**­PHD RESEARCH PROPOSAL**

Proposed by:

**Rudi Primorac**

School of Electrical, Electronics and Computer Engineering School of Computer Science and Software Engineering Faculty of Engineering, Computing and Mathematics The University of Western Australia

**A. PROJECT TITLE AND SUMMARY**

**A.1 Project Title**

Audio-visual biometric recognition

**A.2 Summary**

The research will investigate algorithms for audio-visual biometric recognition, with emphasis on methods based on sparse representation. Algorithms based on sparse representation have been successfully used for face and speaker recognition separately, but, to the best of our knowledge haven’t been utilized for audio-visual recognition. The aims of this project are as follows:

* Introducing joint sparsity -based algorithms into audio-visual field. This includes modifying already existing general joint sparsity – based algorithms to accommodate better for audio data, as well as developing general frameworks for multimodal biometrics using joint sparsity. Comparison with systems which use sparse-based representation separately on each modality will be given.
* Developing fusion strategies based on the quality of each modality. This includes a possible introduction of quality based strategies where quality for each modality might change over time

**B. RESEARCH PROJECT**

**B.1 Background, Research Significance and Motivation**

Biometric recognition is defined as identifying a person based on their physiological and behavioural characteristics [2], [3], [38]. Physiological characteristics are related to the shape of the body, such as face, fingerprints, iris etc. Behavioural characteristics are related to the pattern of behaviour of a person, and include voice, gait, typing rhythm etc.

A biometric system can work in *verification* or *identification* mode [2], [3], [38]. In verification mode, the system verifies the identity of a person by comparing its biometric data with their own data already stored in the database. In identification mode, the system determines the identity of a person by comparing its biometric data with data of all persons enrolled in the system.

Biometric recognition systems are used in a wide range of applications, such as:

* *Government applications -*Passports, ID cards, Driver’s licences
* *Commercial applications -* Computer network login, cellular phones, credit cards, ATM
* *Forensic applications -*Corpse identification, criminal investigations

Biometric systems can be unimodal or multimodal [2], [3], [38]. Unimodal systems operate on one biometric trait only, such as face or voice. Those systems are sensitive to noise and inter-class variations. For example, a system based only on voice is sensitive to channel noise. On the other hand systems based on face are sensitive to illumination changes, pose changes, occlusion.

Multimodal systems use several biometric traits, which mitigates the problems unimodal systems face. Several sensors are used and the information from all of them is then fused to reach the final decision. Fusion can be done on 3 different levels:

* *Feature level fusion* – Features from each modality are fused and used as a single feature vector
* *Score level fusion –* A score is calculated for each modality and those scores are combined to reach the final decision
* *Decision level fusion –* A decision is made for each modality separately, and those decisions are then combined to reach the final decision

The performance of systems working in the verification mode is usually evaluated using *False Accept Rate (FAR), False Reject Rate* (FRR), *Equal Error Rate (*EER), *Half Total Error Rate (*HTER***)*** and/or Failure to Enrol Rate (FTR). The performance of systems working in the identification mode is evaluated using *Recognition Rate (RR).* [2], [3], [38]

**B.2 Literature Review**

This section can be divided into two parts. The first part will give a short overview of existing audio-visual systems. In the second part, an overview of using sparse-representation based techniques in biometrics will be given.

