A New Multimodal Database for Performance Evaluation in System Level

Mi-young Cho[1]*, Young-Sook Jeong[1], Hyun-suh Kim[2], and Howard Kim[2]
[1]Electronics and Telecommunications Research Institute, Daejeon, Korea
{mycho, ysjeong}@etri.re.kr
[2]Color Technology Laboratory, Suwon, Korea
{hyun, howard}@colortechlab.com

Abstract

Multimodal technology has become an important area of research because it holds great potential for overcoming certain problems of traditional speech recognition and face recognition methods. In spite of a variety of multimodal applications, it is difficult to guarantee of performance of the applications due to insufficient test in system level. To evaluate of performance in multimodal system level, we construct a new multimodal database which is calibrated by reference color chart.

Keywords: multimodal DB, delta E

1 Introduction

Biometrics is the technology of establishing the identity of an individual based on one or more intrinsic physiological or behavioral characteristics, such as faces, fingerprints, irises, and voices. Face and speech recognitions are a widely used biometric technology because it is more direct, user friendly, and convenient to use than other biometric approaches. Recently, Face and speech recognition technologies are now significantly advanced, has great potential in the application systems. However, currently the scope of the application is still quite limited. Speech recognition technology is still not robust enough, especially in noisy environments, and recognition accuracy is still not acceptable. Face recognition technology also has the problems that lighting changes, pose changes and time difference between the probe image and the gallery image(s) further degrade the performance. So, audio-visual user recognition has become an important area of research because it holds great potential for overcoming certain problems of traditional speech recognition and face recognition methods.

The multimodal technology has created an enormous interest in a variety of applications such as lip reading [6], audio-visual automatic speech recognition[10][5]. Performance of voice recognition systems in a car environment is poor due to the number of environmental factors such as acoustic noise. Visual information from a speakers lip movement is unaffected by acoustic noise. Rajitha Navarathna [8] proposed the lip reading system using visual information in conjunction with the audio channel has the potential to improve the performance of speech recognition in vehicles. Takami Yoshida [12] proposed audio-visual speech recognition in noisy environment for natural interaction between human and robot. He proposed a new VAD algorithm taking ASR characteristics into account, and a linear-regression-based optimal weight estimation method. Recently, Georgios Galatas proposed audio-visual speech recognition using facial depth information captured by the kinnect[4].

IT CoNvergence PRActice (INPRA), volume: 3, number: 3 (September 2015), pp. 11-17
*Corresponding author: Intelligence and Robot System Research Section, Electronics and Telecommunications Research Institute(ETRI), Daejeon, Republic of Korea, Tel: +82-42-860-6453

(a) XM2VTS database (b) CUAVE database

Figure 1: Samples of existing multimodal database

Unfortunately, it is difficult to guarantee of performance of most multimodal applications due to insufficient test in system level. The best test is direct evaluation from human subjects in real environment. However, in this case, it would be considered impossible to consistently obtain the same way for a lengthy period of time a certain number of persons. That is, it's difficult to guarantee objectivity and reproducibility. Recently, with develpoment of capturing and display devices, test method in system level using high-definition display device is proposed[2]. In this paper, we propose a new multimodal database that is calibrated video data from reference color chart and introduce test method using a new database.

2 Existing databases

There are several multimodal databases available for researchers to be able to directly evaluate their algorithms such as XM2VTS database[7], CUAVE[9]. The XM2VTS database is a multimodal database consisting of face images, video sequences and speech recordings taken of 295 subjects at one month intervals. Each recording contains a speaking head shot and a rotating head shot. Sets of data taken from this database are available including high quality colour images, 32 KHz 16-bit sound files, video sequences and a 3d Model. Since the data acquisition was distributed over a long period of time, significant variability of appearance of clients, e.g. changes of hair style, facial hair, shape and presence or absence of glasses, is present in the recordings. The XM2VTSdatabase contains 4 sessions. During each session two head rotation and "speaking" shots were taken. From the "speaking" shot, where subjects are looking just below the camera while reading a phonetically balanced sentence, a single image with a closed mouth was chosen. Two shots at each session, with and without glasses, were acquired for people regularly wearing glasses.

