MMEDIA 2018

The Tenth International Conferences on Advances in Multimedia

ISBN: 978-1-61208-627-9

April 22 - 26, 2018

Athens, Greece

MMEDIA 2018 Editors

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany

MMEDIA 2018

Forward

The Tenth International Conference on Advances in Multimedia (MMEDIA 2018), held between April 22, 2018 and April 26, 2018 in Athens, Greece, continued a series of events for presenting recent research results on advances in multimedia, mobile and ubiquitous multimedia and to bring together experts from both academia and industry for the exchange of ideas and discussion on future challenges in multimedia fundamentals, mobile and ubiquitous multimedia, multimedia ontology, multimedia user-centered perception, multimedia services and applications, and mobile multimedia.

The new technological achievements in terms of speed and the quality of expanding and creating a vast variety of multimedia services like voice, email, short messages, Internet access, m-commerce, to mobile video conferencing, streaming video and audio.

Large and specialized databases together with these technological achievements have brought true mobile multimedia experiences to mobile customers. Multimedia implies adoption of new technologies and challenges to operators and infrastructure builders in terms of ensuring fast and reliable services for improving the quality of web information retrieval.

Huge amounts of multimedia data are increasingly available. The knowledge of spatial and/or temporal phenomena becomes critical for many applications, which require techniques for the processing, analysis, search, mining, and management of multimedia data.

The conference had the following tracks:

- Multimedia applications
- Perception and cognition for multimedia users

We take here the opportunity to warmly thank all the members of the MMEDIA 2018 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated their time and effort to contribute to MMEDIA 2018. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also gratefully thank the members of the MMEDIA 2018 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that MMEDIA 2018 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the field of multimedia. We also hope that Athens, Greece, provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

MMEDIA 2018 Chairs

MMEDIA Steering Committee

Jean-Claude Moissinac, TELECOM ParisTech, France
Daniel Thalmann, Nanyang Technological University, Singapore

MMEDIA Industry/Research Advisory Committee
Trista Chen, Trista Chen Consulting, USA
Alexander C. Loui, Kodak Alaris Inc., USA
Dimitrios Liparas, Information Technologies Institute (ITI) - Centre for Research & Technology Hellas (CERTH), Greece
Siyu Tang, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Giuseppe Amato, CNR-ISTI, Italy

MMEDIA 2018
Committee

MMEDIA Steering Committee
Jean-Claude Moissinac, TELECOM ParisTech, France
Daniel Thalmann, Nanyang Technological University, Singapore

MMEDIA Industry/Research Advisory Committee
Trista Chen, Trista Chen Consulting, USA
Alexander C. Loui, Kodak Alaris Inc., USA
Dimitrios Liparas, Information Technologies Institute (ITI) - Centre for Research & Technology Hellas (CERTH), Greece
Siyu Tang, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Giuseppe Amato, CNR-ISTI, Italy

MMEDIA 2018 Technical Program Committee
Vladimir Alexiev, Ontotext AD, Bulgaria
Giuseppe Amato, CNR-ISTI, Italy
Stylianos Asteriadis, University of Maastricht, Netherlands
Ramazan S. Aygun, University of Alabama in Huntsville, USA
Jenny Benois-Pineau, University of Bordeaux, France
Fernando Boronat Seguí, Universitat Politecnica de Valencia, Spain
Pierre Boulanger, University of Alberta, Canada
Dumitru Dan Burdescu, University of Craiova, Romania
Nicola Capuano, University of Salerno, Italy
Shannon Chen, Facebook, USA
Trista Chen, Trista Chen Consulting, USA
Luis A. da Silva Cruz, University of Coimbra, Portugal
Maaike de Boer, TNO, Netherlands
Jana Dittmann, Otto-von-Guericke-University Magdeburg, Germany
Vlastislav Dohnal, Masaryk University, Czech Republic
Marcio Ferreira Moreno, IBM Research, Brazil
Daniela Giorgi, Institute of Information Science and Technology - National Research Council of Italy, Italy
Nikolaos Gkalelis, Information Technologies Institute - Centre for Research and Technology Hellas, Greece
William Grosky, University of Michigan-Dearborn, USA
Jung Hyun Han, Korea University, Korea
Masaharu Hirota, Okayama University of Science, Japan
Jun-Won Ho, Seoul Women's University, South Korea
Yin-Fu Huang, National Yunlin University of Science and Technology, Taiwan
Eenjun Hwang, Korea University, South Korea
Hiroshi Ishikawa, Tokyo Metropolitan University, Japan

Ligang Zhang, Centre for Intelligent Systems - Central Queensland University, Brisbane, Australia

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

A Comparison of Face Verification with Facial Landmarks and Deep Features

Giuseppe Amato*, Fabrizio Falchi*, Claudio Gennaro* and Claudio Vairo*
*Institute of Information Science and Technologies of the National Research Council of Italy (ISTI-CNR)
via G. Moruzzi 1, 56124 Pisa, Italy
Email: {giuseppe.amato, fabrizio.falchi, claudio.gennaro, claudio.vairo}@isti.cnr.it

***Abstract*—Face verification is a key task in many application fields, such as security and surveillance. Several approaches and methodologies are currently used to try to determine if two faces belong to the same person. Among these, facial landmarks are very important in forensics, since the distance between some characteristic points of a face can be used as an objective measure in court during trials. However, the accuracy of the approaches based on facial landmarks in verifying whether a face belongs to a given person or not is often not quite good. Recently, deep learning approaches have been proposed to address the face verification problem, with very good results. In this paper, we compare the accuracy of facial landmarks and deep learning approaches in performing the face verification task. Our experiments, conducted on a real case scenario, show that the deep learning approach greatly outperforms in accuracy the facial landmarks approach.**

Keywords–Face Verification; Facial Landmarks; Deep Learning; Surveillance; Security.

I. INTRODUCTION

Face verification is getting higher importance recently. Face verification consists in determining if two faces in two different images belong to the same person or not. Face recognition, on the other hand, aims at assigning an identity to the person the faces belong to. In this paper, we are interested in the face verification problem.

To address the face verification problem, several approaches and techniques have been proposed. Some approaches are based on local features of the images, such as Local Binary Pattern (LBP) [1]. Some other approaches are based on detecting the facial landmarks from the detected face and on measuring the distance between some of these landmarks. Recently, Deep Learning approach and Convolutional Neural Networks (CNNs) have been proposed to address the face verification problem, such as [2]. Facial landmarks are particularly useful when forensics cases have to be discussed in court since they provide objective measures that can be presented to discuss face verification. However, as we will show in the paper, face verification with distances of automatically extracted facial landmarks, is outperformed by methods based on Deep Learning. Facial landmarks should be used after verification is executed using Deep Learning approaches, to provide objective motivation to the decision.

In this paper, we compare the results of performing the face verification with facial landmarks and a Deep Learning based approach. We validated our comparison by analyzing some videos taken in a real-scenario by surveillance cameras placed in the Instytut Ekspertyz Sdowych in Krakow [3]. To this purpose, we used the Labeled Faces in the Wild (LFW) dataset [4] as confusion dataset. In particular, we used the faces detected in these videos as queries to perform a Nearest Neighbor (NN) search with a joined dataset comprising both LFW and the test set videos, in order to classify the persons according to their face similarity.

The rest of the paper is organized as follows: Section II gives a brief overview of the current approaches to the face verification problem. In Section II-A, we describe the features obtained from the facial landmarks that we analyze and compare in this work. In Section II-B, we present the deep feature that we compare to the facial landmarks features. Section III presents an analysis on some of the facial landmarks features and the experiments on the accuracy of all the features considered. Finally, Section IV concludes this work.

II. FACE VERIFICATION

The use of face information to verify the identity of a person is a research area experiencing rapid development, thanks to recent advances in deep learning. This approach falls under the umbrella of the more general identity verification problem [5]. Among the various types of facial information that can be used a fairly obvious one is that coming from the facial landmarks [6]–[9]. Deep Features learned from convolutional networks have shown impressive performance in classification and recognition problems. For instance, 99.77% accuracy of LFW under 6.000 pair evaluation protocol has been achieved by Liu at al. [10] and 99.33% by Schroff et al. of Google [11]. As in our proposed approach, approximate nearest neighbor search methods can be used to improve scalability and works very well as a lazy learning method [12], [13] and also a full-text search engine [14].

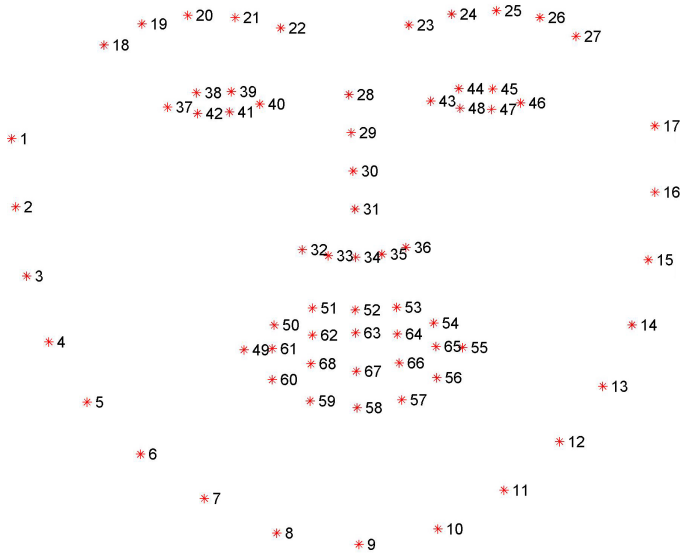
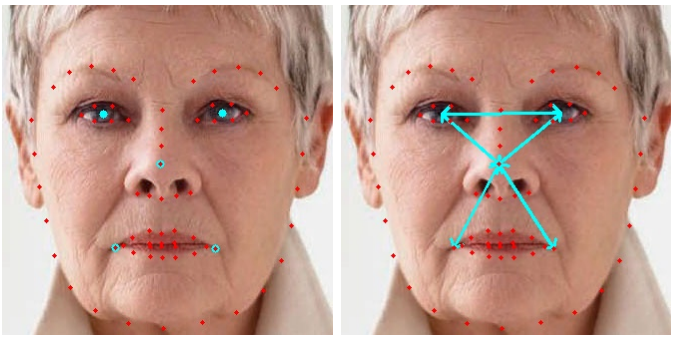


Figure 1. 68 facial landmarks.

 ISBN: 978-1-61208-627-9



(a) Selected 5 nodal points. (b) 5 nodal points distances.

Figure 2. Nodal points and distances used to build the 5-points features.

A. Facial Landmarks Features

Facial landmarks are key points along the shape of the detected face and they can be used as face features to perform several tasks like improve face recognition, align facial images, distinguish males and females, estimate the head pose, and so on.

Key points from landmarks are rarely used as a representation of face verification tasks, typically *facial nodal points* are used instead. As nodal points, we can either use directly some of the facial landmarks or we can compute some new points starting from the facial landmarks. For example, the eyes, the nose, and the mouth are very representative parts of a person's face, so points relative to these parts of the face can be relevant to represent that face. In particular, for example, for the eyes, we can use the centroid of the eye instead of using the facial landmarks that constitute the contour of the eye.

In order to perform the face detection and to extract the facial landmarks from an image, we used the dlib library [15]. In particular, the face detector is made using the Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid, and sliding window detection scheme. The facial landmark detector is an implementation of the approach presented by Kazemi et al. in [16]. It returns an array of 68 points in form of (x,y) coordinates that map to facial structures of the face, as shown in Figure 1. The computational time for extracting the facial landmarks from the image reported in Figure 2 on a MacBook Pro 2013 with an i7 Intel Core 2.5 GHz is about 70 ms.

The distances between nodal points and facial landmarks can be used to build a feature of the face that can be compared with other faces features. In particular, we computed three features based on the distances between nodal points and facial landmarks: the *5-points* feature, the *68-points* feature and the *Pairs* feature. All the distances used to compute these features are normalized to the size of the bounding box of the face. In particular, each distance is divided by the diagonal of the bounding box.

1) 5-points feature: In order to build the 5-points feature, we used five specific nodal points: the centroids of the two eyes, the center of the nose, and the sides of the mouth. The centroids of the two eyes are computed from the six facial landmarks for each eye returned by the dlib library. For the nodal points of the nose and of the mouth, instead, we used directly some of the facial landmarks, respectively the landmark #31 for the nose and the landmarks #49 and #55 for the sides of the mouth (see Figure 2(a)). We used these nodal points to compute the following 5 distances (see Figure 2(b)):

- left eye centroid - right eye centroid
- left eye centroid - nose
- right eye centroid - nose
- nose - left mouth
- nose - right mouth

This produces a 5-dimensional float vector that we used as 5-point feature of the face.

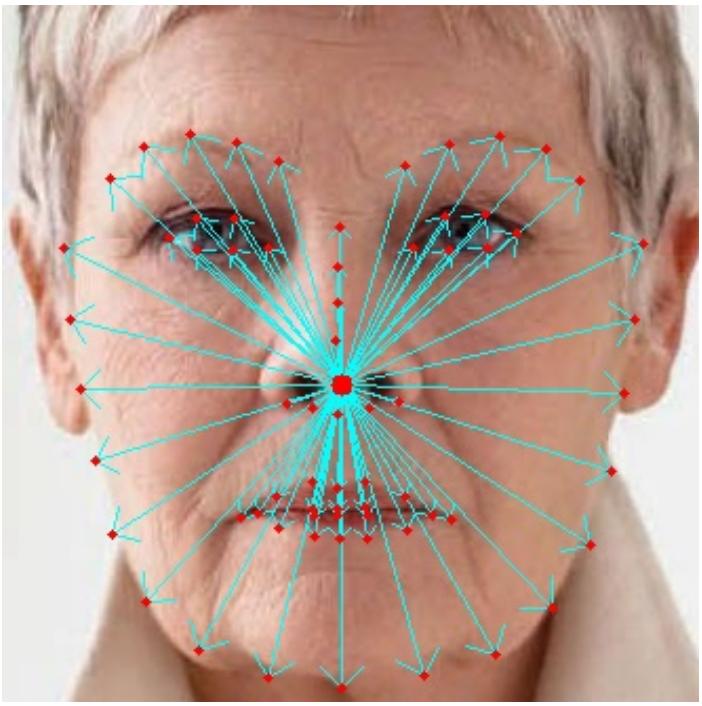


Figure 3. Distances from the centroid of the face to all 68 facial landmarks, used to build the 68-points features.

2) 68-points feature: For the 68-points feature, we computed the centroid of all the 68 facial landmarks returned by the dlib library and we computed the distance between this point and all the 68 facial landmarks (see Figure 3). This produces a 68-dimensional float vector that we used as 68-feature of the face.

3) Pairs feature: The pairs feature is obtained by computing the distance of all unique pairs of points taken from the 68 facial landmarks computed on the input face, as suggested in [9]. This produces a vector of 2.278 float distances that we used as Pairs feature of the face.

B. Deep Features

Deep Learning [17] is a branch of machine learning that uses lots of labeled data to teach computers how to perform perceptive tasks like vision or hearing, with a near-human level of accuracy. In particular, in computer vision tasks, CNNs are exploited to learn features from labeled data. A CNN learns a hierarchy of features, starting from low level (pixels), to high level (classes). The learned feature is thus optimized for the task and there is no need to handcraft it. Deep

 ISBN: 978-1-61208-627-9

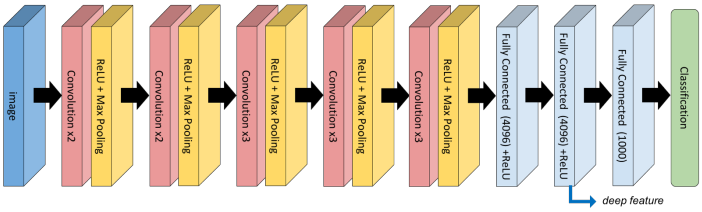


Figure 4. Structure of the VGG-Face CNN used to extract the deep features.

(a) Sample from P1-video2.

(b) Sample from P1-video3.

Figure 5. Samples of videos for Person1.

(a) Sample from P2-video1.

(b) Sample from P2-video2.

(c) Sample from P2-video3.

Figure 6. Samples of videos for Person2.

Learning approaches give very good results in executing tasks like image classification, object detection and recognition, scene understanding, natural language processing , traffic sign recognition, cancer cell detection and so on [18]–[21].

However, CNNs are good not only for classification purposes. In fact, as said before, each convolutional layer of a CNN learns a feature of the input image. In particular, the output of one of the bottom layers before the output of the network, is, in fact, a high-level representation of the input image, that can be used as a feature for that image. We call *deep feature* this representation of the image. This feature can be compared to other deep features computed on other faces, and close deep features vectors mean that the input faces are semantically similar. Therefore, if their distance is below a given threshold, we can conclude that the two faces belong to the same person.

For this work, we used the VGG-Face network [2] that is a CNN composed of 16 layers, 13 of which are convolutional. We took the output of the fully connected layer 7 (FC7) as deep feature, that is a vector of 4.096 floats (see Figure 4). The computational time for extracting the deep feature from the image reported in Figure 2 on a MacBook Pro 2013 with an i7 Intel Core 2.5 GHz is about 300 ms, that is four times the time needed to extract the facial landmarks from the same image.

III. Experimental Evaluation

In this section, we describe the experiments performed to compare the accuracy of the different features described in Sections II-A and II-B in performing the face verification task. We first describe the test set used in our experiments, that is constituted by six videos acquired by surveillance cameras deployed in some of the corridors of the Instytut Ekspertyz Sdowych in Krakow and by the famous face dataset LFW, that we used as confusion set. We then present an analysis of the distances computed over the facial landmarks and, finally, we report some accuracy results obtained by our experiments on the considered features.

A. Test set

We used six videos as test set, provided by the EU Framework Programme Horizon 2020 COST Association COST Action CA16101 [22]. These videos are taken from three different surveillance cameras deployed in the Instytut Ekspertyz Sdowych in Krakow and they capture two different persons (we call them ”Person1” and ”Person2”). Each of them is recorded in all the environments where the cameras are installed. So, we have three videos for Person1 and three videos for Person2. For each video, we analyzed each frame independently. In particular, for each frame, we executed the face detection phase, and for the frames where a face has been detected, we executed the facial landmarks detection algorithm. We then computed the 5-points, 68-points and Pairs features, by exploiting the 68 detected landmarks.