**2.1 Audio-Visual Biometric Systems**

Table 1 lists some of the existing audio-visual systems, and shows what features were used for audio and video agents, which classifier was used, which fusion strategy was utilized, database used (including the number of persons) as well as reported performance.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Author and year | Audio features | Video features | Fusion strategy | Classifier | Database (# persons) | Performance |
| Chibelushi et  al. [18], 1993 | MFCC | Shape-based  (PCA, LDA,  concatenation) | Opinion fusion (weighted  sum) | Neural network | 10 speakers | EER = 1.5% |
| Falavigna  [19], 1995 | MFCC, ∆, ∆∆ | Appearance-based | Opinion fusion (weighted  sum) | VQ | 89 speakers | IR = 98% |
| Dieckmann et  al. [20], 1997 | 2D Fourier Transform | Appearance-based  (gray level  projection and  optical flow  analysis) | Opinion fusion (weighted  sum) | Synergetics [32] | 66 staff | RR = 93% |
| Jourlin et  al. [21], 1997 | LPC, ∆, ∆∆ | Appearance-based  and shape-based | Opinion fusion (weighted  sum) | HMM | M2VTS (37) | IR=100%,  FAR=0.5% |
| Wark et  al. [22],  1999-2001 | MFCC | Shape-based  (PCA, LDA) | Opinion fusion (weighted  sum) | GMM | M2VTS (37) | IR=80% (equal  weight fusion,  12.2dB SNR) |
| Ben-Yacoub  et al. [23],  1999 | LPC | Appearance-based | Post classifier using binary  classifiers (SVM, Bayesian,  FLD, decision tree, MLP) | HMM | XM2VTS (295) | EER<1% |
| Aleksic and  Katsagge-  los [24], 2003 | MFCC, ∆, ∆∆ | Facial Animation  Parameters  (FAPs) | Appending the visual and  audio observation vectors | HMM | CMU (10) | EER=1.71% |
| Chaudhari et  al. [25], 2003 | MFCC | Appearance-based  (DCT) | Feature-level  concatenation, opinion  fusion | HMM | IBM (304) | IR=69.1% |
| Sanderson  and  Paliwal [26],  2004 | MFCC, ∆ | Appearance-based  (PCA) | Adaptive weighted  summation,  concatenation, SVM,  Bayesian classifier | GMM | VidTIMIT (403) | TE vs SNR |
| Erzin et  al. [27], 2005 | DCT | Appearance-based  (eigenface, DCT) | Adaptive cascade with the  ordering based on  reliability of classifier | HMM | MGL-AVD (50) | EER=1.4%(clean),  and 6.3% (5dB  SNR) |
| Fox et  al. [28], 2007 | MFCC | Appearance-based  (DCT) | Opinion fusion (weighted  sum) | HMM | XM2VTS (295) | IR=89.9% |
| Sugiarta et  al. [29], 2010 | Wavelet | DT-CWT,  DT-CWPT, PCA | Feature level fusion of  dual tree complex wavelet  transforms | VQ | VidTIMIT (403) | IR=93.7% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Wong et  al. [30], 2011 | MFCC | Wavelet subbands | Dual optimal multiband  feature fusion | VQ | UNMC-VIER  (123), CUAVE  (36),  XM2VTS(180) | RR=98.4%  (UNMC-VIER),  97.2% (CUAVE),  83.3% (XM2VTS) |
| Sahoo and  Prasanna  [31], 2011 | MFCC, ∆, ∆∆ | Appearance-based  (PCA, LDA) | Sum rule (score level) | GMM | IITG-DIT  M4(94) | EER=5.85% (Co-  hort fed FV) |
| Alam et al [32], 2013 | MFCC, ∆, ∆∆ | Downsampled cropped face | Sum rule | LRC [32] | AusTalk (88) | TE vs SNR |
| McCool et al [1], 2012 | MFCC, ∆, ∆∆ | Local Binary Pattern | Sum rule (logistic regression) | LBP similarity video [1]  PLDA[42] audio | MOBIO (150) | HTER = 11.9% (male)  HTER = 13.3 % (female) |
| Motlicek et al [33] | MFCC, ∆, ∆∆ | 2D Fourier transform | Sum rule | GMM | MOBIO (150) | HTER = 2.6% (male)  HTER = 9.7% (female) |
| Khoury et al [34] | MFCC, ∆, ∆∆ | 2D Fourier transform | Sum rule (logistic regression) | GMM | MOBIO (150) | HTER = 1.9% (male)  HTER = 6.3% (female) |

Table 1: Overview of existing AV systems

Falavigna et al [19] proposed to combine the output scores of face and speaker recognition modules by using a statistical approach. As scores from different modules may lie in different ranges, they highlighted the necessity of normalizing the score. Instead of normalizing the scores, it is also possible to do a majority voting based on several sub-systems’ decisions [20].