The CUAVE(Clemson University Audio-Visual Experiments) database includes two major sections, one of 36 individual speakers and one of 20 speaker pairs. The selection of individuals was not tightly controlled but chosen so that there is a roughly even representation of male and female speakers and also so that different skin tones and accents are present. There are also other features such as glasses, facial hair, and hats. A wide variety of skin and lip tones as well as face and lip shapes are present. The first part including individuals has various recordings of digit strings. Speakers were either asked to remain stationary in the frame of view or move around depending on the task. The frame includes the shoulders and head, and during moving tasks, speakers move side-to-side, front-to-back, and tilt their head. There is also an occasional turn of the head. The recording environment was controlled to produce high-quality, color video and sound. Lighting was controlled, and a green background was used to allow chroma-keying of different backgrounds. This serves two purposes. If desired, the green background can be used as an aid in segmenting the face region, but more importantly, it can be used to add video backgrounds from different scenes, such as a crowd or a moving car to allow for testing of robust feature segmentation and tracking algorithms.

Over the years, a large number of methods have been proposed to analyze biometrics information from images, videos, and recently from depth data. Most methods, however, have been evaluated on

datasets that did not reflect performance in system level. To address these issues, we introduce a new multimodal database that consists of the calibrated video data from reference color chart.

3 A new multimodal database

To help meet the need for the calibrated video data for audio-visual development, ETRI multimodal database was produced. To obtain subject images under various lighting conditions, lightings, diffuser, reflection were used. The locations of lightings are shown in Figure 2.

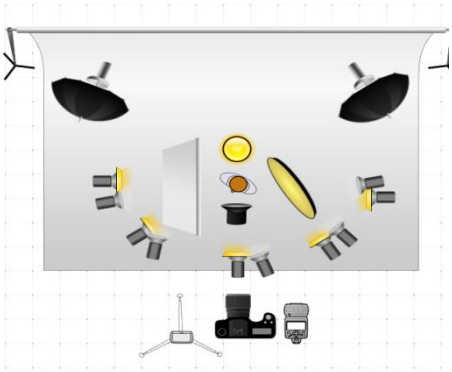

Figure 2: Capturing environment

We took ultra-high definition video clips using a Canon 5D Mark III digital single lens reflex camera with reference color chart so that the face area took up at least two thirds of the whole area of the image. The height of the camera was fixed, and we controlled the height of the chair depending on the subject's height. We captured 198 video clips from 22 subjects, which were captured under nine different lighting directions. Figure 3 shows sample video clips.

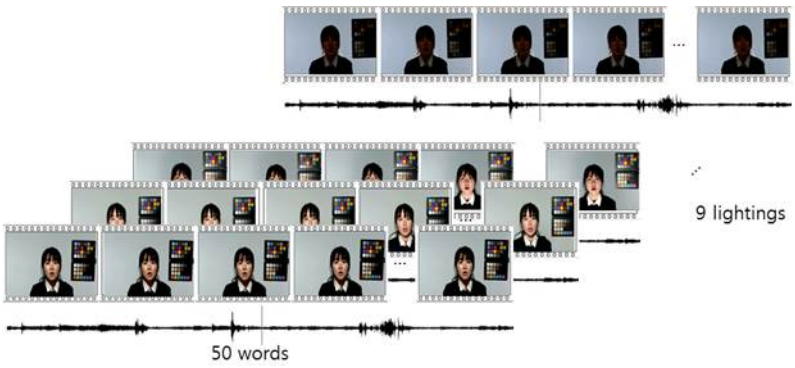

Figure 3: Sample video clips

Before to start calibrating video clips, monitor's white point should be set properly depending on the operating system of computers. Our recorded multimodal database contains images of color reference patterns under controlled lighting. We used DaVinci Resolve 11[1] from BlackMagicDesign to correct color from video clips with standardized color charts, the Datacolor SpyderCheck[3] reference was used. Figure 4 shows to generate an automatic color correction to use to create a neutral grade for the image, to use as a starting point for the rest of grading. Calibrated color values are extracted from Adobe Photoshop. Once a video clip is run and captured, we can sample a calibrated grey color with the color

sampling tool. From here, we choose sampling color to 11 × 11 pixels allow for the each actual sized of color chart target.

Figure 4: Example of spoofing with video

4 Color difference measurement

In this section, we focus on how to measure similarity between real faces and captured video clips. We consider the color difference to solve this problem. Before measuring color difference, we briefly discuss about the color space.