The videos used in our experiments are very challenging because the resolution is low (768x576), and the person is in the foreground of the scene. We have obtained 59 total frames containing faces in all the six videos, that are composed as follows:

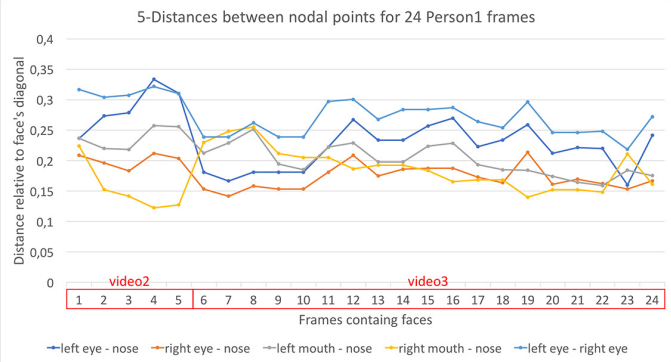
- Person1 (P1):
  - video1: 0 faces detected (the face was never recorded clearly in the video);
  - video2: 5 faces detected;
  - video3: 19 faces detected;
- Person2 (P2):
  - video1: 5 faces detected;
  - video2: 16 faces detected;
  - video3: 14 faces detected;

Figures 5 and 6 show some samples of the Person1 videos and Person2 videos, respectively.

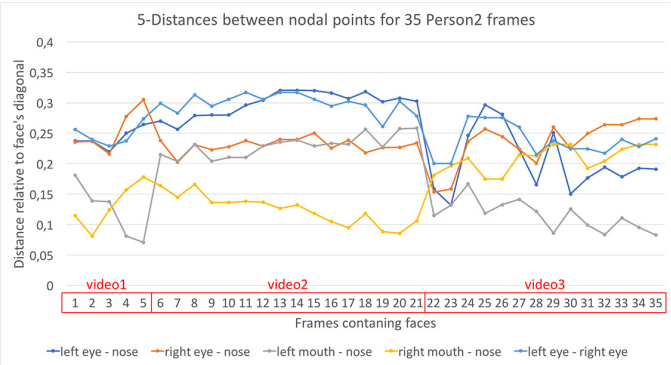
B. Facial landmarks distances measurements

In order to understand if there is a way to better exploit the distance between facial landmark points, we have performed an analysis and computed some measurements on the distances between 5 nodal points and on the distances between the 68 facial landmarks and the centroid, in different frames collected by the sample videos that we used as test set.

 ISBN: 978-1-61208-627-9

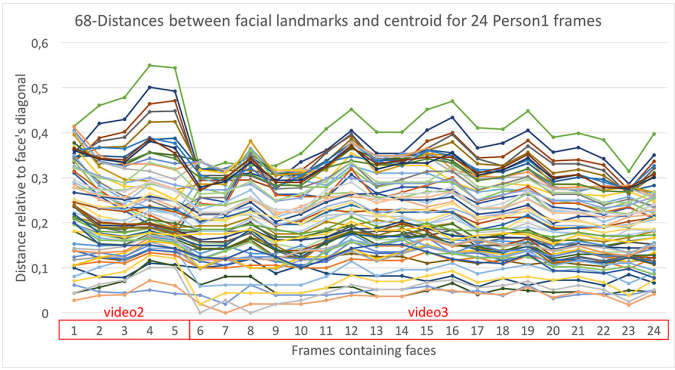


(a) 5-points features for Person1 videos.

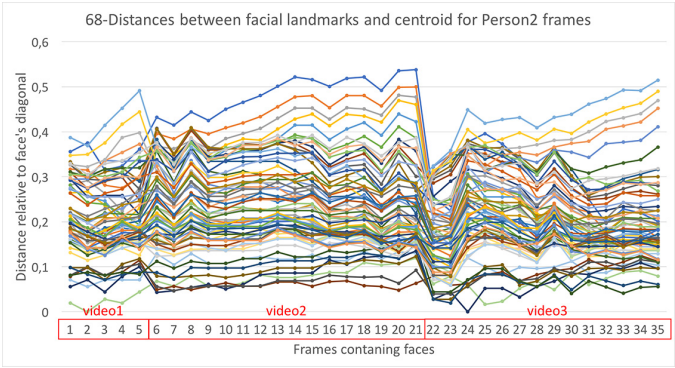


(b) 5-points features for Person2 videos.

Figure 7. Distances between the 5 nodal points in different frames of Person1 (a) and Person2 (b) videos.



(a) 68-point features for Person1 videos.



(b) 68-points features for Person2 videos.

Figure 8. Distances between the 68 facial landmarks and the face centroid in different frames of Person1 (a) and Person2 (b) videos.

Figures 7 and 8 show, respectively, the trend of the components of the 5-points and 68-points features in different frames of the videos, for both persons. Please, recall that Person1 face has been detected in just two videos, while Person2 face has been detected in all three videos. It is possible to notice that, for frames of the same video, the lines of the distances are quite regular, while they have a great difference when moving to another video. This shows that, while a person is seen by the same camera, with the same angle of view, it is possible to use the distance of facial landmarks to recognize a person by its face with good accuracy.

We also computed the average and the variance of the distances between nodal points and facial landmarks reported in Figure 9. In particular, Figure 9(a) reports the average and the variance of the distances between the 5 nodal points and Figure 9(b) reports the average and variance of the distances between the centroid of the face and the 68 facial landmarks. In both cases, the average and the variance are computed on the distance of the same pair of points in all the different frames of Person1 and Person2 videos. The figure shows that the variance is very small in almost every pair of points, and also that the average value of the two persons is quite different in four of five pairs of the nodal points (Figure 9(a)) and in lots of 68 facial landmarks (Figure 9(b)). This means that, by analyzing consecutive frames of a video, when this is feasible, it is possible to increase the possibility to recognize a certain person by using the distance of the facial landmarks.

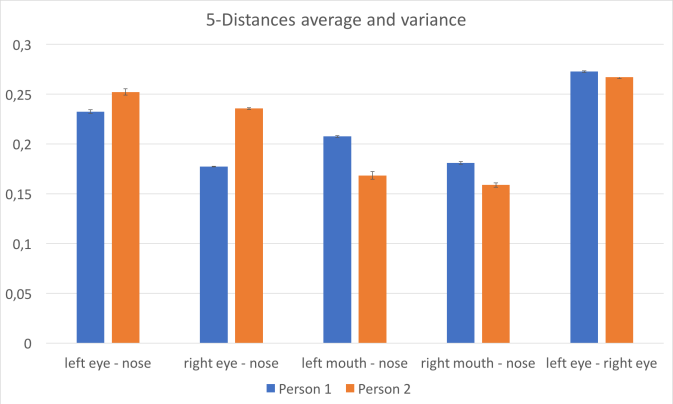
C. Classification Accuracy

We performed some experiments to compare the accuracy in performing the face verification task by using the four different features described above. To this purpose, the faces extracted from the videos were merged with LFW, that has been used as distractor.

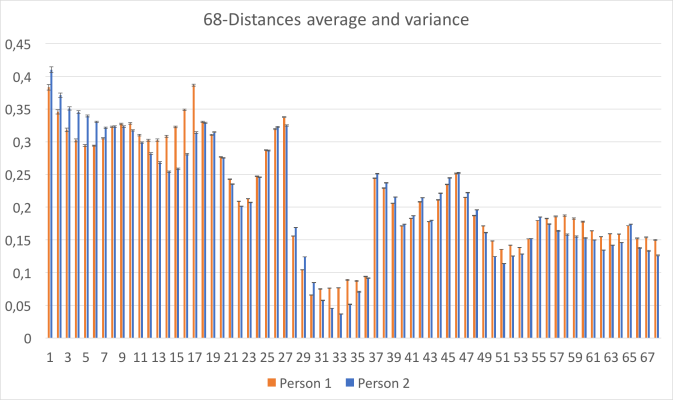
LFW is a very famous face dataset, which contains around 13 thousand faces and 5.750 different identities. All images in LFW are 250x250 pixels and the face is aligned to be in the center of the image. However, there is a lot of background in the images, sometimes capturing also other people faces. This could lead to multiple face detection. Therefore, we cropped each image in the LFW dataset to the size of 150x150 pixels, by keeping the same center, in order to cut the background and avoid multiple face detection. In this case also, we performed the face detection and we computed the facial landmark points by using the dlib library (Figure 10 shows some examples of LFW faces with facial landmarks highlighted). We merged the LFW dataset with the 59 faces that we detected in the test videos and we created a unified dataset. We then extracted the four different features (5-points, 68-points, Pairs and deep features), from all the faces in the new dataset.

We used each of the faces detected in the test set videos as a query for a NN search in the unified dataset. We used the Euclidean distance as dissimilarity measure between features and sorted the entire dataset according to this distance with the given query, from the nearest to the farthest. We discarded the first result of each query since it is the query itself.

 ISBN: 978-1-61208-627-9



(a) 5- points feature average and variance for Person1 videos.



(b) 68-points feature average and variance for Person2 videos.

Figure 9. Average and variance of the 5 and 68 distances for Person1 (a) and Person2 (b).

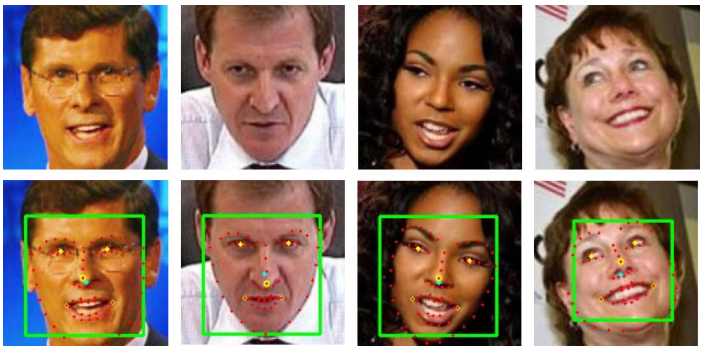


Figure 10. Some examples from LFW dataset and the corresponding detected faces with facial landmarks.

Figure 11 reports some query examples with the Top5 results, for all the features analyzed. For each feature, we report the best and the worst result, in which the biggest number of, respectively, correct and wrong matches in the first five results is obtained. The best result of 5-points feature only got three correct matches in the Top5 results, while all the other features got all correct matches in the Top5 results. The worst result is the same for all the facial landmarks features, that is no correct match in the Top5 results. On the other hand, the deep feature worst result only has one wrong match, that is ranked in the last of the Top5 results.

The different size of the faces detected in the videos is due to the different size of the bounding box of the face computed by the face detector library. This is caused by the different position of the person in the scene with respect to the camera; a bigger face means that the person is closer to the camera.

TABLE I. MEAN AVERAGE PRECISION COMPUTED FOR ALL THE FOUR DIFFERENT FEATURES.

Feature	*mAP*
5-points feature	0.03
68-points feature	0.06
Pairs feature	0.07
Deep feature	0.81

We compared all the four different features by computing the mean Average Precision (mAP) on the results of the queries, so we measured how well the results are ordered according to the query. In particular, for each query, we sum the number of correct results, weighted by their position in the result set, and we divide this value by all the correct elements in the dataset. We then average the precision of all queries, thus obtaining the mean Average Precision for each feature.

The results are reported in Table I. They show that the 68-points feature is two times better than the 5-points feature, and the Pairs feature slightly improves the 68-points feature result. However, the deep feature is more than one order of magnitude better than all the features based on the facial landmarks.

TABLE II. TOP1 AND TOP5 ACCURACY COMPUTED FOR ALL THE FOUR DIFFERENT FEATURES.

Feature	*Top1*	*Top5*
5-points feature	24%	47%
68-points feature	51%	76%
Pairs feature	64%	78%
Deep feature	97%	98%

We also computed the Top1 and Top5 accuracy for all the features considered. The Top1 accuracy counts the percentage of queries in which the first person of the result set is the same person of the corresponding query. The Top5 accuracy considers the first five persons of the result set to check if the correct one is present. Table II shows that 5-points feature works very bad in this scenario with small and low-resolution faces with a Top1 accuracy of only 24% and a Top5 accuracy of 47%. The 68-points feature and the Pairs features, improve the Top1 accuracy of more than twice with respect to the 5-points feature, and up to 78% in case of the Top5 accuracy. Also in this case, however, the deep feature works much better obtaining a 97% Top1 accuracy and a 98% Top5 accuracy.

The facial landmarks have indeed the property of being an accepted proof in trials, and they can be used to classify people in some conditions and with a certain accuracy; they are also faster to be computed with respect to the deep features. However, the deep feature shows much better performance, especially in challenging scenarios with low-resolution faces.

IV. CONCLUSION

In this paper, we presented a comparison between facial landmarks and deep learning approaches in performing the

 ISBN: 978-1-61208-627-9

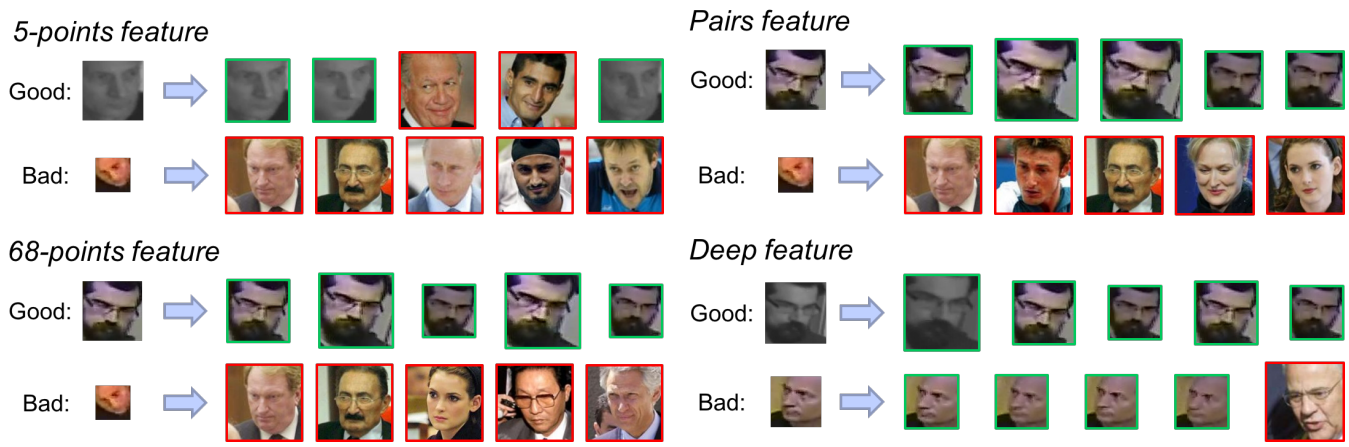


Figure 11. Query examples for the all the kinds of features, with Top5 results. For each feature, the best and the worst results are reported.

face verification task. Facial landmarks are very important in forensics because they can be used as objective proof in trials. We performed our experiments on videos taken in a real scenario and by exploiting the widely used face dataset LFW. Results show that the accuracy of the deep features in verifying whether a face belongs to a given person is much greater than the one of facial landmarks based approach. On the other hand, the deep learning results cannot be used as proof in court. We think, however, that deep features approach should help the forensics process along with facial landmarks. In particular, the latter should be used after the face verification has been executed with deep features, in order to provide an objective measure for the decision.

ACKNOWLEDGMENTS

This work has been partly funded by the "Renewed Energy" project of the DIITET Department of CNR and by the EU Framework Programme Horizon 2020 COST Association COST Action CA16101. Special thanks to Prof. Dariusz Zuba and the Instytut Ekspertyz Sdowych in Krakow for the videos used as test set.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 12, 2006, pp. 2037–2041.

[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in British Machine Vision Conference, 2015.

[3] "Instytut ekspertyz sdowych - krakow," http://ies.krakow.pl/, accessed: 2018-04-13.

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[5] P. Verlinde, G. Chollet, and M. Acheroy, "Multi-modal identity verification using expert fusion," Information Fusion, vol. 1, no. 1, 2000, pp. 17–33.

[6] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in Advances in neural information processing systems, 2014, pp. 1988–1996.

[7] C. Sanderson, M. T. Harandi, Y. Wong, and B. C. Lovell, "Combined learning of salient local descriptors and distance metrics for image set face verification," in Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on. IEEE, 2012, pp. 294–299.

[8] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 365–372.

[9] A. G. Rassadin, A. S. Gruzdev, and A. V. Savchenko, "Group-level emotion recognition using transfer learning from face identification," arXiv preprint arXiv:1709.01688, 2017.

[10] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," arXiv preprint arXiv:1506.07310, 2015.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

[12] J. Park, K. Lee, and K. Kang, "Arrhythmia detection from heartbeat using k-nearest neighbor classifier," in Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on. IEEE, 2013, pp. 15–22.

[13] D. Wang, C. Otto, and A. K. Jain, "Face search at scale," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, 2017, pp. 1122–1136.

[14] G. Amato, F. Carrara, F. Falchi, and C. Gennaro, "Efficient indexing of regional maximum activations of convolutions using full-text search engines," in Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. ACM, 2017, pp. 420–423.

[15] "Dlib library," http://dlib.net/, accessed: 2018-04-13.

[16] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.

[17] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, 5 2015, pp. 436–444.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, 2016, pp. 142–158.

[20] G. Amato, F. Falchi, and L. Vadicamo, "Visual recognition of ancient inscriptions using convolutional neural network and fisher vector," Journal on Computing and Cultural Heritage (JOCCH), vol. 9, no. 4, 2016, p. 21.

[21] G. Amato, F. Carrara, F. Falchi, C. Gennaro, C. Meghini, and C. Vairo, "Deep learning for decentralized parking lot occupancy detection," Expert Systems with Applications, vol. 72, 2017, pp. 327–334.

[22] "Eu framework programme horizon 2020 cost action ca16101," http://www.cost.eu/COST_Actions/ca/CA16101, accessed: 2018-04-13.

