Audiovisual Liveness Detection

Aleksandr Melnikov[2(✉)], Rasim Akhunzyanov[1], Oleg Kudashev[2], and Eugene Luckyanets[1,2]

[1] ITMO University, St. Petersburg, Russia
rasim.akhunzyanov@niuitmo.ru
[2] STC-innovations Ltd., St. Petersburg, Russia
{melnikov-a,kudashev,lukyanets}@speechpro.com

Abstract. Although multi-modal (e.g. voice and face) biometric verification systems were in active development and showed impressive performance they need to be protected from spoofing attacks. In this paper we present methods for verifying face liveness based on estimation of synchrony between audio stream and lips movements track during the pronunciation of passphrase. The passphrase consists of a random set of the predetermined English words that are generated dynamically for each verification attempt. Lip movements extraction is performed by using of so-called Constrained Local Model of face shape. Audio stream is used to determine time intervals of pronounced words by means of automatic segmentation. Estimation of synchrony is done by analysis of lip movements for each word by employing a feedforward neural network and a Gaussian naive Bayes classifier. Finally, liveness score assessment is performed by averaging of individual word predictions during verification phrase utterance. For GRID corpus dataset average EER of 4.38% was achieved.

Keywords: Bimodal · Liveness detection · Anti-spoofing · Voice features · Face features

1 Introduction

In context of increasing advance of biometric security systems the importance of their spoofing attacks reflection is very high. In this paper we consider two biometric modalities: face and voice. Facial biometrics has developed increasingly in recent years. Some state-of-art systems show recognition quality comparable with the human recognition level [28]. However, such systems are susceptible to spoofing attacks that apply user photo or video. There are a large number of works devoted to the detection of such spoofing attacks [5]. It uses a variety of methods: frequency and texture based [10,12,20,21], variable focusing based [13,33], movement of the eyes based [1], optical flow based [3,15,16], blinking based [27], 3D face shape based [19,30], binary classification based [24,29], scenic clues based [23,32], lip movement based [6,17], context based [18].

Development of voice verification systems vulnerability to spoofing attacks has greatly increased recently. A lot of works [14,25] examine effects of voice

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 643–652, 2015.
DOI: 10.1007/978-3-319-23234-8_59

synthesis and voice conversion to speaker recognition performance as well as propose countermeasures to these attacks [22,31].

There are a few works devoted to the bimodal liveness detection. In paper [6] authors determines lip region and mouth fiducial points via color segmentation. MFCC features are extracted from the speech signal. The resulting audiovisual features are classified using GMM. Method requires model parameters optimization for each individual user.

Research of [26] present algorithm that finds the degree of synchronization between the audio and image recordings of a human speaker. It uses canonical correlation to find the best direction to combine all the audio and image data, projecting them onto a single axis. Then it uses Pearson's correlation to measure the degree of synchronization between the audio and image data. However anti-spoofing is out of scope of this paper and authors were not provided tests of described algorithm in sense of ability to determine face spoofing attacks.

In [7] authors combine mouth fiducial points with PCA eigenlips features. Authors of [4] describes the bimodal system for user verification. It uses MFCC features obtained from voice. 2D-DCT textural features are extracted from the lips region. Optical flow estimated as well. These features are fused using reliability weighted summation and then are used in standard approaches to speaker recognition (HMMs for text-dependent and GMMs for text-independent speaker identification). The similar approach is described in [11].

Here we introduce bimodal anti-spoofing system based on markup of voice signal and lips movements. Voice is processed by automatic audio segmentation, that determines boundaries of words in phrase. Lips movements are catched by fiducial points model for facial images. Phrases content was limited by the predefined number of words (so-called dictionary). We used digits because their utterance were provided by GRID dataset [8], but in real-life system it might be expanded as well.

Evaluation results on GRID corpus dataset show good performance of our system. Also we estimated our algorithm on the internal dataset, witch has been collected by frontal camera of several smartphones in real life conditions.