In [18], [19], [21], [22] a weighted fusion is of classifiers is applied. The choice of weights is an open research problem and many different ways have been investigated. For example, within sparsity framework [4], the classifiers are assigned weights depending on the signal quality. On the other hand, in [25] feature concatenation combined with opinion fusion is used; however the recognition rate of only 69.1% is achieved. Bengio et al [43] proposed to use asynchronous hidden Markov model in order to build an audio-visual model. In [1], similarity between histograms of local binary patterns [41] was used to score video frames, while i-vectors [7] and PLDA [42] were used as audio agent scores. In [33], inter-session variability modelling and joint factor analysis were used for both face and speaker authentication, and the scores were later fused using the logistic regression.

Early work on audio-visual authentication was often performed on small, in-house databases. The M2VTS database was created to set a common test bench for researchers, allowing them to evaluate their algorithms using well defined protocols. This database eventually grew into XM2VTS database consisting of 295 subjects [46]. The XM2VTS database was too clean, and not realistic enough for real-life situations. Recently, the challenge of bimodal authentication in mobile phone environments has begun to receive more attention, with the introduction of the challenging MOBIO database [1].

**2.2 Sparse Representations**

The theory of compressive sensing (sparse representation) is a major research area in the field of signal processing. In signal processing theory, by using sparse representation, the signal can be sampled at a much lower rate than the Niquist rate, and still be perfectly recovered. Even though not initially intended for classification purposes, the sparse representation is discriminative in nature, as it selects basis vectors that most compactly represent a certain signal.

In his seminal paper, Wright et al [4] introduced for the first time the use of sparse representation into face recognition field. An unknown face is represented as a sparse linear combination of training faces, where the idea is that the unknown coefficients in the sparse representation will correspond mostly to the training faces belonging to the same class as the test face. The framework introduced is quite general as it also deals with occlusion, and can be used with any kernel.

In [11], sparse representation was introduced for the first time for speaker recognition. Each speaker is represented as a vector of a fixed length, called UBM-GMM supervector. An overcomplete dictionary is then created by concatenating all enrolled speakers. An unknown speaker is that represented as a sparse linear combination of the existing speakers in the dictionary, by solving the l1optimization problem. In [39], instead of using a high-dimensional UBM-GMM supervector, a lower dimensional i-vector [5] was used to create an overcomplete dictionary. In [35], instead of using raw features as atoms in the dictionary, K-SVD [10] algorithm was used to create an appropriate dictionary. In [40], a set of faces was modelled using GMM, and their 2D Fourier transform coefficients were used as atoms in the dictionary.

Recently, joint sparse representation paradigm has been introduced into multimodal biometrics [3], [14]. In that paradigm, a separate dictionary is created for each modality (trait), however the sparse representation is obtained by encouraging all modalities to share a similar sparsity pattern. That way, test data is represented as a sparse linear combination of training data, whilst taking into account correlations as well as coupling information between different modalities. This framework supports quality-based fusion, by assigning quality factor to each modality [4], based on how sparse its representation is. The framework has been used so far in visual field, for combining visual modalities (face, iris, fingerprints). In [13], and extension to JSRC is proposed, based on tree-structured sparsity. It is claimed, that although different sources are correlated, the joint sparsity assumption of JSRC may be too restrictive for some applications in which not all the different modalities are equally correlated and stronger correlations between some groups of the modalities may exist. Tree-structured sparsity model provides a flexible framework to enforce prior knowledge in grouping different modalities by encoding them in a tree, where each leaf node represents an individual modality and each internal node represents a grouping of its child nodes.