 ISBN: 978-1-61208-627-9

Emotion-Aware Design Image Recommendation using Color Image Scale

Dongwann Kang

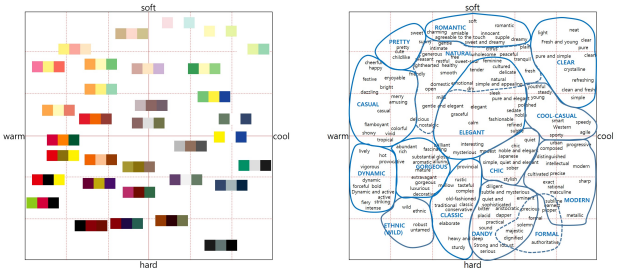
Faculty of Science and Technology
Bournemouth University
Poole, Dorset, BH12 5BB, United Kingdom
Email: dkang@bournemouth.ac.uk

Kyunghyun Yoon

School of Computer Science and Engineering
Chung-Ang University
Seoul, 06974, Korea
Email: khyoon@cau.ac.kr

***Abstract*—Color is one of the visual elements that psychologically affect people's emotion. Although there are slight differences based on culture, several studies in color psychology have found that most single colors generally have meaning or emotion. Therefore, most professional designers use colors in their works to express the emotion. In this paper, we present a novel method that recommends design images using a color combination based on the relation between color and emotion. To achieve this, we estimate emotion based on the color image scale, which is a famous color theory in the field of design, and recommend design images according to the emotion.**

Keywords–color image scale; emotion; design image; color combination.



(a) Three-color combinations (b) The emotion keywords

Figure 1. Three-color combinations and emotion keywords on color image.

I. INTRODUCTION

Color is a visual element that psychologically affects people's emotion. Generally, it is known that single colors have their own meaning or emotion [1]. In addition, the combination of colors also significantly affects emotion [2]. Many people apply these principles knowingly or unknowingly in their daily life, for example, to coordinate clothes, select furniture color, etc.

The color image scale [3] [4] is a theory studied by Shigenobu Kobayashi at Nippon Color & Design Research Institute. In their psychophysical research, they presented over 1000 color combinations to express any emotion, taste, or lifestyle that belongs to 174 semantic keywords on the emotion perceived from color. They labeled each combination of three colors with one of 174 keywords. In addition, they devised a two-dimensional emotion space, the color image scale, which consists of two axes that correspond to the scales cool-warm and soft-hard. On the color image scale, they located every keyword according to its two scales measured by several studies. Figure 1(a) presents the concept that illustrates several examples of three-color combinations, along with their keywords, plotted in the color image scale [4]. In this scale, they also defined 15 categories such that each keyword belongs to one of the categories.

Most professional designers also reflect these principles in their works. To convey intended emotion, they intuitively employ the color combination in their design. At this time, color combinations which are used for an emotion can be different to each other, because it is known that there are lots of available combinations for an arbitrary emotion. Consequently, the colors used in the works depend on designers' knowledge and experiences, so that it is not an easy task for non-experts and beginners to use colors appropriately according to their emotion.

In this paper, we present a novel method that recommends design images using the emotion estimated from images based on the color images scale theory. We establish an emotion prediction model using a machine learning technique. For this, we find the relationship between the emotion and the properties of the color combination in the color image scale. Then, we extract the main colors from the image. Finally, we estimate the emotion of the image via the properties of the main colors extracted. Once the prediction model is ready, any other knowledge is not required in our method, and design images are recommended according to input emotion.

The remainder of this paper is organized as follows. In Section II, we explain our approach for establishing emotion prediction model from color combinations. Then, we present our method for estimating emotion by extracting color combinations from image in Section III. In Section IV, we demonstrate the results of our proposed method and discuss the algorithm used and its limitations. Finally, we conclude this paper in Section V with a summary of our ideas and outline of future work.

II. ESTABLISHING EMOTION PREDICTION MODEL FROM THREE-COLOR COMBINATIONS

To estimate an emotion from an image, we use the three-color combinations surveyed by Kobayashi [4]. His research provides such combinations tagged as the name of the emotion, and thus we can estimate emotion from an image by extracting

 ISBN: 978-1-61208-627-9

(a) (b) (c)

Figure 2. Extracting three-color combinations from input image. (a) Input image, (b) normalized image, and (c) top three colors frequently used.

a color combination from the image. Kobayashi's research also provides the name of each color combination and the emotion position in the color image scale. Because the positions of the emotion keywords are graphically represented in the work [4], we estimate the position by acquiring the centre position of the text in the graph (Figure 1(b)). Consequently, we obtain three-color combinations that include the name of three colors and of the emotion tagged on the combinations, and the emotion position in the color image scale.

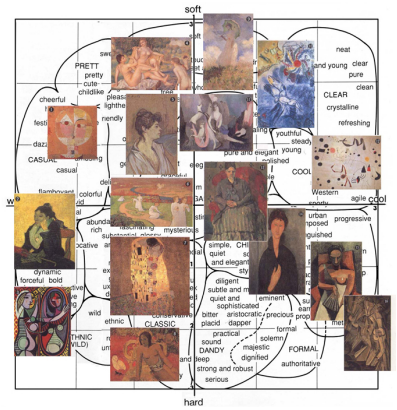
Kobayashi's work [4] did not cover all possible three-color combinations. Thus, estimating an emotion from random color combination is important to find the relationship between each color in his three-color combination. To estimate such relationship, we employ a machine learning technique. First, we extract features from the colors in the combination, such as the hue/saturation/luminance difference between two colors, and the average hue/saturation/luminance value of three colors. Consequently, we obtain a 12-dimensional feature for each three-color combination. Next, we generate data pairs with the features and two-dimensional position of the emotion tagged on the data, three-color combination. Finally, we acquire a prediction function that estimates the emotion coordinates from the random three-color combination using linear regression [5].

For our experiment, we used 936 three-color combinations and 174 emotions. To ignore the order of the colors in the combination, we generated all possible combinations from the given 936 three-color combinations, such that six combinations are generated from each three- color combination. The range of both coordinates in the color image scale is $[-3 : +3]$. In our experiment, the prediction error magnitude was recorded at 0.64. In our analysis, the significant factors seem to be average (avg.) hue, hue difference, avg. saturation, and intensity.

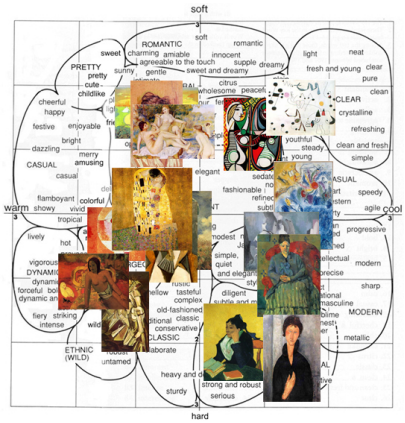
III. Estimating Emotion by Extracting Color Combinations from Image

Once the model for predicting emotion from color combinations is established, it is enabled to estimate the emotion of image by using the color combination of image. In this study, we assume that the three colors used predominantly in an image affect human emotion similarly to three-color combinations. Therefore, we use the three colors most frequently used in an image to estimate emotion.

In general, digital color images have 24-bit depth color. There are too many discrete colors in an image, and thus finding the most frequently used colors is not meaningful. For this reason, we normalize an image by enforcing a limited number of colors. Kobayashi used Hue & Tone 130 system in



(a) Kobayashi's ground truth [4]



(b) Our results

Figure 3. Ground truth color image scale of 16 images used by Kobayashi and our results.

[3] to construct the image scale of three-color combinations, and thus we normalize image colors using the same color system (Figure 2).

After normalizing the colors, we estimate the emotion coordinates in the color image scale of an image using the prediction function described in Section II. Kobayashi showed the coordinates of 16 famous painting images in [4] (Figure 3(a)). Similarly to the emotion names, we acquire the image coordinates by calculating the centre position of each image. For 16 images with ground truth emotion, we estimate emotion as the coordinates in the color image scale (Figure 3(b)). In our experiment, the mean error magnitude was recorded at 2.08.

We then recommend several images of which emotion estimated by our method is closer to the input emotion on the color image scale.

IV. Experimental Results

For machine learning methodology, we used linear regression by using Weka library [5]. We evaluated our prediction performance by using 10-fold cross validation. Figure 4 shows recommended images of given emotion keywords on proposed method.

To evaluate our emotion estimation as described in Section III, we gathered ground truth data of experimental images

 ISBN: 978-1-61208-627-9

(a) ‘provincial’

(b) ‘simple and elegant’

(c) ‘mysterious’

Figure 4. Recommended images of given emotion keywords.

using a Crowdsourced user study, Amazon mTurk [6]. The ground truth annotations of 47 design images were generated by aggregating the study participants’ labels over each image. Figure 5 shows a sample question for labeling the color image scale of a given image. For each image, we asked over 50 participants to select a degree of the two factors, warm-cool and soft-hard, considering color and tone only.

After obtaining the ground truth color image scale of 47 images, we evaluated the performance of our emotion estimation algorithm. In our experiment, the mean errors for warm-cool and soft-hard were measured by calculating the distance between the ground truth and estimated emotion coordinates, and the corresponding values are 0.13 and 0.21.

In our experiment, the performance of emotion estimation from an image is worse than that of the emotion estimation from the three-color combination. In general, digital image colors for the same image differ slightly from each other according to image format and compression rate. Therefore, the prediction performance depends mainly on the color of the image.

V. Conclusion

In this paper, we proposed a novel method that recommends images based on the emotion estimated from the image. For this, we established emotion prediction model by using the color image scale, a well-known theory in design fields, and estimated the emotion of image using top three colors and the model. Then we recommended images of which estimated emotion was closer to input emotion on color image scale. In addition, we conducted crowdsourced user study to evaluate our results.

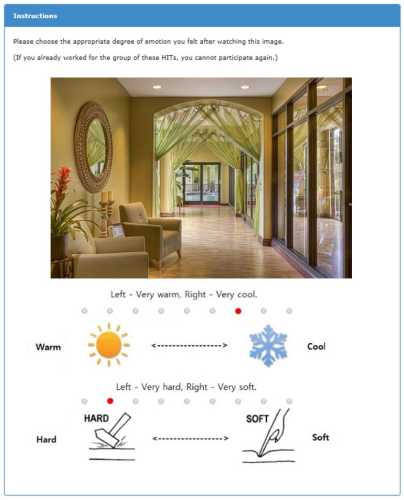


Figure 5. Sample question for labeling color image scale of given image.

Our experiment mainly depended on Kobayashi’s research. Moreover, we obtained the three-color combination from images by naïvely extracting top three colors frequently used; therefore, there is no guarantee that the extracted three-color combination successfully represents the image. Also, it is known that human emotions affected by color can be altered based on era and culture. Consequently, a more robust approach for estimating human emotion is required in our future work.

In this paper, we consider only color. However, the factor that affects the emotion of images is not only color. In our future work, we will study other factors that can affect the emotion of images, such as composition and texture, and improve our emotion estimation by employing these factors.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. NRF-2017R1A2B4007481).

References

[1] B. Wright and L. Rainwater, “The meanings of color,” The Journal of General Psychology, vol. 67, no. 1, 1962, pp. 89–99, pMID: 14008415.

[2] L. Sivik, “Research on the meanings of color combinations,” in Proceedings of the Congress of the Association Internationale de la Coleur (AIC), 1989, pp. 130–132.

[3] S. Kobayashi, “The aim and method of the color image scale,” Color Research & Application, vol. 6, no. 2, 1981, pp. 93–107. [Online]. Available: http://dx.doi.org/10.1002/col.5080060210

[4] ——, Color Image Scale. Kosdansha International, 1991.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: An update,” SIGKDD Explor. Newsl., vol. 11, no. 1, Nov. 2009, pp. 10–18.

[6] P. G. Ipeirotis, “Analyzing the amazon mechanical turk marketplace,” XRDS, vol. 17, no. 2, Dec. 2010, pp. 16–21. [Online]. Available: http://doi.acm.org/10.1145/1869086.1869094

 ISBN: 978-1-61208-627-9

Automatic Electronic Organ Reduction Using Melody Clustering

Daiki Tanaka

Graduate School of Computer and Informaion Sciences
Hosei University
Tokyo, Japan
Email: 17t0016@cis.k.hosei.ac.jp

Katunobu Itou

Faculty of Computer and Information Sciences
Hosei University
Tokyo, Japan
Email: itou@hosei.ac.jp

***Abstract*—Reduction is a method for arranging the scores of a multipart composition to its ensemble. In this study, we propose an arrangement system using the full score to automate the arrangement by reduction. Our target instrument for this study is an electronic organ, which has a score similar to the score of an ensemble. First, we performed clustering based on the rhythm, melodic activity, harmony, sonic richness, and timbre of the instruments to reduce the part number of the melody in the full score. Further, we selected clusters of melodies corresponding to the right-hand, left-hand, and foot parts of the electronic organ. Finally, the musical score was rectified to be able to play the electronic organ. This system was evaluated using four songs. The average values of the right-hand, left-hand, and foot parts were 0.80, 0.71, and 0.77, respectively. These results depicted that the proposed arrangement was suitable for preparing electronic organ scores from the full score.**

Keywords–Music; Arrangement; Reduction; Clustering; Electronic Organ.

I. INTRODUCTION

Piano and electronic organs can play orchestral songs and can be used to express the orchestra performance. However, all songs do not have musical scores. Therefore, we proposed a system that can automatically generate music scores for different songs.

Several studies have been performed on the piano music arrangement; however, the electronic organ music arrangement has not been much studied. Electronic organ can be used to play a wide range of sounds and rhythms. Nonetheless, the number of electronic organ notation is less than the number of piano score. Therefore, this study aims to focus on electronic organ music arrangement using an arrangement method known as reduction. In reduction, the full score (shown in Figure 1) composed of multiple parts is contracted to the main elements by reducing or eliminating the melody. When arranging by reduction, the system must consider the melodies that it must eliminate. These melodies change depending on the instruments for which the music is being arranged; therefore, it is difficult to definitely set any criteria. For example, the piano score comprises two parts: the right and left hands. To reduce a piano score, we can omit a melody like the main melody and the accompaniment. For instruments like the guitar, we eliminate the chord melody. In this way, reduction is performed based on the characteristics of the musical instrument whose scores have to be arranged.

The remainder of the paper is structured as follows. Section II introduces the characteristics of electronic organ and explains the related works. Section III introduces the arrangement method of electronic organ score. Section IV gives the test results of arrangement system. Section V discusses the results. Conclusions are given in Section VI.

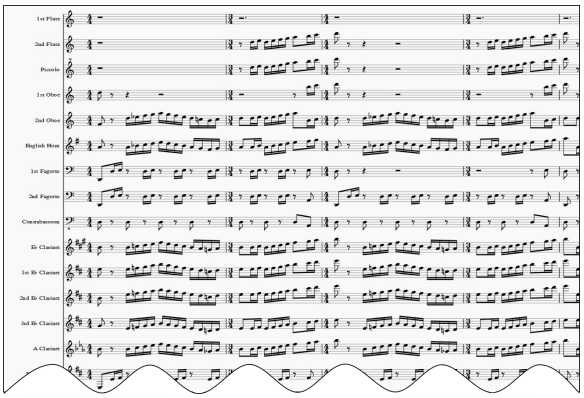


Figure 1. A Multipart Musical Score

II. ELECTRONIC ORGAN REDUCTION

A. Characteristics of the electronic organ

The electronic organ is played using the right and left hands, whereas an electronic organ is played using the right hand, the left hand, and a foot. Additionally, the piano has one keyboard, but the electronic organ has three keyboards. Therefore, the score for an electronic organ is a three-set score (e.g., see Figure 2).

The right hand of the melody often constitutes the main melody. The left hand of the melody often contains the chords with sound numbers of four or less. Foot melodies mainly comprise the root of the chord and are always single notes. The pitch is observed to initially reduce for the right-hand melodies, then for the left-hand melodies, and finally for the foot melodies. When the right-hand and left-hand melodies coincide in time, their range is within one octave.

As a major feature of an electronic organ, it is possible to reproduce timbres for approximately 900 types of musical instruments. Therefore, different timbres overlap on a single keyboard. Additionally, it is possible to switch timbres automatically; therefore, we can continue playing the melodies of different instruments.

Electronic organs contain rhythm boxes, such as percussion instruments. The music rhythms are prepared in advance, and the player plays the organ using this rhythm.

 ISBN: 978-1-61208-627-9

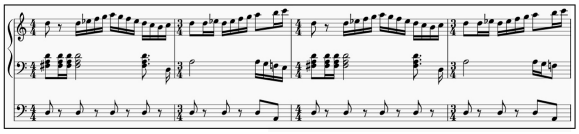


Figure 2. Electronic Organ Score

B. Related Work

Using reduction, the scores of a multipart music can be condensed to the musical score of a single target instrument. Several studies have been performed on reduction by automatic arrangement. Fujita [1] proposed a method for creating piano notations using multiple parts. The method focuses on the average pitch, pronunciation time, pitch, and rhythm pattern and estimates the melody and baseline from the melody of the full score. Additionally, we adopted the melodies for the right-hand and the left-hand parts of the piano. However, we were unable to uniquely determine the melody for the polyphonic songs using this method. Furthermore, the piano does not always continue the melody and baseline at all times. Therefore, this method is insufficient. In this research, we decided to put together the melodies with similar characteristics instead of estimating and scoring the melody. Using these melodies, we chose the melody that sounded similar to that of a piano.

Ito et al. [2] intended to make an ensemble music notation from the full score. First, they performed the clustering of melodies. Further, they eliminated unnecessary melody clusters. However, this clustering did not consider the musical point of view because the researchers considered only the distance of the notes on the score. Matsubara et al. [3] studied clustering in consideration of the musical features. They performed clustering using the features of rhythm, melodic activity, harmony, and sonic richness. In this study, we added a new feature, the timbre quantity, to the clustering proposed by Matsubara et al. Further, using the clustered melodies, we selected a melody cluster that reflected the characteristics of the electronic organ and created an electronic organ score.

