depth
Energy based feature selection	36.91%	18.20%
Intra-frame LDA	41.51%	19.28%
Intra- and inter-frame LDA	43.60%	20.72%

Fusion approach	Video + Depth
Single-stream HMM	41.29%
Two-stream HMM	44.39%

Table 1: Word recognition accuracy, %, at various stages of the visual and depth data processing pipelines (upper table), as well as for the single- and two-stream HMMs, when using video and depth information (lower table).

4. EXPERIMENTS AND RESULTS

In this section we present the experiments conducted in order to test the effectiveness of our system. In the experiments we used the audio, planar video, and depth streams of the BAVCD database. More specifically, we used the English data from 14 subjects in a random 2/3, 1/3 split for training and testing, respectively, to conduct multi-speaker ASR experiments. In order to test our system for robustness under the influence of noise, we corrupted the audio samples with additive babble noise from the NOISEX-92 database at several signal-to-noise ratios (SNRs). Clean audio though was used for model training. As already mentioned, all data utterances were connected digit sequences. The vocabulary size was 11, as both "zero" and "oh" were used for digit "0".

The first set of experiments presents the effects of using the depth stream in conjunction with the video stream, as well as the beneficial effects of feature transformation using the two-stage LDA. More specifically, from the data in Table 1 it is obvious that two-stage LDA improves performance not only compared to the energy based feature selection method, but also compared to intra-frame LDA alone. Furthermore, depth information does improve accuracy when compared to video-only recognition. Finally, the use of multi-stream HMMs improves significantly the ASR accuracy, when compared to baseline feature fusion (single-stream HMMs).

The second set of experiments considers the effects to ASR of fusing the planar video and/or depth data with the traditional audio stream. In Fig. 3 we compare the performance of audio-only ASR to decision fusion based ASR employ-

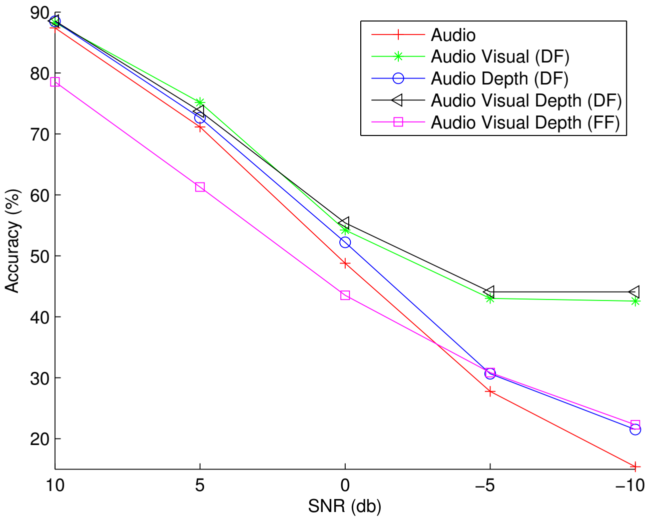

Figure 3: Connected digits speech recognition performance, measured in word accuracy, %, employing audio-only, audio-visual, audio-depth, and audio-visual-depth data, considered at various audio SNR levels. All stream combination results are obtained using decision fusion (DF), with the exception of the audio-visual-depth system, where feature fusion (FF) results are also depicted.

ing two-stream audio-visual HMMs, two-stream audio-depth HMMs, and three-stream audio-visual-depth models. From the results we can see that depth information is beneficial to ASR performance, especially for medium and low SNR values. We can also observe (in the case of audio-visual-depth based ASR) that decision fusion yields significant improvements over the baseline single-stream HMMs (feature fusion).

5. CONCLUSIONS

We have presented a novel multimodal speech recognition system that uses facial depth information, captured by the Kinect, in addition to the audio and video modalities, in order to boost ASR performance and robustness to noise. A two-stage LDA was applied to the visual and depth features, first intra- and subsequently inter-frame, resulting in a considerable increase in recognition accuracy. The depth modality improved ASR performance over audio-only and traditional audio-visual systems, when incorporated into them under the multi-stream decision fusion framework. It is our belief that this benefit will be further increased in future work by improving feature extraction in the depth data stream.

6. ACKNOWLEDGMENTS

G. Potamianos would like to acknowledge partial support from the European Commission through FP7-PEOPLE-2009-RG-247948 grant AVISPIRE.

This material is based upon work supported by the National Science Foundation under Grants No. NSF-CNS 1035913, NSF-CNS 0923494.

7. REFERENCES

[1] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," in *Proc. ICSLP*, 2000, vol. 3, pp. 20–24.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[3] R. Goecke and B. Millar, "The audio-video Australian English speech data corpus AVOZES," in *Proc. ICSLP*, 2004, pp. 2525–2528.

[4] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler, and R. Orglmeister, "WAPUSK20 – a database for robust audiovisual speech recognition," in *Proc. LREC*, 2010.

[5] A. Ortega, F. Sukno, E. Lleida, A. Frangi, A. Miguel, L. Buera, and E. Zacur, "AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition," in *Proc. LREC*, 2004, vol. 3, pp. 763–767.

[6] G. Galatas, G. Potamianos, D. Kosmopoulos, C. McMurrough, and F. Makedon, "Bilingual corpus for AVASR using multiple sensors and depth information," in *Proc. AVSP*, 2011, pp. 103–106.

[7] M. Zamalloa, L.J. Rodriguez, M. Penagarikano, G. Bordel, and J.P. Uribe, "Comparing genetic algorithms to principal component analysis and linear discriminant analysis in reducing feature dimensionality for speaker recognition," in *Proc. GECCO*, 2008, pp. 1153–1154.

[8] G. Potamianos and H.P. Graf, "Linear discriminant analysis for speechreading," in *Proc. MMSP*, 1998, pp. 221–226.

[9] G. Potamianos and C. Neti, "Improved ROI and within frame discriminant features for lipreading," in *Proc. ICIP*, 2001, vol. 3, pp. 250–253.

[10] *The Primesensor Reference Design*, [Online] Available at: `http://www.primesensor.com/?p=514`

[11] C.C. Liebe, C. Padgett, J. Chapsky, D. Wilson, K. Brown, S. Jerebets, H. Goldberg, and J. Schroeder, "Spacecraft hazard avoidance utilizing structured light," in *Proc. of IEEE Aerospace Conference*, 2006, pp. 10.

[12] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*, O'Reilly Media, 2008.

[13] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, 2006.

[14] *The HMM-Based Speech Synthesis System (HTS)*, [Online] Available at: `http://hts.sp.nitech.ac.jp/`