**2.3 Scholars in the Field**

Some of the research experts that work in audio – visual field, as well as experts in sparse representation/ compressive sensing are listed below:

1. Prof. Anil K. Jain, Department of Computer Science and Engineering, Michigan State University, USA. Email: [jain@cse.msu.edu](mailto:jain@cse.msu.edu)

2. Prof. Aggelos K. Katsaggelos, The Image and Video Processing Laboratory, Department of Electrical Engineering and Computer Science, Northwestern University, USA. Email: [aggk@eecs.northwestern.edu](mailto:aggk@eecs.northwestern.edu)

3. Daniele Falavigna, Instituto per la Ricerca Scienti\_ca e Technologica, Trento, Italy. Email: [falavi@irst.itc.it](mailto:falavi@irst.itc.it)

4. Roberto Brunelli, Foundazione Bruno Kessler, Trento, Italy. Email: [roby.brunelli@gmail.com](mailto:roby.brunelli@gmail.com)

5. Ulrich Dieckmann, BioID Research and Development Department, Dialog Communications Systems AG, Erangen, Germany. Email: [ud@dcs.de](mailto:ud@dcs.de)

6. Prof. Michael Elad, Department of Computer Science, Technion – Israel Institute of Technology, Email: [elad@cs.technion.ac.il](mailto:elad@cs.technion.ac.il)

**B.3 Research Plan**

**3.1 Conceptual framework**

The research will focus on developing efficient algorithms that receive a video sequence of a speaking person as an input and either verify that the person is who they claim to be (verification mode) or determine who the person is (identification mode). The emphasis will be on algorithms based on sparse representation. A general block-diagram of the system is shown in Figure 1:

Figure 1: Block diagram of the proposed system

The system consists of several modules that are listed below:

* Audio voice activity detector + feature extraction: Voice activity detector is applied in order to eliminate frames without voice activities (i.e. silence). The goal of extracting features from an audio signal is obtaining a representation that is compact and less redundant, and as such, more suitable for statistical modelling. In this research we use standard normalized MFCC features and its derivatives. We use those features to build a UBM-GMM model, from which we extract the i-vector as a feature for that audio sequence
* Video feature extraction: In the first stage of the project, for each frame the face will be cropped using Viola Jones algorithm [12], and the cropped region will be downsampled to a fixed size, which will be a feature vector for that frame [4]. In later stages, we will also consider different feature extraction techniques, such as 2D Fourier Transform as used in [40], which will allow probabilistic modelling of the whole video sequence using a GMM framework and i-vectors.
* Joint sparse representation: By using the framework described in [14], we will first choose the best video frame (s), by having one modality (video) with a number of different observations. After having the video sequence described by only one feature vector, we will represent both audio and video as a sparse combination of training feature vectors, by utilizing correlations between them within the joint sparsity constraint.
* Fusion: The framework [14] naturally assigns quality measures to each modality, based on how sparse their representation is. The fusion is then performed using weighted sum of scores for each modality

Within this research project we intend to do the following:

* Implementing the basic audio-visual identification system using the joint sparsity framework and evaluating its performance in comparison to other AV systems that use the MOBIO database, as well as compare it to the system that doesn’t use the joint sparsity constraint
* Evaluate the use of different dictionaries, created for example by using the K-SVD algorithm. The major advantages of learned dictionaries are the increased flexibility and the ability to adapt to specific data. K-SVD based classification was first introduced into speaker recognition field by [35], while firstly proposed by [36] for face and object categorization
* Extending the existing general JSRC framework to accommodate better for audio data features. This can be done, for example, by modifying the existing cost function and solving the new optimization problem
* Developing new quality-based fusion methods. One of the ideas to investigate is the notion that the quality of a certain modality may vary with time. We could then take into account only parts of the signal that are of good quality, while discarding the rest of it

**3.1.2 Work done so far**

A comprehensive literature review beyond state of art has been completed. Basic feature extraction modules for the audio agent (based on MFCC and its derivatives) as well as for video agent are implemented, mostly using already existing MATLAB libraries (MSR Identity Toolbox, VOICEBOX Toolbox). The MATLAB Computer Vision Toolbox has been used for face detection using the Viola Jones algorithm [12]. Existing libraries were used as well for *l1* problem solving that gives the sparsest solution for a given set of input features. Joint sparse representation cost function optimization algorithm based on Alternating Direction Method of Multipliers has also been implemented. Experiments are currently underway using the MOBIO database.