Rose Curtis [4] stated that the timbre of an instrument depends on the amplitude envelope of the sound, fluctuations caused by vibrato and tremolo, formant structure, perceived volume, duration, and temporal frequency of the component fluctuation.

Therefore, we decided to focus on the formant structure. The common methods used to analyze the formant structure are the Linear Predictive Coding (LPC) [5], LPC cepstrum [6], cepstrum, and Mel Frequency Cepstrum Coefficient (MFCC) [7]. In this study, we analyzed the sounds of different musical instruments and the sound range that could be produced by each instrument. It was difficult to compare within the same note range. In addition, the sound produced by the musical instrument has the property of a harmonic structure in which the frequency of the integral multiples of the fundamental frequency appears strongly. Thus, the LPC for obtaining the formant structure including the portion where the frequency strongly appears is not suitable here. In this study, we analyze the timbre of musical instruments using MFCC as an analysis method to investigate the formant structure that is not considerably affected by the pitch.

III. ARRANGEMENT METHOD OF ELECTRONIC ORGAN SCORE

This system inputs the full score expressed in the MusicXML [8] format. The system must obtain the instrument name, bar number, octave, note type,pitch, duration, and metrical information from the MusicXML file. Based on this information, we prepared the five features (the pronunciation time, pitch change, harmony, persistent pronunciation pattern, and timbre of the instrument). The full score of the melody was clustered using these features and was further grouped into multiple melody groups having similar characteristics. After clustering, we selected the three groups corresponding to the right hand, left hand, and foot parts of the electronic organ score. Finally, the score was corrected so that it could be played on an electronic organ. The electronic organ score arranged by the system generated the musical score having three parts and was written out in the MusicXML format. We used the MuseScore [9] scorewriter software for creating the MusicXML format.

A. Melody Grouping

1) Features: The melodies are classified based on five features: rhythmic activity, melodic activity, consonance activity, sonic richness, and instrument timbre.

Based on the literature [3], we prepared four features: rhythmic activity, melodic activity, consonance activity, and sonic richness. The timbre feature of the instrument is a newly defined quantity in this research.

Rhythmic Activity: This feature considers the amount of rhythm in the musical score. When the two phrases in a melody are pronounced at the same time, they possess the same rhythm. For example, the rests after the beats in Figure 3(b) generate the same rhythm as depicted in Figure 3(a). As the basic unit, we used the shortest note per unit time. This unit was assigned the value 1 when the note was sounded, and the value 0 was assigned when the note was not sounded. This generated a binary vector RA with n elements.

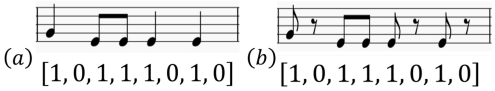


Figure 3. Feature vector of rhythmic activity

Melodic Activity: For most melodies, the pitch varies throughout the melody. Therefore, as the harmony often shares the same movements as the melody, the harmonies and melodies are often considered to be the same phrase. For example, the melodies in Figure 4(a) and (b) can be regarded to be similar melodies with different pitches. In each unit time, a transition between two notes is set to 1 when the note is higher than the preceding note. The transition is set to 0 when the pitch does not change and to -1 when the note is lower than the preceding note. This process generates a vector MA of melodic activity. The vector MA compares each note with its previous neighbor. There is no sound previous to the head sound. Therefore, the comparison starts from the note that immediately follows the head note. This means that the vector is $n-1$, i.e., the number of elements minus one.

 ISBN: 978-1-61208-627-9

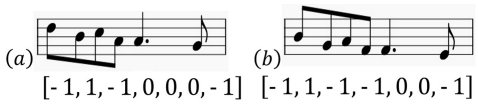


Figure 4. Feature vector of melodic activity

Consonance Activity: Even when the pitches follow similar movements, many non-harmonious sounds can appear in the parts of the melody. These harmony-like accompaniment parts are recognized as different phrases. Therefore, we calculated the top three sounds from all parts of the phrase per unit time and approximately replicated the harmony sound, as depicted in Figure 5(a). In each unit time, we set the inclusion or exclusion of harmonious tones to 1 and 0, respectively, generating a binary vector CA of consonance activity with n elements. Therefore, it is possible to distinguish the melody of Figure 5(c), which contains many melodies and harmonies, from the melody of Figure 5(b), which contains few harmonious sounds.

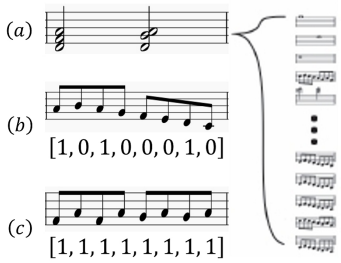


Figure 5. Feature vector of consonance activity

Sonic Richness: This feature quantifies the sound of the instrument. It is possible to distinguish the melody in Figure 6(b), which includes the melody and the rest elements;these elements can be easily distinguished from Figure 6(a) that does not include any rest elements. In each unit time, the presence or absence of sound is assigned as 1 and 0, respectively, generating a binary vector SR of sonic richness with n elements.

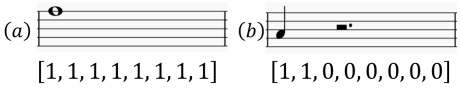


Figure 6. Feature vector of sonic richness

Instrument Timbre: Musical instruments of a similar timbre often play a similar melody. In an electronic organ, you can play with the timbres of multiple instruments stacked on top of each other. Therefore, it is common to play such melodies on the same keyboard. To play a melody, the feature vector IT (Instrument Timbre) is used to measure the similarity between the timbres of each instrument.

First, we modeled the timbres. To extract the features of the timbre, we prepared Musical Instrument Digital Interface (MIDI) sound sources of 29 musical instruments. These instruments are usually used in orchestras and brass performances. The MIDI sounds that we have used were the sounds produced for a time period of approximately 3 s. In addition, the timbre of the instrument could be changed by the pitch using a pitch of approximately two octaves of the note range that were used in the full score.

Further, we considered the section used by the MIDI sound source. The timbre is observed to alter from the beginning to the end of the sound produced by the instrument. Therefore, we decided to use the section corresponding to the Attack-Decay-Sustain-Release (ADSR) envelope [4] separately. ADSR is defined by dividing the amplitude envelope of the time waveform into four sections. The time required for the sound to suddenly move from nil to a peak is known as the attack. Decay is the interval in which the sound decays from the attack level to the designated sustain level. Sustain is the level at which the sound continues for as long as the vocalist's breath continues or while the piano/organ key is pressed. Finally, the release time is the time that is required to decay from the sustain level to zero. In this study, we used 1.5 s as the sustain interval, which does not change much with timbre and is relatively stable as the feature quantity. The formant structure was obtained using a 15-dimensional MFCC having a frame length of 256 points and a shift length of 128 points. This is used as a feature quantity of the timbre model.

We modeled the timbre using the obtained feature quantity. For modeling, we used the mixed Gaussian model called Gaussian Mixture Model (GMM) [10]. We designed the GMM with a mixture ratio of 8.

Further, we measured the similarity of the timbre. The timbre is modeled on the sum of multiple multidimensional normal distributions using GMM. In this study, the similarity of the timbre model is measured by comparing the shape of the probability density function. To calculate the similarity, we used the Kullback-Leibler (KL) divergence [11] for measuring the difference in the probability distribution. When the probability distributions of the timbres of the instruments of parts i and j are P and Q, which are continuous probability distributions, Equation (1) expresses the similarity using the KL divergence. Then, $p(x)$ and $q(x)$ are the probability density functions of the timbre of each musical instrument. A property of the KL divergence is that it is always 0 or more, and it is 0 when the two distributions are equal. The KL divergence is asymmetric (Equation (2)); therefore, we use the average of the KL divergence values in this study.

$$D_{KL}(i,j) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \quad (1)$$

$$D_{KL}(P||Q) \neq D_{KL}(Q||P) \quad (2)$$

The feature quantity of the timbre is a pairwise value between musical instruments. However, the feature quantity defined in [3] is the feature quantity of each musical instrument. Therefore, it is necessary to convert the timbre feature quantity to the feature quantity of each musical instrument. In this study, the value of the KL divergence was converted into the value of each instrument using Multi Dimensional Scaling (MDS) [12]. We converted the divergence to 5-dimensional features. In this study, the IT vector is a normalized vector of these feature vectors.

Figure 7 depicts the IT vector using an isomap in a 2-dimensional musical instrument space; it was visualized to

 ISBN: 978-1-61208-627-9

make the timbre distances of each instrument easier to understand. For example, a brass instrument trumpet and cornet are similar instruments; therefore, they have similar timbres. Therefore, their timbre values are located close to each other. Other instruments having similar timbres are a cello and viola among stringed instruments and an alto and tenor sax among woodwind instruments. Thus, the model created using tone colors can be the correct model.

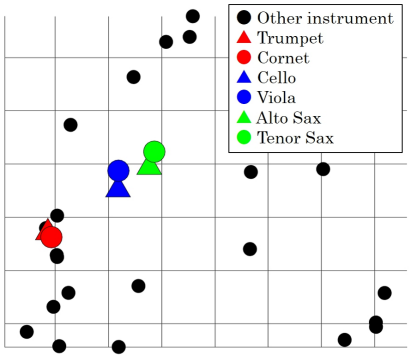


Figure 7. Visualization of the distance of timbre in musical instruments in instrument space

2) Similarity: The scale used for clustering is defined. For distance clustering, we adopted cosine similarity, which groups similar melodies. In the classification, the cosine similarity was subtracted from 1.

The total of all the weights $w_k(k = 1, 2, 3, 4, 5)$ is 1. $D_{RA}(i, j)$ defines the distance between the parts i and j using the feature vector RA. Similarly, $D_{SR}(i, j)$, $D_{MA}(i, j)$, $D_{CA}(i, j)$, and $D_{IT}(i, j)$ defines the distance using the feature vector SR, MA, CA, and IT. We performed clustering using the value of D given in Equation (3).

$$D(i,j) = w_1 D_{RA}(i,j) + w_2 D_{SR}(i,j) + w_3 D_{MA}(i,j) + w_4 D_{CA}(i,j) + w_5 D_{IT}(i,j) \quad (3)$$

3) Melody clustering: We performed clustering by the k-means method using the above five features. The clustering was performed in units of two measures so that the number of simultaneous sounds remained below 5. Melody clustering was implemented as follows.

1) We set the cluster number to $K = 5$, and the part number as N. We acquired the note data for each part in two measures.
2) We prepared the feature vector defined in III-A1 from the note data of each extracted part and applied the k-means method with the cluster number K, using the scale defined by Equation (3).
3) We obtained the clustering results. The clustering terminated when there were four or fewer note names in each group simultaneously or when $K = N$; otherwise, we set $K = K + 1$ and returned to the first step.

The initial value was determined randomly. From the second time onward, the initial value was the value immediately preceding the label number. As the number of clusters differed for each measure, more clusters were assigned than the number of parts. After the experimental exploration, the initial number of clusters was set to 5.

B. Selection of classified groups

Among the classified groups, we identified the melodies corresponding to the right-hand, left-hand, and foot parts of the electronic organ score. We first selected the foot part (the most dissimilar playing method); this was followed by the right-hand and left-hand parts.

The main objective of the foot part was to supplement the whole song with bass melodies. Therefore, in each of the classified groups, melodies with the largest rate of notes with a pitch of C3 or less were selected for the foot part. The rates of bass melodies were calculated by dividing the number of notes below C3 by the number of notes per unit time.

Further, we selected the group corresponding to the right hand. The right hand produces higher sounds than the other keys; therefore, this selection considers the proportion of the treble notes. Furthermore, the right-hand part often controls the melody with a large number of pronunciations (the main melody); therefore, we also considered the rate of the number of pronunciations. The high pitch rate HR(High Rate) was calculated by dividing the number of notes above C5 by the number of notes per unit time. The sound-number ratio PR(Pronunciation Rate) was calculated by dividing the number of sounds in the measures by the duration (when the smallest note was 1). The group with the largest value of these weighted sum was chosen to be the right-hand part.

Finally, we selected the group corresponding to the left-hand part. The left hand often plays a chord; therefore, the group with the highest proportion of chords was selected as this part. The chord rate was calculated by dividing the number of chord sounds by the total number of sounds.

C. Correction based on restriction of electronic organ

The pitches of notes in each group correspond to particular musical instruments. In some cases, these notes cannot be played on the keyboard. Therefore, they must be corrected for compatibility with the electronic organ. To rectify the musical score, we simultaneously examined the right-hand, left-hand, and foot parts in each bar.

We initially rectified the foot-part melody. The foot part cannot play two sounds at the same time; therefore, the appearance of two sounds in the foot part must be corrected. For this purpose, we combined two sounds having the same note name into one sound. Sound combinations are based on the number of pronunciations of the musical instruments. When two notes had different sound names, we retained the sound corresponding to the root sound of the chord and erased the other sounds. When there was no sound corresponding to the root of the chord, we left a melody with a large number of pronunciation instruments. In addition, we modified the pitch range to that of the foot keyboard (i.e., approximately 1.5 octaves from C2 to G3).

Further, we rectified the right-hand and left-hand scores. The lowest and highest notes played at the same time by the right and left hands differed by more than one octave. Similar to the foot part, we rectified the sounds with the same note name if the sounds were observed to correspond to a large number of musical instruments. When two notes possessed

 ISBN: 978-1-61208-627-9

different sound names, they were consolidated such that they never differed more than one octave. The intersection of the left and right hands indicated an infeasible playing situation. In such cases, the sound was consolidated to range from C3 to C4.

IV. EXPERIMENT

A. Experimental method

We evaluated the atmosphere of the performances before and after arranging the music and determined whether the music score was performed by considering the characteristics of the electronic organ. The existing electronic organ score maintained the musical atmosphere and was considered to be easy to play. Therefore, this score was evaluated to be the correct arrangement. On comparing the musical score of the arranged music with that of the correct answer, we were able to evaluate the similarity. In this way, we were able to judge whether the arrangement system was good or bad. Here, the evaluation value was calculated for the right hand, left hand, and foot parts.

We estimated the evaluated value using Equation (4). We subtracted the deletions (i.e., the number of missing notes), insertions (i.e., the number of lacking notes), and substitutions (i.e., the number of mistaken pitches or note names) from the total number of notes in the correct musical score. We divided this value by the total number of notes.

$$\text{The evaluation value} = \frac{\text{The total notes} - (\text{deletions} + \text{insertions} + \text{substitutions})}{\text{The total notes}} \quad (4)$$

The full score that was used for the arrangement is different from the number of parts and measures. We used the following four songs for the evaluation of our model: 1. Ave verum corpus (K.618 v, motet in D major) 2. Pictures at an Exhibition (a suite of 10 pieces having a varied promenade) 3. Salut d'Amour, and 4. Symphony No.36 "Linz" K.425 3rd movement Menuett.

B. Results

Ave verum corpus shows the result of the reduction. Figure 8 depicts the score for four measures in the song. The Figure 8(a) is the existing electronic organ score, whereas Figure 8(b) is the arranged score. The missing notes are marked in blue, and the incorrectly inserted notes are marked in red. Table I depicts the evaluation values of each song. The evaluation values of this song are 0.78, 0.73, and 0.72 for the right-hand, left-hand, and foot parts. While listening to the performance of the arranged score, the evaluation values from 0.7 to 1.0 had a good score and were arranged with the image of the original song. When the evaluation values ranged from 0.5 to 0.7, it was difficult to play using an arranged score. However, the score was arranged to maintain the image of the song. When the score was less than 0.5, it was difficult to play using the arranged score.

The evaluation values of the right-hand, left-hand, and foot parts of Ave verum corpus were more than 0.7; therefore, this is an example of how to arrange accurately. From the Table I, we can observe that the average of the evaluation values of the right-hand, left-hand, and foot part are 0.80, 0.71, and 0.77, respectively. The evaluation of each part has a value of 0.7 or more; therefore, this can be considered to be a good system. Also, this result showed that the performance was improved compared to when do not use instrumental timbre feature amount was added.

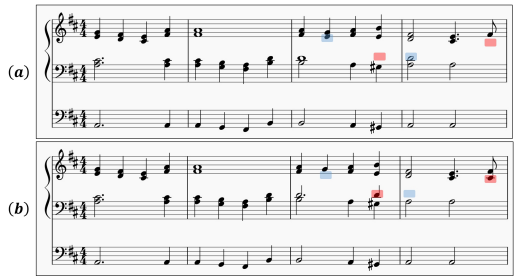


Figure 8. Arrangement results of the (a) existing electronic organ score and (b) proposed system

TABLE I. EVALUATION RESULTS OF EACH SONG

Songs name	Parts	Measures	Right hand	Left hand	Foot
Ave verum corpus	13	46	0.78	0.73	0.72
Pictures at an Exhibition	20	24	0.75	0.66	0.83
Salut d'Amour	6	99	0.81	0.67	0.85
No.36 "Linz" K.425	14	57	0.89	0.70	0.69
Averages	—	—	**0.80**	**0.71**	**0.77**

V. DISCUSSION

A. Different arrangement from the existing electronic organ score

Several measures were arranged unnaturally. If the number of musical instruments of the original score or the number of parts being played is small, only two groups may remain after clustering. If we choose a group with the foot, right-hand, and left-hand parts in that order, there will be a melody for the right hand and foot, and a musical score with no melody will be produced for the left hand. In the electronic organ, it is difficult to play using the foot part; therefore, ideally we do not play using the foot part when we compliment the melody with the right- and left-hand parts. The melody was cut off and connected unnaturally from the right to the left hand. Therefore, it was necessary to reexamine the selection method of the melody group. In the current method, we only look at different measures. We considered it better to select a melody group by considering the connection afresh using the before and after measures of the melody. In the previous method, N-gram [13] of the language model and Hidden Markov Model (HMM) [14] of the probabilistic model were used to consider connecting the melody. Using these methods, we can select an appropriate melody. We can also expect that only the left hand may have rests suddenly and that the melody will be unnaturally connected from the right to the left hand.