The remainder of this paper is organized as follows. Section 2 describes bimodal liveness detection. Experimental work is described in Section 3. Finally, our conclusions are presented in Section 4.

2 Text-dependent Bimodal Liveness Detection

The general audio-video synchronization task can be complicated. In order to simplify it we limit contents of passphrase by predefined dictionary containing several words. Then password phrase needed for biometric verification randomly assembled from dictionary. If phrase has enough length it is suitable for audio verification. Potentially this approach can reduce complexity of video processing and simplify synchrony detection. However, such method has several drawbacks:

1. it is required to gather new bimodal dataset for training that might be time consuming;

2. training and adjustments are needed when new language is added;
3. it is possible that liveness performance varies with language.

2.1 Audio Segmentation

The audio segmentation task consists of automatic time mapping between audio stream and pronounced phrase. We used Hidden Markov Models (HMM) to solve this task. Each word of the target phrase was represented as a set of hidden states. Each state had 0.04 sec. average length and was defined by single diagonal Gaussian. The 12 first MFCC without energy were used for signal parametrization. Also we used two additional hidden states, "pause" and "mean speech", to represent audio segments which are not a part of the target phrase. The Viterbi algorithm was used to decode hidden states and to construct segmentation. This method provides high accuracy for pronounced phrase correctness estimation. Thus, we consider only cases with correct pronounces.

Time boundaries of the each state were used for further audio-visual features construction.

2.2 Visual Features

It is necessary to determine the consistency of the facial movements with the utterance of a passphrase. In Audio-Visual Automatic Speech Recognition (AVASR) systems similar problem is solved by analyzing the movements of the lips. Our problem is easier, so we decided to use simple features — the relative change in the shape of the lips. We use anthropometric points detector (landmark detector), which is based on the face points distribution model (PDM). Such models were under active development in recent years. One of the most successful methods presented so far is the Constrained Local Model (CLM) initially proposed by Cristinacce and Cootes [9].

Constrained Local Model for Landmark Detection. CLM mainly consists of three components: PDM, Patch Experts (PE) and algorithm for PDM parameters fitting. Point Distribution Model describes non-rigid shape variations and global rigid transformation. After parameters of PDM are estimated it's possible to compute location $\mathbf{x}_i = [x_i, y_i]$ of each facial landmark:

$$\mathbf{x}_i = sR(\overline{\mathbf{x}_i} + \Phi_i q) + t,$$

where s, R and t terms are rigid parameters responsible for global shape scaling, rotation and translation accordingly and a set of non-rigid parameters q which are responsible for deformation of mean-shape (points $\overline{\mathbf{x}_i}$).

For details about CLM implementation please refer to [2]. Further in this section we concentrate only on details specific to our work.

Facial Landmarks. Liveness detection was designed to work mainly on mobile platforms with near to real-time performance. However image processing is computationally-intensive by its nature and minimization of computation complexity is crucial due to low energy, memory and processing platform capacities.

It's possible to sacrifice flexibility of CLM by reducing the number of points in the model. But we can't discard all points that are not related to the mouth because it reduce the reliability of detection. We conducted series of experiments and reduced the number of points as shown on Fig. 1. Left picture presents original model, while right introduces reduced one.

Thus, the frame processing time is decreased by:

- reducing of the computation of patch experts feedback about twice;
- simplification of the model to reduce the number of local parameters of PDM, which simplifies the optimization problem.

Such model stably and accurately describes facial geometry, while the temporary frame processing costs are reduced almost twice.