**3.1.3 Database**

All experiments in this research will be performed using the MOBIO database [1].This database is unique because it was captured almost exclusively using mobile phones. It consists of over 61 hours of data with 12 distinct sessions. In total, there are 192 audio-video sequences for each of the 150 participants. This database is quite challenging since it was captures in uncontrolled environments, which leads to high variability of pose and illumination conditions, high variability in speech quality and variability in acquisition environment.

**3.2 Task specification timeline**

|  |  |  |
| --- | --- | --- |
| **Phases** | **Dates** | **Detailed Task Specification** |
| Literature review  Building a basic working system | Oct 2014 -Apr 2016 | Critical review of existing literature in the field of audio-visual biometrics and sparse representation / classification |
| Building a basic audio-visual recognition system based on joint sparsity representation |
| Submitting a conference paper |
| Building an advanced working system | Jan 2016 –Dec 2016 | Submitting a journal paper |
| Expanding joint sparsity framework | Jan 2017 – Mar 2018 | Investigating methods to extend joint sparsity framework for audio – signals |
| Submitting conference and journal papers |
| Developing quality-based fusion methods | Apr 2018 – June 2019 | Developing quality based fusion methods based on sparse representation |
| Submitting a conference paper |
| Thesis preparation | July 2019 – June 2020 | Submitting a journal paper based on our final system |
| Collect and compile writing from various phases of the project |
| Thesis review and submission |

**3.2 References**

[1] C. McCool et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In Proc. IEEE ICME Workshop on Hot Topics in Mobile Multimedia, Melbourne, July 2012

[2] A. Jain and S. Aggarwal. Multimodal biometric system: A survey. reason, 1(1):58-63, 2012.

[3] A. Jain, A. Ross, and S. Pankanti, “Biometrics: A Tool for Information Security,” IEEE Trans. Information Forensics and Security, vol. 1, no. 2, pp. 125-143, June 2006.

[3] S. Shekhar, V, M. Patel, N M. Nasrabadi, and R Chellappa, “Joint Sparse Representation for Robust Multimodal Biometrics Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 1, pp. 113-126, January 2014.

[4] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, “Robust Face Recognition via Sparse Representation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210-227, Feb. 2009.

[5] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," Tech Report Online: http://www.crim.ca/perso/patrick.kenny, 2005.

[6] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, pp. 788-798, 2011.

[7] A. Nefian, L. Liang, T. Fu, and X. Liu. A bayesian approach to audio-visual speaker identification. In Audio-and Video-Based Biometric Person Authentication, pages 1056 -1056. Springer, 2003.

[8] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, vol. 17, pp. 91-108, 1995.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in Digital Signal Processing, 2000, pp. 19-41.

[10] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation,”*IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[11] I. Naseem, R. Togneri, and M. Bennamoun, “Sparse Representation for Speaker Identification,” in *ICPR*, 2010, p. 4460.

[12] P. Viola and M.J. Jones. Robust real-time face detection. International journal of computer vision, 57(2):137{154, 2004.

[13]. S. Bahrampour, A. Ray, N. M. Nasrabadi, and W. K. Jenkins. Quality-based multimodal classification using tree-structured sparsity. In Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR), pp. 4114–4121 (2014).

[14] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In CVPR, pages 3493–3500, 2010

[15] Chen, Y.C., Patel, V., Shekhar, S., Chellappa, R., Phillips, P.: Video-based face recognition via joint sparse representation. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. (2013)

[16] [BC Haris](https://scholar.google.com.au/citations?user=oda11fEAAAAJ&hl=en&oi=sra), [R Sinha](https://scholar.google.com.au/citations?user=TnFe6usAAAAJ&hl=en&oi=sra), “Speaker verification using sparse representation over ksvd learned dictionary,” in Proc. 18th National Conference on Communications 2012 , Feb. 2012.