Further, the system performance was lowered because of the process of correcting the musical score. Additionally, a music score that was difficult to play, such as a sound that overlapped three notes, suddenly appeared in the melody of a single tone. As depicted in the yellow portion of Figure 9(b), extra sounds were inserted, and it was difficult to play using a musical score. This was because we had not considered

 ISBN: 978-1-61208-627-9

the ease of playing instruments. It is possible to improve by considering the relation of the note numbers and the movement of fingers. In the case of Figure 9, the extra inserted sound in (b) was emitted in another part (in this case, the left-hand part) as compared with the existing electronic organ score in (a). In this way, when a mistakenly inserted sound is emitted by another part, it is better to perform processing, such as sound elimination, because the sound is supplemented by other parts.

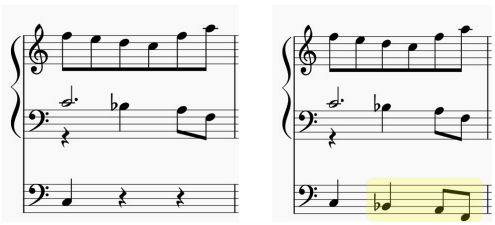


Figure 9. Example of incorrect arrangement of (a) existing electronic organ score and (b) arrangement result of the proposed system

B. Weights of feature quantities used for clustering

For each song used in the experiment, the weight of the feature quantity used for clustering is illustrated in Table II. From Table II, we can observe that the feature quantities RA of the pronunciation time, and MA of the pitch change works well. The electronic organ of the keyboard instrument is the target of this study; therefore, these features are effective. In keyboard instruments, while playing multiple melodies alone, it is difficult to play unless it is the same movement or a melody that can be pronounced simultaneously. Therefore, this system should be capable of reflecting the characteristics of the keyboard.

In future, we propose to arrange other instruments also. The results of this study depicted that we can deal with other instruments by assigning weights to the feature quantities according to the characteristics of each instrument. In case of arranging for a guitar, it would be better to increase the weight of the harmony feature quantities CA because mainly the chords are played on the guitar. Therefore, it is possible to create a guitar melody because it is possible to consider into account the clusters of chords. Furthermore, during the time of cluster selection, the musical score of the guitar can be created by choosing the melody including the chord and melody of the bass part.

TABLE II. WEIGHT USED FOR CLUSTERING

Songs name	RA	MA	CA	SR	IT
Ave verum corpus	0.3	0.4	0.1	0.1	0.1
Pictures at an Exhibition	0.3	0.3	0.2	0.1	0.1
Salut d'Amour	0.3	0.4	0.1	0.1	0.1
No.36 "Linz" K.425	0.4	0.2	0.2	0.1	0.1

VI. CONCLUSION

In this study, we proposed an arrangement system for reducing the full score of the electronic organ. First, melodies were clustered using the pronunciation time, pitch change, harmony, persistent pronunciation pattern, and timbre of instruments to summarize the melodies having similar characteristics from the full score. Next, we selected the melodies corresponding to the right-hand, left-hand, and foot parts. Finally, we modified the melodies so that they could be played using an electronic organ. In clustering, it became easier to organize the melody of musical instruments with similar timbres by considering the timbre features afresh; it also became possible to arrange according to the musical score of the electronic organ. Many errors occurred while selecting the clusters because the melody was selected only within the measures where the notes were played at the same time. In addition, the melodies were chosen without considering the connection between the before-and after-melodies. Using the N-gram or HMM, we intend to select appropriate clusters by considering the ease of connecting melodies.

In future, we plan to arrange music for other instruments also. In this preliminary step, we made an ensemble score covering musical instruments, such as the violin, saxophone, guitar, and bass. Keyboard instruments and saxophones were observed to score well. However, stringed instruments were difficult to play using the arranged score. Therefore, in our future studies, we aim to investigate the musical instrument characteristics and performance methods for other instruments and further arrange them.

REFERENCES

[1] K. Fujita, H. Oono, and H. Inazumi, " A proposal for piano score generation that considers proficiency from multiple part ", IPSJ Special Interest Group on Music and Computer, 2008, pp. 47-52.

[2] S. Ito, S. Sakou, and T. Kitamura, " Automatic Arrangement of Ensemble Score by Contraction of Orchestra Score Considering Importance of Part ", Trans.IPS.Japan, 2013(1), pp. 291-292.

[3] M. Matsubara et al., " Scoreilluminator: Automatic illumination of orchestra scores for readability improvement", Proceedings of the 2009 International Computer Music Conference, ICMC 2009, pp. 113-116.

[4] C. Roads, Computer music—History・Technology・Art, Tokyo Denki University Press, 2001.

[5] J. Makhoul, " Linear prediction: A tutorial review ", Proceedings of the IEEE, Volume:63, Issue:4, April 1975, pp. 561-580.

[6] S. Yoshii, Digital audio processing: Tokai University Press, 1998.

[7] T. Kobayashi, " Cepstral and Mel-Cepstral Analysis of Speech ", The Institute of Electronics, Information and Communication Engineers, 1998, pp. 33-40.

[8] MusicXML[Online]. http://www.musicxml.com/ [17 Mar. 2018].

[9] Musescore—Music Composition and Notation Software[Online]. http://musescore.org/ [17 Mar. 2018].

[10] Y. Dong and D. Li, Automatic speech recognition:a deep learning approach, Springer-Verlag London, 2015.

[11] J. R. Hershey and P. A. Olsen, " Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models ", Acoustics, Speech and Signal Processing, 2007, pp. IV317-IV320.

[12] J. Edwards and P. Oman, Dimensional Reduction for Data Mapping-A practical guide using R, R News, Vol. 3/3, 2003, pp. 2-7.

[13] M. Tomari, M. Sato, and Y. Osana, " Automatic composition based on genetic algorithm and N-gram model ", Proceedings of IEEE Conference on System, Man and Cybernetics, Singapore, 2012, pp. 202-207.

[14] T. Kathiresan, " Automatic melody generation ", Master's thesis, KTH Royal Institute of Technology, 2015, pp. 25-43.

 ISBN: 978-1-61208-627-9

Effects of Lower Frame Rates in a Remote Tower Environment

Jörn Jakobi
Institute of Flight Guidance, Dep. of Human Factors,
German Aerospace Center, DLR e.V.
Braunschweig, Germany
email: joern.jakobi@dlr.de

Maria Hagl
SHS - Sciences de l'Homme et de la Société Psychologie
Université Grenoble-Alpes
Grenoble, France
email: maria.hagl@univ-grenoble-alpes.fr

***Abstract*— In the field of aviation, "Remote Tower" is a current and fast-growing concept offering cost-efficient Air Traffic Services (ATS) for aerodromes. In its basics it relies on optical camera sensor, whose video images are relayed from the aerodrome to an ATS facility situated anywhere, to be displayed on a video panorama to provide ATS independent on the out-of-the-tower-window view. Bandwidth, often limited and costly, plays a crucial role in such a cost-efficient system. Reducing the Frame Rate (FR, expressed in fps) of the relayed video stream is one parameter to save bandwidth, but at the cost of video quality. Therefore, the present article evaluates how much FR can be reduced without compromising operational performance and human factor issues. In our study, seven Air Traffic Control Officers watched real air traffic videos, recorded by the Remote Tower field test platform at the German Aerospace Center (DLR e.V.) at Braunschweig-Wolfsburg Airport (BWE). In a passive shadow mode, they executed ATS relevant tasks in four different FR conditions (2 fps, 5 fps, 10 fps & 15 fps) to objectively measure their visual detection performance and subjectively assess their current physiological state and their perceived video quality and system operability. Study results have shown that by reducing the FR, neither the visual detection performance nor physiological state is impaired. Only the perceived video quality and the perceived system operability drop by reducing FR to 2 fps. The findings of this study will help to better adjust video parameters in bandwidth limited applications in general, and in particular to alleviate large scale deployment of Remote Towers in a safe and cost-efficient way.**

Keywords-Remote Tower; air traffic control; low frame rate; video update rate; detection performance; physiological stress; perceived video quality; perceived system operability.

I. INTRODUCTION

According to several authors, operating regional airports in Germany [29] or elsewhere, e.g., in Australia [4], is already outdated due to high financial deficits. However, a promising alternative, called Remote Tower, invented in 2005 at German Aerospace Center (DLR e.V.) in Braunschweig [13] [14] is already on its way. The main idea consists in enabling ATS decoupled from the Out-of-The-Window (OTW) view from a conventional aerodrome tower. Video cameras capture the aerodrome scenery and relay the video stream to an ATS facility where it is displayed on a video panorama presentation. The new ATS facility can still be located at the Tower building but actually is independent on that location and can be sited anywhere. The gained advantages are manifold. Cameras can provide additional view points of the aerodrome, supplementary infrared cameras can look through fog or darkness or new augmentation features improve the former OTW view, which contributes to safety. Costly constructions of new Towers or maintenance of existing Tower buildings can be saved. The core idea, however, is that more than one aerodrome can be connected to this remote ATS facility. A so called Remote Tower Center (RTC) has the advantage that the ATCOs can switch between aerodromes or can provide ATS to more than one aerodrome simultaneously providing their service when and where it is actually needed. As a main effect, their working time would be exploited much more efficiently [15] and, as a side effect, human errors due to underutilization at work would be diminished [40].

In 2015 the first RTC went in operation. Swedish ATCOs control air traffic of Sundsvall and Örnsköldvik airport from the RTC in Sundsvall [35]. Despite this first success, ambitions to improve the Remote Tower concept run high. Thus, new modalities for controlling a pan-tilt-zoom camera [16] or to augment the video panorama vision [17] are developed and adapted to various operational needs dependent on the operational context. For instance, an ATCO without any approach radar support would need a very high video resolution to detect traffic at far distances. Instead, an ATCO who controls traffic movements on the aerodrome maneuvering area would probably need a sufficient Frame Rate (FR) to precisely judge about the velocity of the traffic. In fact, both, resolution and FR are important operational quality parameter but also bandwidth consuming parameters and therefore cost-driving factors in Remote Towers systems. Thus, new Remote Tower implementations aim to optimize these parameters to a better benefit-cost ratio. With this in mind, we focus on the effects of reduced FR in a Remote Tower context. Certainly, FRs below the critical flicker frequency (CFF) could contribute to a perceived loss of movement fluidity, which might result in perceived loss of video quality. But does lower FRs also evoke negative effects, such as reduced ability to detect traffic movements on the displayed video panorama or even cause physiological stress in ATCOs and lower system operability?

 ISBN: 978-1-61208-627-9

This paper addresses therefore the following research question: What are the effects of lower FRs in a Remote Tower environment on:

1) Visual detection performance,
2) Physiological stress,
3) Perceived video quality,
4) Perceived System operability?

The paper is structured in the following parts: Section II aims at presenting theoretical background concerning the perception of movement, distortions that can appear during video transmissions and a review of scientific literature about the impact of low frame rates on the operator. Linking these three aspects together allows us to specify the research question and the hypotheses. Section III explains the chosen methods and the procedure of the study. Then, in Section IV we will present the obtained results in descriptive and inferential statistics. In Section V these results will be explained and discussed based on theoretical findings and the research question. Section VI draws explicit conclusions by illustrating how the results of the conducted study contribute to science and future Remote Tower implementations.

II. THEORETHICAL BACKGROUND

A. Motion perception

In order to understand better the meaning of reduced FR for humans, we will firstly explain the importance for human beings to perceive motion and, secondly, explain how fluid motion is perceived by humans.

1) Importance of motion perception for human beings

The perception of moving objects is a phenomenon that humans take for granted. In fact, since the earliest childhood, a baby's attention is guided towards moving objects [26]. According to [33], motion perception permits humans to anticipate what he calls "collision time" to estimate the velocity of stimuli. Furthermore, he suggests that the utility of perceiving motion leads to perceiving objects in a tridimensional environment. Other reasons that underline the importance of motion perception consist in distinguishing a stimulus from its background and understanding different textures of objects [33]. For instance, if a gray airplane is in front of a gray cloud, it might be difficult to distinguish the flying object from its background. A light penetration from a different angle can be perceived when the plane moves. In conclusion, we can state that motion perception permits the observer to get to know more about the details of the environment s/he's in. In order to understand to which extent a movement appears to be fluid, some basics of psychophysics and cinematography are necessary and will be explained in the following section.

2) Fluid motion perception

In psychophysics, psychologists refer to absolute threshold if the minimal intensity necessary to perceive a stimulus is perceived by 50% of the observers [21]. As in [39], the CFF is described as the frequency at which the flickering of a flash is not distinguishable from a constant light source. In reality, this threshold can vary by the luminosity of the discontinuous light [21]. According to [24], the sensibility of CFF can also depend on the contrast between the stimulus and its environment. Therefore, the human eye is more sensible to temporal frequencies in high contrast situations between 15 and 20 Hz. The idea of a CFF is also used in cinematography. In cinematographic history, 13 presented images per second were identified as being critical for creating the sensation of fluid movement [23]. Concerning the first movies, 16 frames per second (fps) were not sufficient for showing fluid movements because of the visually perceived intermittent time between each frame. Therefore, cinematographs found a solution by showing the same image two or three times in a successive manner. In total, this means a presentation of 32 or 48 images per second [23] from which 16 are different. More precisely, movies were presented at 16 fps with a refresh rate of 32 Hz or 48 Hz. It's important not to confuse these two notions. Nowadays, the regular FR in cinemas and TV is either 24 fps or 30 fps [23]. FR and refresh rate are two important notions to understand the meaning of human perception of fluid movements in virtual environments. However, perceived fluidity of movements is not the only factor that contributes to an almost perfect presentation of the outer world when it comes to cinematography. Therefore, the next section will treat distortions likely to appear during tele transmissions.

B. Reality distortions through tele transmission

Despite the similarities between an optical sensor camera and the human eye, no camera can represent what we see with our proper eyes. Perceiving the world around us in a stereoscopic manner is already a limit for most conventional cameras that render a monoscopic image. Further, image resolution plays an indispensable role [2]. It allows us to perceive objects from a far distance in a detailed manner. The higher the image resolution, the better we can discriminate stimuli at bigger distances. The human eye has a visual acuity of ca. 1 arc minute [25]. In other words, from a distance of 1 km, the human eye can discriminate two points with a distance of 28 cm. However, conventional Remote Tower camera systems dispose of a medium image resolution [36] lower than 1 arc minute. With 2 arc minutes for instance, a camera could only discriminate two pixels with a distance of 56 cm from a distance of 1 km.

Latency or lag can be another distortion appearing in real-time tele-transmission systems caused by different sources (e.g., transmission problems, data conversion problems). They are expressed by a temporal delay between the input of information into a system and the output as a presentation of the information to the operator [5]. When we face different latency times in between of the presented frames, we talk about jitter.

Finally, the presented FR can result in a distortion of reality. By its reduction, the fluid perception of the movement drops as well. As we have seen it in the previous section, 13 fps are judged as being necessary in order to

 ISBN: 978-1-61208-627-9

perceive fluid movement. This estimation is not absolutely correct, since the threshold can vary between several parameters, for instance the radial velocity of the perceived object. Another distortion related to FR refers to frequency interferences [2], like the well-known wagon-wheel-effect. In a Remote Tower environment, this effect can appear wherever periodic movements are faced, e.g., rotor blades of an aircraft or blinking lights. Some blinking lights need time to light up and are only at their maximum of luminance for a few instants. Hence, by reducing FR, the probability to capture an image during the maximum of luminosity diminishes as well. This could be critical particularly at night or low visibility.

C. Review of low FR effects on the operator

Limited bandwidth made several concerned parties study impacts of low FR on operators [7] [36] [38]. Due to high data transmission costs, researchers have investigated several parameters in order to reduce bandwidth. In the next section, we will present studies that focus on the effects of low FR on performance, psycho-physiological health of operators, as well as on perceived video quality.

1) Effects of low FR on operator visual detection performance

In the context of a Remote Tower environment, in a research study DLR investigated effects of low FR in a real-time tower simulation scenario: 6 fps, 12 fps and 24 fps were tested in order to evaluate the performance of ATCOs to visually discriminate and predict in real-time if aircraft after touchdown is in danger of a runway overrun due to low braking. Then, they were asked about how sure they were about their answer. Results show that by reducing FR, the level of certitude decreases. Further, the authors recommend a FR greater than 30 to achieve a maximum visual discriminability for dynamic events [10] [18]. Another study about unmanned ground vehicles and aircraft showed that the performance of detecting obstacles does not decrease by reducing the FR from 30 fps to 5 fps [7]. Reference [20] obtains similar results in a study about target detection in a 2 fps and 25 fps condition. By reducing FR, the participants' performance did not change in a significant manner. Thus, divergent results can be found. One possible explanation is that not only FR is a factor affecting performance. By a meta-analysis on existing studies in the field of effects of low FRs on performance [6] it was concluded that the effect of low FR on performance of participants is above all task-dependent. Moreover, they have identified an interaction of FR with image resolution. The authors suggest that the right balance between FR and image resolution can help to perceive depth accurately and hence increase the perception of movement in the areas that are farer away from the observer. They also suggest that performance depends on the participants' characteristics, too.

Thus, experienced participants in virtual environments might be less affected by FR reduction. However, not only performance is an important factor to consider when reducing FR. The operator's well-being is an essential parameter to study before lower FRs can be applied. Therefore, the next section will treat impacts of low FRs on psycho-physiological health of operators.

2) Effects of low FR on psycho-physiological health of operators

Regarding to psycho-physiological health, the effects of low FR have been evaluated very rarely. Reference [7] refers to a study at which they tested physiological stress in terms of cyber sickness about several unmanned ground vehicles and aircraft in different FR conditions. The used questionnaire (Simulator Sickness Questionnaire, SSQ) has been validated within a sample of 4000 pilots who participated in trainings in different flight simulators [27]. Nowadays, the SSQ is also used for evaluating cyber sickness in other virtual environments [28]. In study [7] the effects of low FR were non-significant. Thus, participants did not feel sicker in a simulation at 5 fps than at 30 fps. Another study [9] tested spatial stability in a virtual environment by varying the FR (6 fps, 12 fps & 20 fps). As a result, more participants felt sick by reducing FR. Reference [20] did not find significant results. Even though participants at 2 fps expressed higher workload and frustration, the expressed psycho-physiological stress was not significantly higher than at 25 fps. Furthermore, adverse health effects associated with low FRs do not appear in the occupational disease lists [3]. Taking into account these outcomes, divergent results can be found again.