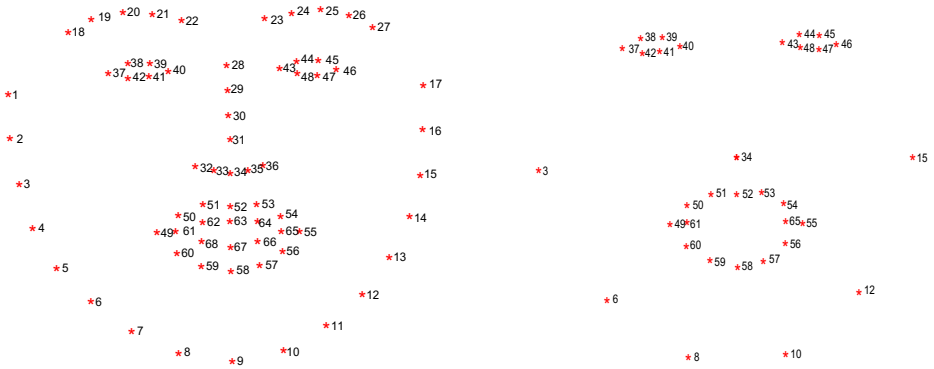

Fig. 1. Full 2D face shape model used in [2] and face shape model with reduced set of points.

2.3 Audio-visual Features Extraction

Given a video of N frames length, facial landmarks are extracted from each frame. Since the shape of lip contour looks like an ellipse, it was decided to use ellipse semi-diameters as characteristic features. Euclidean distances between two pairs of landmarks points are calculated as shown at Fig. 2 and stored in vector of length N. Since time of word utterance varies across sessions, vectors may have different length. Therefore vector lengths are aligned to number of hidden states of audio segmentation results by performing linear interpolation as shown on Fig. 3. After that, each word of utterance represented by fixed-length vector of semi-diameters W_x, where $x \in \{\text{'zero'}, \ldots, \text{'nine'}\}$ and $N_x = \text{length}(W_x)$. This vectors are used as audio-visual features for further classification.

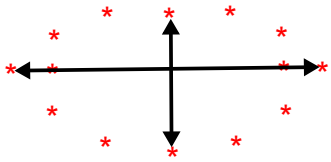

Fig. 2. Vertical and horizontal distances used as video features.

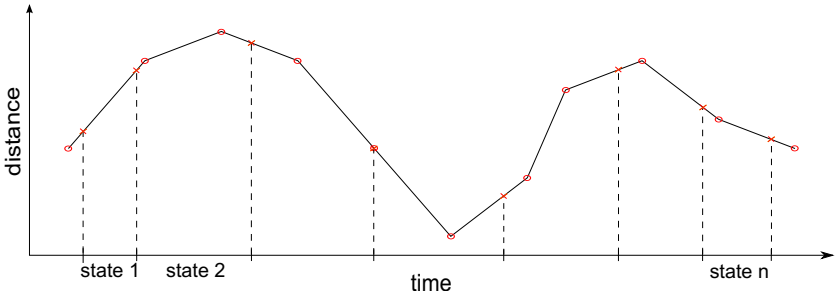

Fig. 3. State interpolation for word 'seven'.

2.4 Decision System

Classification. Training of the system requires two datasets: dataset for training and dataset for fusion optimisation. Train dataset consists of target (synchrony) and impostor (asynchrony) subsets. Impostor sessions were created from target ones by mixing audio and visual features from different utterances. From the train dataset the audio-visual features W_x^{target} and W_x^{impostor} were collected for each word in dictionary.

Obtained dataset is used for training of two classifiers: neural network and gaussian naive Bayes. We used a neural network with N_x inputs and scalar output. Such binary classifier distinguishes visual word for a specific digit from the rest.

For gaussian naive Bayes classifier two hypotheses for audiovisual features W_x are considered: H_x^s — the audio and video are synchrony, H_x^d — the audio and video are asynchrony. As the decision score a log of likelihood ratio $\ln P(W_x|H_x^s) - \ln P(W_x|H_x^d)$ is used, where $P(W_x|H_x^s)$ and $P(W_x|H_x^d)$ are represented by a single gaussian with full covariance matrices.

Resulting Score. Passphrase consists of N_w words pronunciations. For each word from passphrase prediction score is obtained by corresponding classifier. As result we have N_w scores s_k, one for each word. Using these scores we estimate equal error rate threshold T_k and the standard deviation of the impostors distribution σ_k.