[18] Chibelushi Deravi, , Southampton U. K, C. C. Chibelushi, C. C. Chibelushi, F. Deravi, F. Deravi, J. S.Mason, and J. S. Mason. Voice and facial image integration for person recognition, 1993.

[19] R. Brunelli and D. Falavigna. Person identification using multiple cues. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 17(10):955 - 966, 1995.

[20] U. Dieckmann, P. Plankensteiner, and T. Wagner. Sesam: A biometric person identification system using sensor fusion. Pattern recognition letters, 18(9):827-833, 1997.

[21] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. Pattern

Recognition Letters, 18(9):8530858, 1997.

[22] T. Wark and S. Sridharan. Adaptive fusion of speech and lip information for robust speaker identification. Digital Signal Processing, 11(3):169-186, 2001.

[23] S. Ben-Yacoub, J. Luttin, K. Jonsson, J. Matas, and J. Kittler. Audio-visual person verication. In

Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., volume 1.

IEEE, 1999.

[24] PS Aleksic and AK Katsaggelos. An audio-visual person identification and verification system using FAPSs as visual features. In ACM Workshop on Multimodal User Authentication, pages 80{84, 2003.

[25] U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos, and C. Neti. Audio-visual speaker recognition using time-varying stream reliability prediction. In Acoustics, Speech, and Signal Processing, 2003.Proceedings.(ICASSP'03). 2003 IEEE International Conference on, volume 5, pages V-712. IEEE, 2003.

[26] C. Sanderson and K.K. Paliwal. Identity verification using speech and face information. Digital Signal Processing, 14(5):449-480, 2004.

[27] E. Erzin, Y. Yemez, and A.M. Tekalp. Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. Multimedia, IEEE Transactions on, 7(5):840{852, 2005.of afective speech. BIWI technical report no. 270, Computer Vision Lab, ETH Zurich, 2010.

[28] N.A. Fox, R. Gross, J.F. Cohn, and R.B. Reilly. Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts. Multimedia, IEEE Transactions on, 9(4):701-714,2007.

[29] YBG Sugiarta, R. Bambang, et al. Feature level fusion of speech and face image based person identification system. In Computer Engineering and Applications (ICCEA), 2010 Second International

Conference on, volume 2, pages 221 -225. IEEE, 2010.

[30] Y.W.Wong, K.P. Seng, and L.M. Ang. Audio-visual recognition system in compression domain. Circuits and Systems for Video Technology, IEEE Transactions on, 21(5):637-646, 2011.

[31] S.K. Sahoo and S.R.M. Prasanna. Bimodal biometric person authentication using speech and face under degraded condition. In Communications (NCC), 2011 National Conference on, pages 1-5. IEEE, 2011.

[32] M. Alam, R. Togneri; F. Sohel; M. Bennamoun; I. Naseem,Linear Regression-based Classifier for audio visual person identification, 2013 1st International Conference on Communications, Signal Processing and Their Applications, ICCSPA 2013. Vol. 1 USA : IEEE, 2013. p. 5pp.

[33] P. Motlicek, L. El Shafey, R. Wallace, C.McCool, S. Marcel Bi-Modal Authentication in Mobile Environments Using Session Variability Modelling

[34] E. Khoury, L. El Shafey, C. McCool, M. Gunther,S. Marcel Bi-Modal Biometric Authentication on Mobile Phones in Challenging Conditions

[35] C. Tzagkarakis and A. Mouchtaris, Sparsity Based Noise Robust Speaker Identification Using a Discriminative Dictionary Learning Approach,

[36] Z. Jiang, Z. Lin, and L. S. Davis, “Learning a discriminative dictionary for sparse coding via label consistent K-SVD,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Maryland, USA, June 2011, pp. 1697–1704.

[37] H. Hermansky and N. Morgan. Rasta processing of speech. Speech and Audio Processing, IEEE Transactions on, 2(4):578-589, 1994.