3) Effects of low FR on perceived video quality

In scientific literature, we found several studies evaluating perceived video quality in different FR conditions. The perceived quality is often evaluated from acceptability and personal preference. In the context of a study concerning the performance in first person ego shooters, a study varied FR (3 fps, 7 fps, 15 fps, 30 fps & 60 fps) and image resolution (320x240 pixels, 512x384 pixels & 640x480 pixels) [8]. The results clearly indicate a significant preference for higher FRs and even more for a higher image resolution. Surprising effects have been found in a study that aimed to evaluate the video quality under different FR conditions (6 fps, 10 fps, 12 fps, 15 fps, 18 fps, 20 fps & 24 fps) and two image resolution conditions (low & high). A significant difference of video acceptability was not found between different FRs. Moreover, participants preferred higher image resolution to higher FRs [32]. In a study testing the video acceptability in different FR conditions (5 fps, 10 fps & 15 fps), it was stated that video acceptability decreases by reducing FR [1]. In reference [31] the type of motion is stressed: *"The type of motion in a sequence was important when considering the effects of FR on subjective quality".*

To conclude, it is difficult again to find an appropriate FR threshold to guarantee the spectator's satisfaction in terms of video quality. As the previous studies already have suggested, it is very likely that not only FR plays a determinant role for acceptance of video quality.

 ISBN: 978-1-61208-627-9

4) Effects of low FR on the perceived operability

Until now, we have presented studies that examined effects of lower FR on performance, operator health and perceived video quality. However, these three parameters seem to be insufficient to evaluate if a Remote Tower system can be operated in a safe and efficient manner. If the user is not convinced of the system operability, errors can emerge by expressed mistrust in the system. According to [30], confidence in a system and emerging risks can play a mediator role in the system reliability. A system can seem to be perfect but is not if the user does not have a good feeling about it.

D. Research question

The general aim shared by all Remote Tower actors is to develop a system that allows remote air traffic control in the best cost-efficient ratio. Regarding this aim, a known limit is bandwidth. Nowadays, data transmission is still expensive and can be a financial threat if resources are not used efficiently. According to [2], crucial factors concerning bandwidth are field of view, image resolution, color depth, FR and data compression, which seem to be widely accepted, but opinions diverge largely when it comes to image resolution and FR. Some stakeholders believe that higher FR is preferable to higher image resolution. In fact, they believe that low FRs can decrease performance and operator health. However, so far there is no scientific proof that justifies these two presumptions. As it has already been expressed in the theoretical part, effects of low FRs on performance are likely to be task dependent [6] and do not give us clear information about operator health. Yet, impact of low FRs in Remote Tower environments has never been thoroughly tested with an experimental design. On the basis of context analysis and preexisting scientific literature, we will now propose six hypotheses.

E. Hypotheses

$H_{1,1}$: By reducing the FR from 15 fps to 10 fps, 5 fps or 2 fps, the adequate assessment of moving objects by ATCOs decreases.

$H_{0,2}$: By reducing the FR from 15 fps to 10 fps, 5 fps and 2 fps, the operator's performance in visual detection tasks will not decrease.

$H_{0,3}$: ATCOs' performance in visual tracking tasks does not decrease significantly by reducing the FR from 15 fps to 10 fps, 5 fps or 2 fps.

$H_{0,4}$: The physiological stress of operators does not increase significantly when FR is reduced from 15 fps to 10 fps, 5 fps or 2 fps.

$H_{1,5}$: The ATCOs' perception of the video quality will decrease when the FR is reduced from 15 fps to 10 fps, 5 fps or 2 fps.

$H_{1,6}$ The ATCOs' perceived system's operability will decrease when the FR is reduced from 15 fps to 10 fps, 5fps or 2 fps.

III. METHODS

A. Tested variables

The independent categorical variable corresponds to the chosen FR that will be presented in a video at four modalities: 2 fps, 5 fps, 10 fps, and 15 fps. Concerning the dependent variables, the first is to measure the participants' visual performance in three different dimensions: "Adequate Assessment of Moving Objects" (AAMO), "Visual Detection Tasks" (VDT) and "Visual Tracking Tasks" (VTT). The second dependent variable evaluates the participants' "physiological stress", the third one the ATCOs' "perceived video quality", and the fourth one the ATCOs' "perceived operability of the low FR system".

B. Participants

Seven male ATCOs between 31 and 58 years (M = 41.7, SD = 12.0) and five pseudo ATCOs (four men, one woman) between 26 and 52 years (M = 44.0, SD = 10.42) took part in the experiment. Their nationalities were German, English, Hungarian, Norwegian, Romanian and Swedish. We chose pseudo ATCOs as a non-expert control group in order to control potential motivational bias concerning physiological stress. ATCOs were directly invited by an invitation letter. All ATCOs and pseudo-ATCOs were familiar with the Remote Tower concept.

C. Preparation of the study equipment

1) Video material collection, selection and edition

For the experiment the DLR Remote Tower field test platform at Braunschweig-Wolfsburg Airport (BWE) was used. Fig. 1 shows the camera sensors on the roof of the DLR building surveying BWE aerodrome (left). On the right hand side the ATCO working position is depicted. The research prototype is operated with 2 arc minute image resolution and a FR of 30 fps.

Several hours of audio and video material have been recorded via the platform, assessed and selected. Firstly, only records which complied with EUROCAE [12] standard test conditions were selected. Further, it was checked for broad traffic diversity and other relevant visual occurrences (e.g., flock of birds). This step was supported by four ATCOs. In a third step, the final 30-fps video stream was computed to four content-identical streams with 2 fps, 5 fps, 10 fps and 15 fps and a length of 80 minutes each. 2, 5, 10 and 15 fps were chosen as 30 fps is a multiple of them, which helps to avoid the maximum of jitter. 2 fps as the lowest FR was chosen since this FR corresponds to the minimum standard of FR tolerated in a Remote Tower environment [12]. To complete the construction, the video material had to synchronize with the external sound and the radio transmissions. Finally, the jitter was measured in each experimental condition to ensure that it lies below the maximum tolerated value of 0.5 seconds [19].

 ISBN: 978-1-61208-627-9

Figure 1. DLR Remote Tower field test platform at Braunschweig-Wolfsburg Airport (BWE) (left: Camera sensors; right: ATCO working position).

2) Construction of the mid-run visual performance evaluation grid

In a first step, we chronologically listed events that refer to ATC relevant visual tasks, and associated them with the visual requirements stated by the interviewed ATCOs and those in the requirements for EUROCAE Remote Tower specifications [12]. These events were divided into three different categories of questions: the AAMO, VDT, and VTT. As for the AAMO, we mainly took into account the ATCOs' fears of not being able to properly assess the velocity or the stimulus' movement direction. Thus, an exemplary task is to evaluate whether a flashing light can be perceived in a safe and efficient manner. Other tasks include the assessment of flying birds' direction, wind direction and movements of aircraft propellers and human beings on the aerodrome. An exemplary task for VDT consists in detecting an aircraft in the final approach area or in the traffic pattern as soon as possible. Perceiving an aircraft in those positions represents visual requirements according to the interviewed ATCOs. Regarding VTT, the instruction consists in following an aircraft during the take-off phase and hitting a buzzer when it was not noticeable anymore. After classifying all possible tasks, we created and selected a list of possible questions that follows the chronology of occurrences.

3) Construction of the post-run questionnaire

In order to measure the physiological stress of the participants, we concentrated on mentioned symptoms in the interview, such as fatigue, nausea, headache, eye strain or dizziness, which are consistent with the items in a SSQ questionnaire to evaluate cyber sickness [27]. It contains 16 items and the scale is divided into three subscales which measure the dimensions "Nausea", "Oculomotor" and "Disorientation".

The second part of the post-run questionnaire consists in rating the perceived video quality and the perceived operability on a 7-point Likert scale.

4) Pretest

A pseudo-ATCO and two ATCOs participated in the pretest to verify that the scenario did not contain inconsistencies that the tasks relate to an ATCO's daily routine, and that the questionnaires are comprehensible. They accepted the setting and confirmed that the number of tasks was enough not to be bored and that the variety of tasks corresponds well to the different visual requirements that ATCOs have to face during their daily work.

D. Experimental Procedure

The study took place between May 15th 2017 and June 12th 2017. The procedure of the study was structured in two parts. The briefing phase represented the part in which ATCOs were informed and prepared for the actual experiment. The experimental phase corresponded to the video session and the completion of the post-run questionnaire. The written and spoken language was English. The participants were informed that they will see the same video four times at four different FRs. They were left unaware of the FRs to be tested in order to avoid potential effects of previously formed attitudes. They were explained that the order of the videos was randomized for methodological reasons and that they had to complete the SSQ questionnaire before the actual experiment to avoid methodological biases. The ATCO's eyes' position was 2.1 m distance from the 56'' HD screens in order to standardize experimental conditions and to guarantee the necessary visual acuity. The experimenter sat at the participant's right side. The different FR modalities were ordered in a Latin square, in order to randomize the observations. After the last session, the participants answered a supplementary questionnaire in which they gave demographic information about themselves and classified the watched videos in order of preference. Finally, they were asked to give their general opinion on Remote Tower in order to reduce potential motivational biases followed by a general debriefing session.

IV. RESULTS

A. Results concerning visual performance

1) Adequate assessment of moving objects

In order to evaluate AAMO, the ATCOs' answers were coded as "1" when the movement is perceived "safe and efficient" and vice versa as "0" when the movement was perceived as "neither safe, nor efficient". It was observed that the movement of five objects was perceived as being "safe and efficient" by all ATCOs in each of the four FR modalities. These objects correspond to the propeller of three different aircraft on the apron, to a flag from which the ATCOs had to assess the wind direction and the direction of a flock of birds. Concerning the flock of birds, the ATCOs added that it was easy to identify the objects as birds and to deduce their direction.

The other category of objects corresponds to the flashing lights of three vehicles: a fuel truck, a black airport vehicle, and a follow me car. Concerning the fuel truck and the black vehicle, we observed that most ATCOs judged the visibility of the flashing light as being perceivable safe and efficient in the 5 fps, 10 fps and 15 fps conditions but not in the 2-fps condition. This tendency appeared especially regarding to the

 ISBN: 978-1-61208-627-9

follow-me vehicle's flashing lights. In the 2-fps condition, a safe and efficient perception is only admitted one time out of 21 instances over all ATCOs. By comparing all means of all flashing lights instances (35 in total per FR over all 7 ATCOs), we observe that the flashing lights are perceived as being least visible in a safe and efficient manner in the 2-fps condition (M = 0.17, N = 7, SD = 0.21), followed by the 10-fps condition (M = 0.77, N = 7, SD = 0.34) and the 5-fps condition (M = 0.8, N = 7, SD = 0.31). In the 15-fps condition, flashing lights were perceived as the being most visible in a safe and efficient manner (M = 0.85, N = 7, SD = 0.25). A chi-square test supports this tendency: The perception of flashing lights decreases significantly when the FR drops from 15 fps to 10 fps, 5 fps and 2 fps ($\chi^2_{(df = 3, N = 7)}$ = 1, $p < .01$). But these results can only be found for flashing lights.

Thus $H_{1,1}$ is only partially assumed, i.e., by reducing the FR from 15 fps to 10 fps, 5 fps or 2 fps, the adequate assessment of flashing lights decreases but others remain unaffected.

2) Visual Detection Tasks

The mean detection times, centered on the mean to each of the four FR conditions, show that ATCOs take on average less time detecting an aircraft in the approach area at 10 fps (M = -1.4, N = 7, SD = 6.33) than at 15 fps (M = -0.04, N = 7, SD = 5.85), at 2 fps (M = 0.21, N = 7, SD = 5.36) or at 5 fps (M = 1.5, N = 7, SD = 4.32). However, a Friedman test did not show significant difference between ATCOs' reaction time at the four FR conditions ($\chi^2_{(df = 3, N = 7)}$ = 2.14, p = .54). Thus, the reduction of FR does not appear to decrease the ATCOs' performance to detect an aircraft in the final approach area which supports our $H_{0,2}$ to retain the H_0.

Furthermore, all aircraft in different traffic pattern positions, as well as all human beings on the movement and maneuvering area were perceived by ATCOs in each FR condition. In addition, some ATCOs add that the jerky movements perceived in the 2-fps and 5-fps condition helped them to detect the aircraft quicker. According to them, the jerky movements cause a blinking effect and thus attract more attention than an aircraft that moves more smoothly at 10 fps or 15 fps.

3) Visual Tracking Tasks

The measured times of VTT, again centered on the mean of each of the four FR conditions, indicate that on average, ATCOs could visually track departing aircraft longer at 15 fps (M = 1.48, N = 7, SD = 3.95) than at 2 fps (M = -0.05, N = 7, SD = 4.98) or at 5 fps (M = -0.35, N = 7, SD = 8.44), worst at 10 fps (M = -1.06, N = 7, SD = 4.75). Again, the Friedman test could not reveal significant difference of ATCOs' performance to visually track aircraft in all tested four different FRs ($\chi^2_{(df = 3, N = 7)}$, p = .62), which supports our $H_{0,3}$ to retain the H_0.

B. Results concerning physiological stress

After each test run, participants answered the 16 SSQ items on a Likert scale ranging from "0 = none", "1 = slight", "2 = moderate" to "3 = severe".

At the beginning, we checked whether the results of the experimental ATCO group differ significantly from the pseudo-ATCO control group to exclude systematically effecting variance in terms of possible motivational biases on behalf of the ATCOs caused by their general attitude towards the Remote Tower concept. A t-test for independent samples for the Total Sickness Score (TSS) shows that there is no significant difference found between both groups ($t_{(10)}$ = 0.59, p = .56). Thus, only results of the expert group are taken into account for all analyses. Fig. 2 depicts that with a mean TSS of 1176 we observed only under averaged TSS means for all four test conditions: 15 fps (M = 24.3, N = 7, SD = 19.56), 10 fps (M = 37.5, N = 7, SD = 44.66), 5 fps (M = 50.03, N = 7, SD = 71.94), 2 fps (M = 147.06, N = 7, SD = 213) (for the exact calculation of the TSS refer to [27]).

Still marginal, but the highest TSS is measured at 2 fps (see Fig. 2). A Friedman test could not reveal significant difference between "base" and the four test conditions neither for the TSS ($\chi^2_{(df = 4, N = 7)}$ = 5.89, p = .21), nor for the subscale "Nausea" ($\chi^2_{(df = 4, N = 7)}$ = 8.88, p = .06), the subscale "Oculomotor" ($\chi^2_{(df = 4, N = 7)}$ = 3.93, p = .42), or for the subscale "Disorientation" ($\chi^2_{(df = 4, N = 7)}$ = 5.29, p = .26). As postulated, $H_{0,4}$ is to be retained: The psychological stress of operators did not decrease significantly when FR is reduced from 15 fps to 10 fps, 5 fps or 2 fps.

C. Results concerning the perceived video quality

Via a 7-point Likert scale ranging from "1 = totally unacceptable", "2 = unacceptable", "3 = slightly unacceptable", "4 = neutral", "5 = slightly acceptable", "6 = acceptable" to "7 = perfectly acceptable", ATCOs perceived

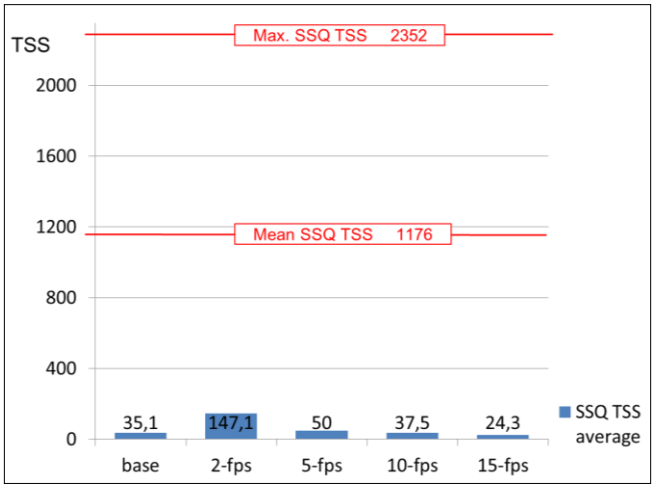


Figure 2. Total Sickness Scores before (base) and for four different FR test conditions.

the video quality as being more acceptable at 15 fps (M = 5.71, N = 7, SD = 1.25, Min = 3, Max = 7) than at 10 fps (M = 4.86, N = 7, SD = 2.67, Min = 1, Max = 7), 5 fps (M = 4.29 , N = 7 , SD = 1.8 , Min = 1 , Max = 6) or at 2 fps (M = 3, N = 7, SD = 1.53, Min = 1, Max = 5).

The more the FR is reduced, the more ATCOs judge the quality of the video as being less acceptable. In the 2-fps condition, the quality was even rated below the average "neutral". This tendency is supported by a Friedman test that revealed significant difference ($\chi^2_{(df = 3, N = 7)} = 12.05$, $p < .01$). As postulated in $H_{1,5}$, the perceived video quality in terms of FR decreased with the reduction of the FR.

After each test run, the ATCOs were asked to estimate the FR of the just watched video. Surprisingly, ATCOs always believed that the FR is superior to what it actually was: Answers after the 2 fps conditions referred to 3.14 fps by average, 5 fps to 6.29 fps, 10 fps to 15.29 fps, and 15 fps to 20.14 fps. By average they overjudged the FR by 53.9%.

After having completed all test runs, the ATCOs were asked to rank the watched videos in their order of preference. Most of them ranked 15 fps at the top. The second rank is mostly shared by videos at 5 fps or 10 fps. The last rank is notably reserved for the 2 fps.