The International Commission on Illumination(CIE) standardized color order systems by specifying the light source (or illuminants), the observer and the methodology used to derive values for describing color. When a color is expressed in CIELAB color space, $L^\star$ defines lightness, $a^\star$ denotes the red/green value and $b^\star$ the yellow/blue value. $L^\star$ represents the darkest black at $L^\star = 0$, and the brightest white at $L^\star = 100$. The color channels, $a^\star$ and $b^\star$, will represent true neutral gray values at $a^\star = 0$ and $b^\star = 0$. The red/green opponent colors are represented along the $a^\star$ axis, with green at negative $a^\star$ values and red at positive $a^\star$ values. The yellow/blue opponent colors are represented along the $b^\star$ axis, with blue at negative $b^\star$ values and yellow at positive $b^\star$ values.

Common definition of color difference or distance between two colors make use of the Euclidean distance in a device independent color space(CIE calls their distance metric $\Delta E^\star ab$). The CIEDE2000 formula was published by the CIE in 2001. The formula provides an improved procedure for the computation of industrial color difference. The CIEDE2000 formula is considerably more sophisticated and computationally involved than its predecessor color-difference equations for CIELAB $\Delta E^\star$ and the CIE94 color-difference$\Delta E94$[11].

The color difference, or ΔE, between a sample color $L_2a_2b_2$ and a reference color $L_1a_1b_1$ is:

$$\Delta E = \sqrt{\left(\frac{\Delta L'}{K_L S_L}\right)^2 + \left(\frac{\Delta C'}{K_C S_C}\right)^2 + \left(\frac{\Delta H'}{K_H S_H}\right)^2 + R_T \left(\frac{\Delta C'}{K_C S_C}\right)\left(\frac{\Delta H'}{K_H S_H}\right)} \tag{1}$$

Table 1 shows changes of $\Delta E^\star 00$ between color patch reference and sample video clip under front lighting condition depending on calibration. Average $\Delta E^\star 00$ of video clip after calibration is lower than that before color adjustment. As a result, we can get a multimodal database similar with real environment.

5 Conclusion

In this paper, we introduce a new multimodal database for performance evaluation in system level and present video clips similar with real image by measuring $\Delta E^\star 00$. In the future work, we apply to performance evaluation of in multimodal system level.

Table 1: $\Delta E^{\star}00$ between reference and video clips before or after calibtation

No.	Name	Color Patch Reference			Uncalibrated video clip				Calibrated video clip			
		$L^{\star}$	$a^{\star}$	$b^{\star}$	$L^{\star}$	$a^{\star}$	$b^{\star}$	$\Delta E^{\star}00$	$L^{\star}$	$a^{\star}$	$b^{\star}$	$\Delta E^{\star}00$
1E	Card White	96	2	3	93	0	2	2.90	99	-1	4	2.95
2E	20 (%) Gray	80	1	2	81	0	2	1.08	85	-2	4	4.26
3E	40(%) Gray	66	1	2	66	1	0	1.13	69	-3	2	4.33
4E	60(%) Gray	50	1	2	48	0	-1	2.17	47	-3	-1	4.35
5E	80(%) Gray	34	0	1	25	0	-1	6.68	32	-3	1	3.55
6E	Card Black	17	1	-1	5	0	0	7.84	15	-1	1	2.40
1F	Primary Cyan	47	-33	-29	54	-12	-33	9.19	49	-18	-32	4.68
2F	Primary Magenta	50	53	-14	56	53	-5	5.72	49	53	-17	1.69
3F	Primary Yellow	84	3	87	82	-5	65	5.66	85	3	77	2.33
4F	Primary Red	41	61	31	44	60	37	3.11	36	60	42	5.22
5F	Primary Green	54	-41	35	52	-39	23	3.92	56	-33	33	2.93
6F	Primary Blue	25	14	-49	17	24	-48	6.46	15	20	-50	6.97
1G	Primary Orange	61	38	61	64	31	61	3.33	59	40	67	2.20
2G	Blueprint	38	7	-43	38	17	-45	2.56	38	7	-48	1.94
3G	Pink	50	49	16	53	50	23	3.82	45	55	20	5.22
4G	Violet	29	19	-24	25	25	-17	4.39	27	22	-24	1.74
5G	Apple Green	72	-24	60	69	-26	46	4.78	74	-20	58	1.63
6G	Sunflower	72	24	72	67	15	61	5.12	65	23	68	5.35
1H	Aqua	70	-32	2	75	-23	-5	5.95	72	-24	-2	4.11
2H	Lavender	54	9	-26	57	14	-24	3.40	58	6	-26	3.57
3H	Evergreen	42	-16	23	38	-16	18	4.05	36	-13	23	5.35
4H	Steel Blue	49	-5	-23	48	-1	-26	1.81	49	-7	-27	2.17
5H	Classic Light Skin	65	18	19	63	16	16	2.42	62	16	17	2.93
6H	Classic Dark Skin	36	14	16	27	17	17	7.39	34	19	19	3.46
	Average							4.37				3.56

Acknowledgments

This work is supported partly by the R&D program of the Korea Ministry of Trade, Industry and Energy(MOTIE) and the Korea Evaluation Institute of Industrial Technology (KEIT). (Project: Technology Development of service robot's performance and standardization for movement/manipulation/HRI/Networking, 10041834).