[38] A.K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. Circuits and Systems for Video Technology, IEEE Transactions on, 14(1):4-20, 2004.

[39] Ming Li, Xiang Zhang, Yonghong Yan, Speaker Verification using Sparse Representations on Total Variability I-Vectors

[40] Ming Li, Shrikanth Narayanan Robust Talking Face Video Verification Using Joint Factor Analysis And Sparse Representation On GMM Mean Shifted Supervectors

[41] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971– 987, 2002.

[42] S. J. D. Prince and S. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007.

[43] S. Bengio. Multimodal authentication using asynchronous HMMs. In *International Conference*

*on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, volume 2688, pages

770–777. Springer-Verlag, 2003. ISBN 978-3-540-40302-9. doi: 10.1007/3-540-44887-X\_89

[44] R. Vogt, S. Sridharan, Explicit modelling of session variability for speaker verification, Computer Speech & Language 22 (1) (2008) 17–38.

[45] R. Wallace, M. McLaren, Total variability modelling for face verification, IET Biometrics (2012) 188–199.

[46] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maître. XM2VTSDB: The extended M2VTS

database. In *International Conference on Audio and Video-based Biometric Person Authentication*

*(AVBPA)*, pages 72–77, 1999.

**C RESEARCH PROJECT DETAILS**

**C.1 Confidential/Sensitive Information**

This project contains no confidential/sensitive information.

**C.2 Intellectual Property Information**

Any intellectual property issues that arise during this study will be conveyed to the Research Committee in due course

**C.3 Fieldwork Information**

No fieldwork is involved for this research

**C.4 Facilities**

Private computer with MATLAB installed

**C.5 Statistical Component**

If required, statistical advice will be requested from the Statistics Clinics conducted by the UWA

**C.6 Skills Audit**

Table 2 shows the summary of the skills required for this project

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Student name: Rudi Primorac** | **Personal rating** | | | | **Evidence** |
| **Professional and research skills** | **None** | **Basic** | **Competent** | **Proficient** |
| Understanding and application of software (MATLAB) required for the research |  |  |  | x | More than 5 years of experience, both in industry and academy |
| Identifying and accessing appropriate bibliographic resources |  |  |  | x | Wrote my undergraduate and MSc thesis as well as a conference paper |
| Use of information technology relevant for the research (referencing software, spreadsheets and databases) |  |  |  | x | Gained experience during graduate studies |
| Familiar with concepts of academic writing |  |  |  | x | Wrote my undergraduate and MSc thesis as well as a conference paper |
| Ability to present and defend research outcomes at seminars and conferences |  |  |  | x | Presented my MSc research at the faculty seminar, poster presentation on an international conference |
| Understanding the theory of signal processing and computer vision |  |  |  | x | Took several undergraduate and graduate courses, several years of experience as a researcher in industry |

Table 2: Skills required for this research project

**C.7 Research Project Communication**

The research project will be published in related conferences and journals. Table 3 shows anticipated research communication.

|  |  |
| --- | --- |
| **Designation Task** | **Anticipated Date** |
| Seminar presentation at the school | December 2015 |
| Conference paper (basic JSRsystem) | April 2016 |
| Journal paper (advanced JSRsystem) | December 2016 |
| Conference paper (generalized JSR framework) | June 2017 |
| Journal paper (generalized JSR framework) | March 2018 |
| Conference paper (quality based fusion) | December 2018 |
| Journal paper – final system | December 2019 |

Table 3: Anticipated research communication

**C.8 Approvals**

No special approvals are required.

**C.9 Data Management**

All data will be stored on a personal laptop and will be back up by personal cloud storage.

**C.10 Research Project Plan**

The research project timeline is presented in Table on page 15.