D. Results concerning the perceived operability of a low FR system

On a 7-point Likert scale from "1 = totally disagree", "2 = disagree", "3 = somewhat disagree", "4 = neither agree nor disagree", "5 = somewhat agree", "6 = agree" to "7 = strongly agree", the ATCOs should answer the following statement: "I would be able to control the air traffic with the given FR.". The perceived operability increased with the increase of FR. Thus, the system operability was perceived least at 2 fps (*M* = 2.86, *N* = 7, *SD* = 1.57, *Min* = 1, *Max* = 5). It increases over-averaged with 5 fps (*M* = 4.14, N = 7, *SD* = 1.87, *Min* = 1, *Max* = 6), 10 fps (*M* = 4.86, *N* = 7, *SD* = 1.57, *Min* = 1, *Max* = 7) and finally with 15 fps (*M* = 5.71, *N* = 7, *SD* = 1.25, *Min* = 3, *Max* = 7). A Friedman test revealed this difference as significant ($\chi^2_{(df = 3, N = 7)} = 12.68$, $p < .01$). Even though only the 2-fps condition is judged below acceptable, $H_{1,6}$ is to be assumed: The lower the FR, the less ATCOs consider the system as being operable.

V. DISCUSSION

The effects of lower FRs on the performance of an adequate assessment of moving object tasks are multilayered and cannot be judged generically. Surprisingly, all propeller movements, the wind flag and the flock of birds, as well as the movement of human beings on the aerodrome were perceived by all ATCOs in all four FR conditions in a safe and efficient manner. Most ATCOs commented that the rapid disappearance of the bird flock over the runway made them worry much more than their jerky movement, which refers rather to an image resolution problem than to a lower FR. Concerning the flashing lights, most ATCOs judge flashing lights to be perceivable safely and efficiently down to 5 fps but when further reduced down to 2 fps the capturing of the rotating beacon at its full brightness decreased and the perception was no longer perceived as being safe and efficient by the majority of ATCOs. Those negative effects in 2 fps was expected and could be covered by using flashing lights with obscure/luminous phases that interfere less with the chosen FR.

As postulated, with respect to the performance in visual detection tasks inferential statistics do not indicate a significant difference between the four FR conditions. Apart from the impression that aircraft seem to move jerkier at lower FRs, especially when they are close to the camera, FR does not seem to play an essential role in detection tasks. In particular planes in the final approach or departure area do not have great lateral movements at all. ATCOs therefore perceive only a point that grows bigger when the plane approaches or shrinks at departure. The concern of not visually detecting an approaching or departing aircraft or an aircraft right downwind due to lower FRs seems therefore be unjustified. In addition, it seems more logical to detect an approaching aircraft earlier or to see an aircraft leaving the aerodrome longer by increasing the image resolution instead of higher FRs. Moreover, even if the movement seemed jerky at times, several ATCOs noticed that the "jumpy" aircraft even attracted their attention.

Physiological stress was tested via the SSQ after each run. As presumed, the inferential results show that no negative effects with respect to physiological stress could be measured. All TSS were under averaged low. Only in the 2 fps condition the severity of the symptoms increased slightly for some ATCOs, but far from any significance. Someone could argue that these findings are biased by very positive beliefs or attitudes towards lower FR system. This potential side effect could be mitigated by using a pseudo-ATCO control group who performed the entire experiment but by definition had a neutral attitude towards lower FRs since they were not involved in the Remote Tower business: Both groups did not significantly distinguish in their SSQ scores. Thus, a systematical effect of bias for the experimental group could be excluded. For the correct interpretation of these results, it is also important to note that the study was dealing with a small sample one can refer to as a sample of experts [11]. In other words, they share some personality characteristics and very specific professional skills, as well as specialized selection and education criteria. Thus, it is very likely to transfer the results found in the inferential statistics to other ATCOs. For an implementation of a Remote Tower with a medium image resolution and low FRs from 2 fps to 15 fps, it can be stated that effects expressed by physiological stress will most likely not appear.

As expected, the perceived video quality decreased significantly with the reduction of FR. These results are not that surprising since air traffic control requires high visual performance and reducing the FR is an obvious loss in terms of video quality. But in the real Remote Tower implementation world, this obvious loss of video quality could be compensated by an increase of image resolution. Since lower FRs seem not to impair detection performance nor induce physiological stress, this trade-off between FR and image resolution seems to be a valid approach to keep

 ISBN: 978-1-61208-627-9

bandwidth consumptions low but better adapt the visual presentation to the air traffic service operators' task: For detecting small aircraft in a far-view distance, high image resolution is needed and FR is not this important. To assess the velocity of aircraft in a near-view distance on the taxiways or apron, higher FRs are essential and image resolution would not play such a significant role. As stated before, this compensation approach could not be realized in the experimental setting, but it can be assumed that the ATCOs' perceived video quality would have been more balanced over the different FR conditions if have done so.

Similar to the results of perceived video quality are the ones concerning the perceived system operability. By no surprise, also a significant difference between the four FR conditions was found. The average of ATCOs "disagree" or "somewhat disagree" about thinking to be able to handle air traffic at 2 fps. At 5 fps and 10 fps, ATCOs expressed to "slightly agree" being able to manage air traffic and at 15 fps, they expressed to "agree". Like already stated above, the experimental setting neglected compensation in terms of image resolution which would probably have balanced the ATCOs' attitude as well.

To conclude the discussion on our findings, we can affirm that according to our results, a system at lower FR is justifiable at least starting from 5 fps. Thus, between 5 fps and 15 fps, the air controllers' visual performance is maintained at the same level. If one wants to set up a lower FR system, one should pay particularly attention to the used flashing lights at the aerodrome in order to choose some which do not interfere with the FR.

Concerning physiological stress, we did not find a significant increase of the scores when the FR is reduced from 15 fps to 2 fps. However, the comparison of the means in the descriptive statistics suggests a slight increase in the TSS at 2 fps. To avoid physiological stress at a system similar to the one at BWE, we recommend rather 5 fps, 10 fps or 15 fps.

With respect to the perceived video quality, the ATCOs preferred higher FRs to lower FRs. They were mostly opponent against 2 fps. In summary, if one wants to operate Remote Tower at a low FR, it is important to develop convincing strategies to increase the tolerance towards low FR. From a psychological point of view, it is not advisable to put ATCOs in front of a 2-fps system hoping that they will accept it. The user centered approach teaches us how important it is for users to experience positive emotions in order to raise acceptability for a new product [34]. Once the video quality of the low FR system is accepted by the ATCOs, the fear of getting sick could be taken away from them and self-efficiency for performance could rather be perceived. By consequent, it is likely that the attitude towards the perceived system operability is expressed more positively.

Further operational simulation and field trials with the operator in the loop are recommended to increase confidence in low FR systems and to gain additional feedback from ATCOs to develop bests designed Remote Tower solutions for the given operational environment.

VI. CONCLUSION

The optimal FR in Remote Tower environments is debated amongst many actors of the Remote Tower community: It must not be too low to endanger safety or operators' health, but also not be too high to increase the consumption of bandwidth or to compromise other parameters like image resolution. The results of this study can mitigate the concerns regarding lower FR settings. The major conclusion of this study is that the visual performance and physiological stress were not affected by lower FRs in between of 15 down to 2fps. In particular, these findings will allow more degrees of freedom in the design process of a Remote Tower implementation to best adapt a local solution to their operational environment. In future research, it remains to be studied how a trade-off between lower FRs and compensation by higher image resolution would be judged by the ATCOs.

REFERENCES

[1] R. T. Apteker, A. A. Fisher, V. S. Kisimov and H. Neishlos, "Distributed multimedia: User perception and dynamic QoS," Conference of International Symposiom on Electronic Imageing: Science and Technology, San José, USA, pp. 226-234, 1994.

[2] B. O. Bakka, Remote Tower Optical Presentation, [Power Point Slides], Kjeller: Kongsberg, 2017, unpublished.

[3] Bundesgesetzblatt. *BGBI*: *List of occupational diseases, 3rd proclamation for amendment of occupational diseases.* [Online]. Available from: http://www.theaustralian.com.au/business/aviation/regional-airports-under-strain-acil-aallen-report-for-aaa/news-story/e95d864f60135d044b5212865ccabae4 2018.03.13

[4] M. Bingemann, Regional airports under strain, ACIL Aallen report for AAA, The Australian. [Online]. Available from: www.theaustralian.com.au/business/aviation/regional-airports-under-strain-acil-aallen-report-for-aaa/news-story/ , 2017.05.05

[5] S. Bryson, Effects of lag and FR on various tracking tasks, Paper presented at SPIE 1915, Stereoscopic Displays and Applications III, San José, USA, pp. 155-166, 1993.

[6] J. Y. C. Chen and J. E. Thropp, "Review of Low FR Effects on Human Performance," in IEEE Transactions on Systems Man and Cybernetics – Part A Systems and Humans 37(6), pp. 1063-107, 2007.

[7] J. Y. C. Chen, P. J. Durlach, J. A. Sloan, and L. D. Bowens, "Robotic operator performance in simulated reconnaissance missions," Technical report ARL-TR-3628, Hum. Res. Eng. Dir., Army Research Laboratory, Abberdeen Proving Ground, Maryland, 2005.

[8] K. T. Claypool and M. Claypool, "On FR and player performance in first person shooter games," in Springer Multimedia Systems Journal (MMSJ) 13(1), 2007, pp. 3-17.

[9] S. R. Ellis, B. D. Adelstein, S. Baumeler, G. J. Jense and R. H. Jacoby, "Sensor spatial distortion, visual latency, and update rate effects on 3D tracking in virtual environments," IEEE Conference on Virtual Reality, Houston, Texas, pp. 239-265, 1999.

[10] S.R. Ellis, N. Fürstenau. N., M. Mittendorf, "Frame Rate Effects on Visual Discrimination of Landing Aircraft Deceleration: Implications for Virtual Tower Design and Speed Percep-tion", Proceedings Human Factors and Ergonomics Society, 55th Annual Meeting, Sept. 19-23, 2011, Las Vegas, NV USA, pp. 71-75.

 ISBN: 978-1-61208-627-9

[11] I. Etikan and K. Bala, "Sampling and Sampling Methods," Biometrics & Biostatistics International Journal 5(6), 00148, 2017.

[12] European Organisation for Civil Aviation Equipment. *EUROCAE*: *Minimum aviation system performance specification for remote tower optical systems*, ED-240, EUROCAE, Malakoff: September 2016.

[13] Fürstenau, N., Rudolph, M., Schmidt, M., Werther, B., Hetzheim, H., Halle, W., & Tuchscheerer, W. (2008, 12). Flugverkehr-Leiteinrichtung (Virtueller Tower). European Patent EP1791364, application date 2005.

[14] N. Fürstenau, Introduction and Overview, In N. Fürstenau (Ed), Virtual and Remote Control Tower (p. 6), Cham (ZG): Springer International Publishing Switzerland, 2016a.

[15] N. Fürstenau, Preface, In N. Fürstenau (Ed), Virtual and Remote Control Tower (p. xii). Cham (ZG): Springer International Publishing Switzerland, 2016b.

[16] N. Fürstenau, Introduction and Overview, In N. Fürstenau (Ed), Virtual and Remote Control Tower (pp. 14-15). Cham (ZG): Springer International Publishing Switzerland, 2016c.

[17] N. Fürstenau, M. Schmidt, M. Rudolph, C. Möhlenbrink and W. Halle, In N. Fürstenau (Ed), "Virtual and Remote Control Tower," (p. 17). Cham (ZG): Springer International Publishing Switzerland, 2016.

[18] N. Fürstenau, M. Mittendorf and S. R. Ellis, "Videopanorama FR requirements derived from visual discrimination of decerleration during simulated aircraft landing," In: N. Fürstenau (Ed), Virtual and Remote Control Tower (pp. 115-137), Cham (ZG): Springer International Publishing Switzerland, 2016.

[19] N. Fürstenau and M. Schmidt, "Remote Tower Experimental System with Augmented Vision Videopanorma," In N. Fürstenau (Ed), Virtual and Remote Control Tower (p. 180), Cham (ZG): Springer International Publishing Switzerland, 2016.

[20] V. Garaj, Z. Hunaiti and W. Balachandran, "Using Remote Vision: The Effects of Video Image FR on Visual Object Recognition Performance," IEEE Transactions on Systems, Man and Cybernetics Society 40(4), pp. 698-707, 2010.

[21] E. Greene, Evaluating letter recognition, flicker fusion and the Talbot-Plateau law using mcrosecond-duration flashes, PLoS One 10(4), 2015.

[22] H. Hagendorf, J. Krummenacher, H.-J. Müller and T. Schubert, "Psychophysics," in H. Hagendorf, J. Krummenacher, H.-J. Müller & T. Schubert (Eds). Allgemeine Psychologie für Bachelor: Wahrnehmung und Aufmerksamkeit (pp. 43-45), Berlin Heidelberg: Springer Verlag, 2011.

[23] J. Hess, The History of FR for Film, 2 February 2015 [Online], Available from: https://www.youtube.com/watch?v=mjYjFEp9Yx0 2018.03.13.

[24] M. Kallonatis and C. Luu, "Temporal Resolution," In H. Kolb, E. & R. Nelson (Eds), The Organization of the Retina and Visual System. Salt Lake City (UT): University of Utah Health Science Center, 2007a.

[25] M. Kallonatis and C. Luu, "Visual Acuity," In H. Kolb, E. & R. Nelson (Eds), The Organization of the Retina and Visual System, Salt Lake City (UT): University of Utah Health Science Center, 2007b.

[26] P. J. Kellman, Ontogenesis of space and motion perception, In W. Epstein & S. Rogers (Eds), Handbook of perception and cognition (Vol. 5 p. 394), New York: Academic Press, 1995.

[27] R. S. Kennedy, N. E. Lane, K. S. Berbaum and M. G. Lilienthal, "Simulator Sickness Questionnaire: An enhanced Method for Quantifying Simulator Sickness" The International Journal of Aviation Psychology, 3(3), 203-220, 1993.

[28] E. M. Kolasinski, Simulator sickness in virtual environments, (Technical Report 1027: Army Project Number 2O262785A791 – Education and Training Technology). Alexandria, Virginia: United States Army Research Institute, 1995.

[29] T. Kotowski, The mistake with regional airports, Frankfurter Allgemeine. [Online]. Available from: http://www.faz.net/aktuell/wirtschaft/unternehmen/kleine-flughaefen-scheitern-wirtschaftlich-14302627.html 2016.06.23

[30] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," Human Factors 46(1). 50-80, 2004.

[31] M. Masry and S. S. Hemami, "An analysis of subjective quality in low bit rate video," Paper presented at International Conference on Image Processing, 2001, Thessaloniki, Greece, pp. 465-468, 2001.

[32] J. D. McCarthy, A. M. Sasse and D. Miras, "Sharp or Smooth? Comparing effects of quantization vs. FR for streamed video," Paper presented at Conference on Human Factors in Computing Systems, 2004, Vienna, Austria: 2004.

[33] K. Nakayama, Biological Image Motion Processing: A Review, Vision, 25(5), 625-660, 1985.

[34] M. Richter and M. D. Flückiger, "Usability and UX," in M. Richter & M. D. Flückinger (Eds), Usability and UX kompakt (p.12), Wiesbaden: Springer Vieweg, 2016.

[35] SAAB. *Remote tower revolutionises air traffic management*. SAAB. [Online]. Available from: http://saabgroup.com/Media/stories/stories-listing/2017-02/remote-tower-revolutionises-air-traffic-management/ 2018.03.13.

[36] M. Schmidt, M. Rudolph and N. Fürstenau, "Remote Tower Prototype and Automation Perspective," in N. Fürstenau (Ed). Virtual and Remote Control Tower (p. 180), Cham (ZG): Springer International Publishing Switzerland, 2016.

[37] S. Schwartz, Motion Perception, in S. Schwartz (Ed), Visual Perception: A Clinical Orientation (p.219-220), New York City: McGraw Hill – Education, 2004.

[38] A. Tripathi and M. Claypool, "Improving multimedia streaming with content-aware video scaling," Workshop on Intelligent Multimedia Computing and Networking (IMMCN), Durham, NC, USA, 2002.

[39] E. F. Wells, G. M. Bernstein, B. W. Scott, P. J. Bennett and J. R. Mendelson, "Critical flicker frequency responses in visual cortex," Experimental Brain Research 139(1), 106-110, 2001.

[40] M. B. Weinger, Vigilance, boredom and sleepiness, *Journal of Clinical and Computing 15*(7-8), 549-552, 1999.

 ISBN: 978-1-61208-627-9

User Density Estimation System using High Frequencies in a Specific Closed Space

Myoungbeom Chung
Division of Computer Engineering
Sungkyul University
Anyang City, South Korea
e-mail: nzin@sungkyul.ac.kr

***Abstract*—In this paper, we propose a user density estimation system using high frequencies and the microphone of a smart device in a specific closed space. High frequencies are sent to the closed space by the server speaker of the proposed system, and smart devices located in the space detect the high frequencies. The smart devices detecting the high frequencies send a message to the server system, and the system counts the smart devices that detected the high frequencies in the space. We tested user density with the proposed system, using 10 smart devices to evaluate performance. According to the test results, the proposed system showed 95% accuracy. The system can estimate exact user density in a specific closed space, and it can be a useful technology for protecting people's safety and measuring space use in indoor spaces.**

Keywords-high frequencies; inaudible sound; density estimation; smart device.

I. INTRODUCTION

Recently, along with the developments of Virtual Reality (VR) and Information Technology (IT), many sports services have begun to be offered, which are available in specific closed spaces, such as VR cafés, screen baseball zones, sports experience game spaces, and bowling pubs. In Korea, now, 450 screen baseball zones were built and about 50 VR cafés were opened at offline theme parks. Additionally, as a start-up item for successful businesses, affiliate stores, such as VR Playce [1] and VRIZ by YJM games [2] have been appearing regularly. Therefore, today, a great number of people go to specific closed spaces to enjoy various services, and user density estimation technologies in specific closed spaces are increasingly required to insure the safety of the service users.