References

[1] BlackmagicDesign. Davinci resolve. `https://www.blackmagicdesign.com/products`.

[2] M. Y. Cho, Y. S. Jeong, and B. T. Chun. A new approach for face recognition performance evaluation for robot using led monitor. In *Applied Mechanics and Materials*, volume 548, pages 939–942. Trans Tech Publ, 2014.

[3] Datacolor. Spydercheckr. `http://spyder.datacolor.com/portfolio-view/spydercheckr/`.

[4] G. Galatas, G. Potamianos, and F. Makedon. Audio-visual speech recognition incorporating facial depth information captured by the kinect. In *Proc. of the 20th European Signal Processing Conference (EUSIPCO'12), Bucharest, Romania*, pages 2714–2717. IEEE, February-March 2012.

[5] A. Karpov, A. Ronzhin, K. Markov, and M. Zeleznỳ. Viseme-dependent weight optimization for chmm-based audio-visual speech recognition. In *Proc. of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH'10), Makuhari, Chiba, Japan*, pages 2678–2681. ISCA, September 2010.

[6] H. A. Mahmoud, K. Alghathbar, and F. B. Muhaya. Motion estimation analysis for unsupervised training for lip reading user authentication systems. In *Proc. of the 10th WSEAS International Conference on Automation & information (ICAI'09), Prague, Czech Republic*, pages 80–87. World Scientific and Engineering Academy and Society, March 2009.

[7] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Proc. of the 2nd International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99), Washington, D.C., USA*, volume 964, pages 965–966, 1999.

[8] R. Navarathna, P. Lucey, D. Dean, C. Fookes, and S. Sridharan. Lip detection for audio-visual speech recognition in-car environment. In *Proc. of the 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA'10), Kuala Lumpur, Malaysia*, pages 598–601. IEEE, May 2010.

[9] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *Proc. of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02), Orlando, Florida, USA*, volume 2, pages II:2017–II:2020. IEEE, May 2002.

[10] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.

[11] G. Sharma, W. Wu, and E. N. Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

[12] T. Yoshida, K. Nakadai, and H. G. Okuno. Two-layered audio-visual speech recognition for robots in noisy environments. In *Proc. of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'10), Taipei, Taiwan*, pages 988–993. IEEE, October 2010.

Author Biography

Mi-Young Cho Mi-Young Cho is a senior researcher at the Electronics and Telecommunications Research Institute. She received her BS degrees in information and telecommunication engineering from Chosun University in 2002, and her MS and PhD degrees in computer science from Chosun University in 2004 and 2008, respectively. Since 2009, she has been researching the testing and evaluation of service robots at the Electronics and Telecommunications Research Institute.

Young-Sook Jeong is a principal researcher at Electronics and Telecommunications Research Institute. She received her BS degrees in computer science from the Ewha Womans University in 1988, and her MS degree in electronic engineering from the Chungnam National University in 2001, respectively. She has research experience in the field of the performance evaluation and standardization for service robot.

Hyun-suh Kim received the MPS degree in Department of Digital Photography from School of Visual Arts, New York, U.S.A, in 2012. Currently, she is a Ph.D candidate in Department of Photography at Chung-Ang University, Korea. Her research interest is the performance of graphics hardware and software application, digital color management, interactions between digital displays and digital photography. She teaches subject in digital photography post production in undergraduate program of Chung-Ang University. She works at Color Technology Laboratory as a senior researcher.

Howard Kim is an Adjunct Professor in the Department of Human ICT Convergence at Sungkyunkwan University and the Department of Colorist at YongIn Songdam College, Republic of Korea. He is a Chief at Color Technology Laboratory. He received the BBA and MBA in Department of Business Administration from Sungkyunkwan University. He is a Ph.D candidate in Department of Management of Technology at Sungkyunkwan University. He teaches subjects in digital color management, digital imaging and perceptual computing. He has more than thirteen years of experience in digital color management industry as a regional representative of multinational corporation.