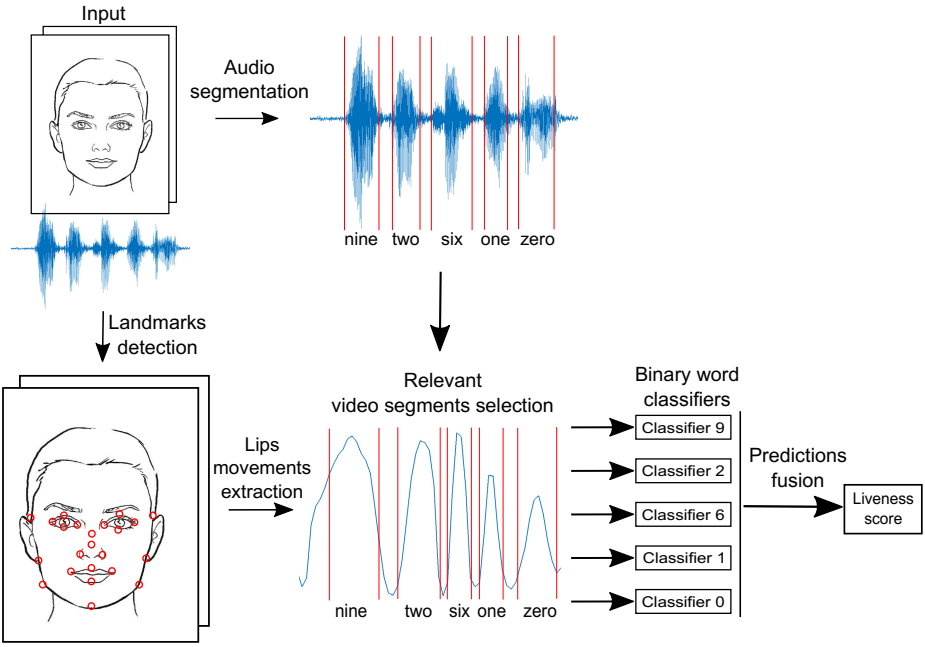

Fig. 4. Method workflow.

Fusion score s was calculated as follows:

$$s = \frac{1}{N_w} \sum_{k=0}^{N_w - 1} \frac{T_k - s_k}{\sigma_k}$$

To select the optimal parameters of T_k simulated annealing optimization method was used. The initial values for this method were obtained from equal error rate threshold described below. EER value is chosen to be a quality function in result fusion system. Overall system workflow is shown on fig.4.

3 Experimental Results

In this section we present experimental results produced for GRID corpus dataset [8]. This dataset consists of 34 speakers, 1000 sessions for each. However, one speaker has no video sessions, so we did not use them. Digits from 'zero' to 'nine' were chosen from dataset to evaluate algorithm. Dataset were splitted into train and test parts by speakers. In oder to increase train dataset size we chose only one speaker for testing at each time. So, we provide 33 train-test cycles with 32 speakers for training and 1 for testing. All results were averaged to obtain final EER result.

We tested two system: gaussian naive Bayes based and neural networks based that were described in section 2.4. Also, two types of speech segmentation were

used: our segmentation and segmentation provided by GRID dataset. In table 1 EER results for different passphrase lengths obtained by concatenation of several speaker sessions are shown. As it can be seen EER decreases with increasing number of digits in passphrase. Neural networks based system shows slightly better results than gaussian naive Bayes based system. It should be noted that our segmentation outperforms ground truth segmentation in current task. Fig. 3 demonstrates bad performance of our system for speaker 26. It happened because CLM algorithm worked bad on this speaker and as a result audiovisual features were wrong.

Our implementation of the facial landmarks extractor allows to achieve necessary performance and use proposed system on modern smartphones with hi-end chipsets in real-time.

We performed one more evaluation on the manually collected dataset in Russian. It contains 3 sessions of 77 male and 76 female subjects, recorded on 3 different mobile devices. Each session consists of several digit phrases. Satisfying results (table 2) approved that our method works good on essentially different datasets.