User density estimation means the estimation of the number of human count in the specific space. Recent technologies for user density estimation are divided into 2 classes. The first class counts the number of people by analyzing video images or by detecting the motion vector of the people [3][4]. Others use radio devices, such as Radio Frequency Identification (RFID) tags, smart devices, sensor nodes, etc. instead of video images, and they count the number of each object for user density estimation [5][6]. However, when only one video image is used, the analysis type using video images cannot count the exact number of people, and it cannot be used in smoggy or dark spaces. Furthermore, because there are some serious issues related to personal privacy risks when video images are used, it is quite difficult to apply video image analysis for user density estimation. The first approach that uses RFID tags must supply the supplement with a RFID tag for user density estimation, and the approach that uses smart devices cannot be used in indoor spaces because of the poor signal from the smart device's Global Positioning System (GPS). Although the above technologies are suitable for outdoor and open spaces, such as subways, soccer stadiums, or baseball stadiums, they cannot be applied in specific closed spaces.

Therefore, in this paper, we propose a new user density estimation system using high frequencies via the speaker of a server system and the microphone of a smart device. The microphone of the smart device can detect an audible frequency range from 20 Hz to 22 kHz, so the smart device can detect specific high frequencies from received sounds via an application [7]. We use a widely available simple speaker for the speaker of the server system, and 2 high frequencies between 18 kHz and 22 kHz. These high frequencies are regularly used in high frequency studies, such as smart information service applications and data transmission using high frequencies; these frequencies have an important feature, which is that people cannot hear them in an indoor space [8][9]. In our system, smart devices located in the same indoor space receive sounds around each device, and the devices send a message to the server when they detect the specific high frequencies by analyzing the received sound. Thus, because the server gathers each message from the smart devices and counts the number of devices, the proposed system can estimate user density in a specific closed space. To evaluate the performance of the proposed application and server system, we developed a high frequency detection application for smart devices and a user density estimation server system, and we conducted an experiment on user density estimation using 10 smart devices. The results show that the proposed system is useful as a user density estimation technology in specific closed spaces because the accuracy of the proposed application and server system is over 95%. Therefore, as the proposed system is a new user density estimation technology using inaudible high frequencies and the microphones of smart devices, it can be a useful technology to protect people's safety in closed spaces, such as VR cafés, screen baseball zones, sports experience game spaces, etc.

 ISBN: 978-1-61208-627-9

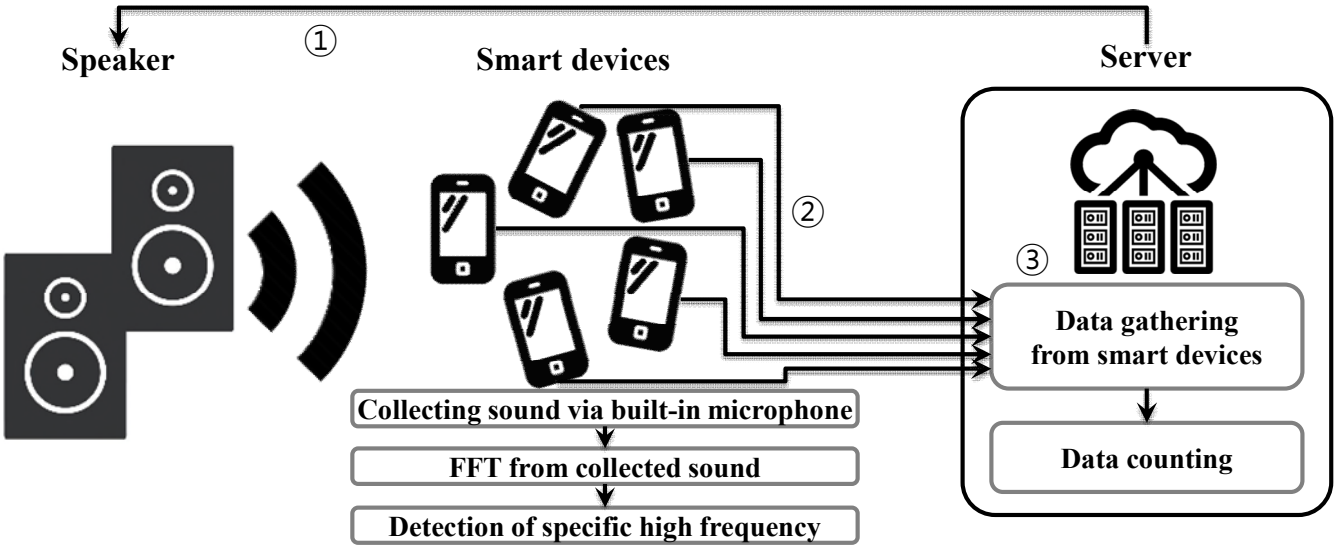


Figure 1. Total flow of the proposed application and server system.

This paper is organized as follows. In Section 2, we describe the proposed application based on smart devices and a server system. In Section 3, we describe an experiment using the proposed application and server system and discuss the results of the experiment regarding performance. Finally, in Section 4, we present the conclusions and our further research.

II. User Density Estimation System Using High Frequencies

In this section, we explain the proposed application based on smart devices and a server system for user density estimation in a specific closed space. The total flow of the proposed system is shown in Figure 1. In Figure 1, the speaker of the server system generates 1 specific pair of high frequencies (over 18 kHz) in a specific closed space over a fixed number of seconds (①), and the smart devices in the space collect the nearby sound via the microphone of each smart device. The collected sounds are converted to frequencies using Fast Fourier Transform (FFT) [10], and each smart device sends the pair of high frequencies and its own GPS information to the server system when it detects the specific pair of high frequencies over 18 kHz (②). The server gathers the data for the pair of high frequencies and the GPS information from each smart device, and then it counts the number of smart devices located in the specific closed space at the same time (③).

The specific pair of high frequencies over 18 kHz is selected as two high frequencies of 100 Hz units between 18 kHz and 22 kHz (total: 41 types of pairs). To avoid interference from other high frequencies, the interval between each high frequency is over 600 Hz. Thus, the pair of high frequencies can be composed of a total 595 types, such as 18.0 kHz–18.7 kHz, 18.0 kHz–18.8 kHz, … , 21.3 kHz–22.0 kHz. The composed pairs of high frequencies are generated by the speaker of the proposed server system and produced ***n*** times over ***k*** seconds. ***k*** is the duration of the pair of high frequencies and ***n*** is their repetition time. The produced type of pair of high frequencies is shown below in Figure 2.

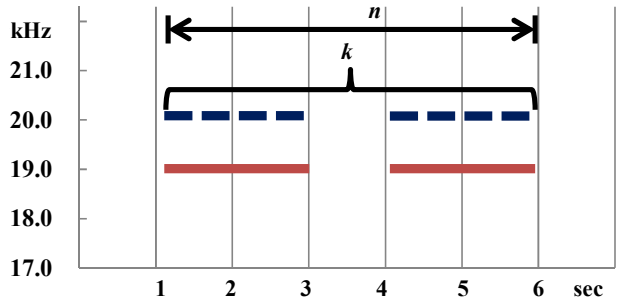


Figure 2. Example of the proposed pair of high frequencies for user density estimation.

In Figure 2, the pair of high frequencies is 19.0 kHz and 20.0 kHz, ***k*** is 5 seconds, and ***n*** is 2 times. The pair of high frequencies is generated by the speaker of the server system, and each smart device located in the specific closed space checks whether the pair of high frequencies consistently exists or not. If a smart device detects the pair of high frequencies, it waits for the fixed ***m*** seconds and detects the pair of high frequencies again to confirm that it is the same pair. Then, if the first and second pairs of high frequencies are the same, the smart device sends the pair's high frequency value and the smart device's GPS information to the server system.

 ISBN: 978-1-61208-627-9

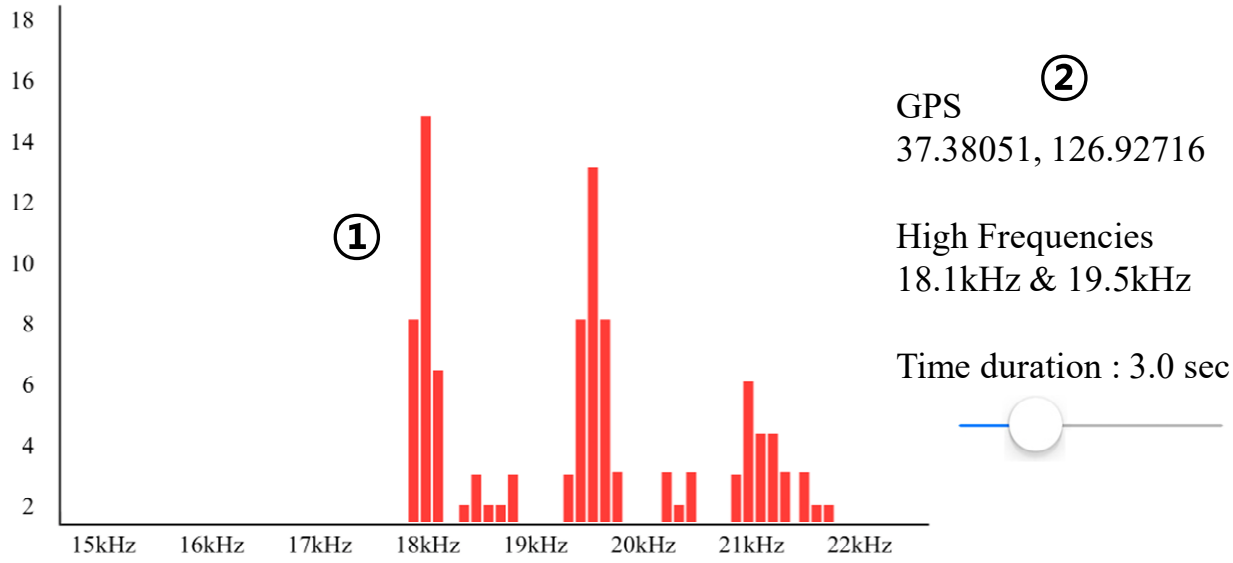


Figure 3. Screen composition of the proposed application for user density estimation.

Next, the server system confirms whether the values of the pair of high frequencies from the speaker and from the smart device are the same or not. If they are the same, the server system calculates each distance from the speaker's GPS coordinates to the smart device's GPS information. If the distance is within a critical distance (***r***) using Euclidean distance, the smart device is located in the specific closed space and the server system counts the smart device. Thus, the proposed application and server system can estimate user density in the specific closed space.

III. EXPERIMENTS AND EVALUATION

This section explains the proposed application based on smart devices for user density estimation. We describe the experiment for user density estimation and analyze the results of the experiment using the proposed application and server system. The screen composition of the proposed application is shown in Figure 3. In Figure 3, the graph located on the left-hand side of the figure shows the bin value of the high frequencies from the collected sound data, and we confirmed that 18.1 kHz and 19.5 kHz stand out among the other high frequencies. The text located on the right-hand side of Figure 3 is the smart device's GPS information, the detected pair of high frequencies, the set duration time (***k*** seconds), and the slider bar for setting duration time ***k***. In Figure 3, because we assumed that ***n*** was 2 times and ***m*** was 1 second, if we set the slider bar to 3, ***k*** is 3 seconds and the application detects the first pair of high frequencies during (***k***–1)/2 seconds. Then, the application waits for 1 second and detects the second pair of high frequencies during (***k***–1)/2 seconds again. For example, if we set 3 as ***k*** seconds, the application checks the first pair of high frequencies for 1 second, waits for 1 second, and checks the second pair of high frequencies again for 1 second.

Next, we continued the experiment using the proposed application server system. The specific closed space was a 7 × 4 m laboratory, and the speaker of the server system was located in the top corner of the laboratory, as illustrated in Figure 4. The space had a table, a hanger, four desks, and four chairs. A speaker which was named Harman Kardon Omni 20+ was located at left top corner of the space.

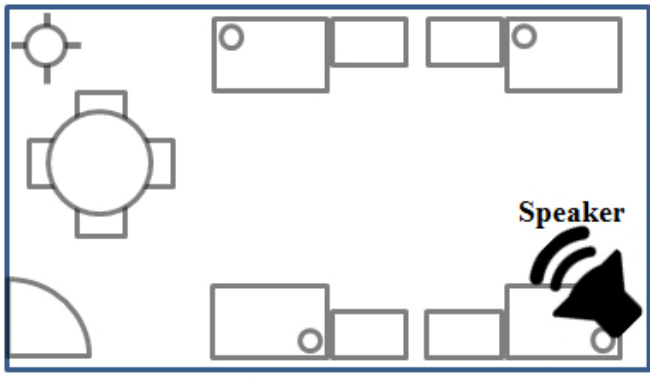


Figure 4. The floor plan of laboratory for experiment

The pair of high frequencies was 18.0 kHz and 19.0 kHz, ***k*** was 3.0 seconds, and ***m*** was 1.0 seconds. We used 10 smart devices of various models, such as iPhone 7, iPhone 6, and Galaxy s7. The server hardware was Intel(R) Core(TM) i5 CPU 750, 8G RAM, and the server environment was Apache 2.2.14, PHP 5.2.12, and MySQL 5.1.39. Each smart device was running the proposed application in background mode, and they were located in various positions around the laboratory, such as on the desk, on the chair, in the inside pocket of a jacket hung from a hanger, in front of a computer monitor, or on the floor. We generated the pair of high frequencies 100 times using the speaker of the server system

 ISBN: 978-1-61208-627-9

in the laboratory; Figure 5 shows the detection results of the pair of high frequencies from each smart device.

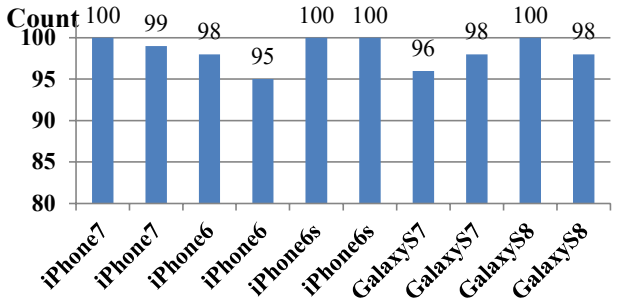


Figure 5. Detection results of the pair of high frequencies from each smart device.

In Figure 5, i7 means iPhone 7, i6 means iPhone 6, i6s means iPhone 6s, G7 means Galaxy s7, and G8 means Galaxy s8. The count does not refer to the detection number of the pair of high frequencies; it refers to the number of times a signal was sent to the server system when each smart device detected the pair of high frequencies. Most smart devices detected the pair of high frequencies over 95 times. We can see that the fourth i6 showed detection 95 times, and the seventh G7 showed detection 96 times. Because these two devices were located in the inside pocket of a jacket, we expected that these devices would have more difficulty than the others in detecting the pair of high frequencies.

Next, using Euclidean distance (***r***: 10 m), we checked the results from the server system to see whether or not all of the smart devices were located in the same place as the data from the pair of high frequencies. Because the fourth i6 and seventh G7 sent the data to the server system 95 and 96 times, respectively, the server system showed 95 times that ten smart devices were located within the same place at the same time. Thus, the proposed application and server system showed 95% accuracy from this experiment, and we believe that the proposed system can be a useful technology for user density estimation in a specific closed space.

IV. Conclusions and Future Work

In this paper, we have proposed a new user density estimation system using pairs of high frequencies from a server system and the microphones of smart devices. In the experiment, the server system generated the pairs of high frequencies in the specific closed space, and various smart devices detected the pairs of high frequencies and sent the frequencies' data and the smart devices' GPS data to the server system. From this process, the server system was able to count the number of smart devices located in the same space at the same time, and it was able to estimate user density in the specific closed space. Therefore, the proposed application and server system could be useful systems to estimate user density in closed spaces, and it could be a useful technology to protect user safety in closed spaces, such as VR cafés, screen baseball zones, sports experience game spaces, and bowling pubs. Because, when an emergency such as a fire occurs or building collapse, the proposed method guides to disaster relief staff the exact location of people in the closed space.

In future research, we will study a user density estimation system for multiple closed spaces in the same building and develop a visualization of the density results from multiple closed spaces from the server system. And, we will compare to our proposed method and the other user density estimation technologies using more smart devices which are moving or stopping in closed spaces such as sports game spaces, bowling pubs, and VR cafes. Moreover, we will study how the accuracy of the proposed application and server system can be improved.

Acknowledgment

This research project was supported in part by the Ministry of Education under Basic Science Research Program (NRF-2013R1A1A2061478) and (NRF-2016R1C1B2007930), respectively.

References

[1] VR Playce, http://www.vrplayce.co.kr/, Online access April 12th 2018.

[2] VRIZ by YJM games, http://english.yonhapnews.co.kr/news/2017/02/15/0200000000AEN20170215009900320.html, Online access April 12th 2018.

[3] V. Lempitsky, A. Zisserman, "Learning to count objects in images," In Advances in neural information processing systems, pp. 1324-1332, 2010.

[4] H. Wang, T. Wang, K. Chen, and J.-K. Kämäräinen, "Cross-granularity graph inference for semantic video object segmentation," International Joint Conference on Artificial Intelligence (IJCAI), 2017.

[5] Y. Cong, H. Gong, S. C. Zhu, and Y. Tang, "Flow mosaicking: Real-time pedestrian counting without scene-specific learning," In Computer Vision and Pattern Recognition (CVPR), June 2009.

[6] F. Li et al., "A reliable and accurate indoor localization method using phone inertial sensors," Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, Sep. 2012, pp. 421-430, ISBN: 978-1-4503-1224-0

[7] M. B. Chung, "An Advertisement method using inaudible sound of speaker," Journal of the Korea Society of Computer and Information, vol. 20, no. 8, pp. 7-13, August 2015.

[8] J. B. Kim, J. E. Song, and M. K. Lee, "Authentication of a smart phone user using audio frequency analysis," Journal of the Korea Institute of Information Security and Cryptology, vol.22, no.2, pp.327-336, April 2012.

[9] M. B. Chung and H. S. Choo, "Near wireless-control technology between smart devices using inaudible high-frequencies," Multimedia Tools and Applications, vol.74, no.15, pp.5955-5971, August 2015.

[10] A. Bellini et al., "High frequency resolution techniques for rotor fault detection of induction machines," IEEE Transactions on Industrial Electronics, vol.55, no.12, pp.4200-4209, 2008.

 ISBN: 978-1-61208-627-9