**D Research Training**

**D.1 Research Training Plan**

Table 4 shows Research Training Plan for this research project.

|  |  |
| --- | --- |
| **Activity** | **Anticipated Date** |
| **Candidature Tasks**   * Enrolment * Compulsory Academic Conduct Essential (ACE) * Research Proposal Due * Confirmation of Candidature Due * Thesis Submission | Completed 17/10/2014  Completed October 2014  28/07/2015  16/10/2016  July 2020 |
| **Research and Academic Tasks**   * GRS Seminar: How to Prepare And Submit a Research Proposal * Independent Study on Speech Processing, Computer Vision and Machine Learning | Completed online April 2015  Throughout the candidature |
| **Annual tasks**  Submitting annual progress report to the Graduate Research School | 16/10 2015, 2016, 2017, 2018, 2019 |

Table 4: Research Training Plan

**D.2 Confirmation of Candidature**

The list of task required for the confirmation of candidature is given at the back of Research Proposal Coversheet.

**D.3 Working hours**

Working hours will be 15 hours/week, in accordance with UWA policy.

**E Budget**

Table 5 shows the estimated costs for this project. I plan to apply for the Graduate Research School Travel Grant to participate in international conferences, where I will present my research.

|  |  |  |
| --- | --- | --- |
| **Description** | **Costs**  **Year 1 Year 2 Year 3 Year 4 Year 5 Year 6** | **Source**  **School GRS** |
| **Administrative and research costs**   1. Workstation 2. Software Licence for MATLAB | $0 $1500 $0 $0 $0 $0  $0 $75 $75 $75 $75 $0 | $1500 $0  $300 $0 |
| **Training costs**  Seminars and workshops by GRS | $0 $0 $0 $0 $0 $0 | $0 $0 |
| **International conferences attendance**  Flights, registration, accommodations | $0 $0 $1850 $0 $0 $0 | $0 $1850 |
| **Sub – total** |  | $1800 $1850 |
| **Total** |  | $3650 |

Table 5: Overview of anticipated costs for this project

**F Supervision**

Principal & Coordinating Supervisor: Professor Roberto Togneri (40%)

* Managing overall research direction
* Providing expertise in speech signal processing
* Reviewing research outputs for thesis and publications
* Providing feedback on project’s progress during regular meetings

Principal Supervisor: Winthrop Professor Mohammed Bennamoun (40%)

* Providing expertise in computer vision
* Reviewing research outputs for thesis and publications
* Providing feedback on project’s progress during regular meetings

Co-Supervisor: Dr Ferdous Ahmed Sohel (20%)

* Providing expertise in computer vision
* Reviewing research outputs for thesis and publications
* Providing feedback on project’s progress during regular meetings

**G Research Project Plan**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| YEAR | YEAR 1 | | | | | | YEAR 2 | | | | | | YEAR 3 | | | | | | YEAR 4 | | | | | | YEAR 5 | | | | | | YEAR 6 | | | | | | |
|  | 2015 | | | | | | 2016 | | | | | | 2017 | | | | | | 2018 | | | | | | 2019 | | | | | | 2020 | | | | | |
| Activity |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Literature review |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| a) Critical review of literature | x | x | x | x | x | x | x | xx | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| b) Determine the focus of research | x | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Proposal Writing and Submission |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Basic JSR system |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| a) Audio and video feature extraction |  | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| b) Basic JSR system development |  |  | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| c) Experiments ,conference paper preparation |  |  |  |  | x | x | x | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Seminar Presentation at the School |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x |  |  |  |
| Advanced JSR System |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| a) System development |  |  |  |  |  |  |  | x | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| b) Technical report preparation |  |  |  |  |  |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| c) Journal paper preparation |  |  |  |  |  |  |  |  |  |  | x | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Generalized JSR Framework |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| a) System development |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x | x | x | x | x | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| b) Preparing a conference paper |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| c) Preparing a journal paper |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x | x |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Quality Based Fusion |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| a) System development |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x | x | x | x | x | x |  |  |  |  |  |  |  |  |
| b) Conference and journal paper preparation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x | x | x | x | x |  |  |  |  |  |  |
| Thesis Writing |  | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| a) Collect all components of research |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x | x |  |  |  |  |  |
| b) Review and final submission |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | x | x | x |  |  |