Table 1. EER results for fusion system on digits, %

num. of digits		2	3	4	5
GRID seg	NeuralNet	14.71	10.44	7.1	5.68
	GNB	15.35	10.74	7.82	5.86
our seg	NeuralNet	12.37	8.29	5.82	4.38
	GNB	13.27	8.75	6.27	4.61

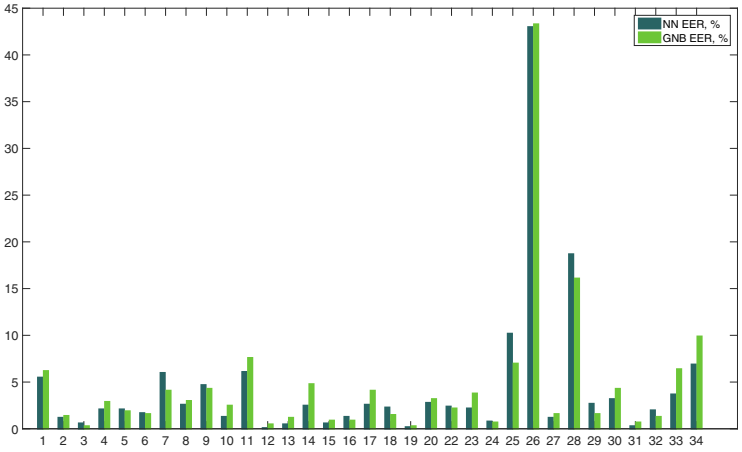

Fig. 5. EER per speaker for our segmentation, %

Table 2. EER results for fusion system on Russian dataset, %

num. of digits		2	3	4	5
our seg	NeuralNet	10.53	7.11	4.32	3.64
	GNB	12.85	8.06	5.51	4.17

4 Conclusion

In this paper we have introduced the new method for liveness detection. Through experiments on GRID dataset, we have shown that it is efficient in resolving liveness detection task with average EER of 4.38%. In future work we intend to use more complex information from facial features, such as optical flow changing, more accurate lip contour tracking and etc. Also, more robust facial PDM algorithm may be used.

Acknowledgments. The authors would like to thank the anonymous reviewers for their comments, which have significantly improved quality of the the manuscript.

References

1. Ali, A., Deravi, F., Hoque, S.: Liveness detection using gaze collinearity. In: 2012 Third International Conference on Emerging Security Technologies (EST), pp. 62–65. IEEE (2012)
2. Baltrusaitis, T., Robinson, P., Morency, L.: 3d constrained local model for rigid and non-rigid facial tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2610–2617. IEEE (2012)
3. Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: International Conference on Image Analysis and Signal Processing, IASP 2009, pp. 233–236. IEEE (2009)
4. Çetingül, H.E., Erzin, E., Yemez, Y., Tekalp, A.M.: Multimodal speaker/speech recognition using lip motion, lip texture and audio. Signal Processing **86**(12), 3549–3558 (2006)
5. Chakraborty, S., Das, D.: An overview of face liveness detection (2014). arXiv preprint arXiv:1405.2227
6. Chetty, G., Wagner, M.: Automated lip feature extraction for liveness verification in audio-video authentication. Proc. Image and Vision Computing, 17–22 (2004)
7. Chetty, G., Wagner, M.: Multi-level liveness verification for face-voice biometric authentication. In: 2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference, pp. 1–6. IEEE (2006)
8. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America **120**(5), 2421–2424 (2006)
9. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In BMVC, vol. 2, pp. 6. Citeseer (2006)
10. Das, D., Chakraborty, S.: Face liveness detection based on frequency and micro-texture analysis. In: 2014 International Conference on Advances in Engineering and Technology Research (ICAETR), pp. 1–4. IEEE (2014)

11. Dean, D., Sridharan, S.: Dynamic visual features for audio-visual speaker verification. Computer Speech & Language **24**(2), 136–149 (2010)
12. Kim, G., Eum, S., Suhr, J.K., Kim, D.I., Park, K.R., Kim, J.: Face liveness detection based on texture and frequency analyses. In: 2012 5th IAPR International Conference on Biometrics (ICB), pp. 67–72. IEEE (2012)
13. Kim, S., Yu, S., Kim, K., Ban, Y., Lee, S.: Face liveness detection using variable focusing. In: 2013 International Conference on Biometrics (ICB), pp. 1–6. IEEE (2013)
14. Kinnunen, T., Wu, Z.-Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H.: Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4401–4404, March 2012
15. Kollreider, K., Fronthaler, H., Bigun, J.: Evaluating liveness by face images and the structure tensor. In: Fourth IEEE Workshop on Automatic Identification Advanced Technologies, 2005, pp. 75–80. IEEE (2005)
16. Kollreider, K., Fronthaler, H., Bigun, J.: Non-intrusive liveness detection by face images. Image and Vision Computing **27**(3), 233–244 (2009)
17. Kollreider, K., Fronthaler, H., Faraj, M.I., Bigun, J.: Real-time face detection and motion analysis with application in "liveness" assessment. IEEE Transactions on Information Forensics and Security **2**(3), 548–558 (2007)
18. Komulainen, J., Hadid, A., Pietikainen, M.: Context based face anti-spoofing. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8. IEEE (2013)
19. Lagorio, A., Tistarelli, M., Cadoni, M., Fookes, C., Sridharan, S.: Liveness detection based on 3d face shape analysis. In: 2013 International Workshop on Biometrics and Forensics (IWBF), pp. 1–4. IEEE (2013)
20. Määttä, J., Hadid, A., Pietikäinen, M.: Face spoofing detection from single images using texture and local shape analysis. IET Biometrics **1**(1), 3–10 (2012)
21. Maatta, J., Hadid, A., Pietikainen, M.: Face spoofing detection from single images using micro-texture analysis. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7. IEEE (2011)
22. Marcel, S., Nixon, M.S., Li, S.Z.: Handbook of Biometric Anti-Spoofing. Springer (2014)
23. Pan, G., Sun, L., Zhaohui, W., Wang, Y.: Monocular camera-based face liveness detection by combining eyeblink and scene context. Telecommunication Systems **47**(3–4), 215–225 (2011)
24. Peixoto, B., Michelassi, C., Rocha, A.: Face liveness detection under bad illumination conditions. In: 2011 18th IEEE International Conference on Image Processing (ICIP), pp. 3557–3560. IEEE (2011)
25. Shchemelinin, V., Topchina, M., Simonchik, K.: Vulnerability of voice verification systems to spoofing attacks by TTS voices based on automatically labeled telephone speech. In: Ronzhin, A., Potapova, R., Delic, V. (eds.) SPECOM 2014. LNCS, vol. 8773, pp. 475–481. Springer, Heidelberg (2014)
26. Slaney, M., Covell, M.: Facesync: a linear operator for measuring synchronization of video facial images and audio tracks. In: NIPS, pp. 814–820 (2000)
27. Sun, L., Pan, G., Wu, Z., Lao, S.: Blinking-based live face detection using conditional random fields. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 252–260. Springer, Heidelberg (2007)
28. Taigman, Y., Yang, M., Ranzato, M.A., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708. IEEE (2014)

29. Tan, X., Li, Y., Liu, J., Jiang, L.: Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 504–517. Springer, Heidelberg (2010)
30. Wang, T., Yang, J., Lei, Z., Liao, S., Li, S.Z.: Face liveness detection using 3d structure recovered from a single camera. In: 2013 International Conference on Biometrics (ICB), pp. 1–6. IEEE (2013)
31. Zhizheng, W., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H.: Spoofing and countermeasures for speaker verification: A survey. Speech Communication **66**, 130–153 (2015)
32. Yan, J., Zhang, Z., Lei, Z., Yi, D., Li, S.Z.: Face liveness detection by exploring multiple scenic clues. In: 2012 12th International Conference on Control Automation Robotics & Vision (ICARCV), pp. 188–193. IEEE (2012)
33. Yang, L.: Face liveness detection by focusing on frontal faces and image backgrounds. In: 2014 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), pp. 93–97. IEEE (2014)